Hierarchical Multilabel Ship Classification in Remote Sensing Images Using Label Relation Graphs

Jingzhou Chen and Yuntao Qian, *Senior Member, IEEE*

***Abstract*— Hierarchical multilabel classification (HMC) assigns multiple labels to each instance with the labels organized under hierarchical relations. In ship classification in remote sensing images, depending on the expert knowledge and image quality, the same type of ships in different remote sensing images may be annotated with different class labels from coarse to fine levels such as merchant ship (MS) or container ship (CTS). In this article, we propose a novel deep network with two output channels and their associated loss functions to learn an HMC classifier using samples labeled at different levels in the hierarchy. In the proposed network, a hierarchy and exclusion (HEX) graph is introduced to model the label hierarchy, which satisfies hierarchical constraints by encoding semantic relations between any two labels. The output nodes of the first channel are organized according to the HEX graph, and its corresponding probabilistic classification loss is built to reflect the hierarchical structure of the HEX graph. On the other hand, the output nodes of the second channel only represent the finest grained (last level in the hierarchy) classes, and its multiclass cross-entropy loss is designed to enhance the discriminative power of the HMC classifier on the last level labels, which is also compatible with constraints in the HEX graph. The combination of these two losses from two output channels can effectively transfer the hierarchical information of ship taxonomy during network training. Experimental results on two commonly used ship datasets demonstrate that the proposed method outperforms the state-of-the-art HMC approaches, and is especially advantageous when trained with fewer fine-grained samples.**

***Index Terms*— Deep learning, hierarchical multilabel classification (HMC), label relation graph, ship classification.**

I. Introduction

CLASSIFICATION of inshore and offshore ships is an essential task for coastal and harbor management [1], [2]. Compared to ship images captured by ground cameras [3], [4] at short range, top view images captured by airborne or satellite platforms [5], [6] often suffer from weather conditions like waves and clouds, similar appearance of ships in top view, and large scale variation of image resolutions, which makes ship classification in remote sensing images a challenging task. Traditional methods adopt handcrafted features, including scale-invariant feature transform (SIFT) [7], local binary patterns (LBPs) [8], and hierarchical multiscale LBP (HMLBP) [9] to distinguish different ships. Recently, the convolution neural networks (CNNs) [10]–[12] have led to breakthroughs on this task. Through a series of convolution and pooling layers, CNN models such as VGGNet [13] and ResNet [14] can extract more robust and discriminative features for ship classification.

Depending on the expert knowledge and image quality, the same type of ships in different remote sensing images may be annotated with different class labels from coarse to fine levels. For example, Liu *et al.* [5] and Zhang *et al.* [6] categorized ships into three levels. Level-1 defines ship and nonship objects. Level-2 classifies ships into coarse-grained labels, e.g., Warcraft (WC) or merchant ship (MS). Level-3 refers to the fine-grained labels, e.g., destroyer (DT) or bulk carrier (BC). Therefore, ship classification methods shall be able to make full use of training samples containing the low-resolution images with coarse-grained labels and high-resolution images with fine-grained annotations.

In conventional classification tasks, a single class label is assigned to a given instance from a set of mutually exclusive class labels [15], [16]. Multilabel classification generalizes the standard classification by allowing the input instance to have multiple labels concurrently. For example, multilabel image classification [17] aims to predict a set of salient objects in an image. In the multilabel text classification [18], each document may belong to several predefined topics simultaneously. We are particularly interested in hierarchical multilabel classification (HMC) [19] which allocates multiple labels to each sample, where labels are organized into a hierarchical structure. Such a structure formalizes the relationship among labels into a tree structure or a directed acyclic graph (DAG).

HMC problems naturally arise in many domains, such as text categorization [20]–[22] and bioinformatics [23]–[25]. Recently more works focus on text categorization [26]–[29] because texts inherently lie in the abstract structured space. In image processing and analysis, HMC systems have been used to annotate medical images [30] and classify diatom images [31]. HMC is also a prevalent problem in object recognition in remote sensing images, especially for ship classification, as the standard taxonomy system of ships is organized by a hierarchical structure. HMC methods exploit semantic relations between two labels in the real world and use

Manuscript received August 6, 2021; accepted August 30, 2021. Date of publication September 15, 2021; date of current version January 31, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0100500 and in part by the National Natural Science Foundation of China under Grant 62071421. *(Corresponding author: Yuntao Qian.)*

The authors are with the College of Computer Science, Zhejiang University, Hangzhou 310027, China (e-mail: 11621038@zju.edu.cn; ytqian@zju.edu.cn).

Digital Object Identifier 10.1109/TGRS.2021.3111117

the training samples with different levels of labels, leading to more accurate and robust classification results. Surprisingly, very few published articles have focused on HMC problem in remote sensing. This motivates us to study HMC based ship classification in remote sensing images and propose an innovative method under the deep learning framework.

HMC has two distinct characteristics: 1) the number of samples per class is usually much smaller at deeper levels of the hierarchy due to the limitations of sample quality, expert knowledge, and distribution of samples along the hierarchy and 2) the predictions must be coherent, i.e., following the hierarchy constraints. Based on these two characteristics, many approaches have been presented from flat-based to hierarchical structure-based classifiers, then to deep neural network (DNN)-based classifiers. Early flat-based methods assume all the labels in a given hierarchy are independent [32]–[34]. A straightforward way is to predict labels at the leaf nodes and heuristically add their ancestor labels according to taxonomy knowledge. However, it neglects all samples labeled at the nonleaf nodes. Other flat-based methods simplify HMC as a standard multilabel classification problem, in which labels at different levels are treated the same, while postprocessing is required to correct label inconsistencies.

In contrast to flat-based methods, hierarchical approaches take the hierarchy of labels into account. They can be categorized into local and global approaches [19]. Local methods generate a hierarchy of local classifiers, which is later used to classify instances following a top-down strategy. There are three ways of using the local information: a local classifier per node (LCN), a local classifier per parent node (LCPN), and a local classifier per level (LCL). LCN approaches [35]–[37] learn one binary classifier for each node of the label hierarchy. LCPN [38], [39] methods train a multiclass classifier to distinguish the child nodes for each parent node. LCL approaches [40]–[43] generate a multiclass classifier for each level of the label hierarchy. Global approaches gather all the levels of labels together and predict them with a single classifier. A variety of classifiers have been used for this purpose, including support vector machine (SVM) [44], logistic regression [45], Bayesian classifiers [46], [47], Markov model [48], and global margin maximization [49].

Recently, DNNs have been used for HMC problems and have shown their superiority over traditional techniques. The studies usually go along two paths to develop DNN based HMC approaches, designing new network architectures [50]–[52] or new loss functions [53], [54], respectively. In the first kind of approach, they attempt to embed the label hierarchy in their architectures. Cerri *et al.* [50] designed a chain of multilayer perceptron (MLP) networks, and each MLP network is responsible for a level of the label hierarchy to extract feature information from the instances at this level. Wehrmann *et al.* [52] fit the DNN network layers to the label hierarchy, where each layer predicts the labels in the corresponding hierarchical level, and the final layer integrates the prediction results of all levels to output the overall hierarchical label structure. These methods design specific network architectures to fuse features of the samples from different levels of the label hierarchy, whereas they always neglect the semantic information embedded in hierarchical class labels.

The second kind of DNN-based method directly considers the hierarchical label relations via loss functions. For example, Giunchiglia and Lukasiewicz [53] modified the binary cross-entropy loss by imposing the hierarchy constraint that enforces the sigmoid output of a label in the hierarchy to be the maximum score of all its subclasses, including itself. The loss function in [54] encodes the relations between any two labels in the hierarchy by hierarchy and exclusion (HEX) graph, in which each output node represents a label, and the edges in the graph capture semantic relationships between two labels. Compared with architecture-based methods, loss function-based methods avoid the network topology selection which normally requires a large number of trials and tuning, and only focus on the HMC based loss function construction. In this article, we follow the path of loss function-based DNN methods to tackle hierarchical ship classification in the remote sensing images. The main contributions are summarized as follows.

1) Limited by the image quality and expert knowledge, objects always are annotated at different levels in a hierarchical class label structure, so HMC becomes to be a hot topic and a challenging task. Ship classification in remote sensing images also faces this problem, but there is scarcely any work on it. This article is the first time to apply HMC for ship classification in the remote sensing community.
2) We propose a new loss function for DNN-based HMC method, which combines the multiclass cross-entropy loss and the probabilistic classification loss. HEX graph-based probabilistic classification loss leverages the hierarchical label structure, so it can transfer hierarchical knowledge during learning. However, when trained with few fine-grained samples, the probabilistic loss struggles to separate fine-grained classes, which impairs its learning capability for fine-grained classes. Conversely, the multiclass cross-entropy loss is widely used for fine-grained image classification, which can make up the deficiency of the probabilistic classification loss in learning fine-grained classes.
3) We conduct experiments on three real-world datasets. The experimental results demonstrate that the proposed method is superior to the state-of-the-art HMC methods especially when trained with fewer fine-grained samples.
4) HMC in remote sensing images is of significance from the aspects of both theory and application. The proposed method is general in nature and can be used for other scenarios beyond remote sensing.

II. Related Work

Recent HMC studies usually adopt neural networks. They either construct appropriate network architectures [50], [52] to extract features according to the label hierarchy or design proper loss functions [53], [54] to follow hierarchy constraints. HMC with local multilayer perceptrons (HMC-LMLP) [50] proposes to train a chain of MLPs each with a single hidden layer. First, an MLP is trained for the labels in the first

hierarchical level, then a second MLP is associated with the next level of the hierarchy, and finally, the last MLP is trained for the last level. The input of each MLP uses the output provided by the previously trained MLP to augment the feature vector of the instance. This supervised incremental greedy procedure continues until the last level of the hierarchy is reached. Hierarchically regularized deep graph convolutional neural network (HR-DGCNN) [51] is developed for HMC of texts, which first converts texts to graph-of-words, and then uses graph convolution operations to filter this graph. To further leverage the hierarchy of labels, a regularization term on the weights of the fully connected (FC) layer is added to represent the dependency among labels. Another network structure, HMC network (HMCN) [52], has two variants: the feed-forward version HMCN-F and the recurrent version HMCN-R. We adopt HMCN-F for comparison in our experiments since HMCN-F obtains better results than HMCN-R [52]. In HMCN-F, information flows in two ways. The global flow starts with the input layer and traverses all FC layers until it reaches the global output. The local flows also begin with the input layer. They not only pass by their respective global FC layers but also go through specific local FC layers, and finally end at the corresponding local output. All local outputs are then concatenated and pooled with the global output to generate a final consensual prediction. These architecture-based HMC methods achieve comparative results by merging features extracted from different levels of the hierarchy. However, their loss functions sum over binary cross-entropy losses corresponding to different hierarchical levels, which assumes each label is independent of each other, causing the implicit hierarchical relations between two semantic labels to be ignored.

The other type of technique explicitly imposes the hierarchical constraints in their loss functions [53], [54]. Coherent HMC neural network (C-HMCNN) [53] revises the binary cross-entropy loss to satisfy the hierarchy constraint. This strategy ensures that no hierarchy violation happens, i.e., for any threshold, when C-HMCNN predicts a sample belonging to a class, this sample also belongs to its parent class. Moreover, C-HMCNN multiplies sigmoid scores of each subclass by their corresponding binary labels to remedy the gradient behavior that wrongly adjusts the scores of the lower level classes and their parents in the standard binary cross-entropy loss. In other words, C-HMCNN can teach the network how to better make the prediction on the upper level classes using the prediction results on the lower level ones. HEX graph based method [54] uses the label relationship between two output nodes of a classification model to construct its loss function, in which each node represents a distinct label in the hierarchy, and edges between two nodes denote three semantic relationships: parent–child, mutual exclusive, and independent. Different from C-HMCNN that only restricts the parent–child correlation, HEX specifies links between any two nodes with three kinds of semantic edges. Its probabilistic classification loss has two good properties. First, the marginal probability of a fine-grained sample relies on the sum of its ancestors' scores. This implies that the scores of parents can impact decisions about their descendants. Second, the marginal probability of a coarse-grained sample also depends on the probabilities of its fine-grained subclasses, i.e., aggregating the knowledge of all its subclasses. Therefore, the HEX graph can transfer hierarchical knowledge during learning, but the probabilistic classification loss fails to distinguish fine-grained classes when trained with few fine-grained samples and relatively more coarse-grained samples.

Compared with architecture-based methods, loss function-based approaches avoid the network topology selection which normally requires a large number of trials and tuning, and focus on the HMC based loss function construction. Inspired by the loss function-based methods, an innovative loss function is proposed in this article. We introduce the HEX graph to represent the hierarchical label structure as it better defines the relationship between any two labels. The proposed loss function combines the probabilistic classification loss with the multiclass cross-entropy loss, in which the probabilistic classification loss encodes the hierarchical label knowledge, while the multiclass cross-entropy loss enhances the discrimination of the last-level (fine-grained) labels.

III. Proposed Method

In this section, we first introduce the formalism of the HEX graph, which allows us to express prior knowledge about the label hierarchy. Then we describe the network architecture that includes two parallel output channels. One output channel corresponds to fine-grained (the lowest level) classification, and the other organizes its output nodes in the way of the HEX graph to reflect the hierarchical structure of ship taxonomy. Finally, we describe the proposed loss function in detail.

A. Formalism of HEX Graph

The HEX graph model [54] encodes the relations between any two labels in the hierarchy by constructing the HEX graph, in which each node represents a label, and the edges in the graph capture semantic relationships between two labels. HEX graph $G = (V, E_h, E_e)$ consists of a set of nodes $V = \{v_1, \ldots, v_n\}$, directed edges $E_h \subseteq V \times V$, and undirected edges $E_e \subseteq V \times V$. Each node $v \in V$ corresponds to a distinct class label. The number of nodes n equals the number of all labels in the hierarchy. A directed edge $(v_i, v_j) \in E_h$ is a subsumption edge, indicating that class i subsumes label j, e.g., "aircraft carrier (AC)" is a parent or superclass of "Nimitz-class aircraft carrier (NT)." An undirected edge $(v_i, v_j) \in E_e$ is an exclusion edge, denoting that classes v_i and v_j are mutually exclusive, e.g., a ship cannot be the AC and MS simultaneously. If two labels are not connected by an edge, it means that these two labels are independent without constraining each other.

An example of the formalism of the HEX graph is illustrated in Fig. 1. We assume that there are two-level ship classes, i.e., coarse-grained classes and their fine-grained subclasses, in which the AC is the superclass of NT and Tarawa-class amphibious assault ship (TA), while the MS is the superclass of the container ship (CTS) and car carrier (CC). All classes in the same level of the class hierarchy are mutually exclusive. Their relationships can be defined by the HEX graph. It should be noted that any two fine-grained classes

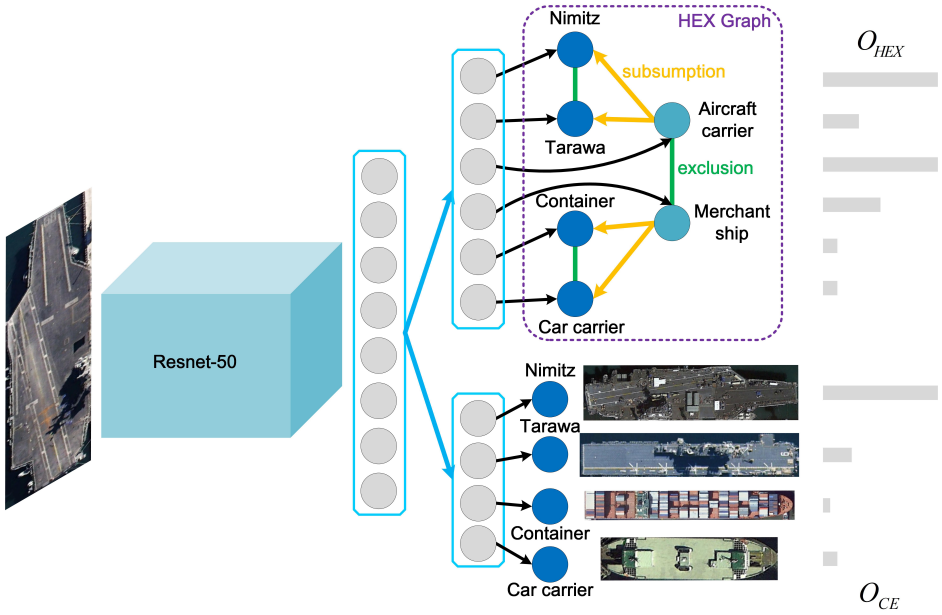


Fig. 1. Network architecture consists of a backbone network (ResNet-50) and two parallel output channels: O_{HEX} and O_{CE}. Each node corresponds to a label in the hierarchy with a directed black edge. The directed blue edges indicate the features extracted from the backbone network and fed to these two output channels. In HEX graph, directed yellow edges represent subsumption relationships, and undirected green edges stand for mutual exclusion.

share an exclusion edge in the HEX graph, but in Fig. 1, we only draw those exclusion edges between two fine-grained classes having the same parent node. If two coarse-grained classes are mutually exclusive, their descendants are mutually exclusive implicitly, so we remove these redundant edges to simplify the graph.

Each class label takes binary values, i.e., $v_i \in \{0, 1\}$, representing whether an object owns this class or not. Each edge then defines a constraint on values two labels of its incident nodes can take. An assignment of $(v_i, v_j) = (0, 1)$ (e.g., a destroyer but not a WC) for a subsumption edge $(v_i, v_j) \in E_h$ is illegal, while $(v_i, v_j) = (1, 1)$ (it is both WC and MS) is also an illegal assignment for an exclusion edge $(v_i, v_j) \in E_e$. Defined by these local constraints of individual edges, a legal global assignment of all labels in the hierarchy is a binary label vector $\mathbf{y} \in \{0, 1\}^n$ for an object. The set of all legal global assignments forms the state space $S_G \subseteq \{0, 1\}^n$ of graph G.

B. Network Architecture

As one of the commonly used CNN architectures, ResNet-50 is employed as the backbone network whose output layer is replaced with two parallel output channels. Fig. 1 shows our network architecture. The first output channel forms the probabilistic classification loss, in which each sigmoid node corresponds to a distinct label in the hierarchy. We adopt sigmoid nonlinearity instead of the softmax because sigmoid nodes signify independent relations, whereas the softmax implies mutual exclusion. We then reflect the hierarchical constraints by organizing these nodes with the HEX graph. The second output channel adopts the multiclass cross-entropy loss, in which the softmax outputs pay close attention to all mutually exclusive fine-grained classes in the last level of the hierarchy. For simplicity, we denote the first and the second output channels as O_{HEX} and O_{CE}, respectively. During the inference, softmax scores from O_{CE} are combined with the sigmoid scores of O_{HEX} to predict the final scores of all class labels in the hierarchy.

C. Loss Function

The goal of this article is to leverage the label hierarchy of ship taxonomy to tackle the HMC problem. By exploiting the hierarchy constraints, our DNN model can improve classification performance and produce coherent label predictions. The proposed loss function combines two forms of losses: the probabilistic classification loss and the multiclass cross-entropy loss. The probabilistic classification loss encodes the hierarchical structure using HEX graph $G = (V, E_h, E_e)$, and the multiclass cross-entropy loss focuses on the discrimination of fine-grained classes. We first describe the probabilistic classification loss $\mathcal{L}_{\text{HEX}}$ of the output channel O_{HEX}, and illustrate its strengths and weaknesses. Then we give our solution to remedy its weaknesses via the multiclass cross-entropy loss $\mathcal{L}_{\text{CE}}$ of the output channel O_{CE}.

1) Probabilistic Classification Loss: In our network, the probabilistic classification loss utilizes sigmoid nodes from O_{HEX}, in which each node corresponds to a label in the HEX graph. Suppose the number of sigmoid nodes in the HEX graph is n, and $\mathbf{y} \in \{0, 1\}^n$ is the binary label vector representing an assignment of all labels. Let $\mathbf{x}$ be an input image, and $\bar{x}_i$ is the sigmoid output of the i-th label node. The joint probability of all sigmoid nodes in O_{HEX} can be computed as

$$\tilde{P}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} \phi_i(\bar{x}_i, y_i) \prod_{i,j \in \{1,\ldots,n\}} \psi_{i,j}(y_i, y_j) \tag{1}$$

where $\tilde{P}(\mathbf{y}|\mathbf{x})$ is the unnormalized probability, and $\phi_i(\bar{x}_i, y_i) = e^{\bar{x}_i[y_i=1]}$. $\psi_{i,j}(y_i, y_j)$ is the constraint defined in HEX graph between any two labels

$$\psi_{i,j}(y_i, y_j) = \begin{cases} 0, & \text{if violates constraints} \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

The probability is then normalized by $\Pr(\mathbf{y}|\mathbf{x}) = (\tilde{P}(\mathbf{y}|\mathbf{x})/Z(\mathbf{x}))$, where $Z(\mathbf{x})$ is the partition function summing over the state space $\bar{\mathbf{y}} \in S_G$ of graph G

$$Z(\mathbf{x}) = \sum_{\bar{\mathbf{y}} \in \{0,1\}^n} \prod_{i=1}^{n} \phi_i(\bar{x}_i, \bar{y}_i) \prod_{i,j \in \{1,\ldots,n\}} \psi_{i,j}(\bar{y}_i, \bar{y}_j). \tag{3}$$

Given an input ship image with the ground truth label i in HEX graph, i.e., $y_i = 1$, its probability of label i is calculated by marginalizing all other labels

$$\Pr(y_i = 1|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \sum_{\bar{\mathbf{y}}: \bar{y}_i = 1} \prod_{i} \phi_i(\bar{x}_i, \bar{y}_i) \prod_{i,j} \psi_{i,j}(\bar{y}_i, \bar{y}_j). \tag{4}$$

In other words, we obtain the marginal probability $\Pr(y_i = 1|\mathbf{x})$ of label i by summing over all legal binary label vectors $\bar{\mathbf{y}} \in S_G$ that include $\bar{y}_i = 1$. The marginal probability of a leaf label in graph G relies on the sum of its ancestors' scores because all its ancestors must be 1 if the label of this leaf node takes value 1, which enables the parents' scores to influence the descendants' decisions. On the other hand, the marginal probability of an internal label is marginalized over all possible states of its descendants, i.e., aggregating the information from all its subclasses.

However, if $Z(\mathbf{x})$ and marginalization are computed with brute force, the computation grows exponentially when the number of labels increases. To speed up this process, Deng *et al.* [54] proposed to modify the standard junction tree algorithm by transforming a HEX graph in two directions. Following this approach, we first generate the minimally sparse graph and the maximally dense graph by removing and adding all redundant edges in the original HEX graph, respectively. Redundant edges are those that can be removed or added without changing the state space. Accordingly, the minimally sparse graph and the maximally dense graph are equivalent to the original HEX graph because they all have the same state space. We then build a junction tree based on the generated minimally sparse graph. For each clique of the junction tree, its state space is listed using the subgraph induced by the same clique on the maximally dense graph. Finally, we run two passes of sum-product message passing on the junction tree, performing computation only on the legal states of each clique. In the work of Deng *et al.* [54], the algorithm performs the message passing using a single sample each time. We improved it using a batch of samples to reduce the computational cost.

In the training process, the observed label can be at any level of the hierarchy, and we maximize the marginal likelihood of the observed groundtruth label. Given m training samples $\mathcal{D} = \{\mathbf{x}^{(l)}, \mathbf{y}^{(l)}, g^{(l)}\}$, $l = 1, \ldots, m$, where $\mathbf{y}^{(l)}$ is the complete ground truth label vector and $g^{(l)} \in \{1, \ldots, n\}$ is the index of the observed label, the probabilistic classification loss is

$$\mathcal{L}_{\text{HEX}}(\mathcal{D}) = -\sum_{l} \ln\left(\Pr\left(y^{(l)}_{g^{(l)}} = 1|\mathbf{x}^{(l)}\right)\right). \tag{5}$$

In many practical applications, the number of samples annotated by coarse-grained labels is much more than the number of samples having fine-grained labels, as the high-resolution and high-quality remote sensing images are harder to capture. The probabilistic classification loss is able to reflect the hierarchical label structure, but when inputting a training sample with a coarse-grained label, it tends to increase the prediction scores of the corresponding class and its subclasses while neglecting the difference in its subclasses. If the training samples with fine-grained labels are very few, the probabilistic loss may cause the learning to neglect the discrimination information in fine-grained classes, and fail to well separate the fine-grained classes. Our experiments in Section IV-D1 also find this problem. One simple but feasible solution is to increase the weight of fine-grained classification loss. Therefore, we add the multiclass cross-entropy loss especially for paying close attention to separate last-level fine-grained classes.

2) Multiclass Cross-Entropy Loss: In fine-grained image classification, the multiclass cross-entropy loss is commonly used to separate classes having subtle intraclass variations. Therefore, we use the multiclass cross-entropy loss $\mathcal{L}_{\text{CE}}$ of the output channel O_{CE} to increase the discriminative power for fine-grained classes. The multiclass cross-entropy loss is defined as

$$\mathcal{L}_{\text{CE}} = -\hat{\mathbf{y}}^{(l)} \cdot \ln \hat{\mathbf{x}}^{(l)} \tag{6}$$

where $\hat{\mathbf{y}}^{(l)}$ is the one-hot vector of the lth sample derived from a leaf fine-grained class, meaning it has 1 on a single position corresponding to the target class and 0's everywhere else, and $\hat{\mathbf{x}}^{(l)}$ is the softmax probability vector of the lth sample.

3) Combinatorial Loss: To mitigate the deficient learning for fine-grained classes, we combine the multiclass cross-entropy loss with probabilistic classification loss. The multiclass cross-entropy loss employs softmax outputs from O_{CE}, in which each node corresponds to a fine-grained leaf label in the hierarchy. Moreover, softmax nodes imply mutually exclusive relations in fine-grained labels, which reflects the hierarchy constraint defined in the HEX graph. The combined loss is defined as

$$\mathcal{L}_{\text{com}}(\mathbf{x}^{(l)}, y^{(l)}_{g^{(l)}}) = \begin{cases} \mathcal{L}_{\text{CE}} + \lambda * \mathcal{L}_{\text{HEX}}, & \text{if } g^{(l)} \text{ is in leaf nodes} \\ \mathcal{L}_{\text{HEX}}, & \text{otherwise} \end{cases} \tag{7}$$

where the λ is the hyper-parameter. Depending on whether $\mathbf{x}^{(l)}$ is discovered at fine-grained leaf labels, the combined loss decides whether it needs to incorporate $\mathcal{L}_{\text{CE}}$ or not. Finally, the total loss on $\mathcal{D}$ is

$$\mathcal{L}_{\text{total}}(\mathcal{D}) = \sum_{l} \mathcal{L}_{\text{com}}\left(\mathbf{x}^{(l)}, y^{(l)}_{g^{(l)}}\right). \tag{8}$$

IV. Experiments

Our approach utilizes ResNet-50 as the backbone network to extract features from images, and we reorganize the nodes in O_{HEX} according to the HEX graph. The whole model is optimized with the proposed loss function that integrates the probabilistic classification loss with the multiclass cross-entropy loss. The adopted optimizer is stochastic gradient descent (SGD) with momentum. Regarding the hyper-parameters, we empirically set the batch size as 32, the momentum as 0.9, and the number of epochs as 40. The initial learning rate is 0.001 and is multiplied by 0.1 every 20 epochs.

In practical applications, few samples may have fine-grained labels due to poor imaging quality and lack of expert knowledge, so more samples are observed at coarse-grained parent classes. Hierarchical classifiers should be able to learn with a mixture of coarse and fine labels and make coherent decisions. Therefore, in the training set, we select 0%, 30%, 50%, 70%, and 100% samples from each fine-grained class and relabel them to their immediate parent classes, respectively. The extreme case 0% represents that no fine-grained samples are relabeled to their parent classes, corresponding to conventional fine-grained classification. 100% denotes that all fine-grained samples from each class are relabeled to their parents. Other cases indicate that part of the fine-grained samples are relabeled to parent classes, and the rest still own fine-grained leaf labels. All images in the test set are tested with fine-grained labels.

We conduct experiments on two real-world ship datasets to validate the effectiveness of the proposed method and examine the improvement of our method over several baselines and the state-of-the-art DNN-based HMC methods. Visualization of prediction scores testifies coherent predictions and demonstrates the improvement made in our combined loss function. To further mimic the aforementioned limitations, we reduce the image resolution of selected fine-grained samples before relabeling to investigate whether HMC methods effectively exploit the hierarchical label structure.

A. Datasets

High resolution ship collection (HRSC) [5] is one of the most commonly used datasets in ship classification. The image size ranges from 300 × 300 to 1500 × 900. The training set contains 617 images, and the test set includes 438 images. Based on practical demands, there are three levels in the hierarchy organized into a tree structure. The root node separates ships and nonship objects. In the internal nodes, ships are classified into three coarse-grained categories, i.e., AC, WC, and MS. Leaf nodes in the third level refer to the fine-grained categories. In our experiments, we adopt the last two levels that comprise three coarse-grained classes and 21 fine-grained subclasses: NT, enterprise-class aircraft carrier (ET), Kuznetsov-class aircraft carrier (KT), TA, midway-class aircraft carrier (MD), Arleigh Burke-class destroyer (AB), Whidbey Island-class dock landing ship (WI), Oliver Hazard Perry-class frigate (OHP), San Antonio-class amphibious transport dock (SA), Ticonderoga-class cruiser (TD), Austin-class amphibious transport dock (AT), command ship (CS), medical ship (MDS), Warcraft-A (WCA), CTS, hovercraft (HC), yacht (YT), cruise ship (CUS), cargo ship (CGS), car carrier-A (CC-A), and car carrier-B (CC-B). Fig. 2(a) illustrates their parent-child relations. Since HRSC provides bounding box annotations, we crop every ship instance in images to avoid interference from the background.

Fine-grained ship classification (FGSC) [6] is a recently released dataset, which contains 4080 images and 22 fine-grained categories. The image size ranges from 40 × 40 to 800 × 800. From each category, 20% images are randomly selected for testing and the rest for training. In order to form the label hierarchy, we choose 21 fine-grained classes and add three coarse-grained parents following the same taxonomy in HRSC. Thus, the label hierarchy of FGSC is a two-level tree structure. Fig. 2(b) displays details of the label hierarchy. The 21 fine-grained classes are TA, general (GE), DT, landing craft (LC), frigate (FG), SA, cruiser (CU), AT, CS, MDS, combat boat (CB), auxiliary ship (AS), CTS, CC, HC, BC, oil tanker (OT), fishing boat (FB), passenger ship (PS), liquefied gas ship (LGS), and barge (BG).

B. Evaluation Metrics

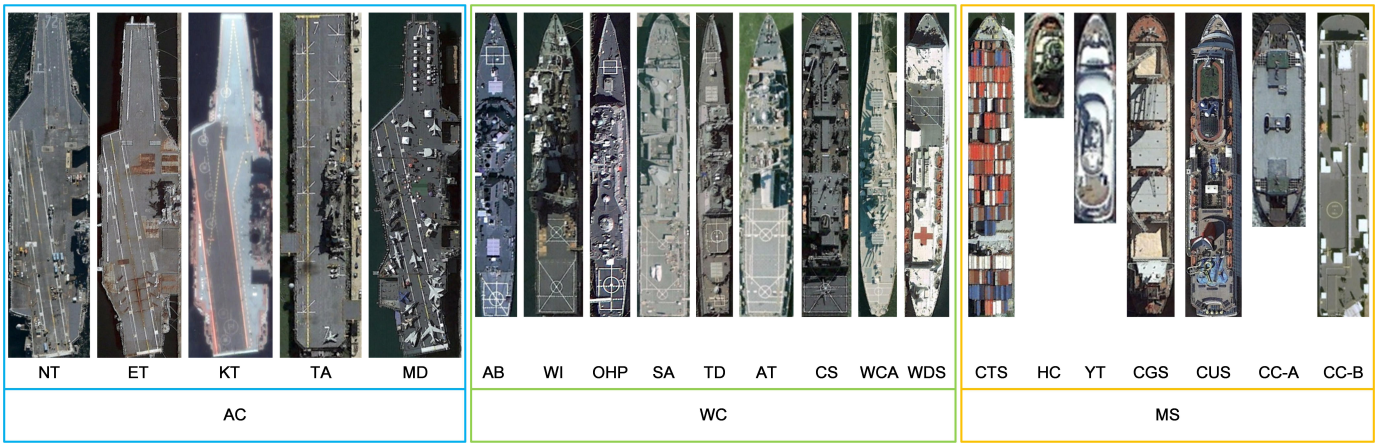
The output of HMC is a probability vector for each class. We employ two evaluation metrics. The first metric, overall accuracy (OA), evaluates fine-grained leaf labels, i.e., fine-grained image classification. We take the maximum value of the output probability vector corresponding to the last-level fine-grained classes as the predicted label and compute OA on the test set. The second metric AU($\overline{\text{PRC}}$) evaluates the output probability vector of all classes in the hierarchy. The final predictions of all classes (binary vector indicating the presence or absence of each class) are generated after the thresholding operator. We want to evaluate our predictive model independently of the threshold, as different contexts may require different threshold settings. Therefore, we follow the standard of HMC research [24], [52], [53], and use precision-recall curve (PRC) as the evaluation criterion. While PRC can be calculated for each individual class in HMC, we utilize an average PRC to quantify the overall performance, which is the same as [24], [52], and [53]. Specifically, for a given threshold value, one point ($\overline{\text{Prec}}$, $\overline{\text{Rec}}$) in the average PRC is computed as

$$\overline{\text{Prec}} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n} \text{TP}_i + \sum_{i=1}^{n} \text{FP}_i}$$
$$\overline{\text{Rec}} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n} \text{TP}_i + \sum_{i=1}^{n} \text{FN}_i} \tag{9}$$

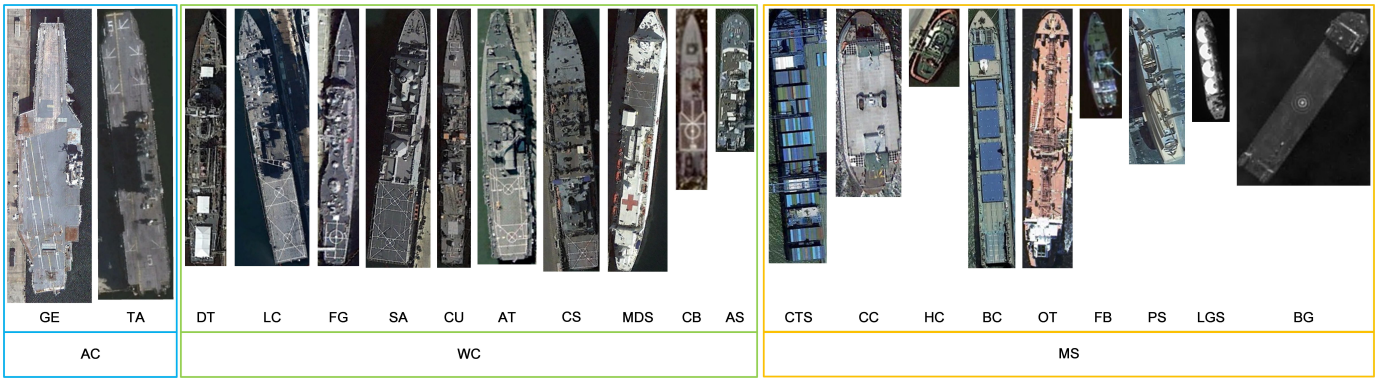
where i ranges over all classes, and TP_i, FP_i, and FN_i are the numbers of true positives, false positives, and false negatives for class label i, respectively. By varying the threshold, an average PRC is obtained and AU($\overline{\text{PRC}}$) denotes the area under this curve.

C. Determination of λ

We choose the appropriate value of λ by analyzing three relabeling proportions on HRSC. Table I gives the results of



(a)



(b)

Fig. 2. Illustrations for the two-level tree structure of the class hierarchy in two ship datasets. (a) HRSC. (b) FGSC.

TABLE I

EXPERIMENTAL RESULTS WITH DIFFERENT VALUES OF λ ON HRSC BY RELABELING SAMPLES AT EACH FINE-GRAINED CLASS TO THEIR IMMEDIATE PARENT CLASSES WITH DIFFERENT PROPORTIONS

Relabeling		0.5	1.0	2.0
0%	OA	94.3	**94.6**	94.5
	$AU(\overline{PRC})$	0.964	0.977	**0.981**
50%	OA	87.5	**88.2**	87.8
	$AU(\overline{PRC})$	0.918	**0.948**	0.932
90%	OA	74.2	75.2	**77.2**
	$AU(\overline{PRC})$	0.773	**0.902**	0.861

OA and AU($\overline{\text{PRC}}$). According to Table I, $\lambda = 1$ achieves the best results in most cases under two evaluation metrics. Thus, we set the λ to 1 in the following experiments.

D. Comparison With Baselines and State-of-the-Art Methods

1) Comparison With Baselines: We set up two baselines for comparison. The first baseline (softmax-leaf) utilizes ResNet-50, and its softmax output layer corresponds to fine-grained leaf labels. Different from the first one, the softmax output layer in the second baseline (softmax-all) includes all labels in the hierarchy. Both baselines are trained with the multiclass cross-entropy loss. As our method is developed from HEX [54], we also include HEX as a baseline. We only record OA results since softmax-leaf and softmax-all are not suitable for the AU($\overline{\text{PRC}}$) metric. Tables II and III give the experimental results on HRSC and FGSC, respectively.

Softmax-leaf only performs the fine-grained image classification with fewer fine-grained training samples, so it does not take other level classes into account. Although softmax-all can utilize the training samples with different level class labels, it organizes all labels in the hierarchy as mutually exclusive nodes in the softmax output layer, ignoring some label relations in the hierarchy. By exploiting hierarchical structure with HEX graph that can represent all hierarchical relation information, HEX surpasses those two baselines. However, it can be found that when more training samples are relabeled to coarse-grained classes, the fine-grained classification performance of HEX degenerates drastically. In contrast, our proposed method consistently outperforms HEX by adding the cross-entropy loss on fine-grained leaf classes, which proves that the proposed combinatorial loss function enhances the learning capability on fine-grained classes.

2) Comparison With State-of-the-Art Methods: We compare the proposed method with four state-of-the-art DNN-based approaches: HMC-LMLP [50], HMCN [52], HEX [54], and C-HMCNN [53]. Moreover, we combine the multiclass

TABLE II

OA(%)/AU($\overline{\text{PRC}}$) RESULTS ON HRSC BY RELABELING SAMPLES AT EACH FINE-GRAINED CLASS TO THEIR IMMEDIATE PARENT CLASSES WITH DIFFERENT PROPORTIONS

Relabeling	softmax-leaf	softmax-all	HEX [54]	C-HMCNN [53]	C-HMCNN-CE	HMC-LMLP [50]	HMCN [52]	ours
0%	91.0/	91.7/	88.1/0.957	88.6/**0.979**	91.8/0.977	87.9/0.976	89.2/0.977	**94.6**/0.977
30%	87.7/	86.6/	63.8/0.778	87.4/0.948	87.9/0.836	80.9/0.950	83.0/**0.955**	**89.4**/0.949
50%	81.7/	81.5/	51.6/0.744	83.7/0.930	85.8/0.895	78.7/0.941	82.2/0.944	**88.2/0.948**
70%	73.6/	69.2/	41.7/0.675	74.6/0.882	78.2/0.702	60.7/0.882	69.8/0.908	**84.8/0.935**
90%	48.8/	42.5/	38.4/0.673	64.3/0.789	72.9/0.716	47.6/0.811	61.4/0.857	**75.2/0.902**
100%	NIL	6.9/	19.6/0.658	5.1/0.573	11.3/0.565	4.8/0.519	21.2/0.574	11.5/0.657

TABLE III

OA(%)/AU($\overline{\text{PRC}}$) RESULTS ON FGSC BY RELABELING SAMPLES AT EACH FINE-GRAINED CLASS TO THEIR IMMEDIATE PARENT CLASSES WITH DIFFERENT PROPORTIONS

Relabeling	softmax-leaf	softmax-all	HEX [54]	C-HMCNN [53]	C-HMCNN-CE	HMC-LMLP [50]	HMCN [52]	ours
0%	87.1/	87.1/	80.0/0.926	87.5/**0.972**	87.5/0.887	66.0/0.866	70.2/0.888	**88.5**/0.940
30%	84.0/	80.7/	55.6/0.726	81.3/0.925	82.5/0.838	54.9/0.797	63.7/0.848	**85.2/0.934**
50%	77.4/	74.6/	51.5/0.723	75.3/0.894	80.6/0.867	49.9/0.793	57.5/0.821	**83.1/0.922**
70%	65.4/	64.1/	37.8/0.698	63.1/0.815	71.9/0.779	47.2/0.752	52.7/0.788	**76.0/0.892**
90%	37.5/	36.5/	32.5/0.628	35.9/0.669	48.8/0.616	36.6/0.714	37.4/0.710	**60.0/0.819**
100%	NIL	6.5/	17.4/0.611	4.4/0.555	7.1/0.394	4.3/0.548	16.8/0.387	17.4/0.595

cross-entropy loss with C-HMCNN for a fair comparison, abbreviated as C-HMCNN-CE. Tables II and III display compared results on two ship datasets with two evaluation metrics.

In Tables II and III, C-HMCNN-CE achieves better OA results than C-HMCNN, which validates that combination with multiclass cross-entropy loss does improve classification accuracy on fine-grained leaf labels. However, AU($\overline{\text{PRC}}$) results of C-HMCNN-CE are inferior to C-HMCNN. We think that the modified binary cross-entropy loss in C-HMCNN implies all labels are independent except for the parent-child constraint, but the multiclass cross-entropy loss does not conform with such a relationship because it assumes mutual exclusion in fine-grained leaf labels. Unlike C-HMCNN-CE, the relationship of mutual exclusion implied in the multiclass entropy loss is consistent with constraints defined in the HEX graph for fine-grained classes. Our method outperforms compared HMC methods in both metrics, especially for fewer training samples of fine-grained labels.

3) Visualization of Coherent Predictions: We validate two critical aspects in our method: 1) testifying whether our method makes coherent predictions in hierarchy and comparing to other related HMC approaches HMC-LMLP [50] and HMCN [52] and 2) intuitively demonstrating the improvement of the proposed method over HEX [54]. We select three fine-grained samples from each dataset and visualize their respective coarse-grained and fine-grained prediction scores. Prediction results of our method, HEX, HMC-LMLP, and HMCN are displayed in Fig. 3, corresponding to results of 0% in Tables II and III. Moreover, we illustrate prediction scores from two output channels, O_{HEX} and O_{CE}, in our model for better analysis. O_{HEX} corresponds to all labels in the hierarchy, and O_{CE} represents fine-grained labels. The ground truth is a hierarchical multilabel vector of coarse-grained classes and their fine-grained subclasses.

In Fig. 3, considering the results of HRSC, O_{HEX} and HEX make coherent predictions on two fine-grained classes, i.e., NT and CTS, and their respective parents, i.e., AC and MS. O_{CE} also correctly classify NT and CTS. However, for MDS in HRSC, HEX only determines its correct parent class (WC) but fails to recognize fine-grained classes. On the contrary, our method accurately makes consistent predictions on two-level class labels, i.e., MDS and WC. Such an improvement demonstrates that our loss function boosts O_{HEX} decisions on fine-grained classes by incorporating the supervision from O_{CE}. Similar results can be found in FGSC. O_{HEX} and HEX wrongly confuse TA with GE since they share the same superclass AC, but O_{CE} can exactly distinguish them. For the other two fine-grained classes, DT and CC, HEX only determines their respective superclasses WC and MS. With the supervision from O_{CE}, O_{HEX} improves fine-grained predictions over HEX and makes consistent decisions. In Fig. 3(a), HMC-LMLP, HMCN, and our approach all successfully make coherent decisions. For a more challenging dataset, FGSC, HMCN fails to recognize two fine-grained classes, i.e., DT and CC, and their respective parents WC and MS. HMC-LMLP only predicts the correct parent classes WC and MS, but it cannot exactly classify their respective subclasses DT and CC.

C-HMCNN replaces the output of a class with maximum scores of its subclasses, thus ensuring coherent predictions in the parent–child relationship. On the other hand, C-HMCNN ignores mutually exclusive correlations in fine-grained leaf

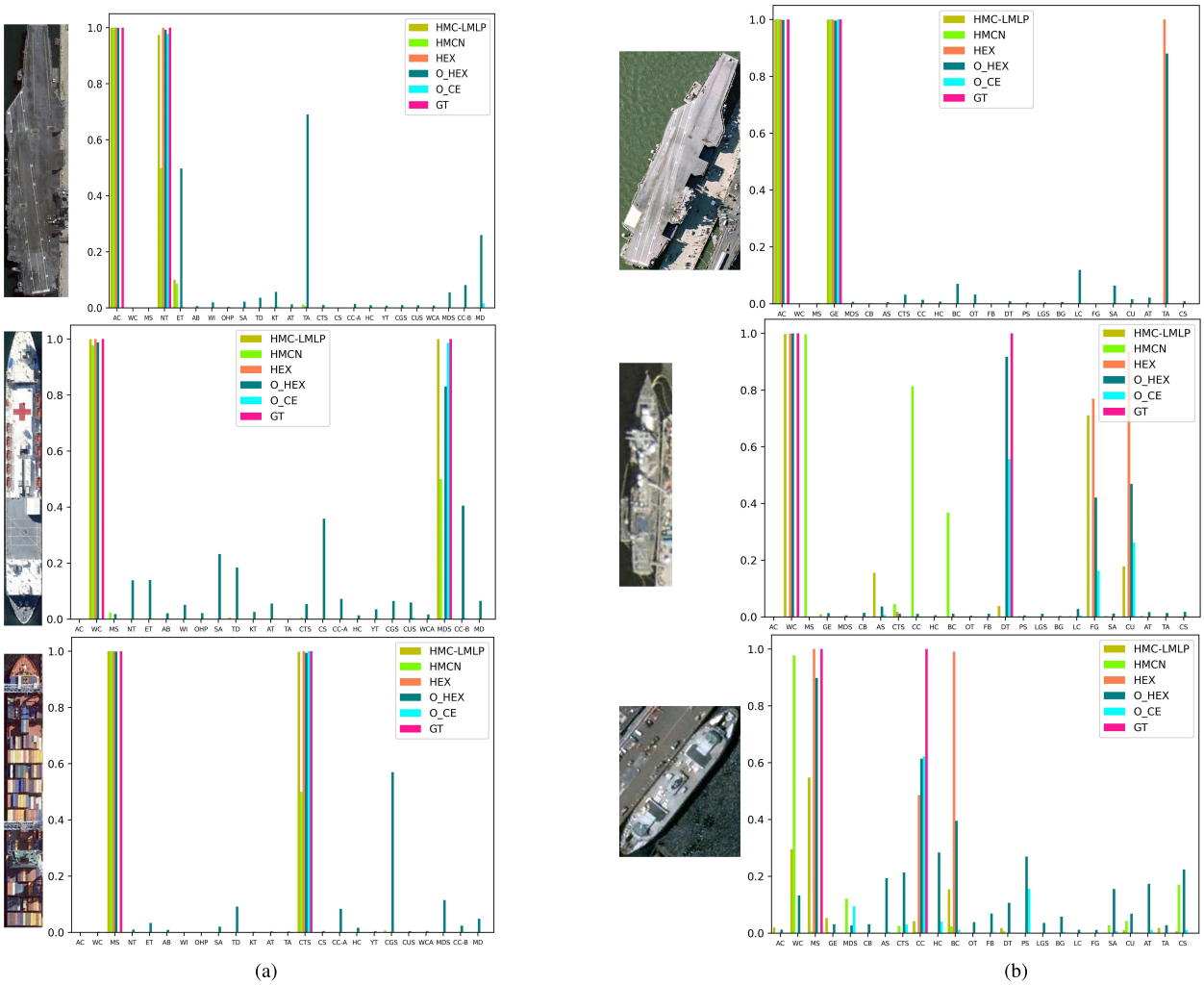


Fig. 3. We select six samples from two datasets and display the coarse-grained and fine-grained prediction scores of O_{HEX} and O_{CE} in our network and the output of HEX, HMC-LMLP, and HMCN. For three ships in HRSC, the ground truth of their fine-grained classes are NT, MDS, and CTS from top to bottom, and their corresponding coarse-grained labels are AC, WC, and MS, respectively. For three ships in FGSC, the groundtruth of their fine-grained classes are GE, DT, and CC from top to bottom, and their corresponding coarse-grained labels are AC, WC, and MS, respectively. The vertical axis represents the prediction score, ranging from 0 to 1. The horizontal axis exhibits abbreviated names of all labels in the hierarchy, in which the left three labels are coarse-grained parent classes, and the rest are fine-grained classes. (a) HRSC. (b) FGSC.

labels. Our approach explicitly encodes such constraints in the HEX graph, further enhanced by the cross-entropy loss. To compare the classification performance of our method and C-HMCNN in fine-grained classes, we plot their confusion matrixes corresponding to results of 0% in Tables II and III on two datasets. In Fig. 4, C-HMCNN behaves poorly on some classes of HRSC and FGSC, while our proposed method achieves better performance on both datasets.

E. Analyze HMC Methods by Reducing Image Resolution

In most cases, the samples captured at low resolution are always annotated with coarse-grained class labels, and the samples captured at high resolution can be annotated with fine-grained class labels. For example, in Fig. 5 two original images In FGSC dataset have AS and HC labels, respectively. If they are transformed into the corresponding low-resolution images, they become to be smaller and blurrier, so that we can only recognize them as coarse-grained classes WC and MS instead of the fine-grained subclasses AS and HC. To mimic this practical scenario, we reduce the image resolution of images with fine-grained labels to generate the images with coarse-grained labels, by the nearest-neighbor interpolation with a factor of 2. There are two different settings in our following experiments: 1) 0% signifies that we copy all fine-grained samples, reduce their image resolution, then relabel these copied samples to their parent classes. Namely, we employ data augmentation by organizing low-resolution images in a hierarchical way and 2) we select 30%, 50%, 70%, 90%, and 100% samples from each fine-grained class, reduce the image resolution of selected ones, and relabel them to their immediate parent classes, respectively. For cases of 30%, 50%, 70%, and 90%, the rest samples still own fine-grained leaf labels without reduction in image resolution. All images in the test sets are tested with fine-grained labels.

We report OA and AU($\overline{\text{PRC}}$) results on two datasets in Tables IV and V. Comparing with other HMC methods

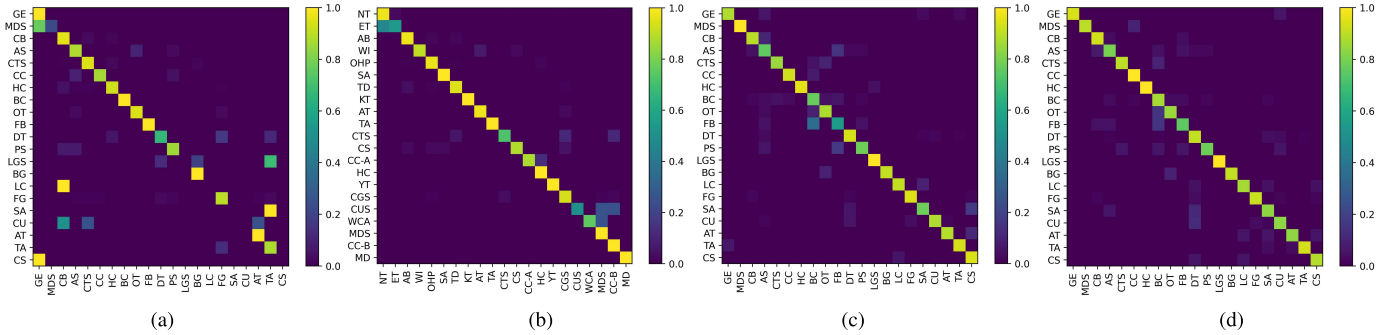


Fig. 4. Confusion Matrixes of our method and C-HMCNN on two datasets. The horizontal axis represents predicted labels, and the vertical axis is true labels. (a) C-HMCNN for HRSC. (b) Our for HRSC. (c) C-HMCNN for FGSC. (d) Our for FGSC.

TABLE IV

OA(%)/AU($\overline{\text{PRC}}$) RESULTS ON HRSC BY REDUCING THE IMAGE RESOLUTION AND RELABELING SAMPLES AT EACH FINE-GRAINED CLASS TO THEIR IMMEDIATE PARENT CLASSES ACCORDING TO DIFFERENT PROPORTIONS

Relabeling	HMC-LMLP [50]	HMCN [52]	C-HMCNN [53]	ours
0%	84.9/0.964	88.5/0.968	90.4/**0.984**	**95.5**/0.979
Diff	-3/-0.012	-0.7/-0.009	1.8/0.005	0.9/0.002
30%	83.3/**0.955**	83.9/0.947	85.8/0.941	**89.9**/0.954
Diff	2.4/0.005	0.9/-0.008	-1.6/-0.007	0.5/0.005
50%	74.9/0.927	81.5/**0.949**	83.4/0.930	**88.3**/0.943
Diff	-3.8/-0.014	-0.7/0.005	-0.3/0	0.1/-0.005
70%	63.9/0.889	74.5/0.910	74.2/0.873	**85.6/0.941**
Diff	3.2/0.007	4.7/0.002	-0.4/-0.009	0.8/0.006
90%	47.2/0.818	62.3/0.855	63.8/0.801	**74.5/0.894**
Diff	-0.4/0.007	0.9/-0.002	-0.5/0.012	-0.7/-0.008
100%	9.3/0.495	12.6/0.574	4.1/0.573	10.3/0.655
Diff	4.5/-0.024	-8.6/0	-1/0	-1.2/-0.002

(a) (b)

Fig. 5. Reducing the image resolution by a factor of 2. Two high-resolution images with fine-grained labels AS and HC, respectively, in FGSC, are transformed to low-resolution images. (a) AS. (b) HC.

HMC-LMLP, HMCN, and C-HMCNN, the proposed approach achieves the best results in most cases, especially for fewer fine-grained training samples, which reflects that our approach effectively uses the supervision from coarse-grained samples to boost the learning of fine-grained classes. Superior results also demonstrate that the proposed loss function successfully transfers hierarchical knowledge during learning. Considering results of 0% case on two datasets, it can be found that the hierarchical data augmentation improves loss-based HMC methods but decreases architecture-based HMC methods, which implies that our method and C-HMCNN better encode hierarchical relationships to utilize augmented coarse-grained samples than architecture-based methods.

For better illustration, we subtract corresponding counterparts in Tables II and III from results in Tables IV and V, respectively. The Diff results corresponding to cases of 30%, 50%, 70%, and 90% are shown in Tables II, III, and Fig. 6. Compared to HMC-LMLP, HMCN, and C-HMCNN, our method reaches more stable outcomes in most cases, which shows that the proposed approach is able to take full advantage of low-resolution images with coarse-grained labels, and is more suitable for real application scenarios.

F. Evaluate HMC Methods in Another Scenario

Both FGSC and HRSC datasets have only two-level hierarchical label relation, we would like to show the proposed HMC method can tackle label hierarchy with more than two levels. More importantly, the proposed method is general in nature, so we will show that its applications are beyond ship classification and remote sensing images. Therefore, we select the popular CUB-200-2011 bird dataset [55] to evaluate the proposed method and conduct comparative experiments with other

TABLE V

OA(%)/AU($\overline{\text{PRC}}$) RESULTS ON FGSC BY REDUCING THE IMAGE RESOLUTION AND RELABELING SAMPLES AT EACH FINE-GRAINED CLASS TO THEIR IMMEDIATE PARENT CLASSES ACCORDING TO DIFFERENT PROPORTIONS

Relabeling	HMC-LMLP [50]	HMCN [52]	C-HMCNN [53]	ours
0%	61.2/0.835	70.4/0.872	**91.0/0.978**	89.0/0.949
Diff	-4.8/-0.031	0.2/-0.016	3.5/0.006	0.5/0.009
30%	60.2/0.823	64.9/0.839	83.4/**0.945**	**86.9**/0.933
Diff	5.3/0.026	1.2/-0.009	2.1/0.02	1.7/-0.001
50%	53.5/0.801	58.1/0.821	77.4/0.907	**81.9/0.917**
Diff	3.6/0.008	0.6/0	2.1/0.013	-1.2/-0.005
70%	44.3/0.759	52.5/0.763	63.4/0.813	**77.9/0.902**
Diff	-2.9/0.007	-0.2/-0.025	0.3/-0.002	1.9/0.01
90%	35.7/0.698	37.8/0.701	37.4/0.693	**57.9/0.813**
Diff	-0.9/-0.016	0.4/-0.009	1.5/0.024	-2.1/-0.006
100%	6.5/0.536	18.2/0.524	5.4/0.570	17.8/0.627
Diff	2.2/-0.012	1.4/0.137	1/0.015	0.4/-0.01

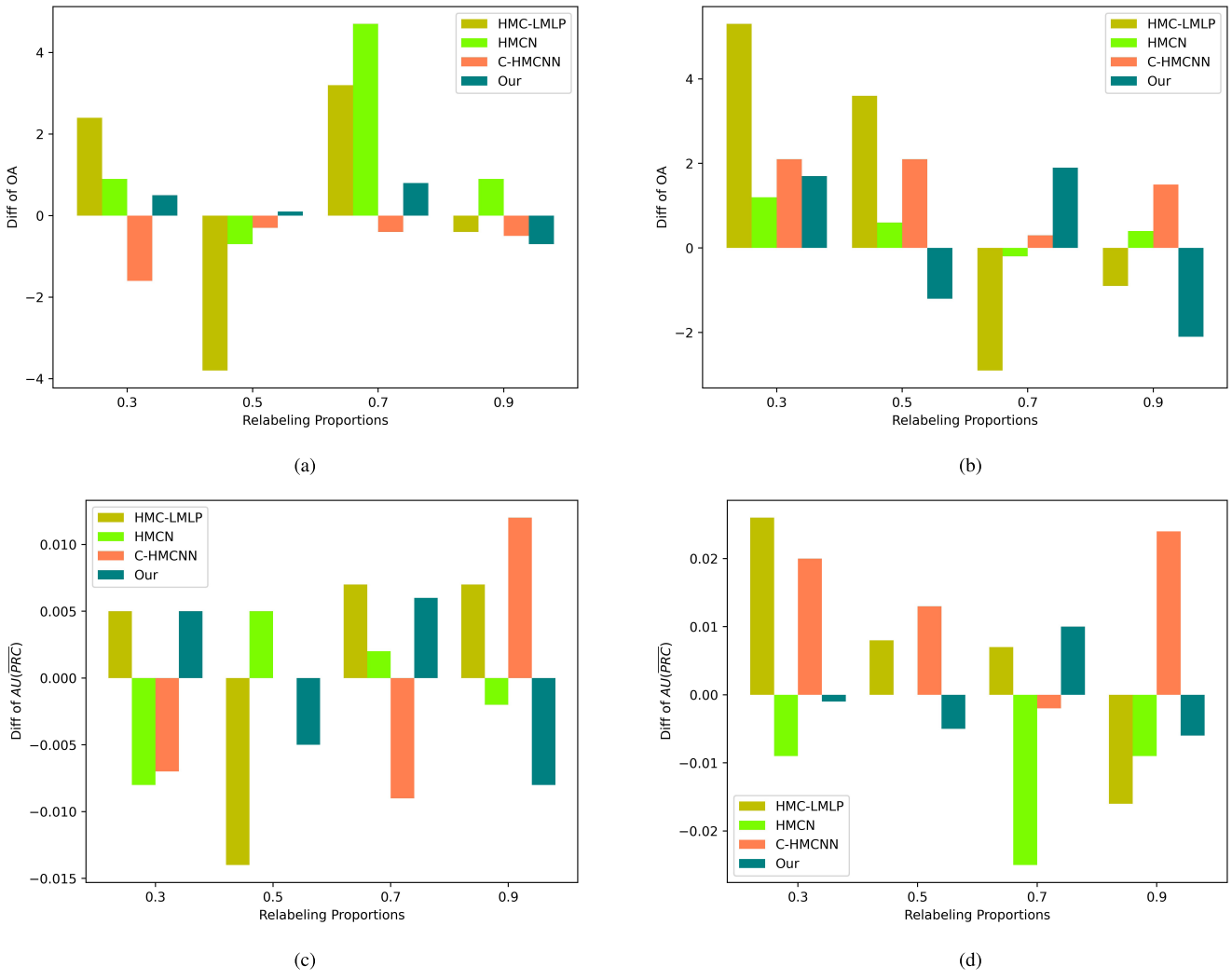


Fig. 6. Diff OA and AU($\overline{\text{PRC}}$) results of compared HMC methods on two datasets. (a) Diff of OA on HRSC. (b) Diff of OA on FGSC. (c) Diff of AU($\overline{\text{PRC}}$) on HRSC. (d) Diff of AU($\overline{\text{PRC}}$) on FGSC.

related approaches. This dataset has three hierarchical levels, with 200 species in the bottom level, 38 families in the second level, and 13 orders in the top level by tracing their parent nodes (superclasses) in Wikipedia. It contains 11 788 images, which are split into 5994 training samples and 5794 test images. The experimental results are displayed in Table VI, which also demonstrate that the proposed approach consistently outperforms those compared HMC methods in most

TABLE VI
OA(%)/AU($\overline{\text{PRC}}$) RESULTS ON CUB-200-2011 BY RELABELING SAMPLES AT EACH FINE-GRAINED CLASS TO THEIR IMMEDIATE PARENT CLASSES WITH DIFFERENT PROPORTIONS

Relabeling	HMC-LMLP [50]	HMCN [52]	C-HMCNN [53]	ours
0%	70.5/0.913	73.7/0.909	73.9/0.932	**77.3/0.936**
30%	63.6/0.889	68.7/0.885	67.1/0.903	**72.5/0.918**
50%	56.4/0.851	63.4/0.861	60.6/0.871	**68.5/0.898**
70%	42.9/0.791	53.1/0.814	45.0/0.809	**59.0/0.840**
90%	22.3/0.694	28.7/0.698	23.8/**0.749**	**38.5**/0.730
100%	1.0/0.0.650	1.1/0.631	1.5/0.588	12.9/0.642

cases. This experiment validates the generalization of our method.

V. DISCUSSION

We aim to model the label hierarchy by designing an appropriate loss function for the HMC problem. The proposed loss function comprises two critical parts. First, it explicitly encodes the hierarchical structure information with the HEX graph. Then it combines the probabilistic classification loss with the multiclass cross-entropy loss imposed on fine-grained leaf labels to remedy the insufficient learning of fine-grained classes. The comprehensive experiments demonstrate the effectiveness of these two parts. Compared to HMC-LMLP and HMCN, our superior results signify that it is necessary to exploit hierarchical constraints in semantic labels. Unlike C-HMCNN that only imposes the parent–child relation, our method further constrains the mutually exclusive correlation and augments this correlation in fine-grained leaf labels with the cross-entropy loss. Experimental results show that our combination improves fine-grained classification performance over HEX and obtains better results than C-HMCNN. Finally, the proposed approach outperforms all compared state-of-the-art HMC methods.

VI. CONCLUSION

In this article, we study the HMC problem for ship classification because various types of ships inherently own hierarchical label structure, and FGSC plays an important role in many applications. Our method can further generalize to other datasets having hierarchical annotations. To encode the hierarchical label structure in ship taxonomy, we introduce the HEX graph that specifically constrains three semantic relationships between any two labels in the hierarchy. Although exploiting the hierarchical structure information, the probabilistic classification loss based on HEX graph cannot well distinguish fine-grained classes when the fine-grained samples are few. We combine it with the multiclass entropy loss imposed on fine-grained leaf labels to improve the learning of fine-grained classes. Moreover, our approach can generalize to more sophisticated networks since we only organize nodes in the output layer according to the HEX graph. We employ two different evaluation metrics. The first metric evaluates fine-grained leaf labels, and the second measures the coherence of predicted results by taking all labels in the hierarchy into consideration. Our method shows superior performance in both metrics compared to other state-of-the-art HMC approaches on three real-world datasets.

REFERENCES

[1] Z. L. Szpak and J. R. Tapamo, "Maritime surveillance: Tracking ships inside a dynamic background using a fast level-set," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 6669–6680, 2011.

[2] X. Yang *et al.*, "Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, p. 132, 2018.

[3] M. M. Zhang, J. Choi, K. Daniilidis, M. T. Wolf, and C. Kanan, "VAIS: A dataset for recognizing maritime imagery in the visible and infrared spectrums," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 10–16.

[4] Z. Shao, W. Wu, Z. Wang, W. Du, and C. Li, "SeaShips: A large-scale precisely annotated dataset for ship detection," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2593–2604, Oct. 2018.

[5] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods*, 2017, pp. 324–331.

[6] X. Zhang, Y. Lv, L. Yao, W. Xiong, and C. Fu, "A new benchmark and an attribute-guided multilevel feature representation network for fine-grained ship classification in optical remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1271–1285, 2020.

[7] K. Rainey and J. Stastny, "Object recognition in ocean imagery using feature selection and compressive sensing," in *Proc. IEEE Appl. Imag. Pattern Recognit. Workshop*, Oct. 2011, pp. 1–6.

[8] Z. Song, H. Sui, and Y. Wang, "Automatic ship detection for optical satellite images based on visual attention model and LBP," in *Proc. IEEE Workshop Electron., Comput. Appl.*, May 2014, pp. 722–725.

[9] Z. Guo, L. Zhang, D. Zhang, and X. Mou, "Hierarchical multiscale LBP for face and palmprint recognition," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 4521–4524.

[10] H. Lin, Z. Shi, and Z. Zou, "Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1665–1669, Oct. 2017.

[11] C. Bentes, D. Velotto, and B. Tings, "Ship classification in TerraSAR-X images with convolutional neural networks," *IEEE J. Ocean. Eng.*, vol. 43, no. 1, pp. 258–266, Jan. 2018.

[12] Q. Shi, W. Li, R. Tao, X. Sun, and L. Gao, "Ship classification based on multifeature ensemble with convolutional neural network," *Remote Sens.*, vol. 11, no. 4, p. 419, Feb. 2019.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[17] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5177–5186.
[18] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 115–124.
[19] C. N. Silla Jr., and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining Knowl. Discovery*, vol. 22, nos. 1–2, pp. 31–72, 2011.
[20] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "RCV1: A new benchmark collection for text categorization research," *J. Mach. Learn. Res.*, vol. 5, pp. 361–397, Apr. 2004.
[21] J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor, "Kernel-based learning of hierarchical multilabel classification models," *J. Mach. Learn. Res.*, vol. 7, pp. 1601–1626, Jul. 2006.
[22] A. Mayne and R. Perry, "Hierarchically classifying documents with multiple labels," in *Proc. IEEE Symp. Comput. Intell. Data Mining*, Mar. 2009, pp. 133–139.
[23] Z. Barutcuoglu, R. E. Schapire, and O. G. Troyanskaya, "Hierarchical multi-label prediction of gene function," *Bioinformatics*, vol. 22, no. 7, pp. 830–836, 2006.
[24] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel, "Decision trees for hierarchical multi-label classification," *Mach. Learn.*, vol. 73, no. 2, pp. 185–214, 2008.
[25] L. Schietgat, C. Vens, J. Struyf, H. Blockeel, D. Kocev, and S. Džeroski, "Predicting gene function using hierarchical multi-label decision tree ensembles," *BMC Bioinf.*, vol. 11, no. 1, pp. 1–14, Dec. 2010.
[26] Y. Mao, J. Tian, J. Han, and X. Ren, "Hierarchical text classification with reinforced label assignment," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 445–455.
[27] W. Huang *et al.*, "Hierarchical multi-label text classification: An attention-based recurrent network approach," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 1051–1060.
[28] Y. Meng, J. Shen, C. Zhang, and J. Han, "Weakly-supervised hierarchical text classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 6826–6833.
[29] B. Chen, X. Huang, L. Xiao, Z. Cai, and L. Jing, "Hyperbolic interaction model for hierarchical multi-label classification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 7496–7503.
[30] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Džeroski, "Hierarchical annotation of medical images," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2436–2449, 2011.
[31] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Džeroski, "Hierarchical classification of diatom images using ensembles of predictive clustering trees," *Ecol. Informat.*, vol. 7, no. 1, pp. 19–29, Jan. 2012.
[32] C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka, "Automated categorization in the international patent classification," *ACM SIGIR Forum*, vol. 37, no. 1, pp. 10–25, Apr. 2003.
[33] B. Hayete and J. R. Bienkowska, "Gotrees: Predicting go associations from protein domain composition using decision trees," in *Proc. Pacific Symp. Biocomput.*, Dec. 2004, pp. 127–138.
[34] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2015, pp. 103–112.
[35] P. N. Bennett and N. Nguyen, "Refined experts: Improving classification in large taxonomies," in *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2009, pp. 11–18.
[36] W. W. Bi and J. Kwok, "Multilabel classification on tree- and DAG-structured hierarchies," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 17–24.
[37] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni, "Incremental algorithms for hierarchical classification," *J. Mach. Learn. Res.*, vol. 7, pp. 31–54, Jan. 2006.
[38] M. Kulmanov, M. A. Khan, and R. Hoehndorf, "DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier," *Bioinformatics*, vol. 34, no. 4, pp. 660–668, Feb. 2018.
[39] C. Xu and X. Geng, "Hierarchical classification based on label distribution learning," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 5533–5540.
[40] M. Ruiz and P. Srinivasan, "Hierarchical text categorization using neural networks," *Inf. Retr.*, vol. 5, no. 1, pp. 87–118, 2002.
[41] R. Cerri, R. C. Barros, and A. C. P. L. F. de Carvalho, "Hierarchical multi-label classification using local neural networks," *J. Comput. Syst. Sci.*, vol. 80, no. 1, pp. 39–56, Feb. 2014.
[42] Y. Li *et al.*, "DEEPre: Sequence-based enzyme EC number prediction by deep learning," *Bioinformatics*, vol. 34, no. 5, pp. 760–769, Mar. 2018.
[43] Z. Zou, S. Tian, X. Gao, and Y. Li, "MlDEEPre: Multi-functional enzyme function prediction with hierarchical multi-label deep learning," *Frontiers Genet.*, vol. 9, p. 714, Jan. 2019.
[44] L. Cai and T. Hofmann, "Hierarchical document categorization with support vector machines," in *Proc. 13th ACM Conf. Inf. Knowl. Manage.*, 2004, pp. 78–87.
[45] S. Gopal and Y. Yang, "Recursive regularization for large-scale classification with hierarchical and graphical dependencies," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2013, pp. 257–265.
[46] C. N. Silla Jr., and A. A. Freitas, "A global-model naive Bayes approach to the hierarchical prediction of protein functions," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Dec. 2009, pp. 992–997.
[47] O. Dekel, J. Keshet, and Y. Singer, "Large margin hierarchical classification," in *Proc. Int. Conf. Mach. Learn.*, 2004, pp. 27–34.
[48] J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor, "Learning hierarchical multi-category text classification models," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 744–751.
[49] X. Qiu, W. Gao, and X. J. Huang, "Hierarchical multi-label text categorization with global margin maximization," in *Proc. Joint Conf. Annu. Meeting Assoc. Comput. Linguist. Int. Joint Conf. Nat. Lang. Process. (ACL-IJCNLP)*, 2009, pp. 165–168.
[50] R. Cerri, R. C. Barros, A. C. P. L. F. de Carvalho, and Y. Jin, "Reduction strategies for hierarchical multi-label classification in protein function prediction," *BMC Bioinf.*, vol. 17, no. 1, pp. 1–24, Dec. 2016.
[51] H. Peng *et al.*, "Large-scale hierarchical text classification with recursively regularized deep graph-CNN," in *Proc. World Wide Web Conf.*, 2018, pp. 1063–1072.
[52] J. Wehrmann, R. Cerri, and R. Barros, "Hierarchical multi-label classification networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5075–5084.
[53] E. Giunchiglia and T. Lukasiewicz, "Coherent hierarchical multi-label classification networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–12.
[54] J. Deng *et al.*, "Large-scale object classification using label relation graphs," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 48–64.
[55] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

Jingzhou Chen received the B.E. degree in computer science and technology from Sichuan University, Chengdu, China, in 2016. He is currently pursuing the Ph.D. degree with the College of Computer Science, Zhejiang University, Hangzhou, China.

His research interests include machine learning, pattern recognition, and remote sensing image processing.

Yuntao Qian (Senior Member, IEEE) received the B.E. and M.E. degrees in automatic control from Xi'an Jiaotong University, Xi'an, China, in 1989 and 1992, respectively, and the Ph.D. degree in signal processing from Xidian University, Xi'an, in 1996.

From 1996 to 1998, he was a Post-Doctoral Fellow with Northwestern Polytechnical University, Xi'an. Since 1998, he has been with the College of Computer Science, Zhejiang University, Hangzhou, China, where he became a Professor in 2002. From 1999 to 2001, in 2006, in 2010, in 2013, from 2015 to 2016, and in 2018, he was a Visiting Professor with Concordia University, Montreal, QC, Canada, Hong Kong Baptist University, Hong Kong, Carnegie Mellon University, Pittsburgh, PA, USA, the Canberra Research Laboratory of NICTA, Macau University, Taipa, Macau, and Griffith University, Nathan, QLD, Australia. His research interests include machine learning, signal and image processing, pattern recognition, and hyperspectral imaging.

Dr. Qian is currently an Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.