Hindawi
Scientific Programming
Volume 2022, Article ID 1822539, 12 pages
https://doi.org/10.1155/2022/1822539

Research Article

Semantic Understandings for Aerial Images via Multigrained Feature Grouping

Dan Lin [1] **and Zhikui Chen** [1,2]

[1]*School of Software Technology, Dalian University of Technology, Dalian, Liaoning 116621, China*
[2]*Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian, Liaoning 116621, China*

Correspondence should be addressed to Zhikui Chen; zkchen@dlut.edu.cn

Received 27 January 2022; Accepted 28 March 2022; Published 25 April 2022

Academic Editor: Boxiang Dong

Aerial images play a key role in remote sensing as they can provide high-quality surface object information for continuous communication services. With advances in UAV-aided data collection technologies, the volume of aerial images has been greatly promoted. To this end, semantic understandings for these images can significantly improve the quality of service for smart devices. Recently, the multilabel aerial image classification (MAIC) task has been widely researched in academics and applied in industries. However, existing MAIC methods suffer from suboptimal performance as objects are located in different sizes and scales. To address these issues, we propose a novel multigrained semantic grouping model for aerial image learning, named MSGM. First, image features presented by the backbone are sent to spatial pyramid convolutional layers which extract the instances in a parallel manner. Then, three grouping mechanisms are designed to integrate the instances from the pyramid framework. In addition, MSGM builds a concept graph to represent the label relationship. MSGM resorts to the graph convolutional network to learn the concept graph directly. We extensively evaluate MSGM on two benchmark aerial image datasets, the commonly used UCM dataset, and the high-resolution DFC15 dataset. Quantitative and qualitative results support the effectiveness of the proposed MSGM.

1. Introduction

Great improvements in computer vision tasks have been achieved in the last few years, especially on image classification [1], object detection [2, 3], semantic segmentation [4, 5], and so on. In the field of remote sensing, vision-based sensors can integrate aerial images for continuous communication services. Thus, these devices are densely developed in applications. With the advances of data collection technologies, huge amounts of aerial panoramic images are monitored and available for academic research. So how to automatically acquire semantic understandings for these aerial images is of great significance. Recently, aerial panoramic image classification provides a solution for this kind of problem automatically.

Traditional models map the given aerial image into a single semantic information, named the single-label aerial image classification (SAIC) task [6]. However, aerial panoramic images with a higher spectral resolution can provide a much wider field of view and are thus associated with abundant content [7, 8]. For example, in Figure 1, *ship*, *water*, and *tree* coexist in the given aerial image. For these aerial images, complex variations in viewpoint, scale, illumination, and occlusion make existing course-grained SAIC methods fail to learn semantic information sufficiently. As a result, multilabel aerial image classification (MAIC) methods are proposed to deal with novel aerial panoramic image understanding tasks. For a given aerial image, MAIC methods aim to build a function that produces multiple labels (objects) of interest inside [9]. As an emerging research field, MAIC has attracted huge attention of researchers and is therefore applied in many applications, such as air quality monitoring [10], urban management [11], and social community detection [12, 13]. An increasing number of multilabel classification frameworks have been designed on aerial images from various angles.

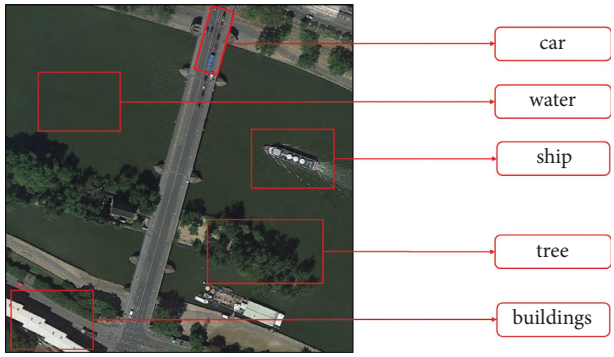


Figure 1: An example multilabel aerial image. Annotations: *car*, *water*, *ship*, *tree*, and *buildings*.

However, existing MAIC methods still suffer from suboptimal performance as it is hard to extract objects located in different areas and scales of aerial panoramic images. It is well recognized that aerial image features provide the most fundamental information in the label predicting process. Early models commonly employ pretrained convolutional neural networks (CNNs) as the backbone to learn images. However, the direct use of the rough features can result in limitations in the effectiveness on the aerial images. As these backbones are designed to extract single core feature for the given image, which cannot handle the complicated multilabel image feature learning. Then, some MAIC methods extract semantic features from aerial images with object localization models. However, object localization models need a large number of irrelevant, redundant proposals, accompanying with high computing costs [14]. Some other methods introduce recurrent neural networks (RNNs) to further build label correlations to guide the final prediction [15]. They learn dependencies sequentially and cannot exploit the content correlations, thus would miss key information when predicting multiple labels.

To address these issues, our motivation is to design novel semantic understanding methods in the multilabel classification task on aerial panoramic images. Inspired by the recently proposed spatial pyramid convolutional (SPC) technologies [16], which learn the multiscale representations of the given image in a parallel way, we designed a multigrained semantic grouping framework for the MAIC task. To do so, we first leverage specifically designed SPC layers to generate multiple feature maps for the raw image. These feature maps are at various scales and contain some semantic instances. Then, we integrate these multigrained maps with three designed semantic grouping mechanisms. Besides, to capture the semantic features from label correlations, we build a concept graph to present the relationships of labels and resort to the graph convolutional network (GCN) to extract the features of the label graph directly.

The main contributions of this article are as follows:

(1) We propose a novel multigrained semantic grouping model for multilabel aerial image understanding, named MSGM. MSGM aims to provide comprehensive semantic classifications for aerial panoramic images by learning multigrained semantic features of images and the concept graph of labels.

(2) To extract more fine-grained features from the given aerial image, we design SPC layers with multiscale feature encoders in a parallel manner. Then, these multidimensional representations are organized into final aerial image features by three designed multigrained semantic grouping mechanisms. To capture self-adapted information of aerial label correlations, we build the label relationship into a concept graph and learn the concept graph structure directly by a designed GCN module with an attention mechanism.

(3) Sufficient experiments at various angles are carried out to verify the performance of the proposed MSGM model on both the UCM and DFC15 aerial image datasets. The results demonstrate the effectiveness of MSGM in not only the research but also the real-world application.

1.1. Related Work. In the existing remote sensing ecosystem, a large amount of data generated by sensors and devices is transferred [17, 18]. Recently, with the advances in real-time data collection technologies, an increasing number of aerial panoramic images have been acquired for scientific research [19]. The MAIC task has become a fundamental problem in the aerial image learning field.

Benefiting from machine learning technologies, aerial image classification has a broad follow-up application prospect [20]. A plethora of work has been carried out from a wide range of angles in this task to achieve higher accuracy in the aerial image label predicting. Among these proposed methods, one strategy, named problem transformation, aims to convert the MAIC problem into existing, well-established learning scenarios, binary relevance classifiers [21], and the K-medoids approach [22], to name a few. Other strategies, named algorithm adaptation, leverage popular learning techniques to deal with multilabel aerial images directly, such as decision trees [23] and neural networks [24]. State-of-the-art aerial image classifiers can be integrated into MAIC problems directly by multibinary classification loss. However, they cannot achieve satisfying performance as they ignore two kinds of crucial information during the pipeline of MAIC tasks: heuristic label correlation features and representative image features.

1.2. Image Feature Extraction. Image feature extraction is the fundamental step during aerial image processing. In the past few years, deep learning-based models have shown a powerful ability to extract representative features. And MAIC methods based on deep learning have shown perspective performance. Usually pretrained on large datasets, these CNN-based models can be applied directly in MAIC tasks in an end-to-end way. Many types of research have been conducted by first extracting features with the CNN decoder and then grafting label predictors such as active learning framework [25] and GCN [26]. However, the CNN encoder is trained on single-label aerial images each of which has only one object of interest. To this end, features from CNN encoders are for the whole image and may result in a suboptimal prediction for aerial images with multiple objects because these objects exist in the image in various locations, sizes, and shapes.

To overcome the abovementioned drawbacks, a series of models were proposed to learn more fine-grained features from the raw aerial images, such as object detection methods by localizing the task-specific regions [14], handling partial occlusions by part detection [27], detecting local information via SPC layers [16], and so on.

1.3. Label Correlations Learning. Extensive methods of image classification tasks in the literature focus on exploring the correlations among labels. Probabilistic graph models are widely utilized to formulate the coexistence of labels in early research [28]. The CGL model builds a conditional label structure learning method within a unified Bayesian framework [29]. Also, label correlation was signified by a low-rank mapping matrix in [30]. A method was designed based on the graph Laplacian regularization to exploit the label correlation in the local neighborhood [31]. Suffering from computational cost, these methods are difficult to apply to reality.

Recently, with the inference ability of RNN, both semantic and spatial label relations can be extracted with only image-level supervision in a sequential way [14, 32]. Furthermore, methods based on the attention mechanism were also proposed for automatically assigning the weight of different label dependencies [33].

Lately, the newly proposed graph convolutional networks, designed for nongrid structured data modeling, have been introduced in classification tasks [26]. Different from the traditional Euclidean-structured CNN models, GCN can learn a non-Euclidean graph structure directly and thus hold the strong ability of correlation inference [9].

1.4. Motivation. The aforementioned methods coped with the two crucial challenges of the MAIC task by different kinds of well-designed frameworks. However, most of them either ignore the label correlations during extracting more representative image features or just utilize rough course-grained image features during building label dependencies in the MAIC task. In addition, the deep learning-based models utilize pretrained backbones to extract image features. However, these backbones are designed to extract a single core feature for the given image, which is suboptimal for multilabel image feature learning. Based on this observation, in this paper, we aim to propose a comprehensive model for the MAIC task that integrates multigrained semantic information from both images and labels. Inspired by the success of the SPC framework, the proposed MSGM model leverages SPC layers in a parallel manner to extract fine-grained image representations. Then, we design different grouping mechanisms to adjoin these representations, each containing several instances. Furthermore, the GCN framework is introduced to learn the semantic network features of the label correlation.

2. Methodology

2.1. Problem Definition. Given the image set X and the label set L, where $X = \{x_1, x_2, \ldots, x_n\}$ represents the n aerial images and $L = \{l_1, l_2, \ldots, l_c\}$ be c labels of this dataset. Each image $x_i \in X$ is annotated with its labels $y_i = \{y^1{}_i, y^2{}_i, \ldots, y^c{}_i\}$, where $y^k{}_i = 1$ if x_i is labelled with l_k, $k \in [1, 2, \ldots, c]$, otherwise $y^k{}_i = 0$. The primary definition of the MAIC problem is to model a function that takes the given input image x_i as the independent variable and outputs its predicting label vector y_i, i.e., $f: x_i \longrightarrow y_i$.

2.2. The Proposed MSGM Model. Figure 2 illustrates the framework of the proposed model including the image feature extractor, label correlation extractor, and multilabel classifier. The image feature extractor is constructed by multiscale SPC layers. In addition, the label correlation extractor constructs the concept graph and utilizes the attention GCN framework to directly extract the concept graph representations. Finally, the multilabel classifier integrates the bilateral information and outputs the predicted labels.

2.2.1. Image Feature Extractor. To extract more task-specific image features, we generate multiple instances of each image. For the given input image x_i, the output feature map from the backbone has a dimension of $2048 \times 14 \times 14$. Then the feature map is filtered by different sizes of kernels on each SPC layer to get a group of feature maps in a parallel manner, which is described in Subsection C. Finally, the feature space is integrated by the grouping mechanisms to get the image-level feature $\boldsymbol{f}_i \in R^m$ $(m = 2048)$.

2.2.2. Label Correlations Extractor. To represent label correlations, we first build the concept graph based on the cooccurrence of labels in the label set L. Then, the attention GCN (which is described in Subsection D) extracts the concept graph topologies and generates the label-level features $G \in R^{c \times m}$.

2.2.3. Multilabel Classifier. The image-level feature $f_i \in R^m$ and label-level features $G \in R^{c \times m}$ are integrated by the multilabel classifier as follows [9]:

$$\widehat{y}_i = \text{sigmoid}\left(G f_i^T\right). \tag{1}$$

The predicted score $\widehat{y}_i^k \in \widehat{y}_i$ is the probability of the corresponding label l_k, $\widehat{y}_i^k \in [0,1]$, with its ground truth value $y^k{}_i \in \{0,1\}$. In this way, if $\widehat{y}_i^k$ is above 0.5, we set the l_k as positive for the image x_i.

2.3. Pyramid Convolution for the Image Feature Extraction. To get more fine-grained features and generate multiple instances for each image, we treat the output from the backbone as a bag. Each 14×14 sized feature vector on dimension of 2,048 represents an instance of the input image. To this end, the bag contains these instances that represent different patches of the raw input image. However, filters with receptive fields of the same size are unlikely to capture all objects in different sizes and scales. To address this problem, we introduced the SPC component to learn the

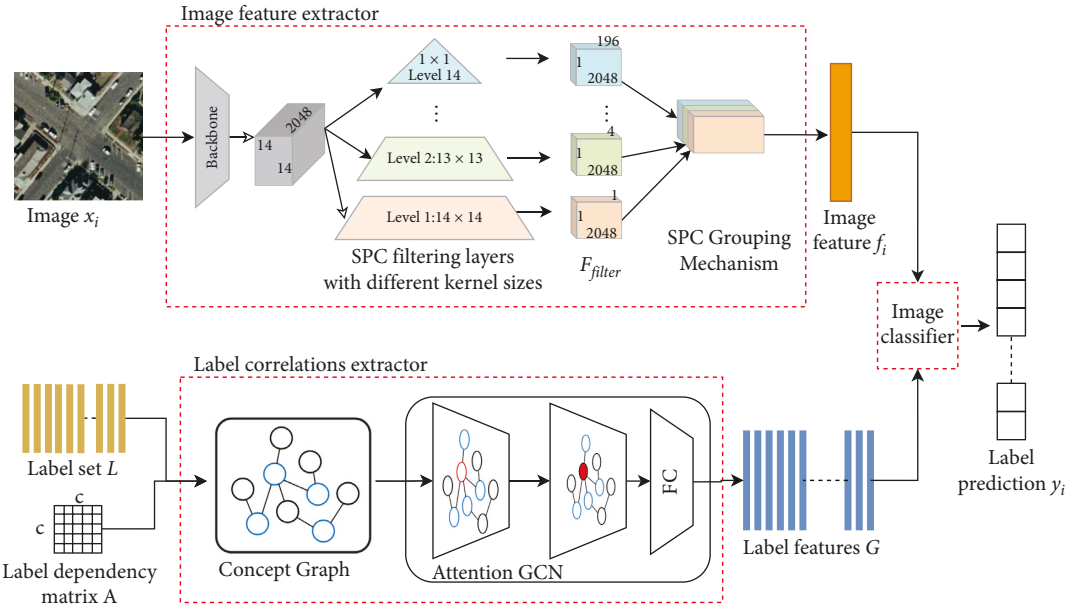


Figure 2: Flow chart of the proposed MSGM framework.

multiscale task-specific features. As is illustrated in Figure 2, SPC components consist of two components: SPC filtering layers and SPC grouping mechanism.

2.3.1. SPC Filtering Layers. SPC filtering layers are a set of parallel convolution operations whose filters are with a range of sizes from 1×1 to $w \times w$. ($w = .14$). With input $\hat{x}_i$ from the backbone, the output of i–th convolutional operation is as follows:

$$f^j_{\text{filter}} = \sigma\left(\hat{x}_i, \theta_j\right), \tag{2}$$

where θ_j indicates model parameters, and σ is the activation function.

The output of each convolution operation is a group of feature maps, denoted as $F_{\text{filter}} = \left\{f^j_{\text{filter}}\right\}^w_{j=1}$, $f^j_{\text{filter}} \in R^{(w-j+1)\times(w-j+1)\times 2{,}048}$, and each of them contains the corresponding instances. A $j \times j$ size filter generates $(w–j+1) \times (w–j+1)$ feature maps and instances. For example, a 2×2 size filter generates the corresponding feature map with the size of 13×13 and instance number of 169. Furthermore, there are 196 instances for a 1×1 size filter. The stride is a hyperparameter, which is empirically set to 1 in the experiments. Since the sizes of the filters are in various scales, the receptive field sizes are different. Each layer contains filters of different sizes and depths, which can capture different levels of features from the scenes. To this end, the SPC filtering layers can learn the features of objects with different sizes, positions, and scales.

2.3.2. SPC Grouping Mechanism. The feature maps F_{filter} generated by SPC filtering layers are in different scales and each of them can have some instances. To this end, the instances of a bag need to be aggregated into final aerial image features for the multiple label predicting. To address the abovementioned issue, we designed three grouping mechanisms at various angles, the feature alignment grouping (FAG): the feature stacking grouping (FSG), and the aligned feature stacking grouping (AFSG), respectively, as shown in Figure 3.

(a) The design of FAG aims to align the different dimensions of feature vectors. To do so, a parallel of fully connected (FC) layers are applied to the feature maps F_{filter}, as shown in Figure 3(a).

The output of the FC layers is unified-scaled feature vectors $F_{\text{FAG}} = \left\{f^j_{\text{FAG}}\right\}^w_{j=1}$, $f^j_{\text{FAG}} \in R^{2{,}048}$. Then, these feature vectors are integrated as the final image feature $\boldsymbol{f}_i$ by the average operation as follows:

$$f_i = \sum_{j=1}^{w} \varphi\left(f^j_{\text{FAG}}, W_{\text{FAG}}\right), \tag{3}$$

where W_{FAG} is the weight matrix in FC layers, $w = 14$ is the number of SPC filtering layers, and $\varphi()$ is the activation function. In this way, FAG provides a solution where instances on each feature map can be learned independently.

(b) Compared with FAG, the idea of FSG is concatenating the feature vectors with different dimensions directly. As shown in Figure 3(b), the group of feature maps F_{filter} from SPC filtering layers are first integrated into a unified feature map $F_{\text{FSG}} \in R^{1\times 2{,}048\times(1+2^2+\cdots+w^2)}$ as follows:

$$F_{\text{FSG}} = \text{conc}\left(f^j_{\text{FAG}}\right) = \left\{f^1_{\text{FAG}}, f^2_{\text{FAG}}, \ldots, f^w_{\text{FAG}}\right\}, \tag{4}$$

where $w = 14$ is the number of SPC filtering layers. Then, a FC layer and a max-pooling layer are utilized to transform the scale of F_{FSG} as follows:

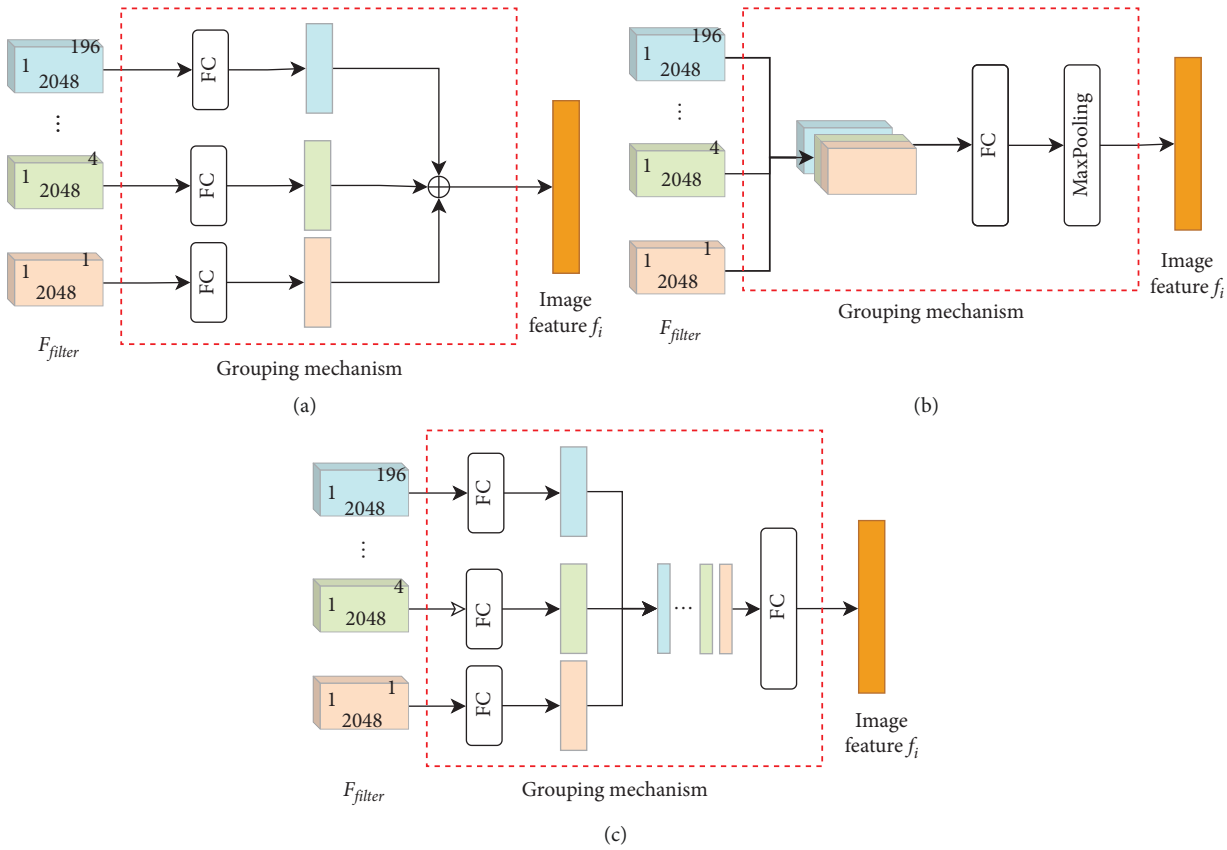


Figure 3: Frameworks of three kinds of feature grouping mechanisms. (a) FAG, (b) FSG, and (c) AFSG.

$$f_i = \varphi\left(\text{conc}\left(f_{\text{FAG}}^j\right), W_{\text{FSG}}\right), \tag{5}$$

where W_{FSG} is the weight matrix in FC layers, and $\varphi()$ is the activation function. As a result, FSG provides a direct approach by absorbing information from every feature representation and treating them as a whole map.

(c) Based on the abovementioned two approaches, AFSG deals with the feature maps in a fine-grained way, which can be the combination of FAG and FSG. As in Figure 3(c), AFSG first utilizes FC layers to get the unified-scaled feature vectors F_{FAG}, same as in FAG. Then, these feature vectors F_{FSG} are stacked into a unified feature space $F_{\text{FSG}} \in R^{14\times 2,048}$, followed by the FC layer to generate image feature $\mathbf{f}_i$ as follows:

$$f_i = \varphi\left(\text{conc}\left(\varphi\left(f_{\text{FAG}}^j, W_{\text{AFSG1}}\right)\right), W_{\text{AFSG2}}\right), \tag{6}$$

where $W_{\text{AFSG1}}, W_{\text{AFSG2}}$, are two weight matrix in the FC layers, and $\varphi()$ is the activation function.

Based on the grouping mechanisms, we get the image-level feature $f_i \in R^m$ from various perspectives. In the experimental part, we will evaluate the three grouping mechanisms, respectively. In the following part of this paper, the AFSG method in Figure 3(c) is set as the default choice without special instructions.

2.4. GCN for the Label Correlations Extraction

2.4.1. Construction of Concept Graph. To build the concept graph, we first represent all label nodes in the label space L as d-dimensional feature vectors L, where each label node $l_i \in L$ is denoted as $l_i \in R^d$. Inspired by ML_GCN [34], for the common label space L, we model the label correlation dependency in the form of conditional probability, i.e., $P(l_j|l_i)$ is the probability of occurrence of label l_j when label l_i appears:

$$P\left(l_j|l_i\right) = \frac{N_{i,j}}{N_i}, \tag{7}$$

where $N_{i,j}$ denotes the frequency of images that contain both label l_i and l_j, and N_i denotes the frequency of images that contain label l_i.

To clarify, the correlation matrix is asymmetrical, where $P(l_j|l_i)$ is not equal to $P(l_i|l_j)$. The concept graph is a direct graph. Based on the coexistence of labels in the correlation matrix, the concept graph is built to represent the label correlations.

2.4.2. Graph Convolutional Network Recapitulation. For a given graph G, the processing idea is to integrate the knowledge from other neighbor nodes to update the central node features. Different from the traditional CNN-based models, GCN aims to learn a function on the given graph G that takes nodes features $H^i \in R^{c\times d}$ and the correlation

matrix $A \in R^{c\times c}$ as inputs, and outputs the features of updated nodes $H^{i+1} \in R^{c\times d'}$. Specifically, for the i-th GCN layer, for the input feature representation $H^i \in R^{c\times d}$, the output $H^{i+1} \in R^{c\times d'}$ can be written as a nonlinear operation [34]:

$$H^{i+1} = g\left(\mathrm{A}H^i W^i\right), \tag{8}$$

where $W^i \in R^{d\times d'}$ is the weight transformation matrix. A$\in$ $A \in R^{c\times c}$ denotes the correlation matrix. The nonlinear function $g(\,)$ is usually acted by LeakReLU [35].

2.4.3. The Attention-Based GCN. From (8), GCN works by propagating information between nodes based on the correlation matrix A. However, in most previous works, A is predefined by the conditional probability and kept fixed during the node feature learning. This kind of fixed matrix is not enough for the complicated correlations of objects. We design the attention-based GCN (attGCN) layer to address this problem in the MAIC task.

The detailed introduction of the attGCN layer can be elaborated as follows: the attGCN layer is designed to learn the label node feature by integrating the typology information from its 1-step neighbors. First, a *DotProduct* attention mechanism is performed on the input of label embeddings and their 1-step neighbors embeddings [36]. Then, the label embeddings are updated by the combination of their attention-weighted neighbor node embeddings. In specific, we first represent all label nodes as d-dimensional feature embeddings. The label node $l_i \in L$ is embedded as $\mathbf{l}_i$ $\in R^d$. The neighbor node set $h_i \in H$ of l_i is denoted as $\mathbf{h}_i = \left\{\mathbf{h}_i^1, \mathbf{h}_i^2, \ldots, \mathbf{h}_i^K\right\}$, in which $\mathbf{h}_i^K \in R^d$, $k\in[1,2, \ldots, K]$, K is the number of neighbors. Then, the attention score between the label node li and its one neighbor node h_i^p is calculated in reference to [36] as follows:

$$u_{ip} = g\left(w^T\left(l_i \cdot h_i^p\right)\right), \tag{9}$$

where $\mathbf{w} \in R^d$ is the weight vector to be learned. $\mathbf{l}_i \in R^d$ is the feature embedding of the label node l_i, and $\mathbf{h}_i^p \in R^d$ is the feature embedding of the neighbor node h_i^p. The operation of $(\cdot)$ represents a dot product operation. The g is a nonlinear function (acted by LeakReLU). The *softmax* function is utilized to normalize the attention scores among different label nodes. Based on the attention scores above, the feature vector of label node l_i is then calculated as the weighted combination of $\boldsymbol{K}$ neighbor nodes [9] by

$$\widetilde{l}_i = \frac{1}{K}\sum_{k=1}^{K} u_{ik}\mathbf{h}_i^k + \mathbf{l}_i, \tag{10}$$

where $\widetilde{l}_i$ represents the updated feature vector of the label node l_i.

2.5. Learning Algorithm. As is introduced above, there are three components in the proposed MSGM framework, namely, an image feature extractor (with the backbone, SPC filtering layers, and the SPC grouping mechanism), a label correlation extractor (with the concept graph and attention-based GCN), and a multilabel classifier. The whole model is trained end-to-end. We utilize the cross-entropy loss function to train the model with the annotation y and the prediction $\widehat{y}$. The loss function *Loss* is as follows [34]:

$$\mathrm{Loss} = \frac{1}{c}\sum_{k=1}^{c}\left[y^k \log\left(\widehat{y}^k\right) + \left(1 - y^k\right)\log\left(1 - \widehat{y}^k\right)\right], \tag{11}$$

where y^k and $\widehat{y}^k$ are the k-th dimension of y and $\widehat{y}$, respectively, $k\in[1,2, \ldots, \mathrm{c}]$, and c is the size of the label set L. In addition, the backpropagation algorithm with a stochastic gradient descent mechanism is utilized to optimize the parameters.

We emphasize the preciseness of this learning procedure. During the training phase, the multilabel classifier generates the final predictions with the image feature $\mathbf{f}_i$ and the label correlation feature G.

3. Experiments and Results

We conducted two kinds of experiments, quantitative and qualitative, to evaluate the proposed MSGM method. Quantitative results are numerical scores of these metrics studied in this manuscript, i.e., scores on EP, ER, EF_1, EF_2, and CP, CR, CF_1, CF_2. Quantitative results show the performance comparison with state-of-the-art methods on the UCM and DFC15 multilabel datasets. Qualitative results are the results of case studies, feature visualization, and histogram. Qualitative results illustrate the effectiveness of the proposed MSGM method. This section will illustrate the experimental results in comparison with the state-of-the-art on UCM and DFC15 multilabel aerial image datasets, respectively. Then, the ablation studies are conducted to evaluate the key aspects of the proposed approach.

3.1. Experiment Details

3.1.1. Evaluation Metrics. Three kinds of metrics are widely used for classification models: precision, recall, and F-score. Specifically, in the multilabel classification task, the method performance can be valued on both example-based and label-based angles [9]. Here, example-based metrics demonstrate the dimension of aerial images. And the label-based metrics evaluate the performance from the perspective of labels [34]. In this way, we calculate the example-based precision (EP), recall (ER), F-scores (EF_1 and EF_2), and the label-based precision (LP), recall (LR), and F-scores (LF_1 and LF_2) as metrics in this paper.

The average label-based and example-based metrics are formalized as follows:

$$\begin{aligned}
\mathrm{EP} &= \frac{\sum_{i-1}^{M} N_i^{cor}}{\sum_{i-1}^{M} N_i^{pre}},\\
\mathrm{LP} &= \frac{1}{M}\sum_{i=1}^{M}\frac{N_i^{cor}}{N_i^{pre}},\\
\mathrm{EP} &= \frac{\sum_{i-1}^{M} N_i^{cor}}{\sum_{i-1}^{M} N_i^{\mathrm{gt}}},\\
\mathrm{LP} &= \frac{1}{M}\sum_{i=1}^{M}\frac{N_i^{cor}}{N_i^{gt}},\\
\mathrm{EF}_n &= \frac{\left(1+n^2\right)\times \mathrm{EP}\times \mathrm{ER}}{n^2\mathrm{EP}+\mathrm{ER}}, n = 1,2,3\\
\mathrm{LF}_n &= \frac{\left(1+n^2\right)\times \mathrm{LP}\times \mathrm{LR}}{n^2\mathrm{LP}+\mathrm{LR}}, n = 1,2,3,
\end{aligned} \tag{12}$$

where M is the scale of label set L. For i-th label l_i, N_i^{cor} represents the number of correctly predicted samples, N_i^{pre} represents the total number of samples predicted as positive, and N_i^{gt} denotes the number of ground truth.

3.1.2. Implementation Details. The details of the model components are as follows: the image encoder module utilizes the Resnet-101 (pretrained on ImageNet) as the backbone. Then, a set of pyramid convolutional layers is stacked in parallel for feature extraction. The label encoder module is composed of two stacked GCN layers (with output sizes of 1,024 and 2,048, respectively). The fusion layer is a dot product operation layer. In the experiment part of this paper, the AFSG method in Figure 3(c) is set as the default choice.

We select the hyperparameters of our model via grid search according to the metrics on the validation set. Specifically, we select the learning rate among {0.0005, 0.001, 0.003, 0.005} and the batch size as {8, 16, 32}. Finally, the learning rate is set to 0.001, and the batch size is 16. We utilize the SGD as the optimizer of the network and LeakyReLU as the nonlinear activation function. The network is trained for 20 epochs in total. All experiments are performed on an NVIDIA GeForce RTX GPU and implemented in *Python* using the PyTorch framework.

3.1.3. Datasets. We employ two multilabel aerial image datasets, UCM [37] and DFC15 [15] multilabel datasets. The number of aerial images and classes in each dataset are shown in Table 1. Rebuilt from the single-labeled UC Merced Land Use Dataset [38], the UCM multilabel dataset is annotated with multiple tags based on visual inspection. There are 2,100 samples in UCM and each sample has 256×256 pixels with a spatial resolution of one foot. The label space consists of *airplane, sand, pavement, buildings, cars, chaparral, court, trees, dock, tank, water, grass, mobile-home, ship, bare-soil, sea, and field*. In this work, we randomly sampled 80% of images evenly from every category for training and the remaining 20% for testing.

Table 1: Statistics of multilabel aerial image datasets.

Dataset	Image	Class	Training	Testing
UCM	2,100	17	1,680	420
DFC15	3,342	8	2674	668

The DFC15 dataset is a newly proposed multilabel dataset. It is rebuilt from the single-labeled dataset (published in the 2015 IEEE GRSS Data Fusion Contest [39]). Compared to the UCM dataset, the DFC15 dataset is more challenging with an extremely higher spectral resolution of 5 cm. There are totally eight labels in the label set, including *impervious, water, clutter, vegetation, building, tree, boat, cars*. The number of images is 3, 342. 80% of them are randomly selected as the training set and 20% for network testing.

3.2. Experimental Results. In this subsection, the experimental results of MSGM on two datasets are illustrated. To clarify, we compare MSGM with other candidates with the backbone of ResNet. In addition, some benchmark comparisons with GCN-based multilabel image classification models are conducted. Besides, we list the annotation case study results to show the effectiveness of MSGM. In addition, results of MSGM with three grouping mechanisms are compared and analyzed.

3.2.1. Results on the UCM Multilabel Dataset. We compare with current multilabel aerial image classification methods, ResNet-50 [40], ResNet-RBFNN [41], CA-ResNet-LSTM [15], CA-ResNet-BiLSTM [15], Image-GCN [42], and ML_GCN [34]. This is because they are trained based on pretrained ResNet. Table 2 lists the scores of different models on each metric that we analyze in this paper. For reading convenience, we mark the highest scores in bold. In general, MSGM achieves superior performance on both example-based and label-based metrics.

For example-based metrics, the scores of MSGM on EP and ER are 83.61% and 85.48%. MSGM surpasses EP by 5.67% over CA-ResNet-BiLSTM which is state-of-the-art. In terms of EF_1 and EF_2, MSGM achieves 84.54% and 85.10%, respectively. Although slightly lower on EF_2 than CA-ResNet-BiLSTM, our model achieves a corresponding improvement on EF_1, showing that our model can obtain high precision while maintaining the recall. Furthermore, MSGM outperforms the GCN-based model from [42] remarkably on every metric. In comparison with ML_GCN [34], MSGM shows a stronger ability on example-based metrics with increases of 3.58% on EF_1 and 3.46% on EF_2.

For label-based metrics, the proposed MSGM achieves 89.98% and 85.07% on LP and LR, which are 3.86% and 0.81% over CA-ResNet-BiLSTM. In addition, the scores of MSGM on LF_1 and LF_2 are 87.46% and 86.01%, much higher

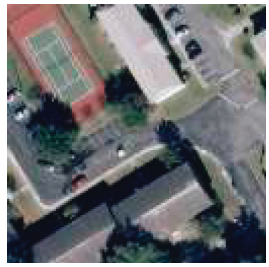
Table 2: Performance comparison with state-of-the-art methods on the UCM multilabel dataset (%).

Model	EP	ER	EF_1	EF_2	LP	LR	LF_1	LF_2
ReNet-50 [40]	80.86	81.95	81.4	81.73	88.78	78.98	83.59	80.76
ResNet-RBFNN [41]	79.92	84.59	82.19	83.61	86.21	83.72	84.95	84.21
CA-ResNet-LSTM [15]	79.9	86.14	82.90	84.82	86.99	82.24	84.55	83.15
CA-ResNet-BiLSTM [15]	77.94	**89.02**	83.11	**86.56**	86.12	84.26	85.18	84.63
Image_GCN [42]	75.00	69.00	71.86	70.12	76.00	69.00	72.33	70.29
ML_GCN [34]	79.86	82.10	80.96	81.64	86.42	80.83	83.53	81.89
MSGM	**83.86**	85.48	**84.54**	85.10	**89.98**	**85.07**	**87.46**	**86.01**

Table 3: Performance comparison with state-of-the-art methods on the DFC15 multilabel dataset (%).

Model	EP	ER	EF_1	EF_2	LP	LR	LF_1	LF_2
ReNet-50 [40]	84.89	75.64	80	77.33	81.5	59.99	69.11	63.33
ResNet-RBFNN [41]	82.64	78.76	80.65	79.51	72.01	69.85	70.91	70.27
CA-ResNet-LSTM [15]	85.66	75.84	80.45	77.62	83.83	60.05	69.97	63.66
CA-ResNet-BiLSTM [15]	91.93	79.12	85.05	81.39	**94.35**	62.35	75.08	66.89
ML_GCN [34]	93.52	**93.76**	93.64	**93.71**	91.15	90.02	90.58	90.24
MSGM	**94.61**	92.71	**93.65**	93.08	91.42	**90.70**	**91.06**	**90.84**

Table 4: Example predictions on the UCM multilabel dataset.

Images in UCM dataset				
	(a)	(b)	(c)	(d)
Ground truth	*bare-soil, trees, buildings, pavement, cars, court, grass*	*cars, bare-soil, pavement, buildings, trees, grass*	*grass, trees, water*	*buildings, pavement*
Predictions	*bare-soil, trees, buildings, pavement, cars, court, grass*	*cars, bare-soil, pavement, buildings, trees, grass*	*grass, sand, trees, water*	*bare-soil, buildings, grass, pavement*

than the second place. Specifically, MAGM improves ML_GCN by over 3.93% on LF_1 and 4.12% on LF_2.

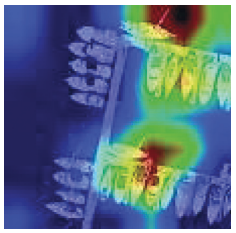
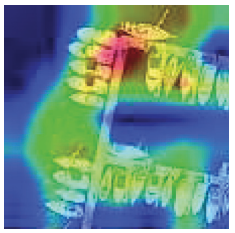
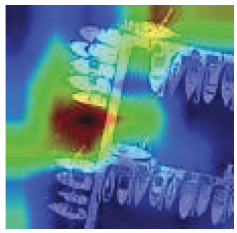
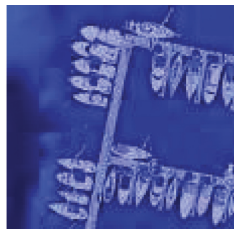
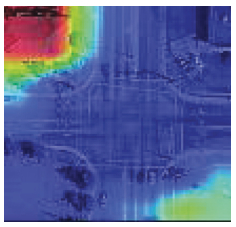
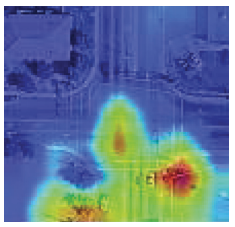
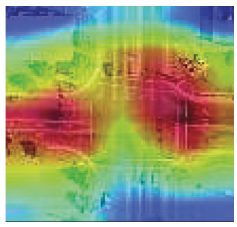
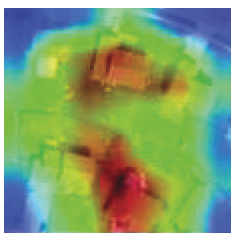
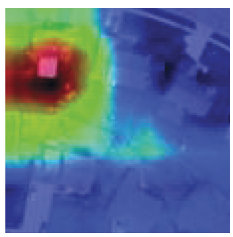
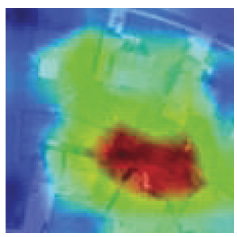
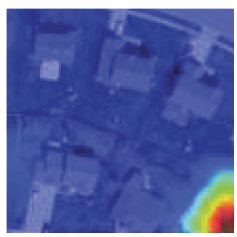
The experimental results on the UCM dataset verify the effectiveness of MSGM with the SPC layers and the concept graph. In the image feature learning phase, the SPC-based extractor can learn more fine-grained representations to help the model understand the image. During label predicting, the concept graph provides significant semantic information from label correlations.

3.2.2. Results on the DFC15 Multilabel Dataset. On theDFC15 dataset, we compare with existing MAIC methods, the ResNet-50 [40], ResNet-RBFNN [41], CA-ResNet-LSTM [15], CA-ResNet-BiLSTM [15], and ML_GCN [34]. Quantitative scores are organized in Table 3. And as above mentioned, we mark the highest scores on each metric in bold.

On example-based metrics, the proposed MSGM method obtains 94.61% and 92.71% on EP and ER and 93.65% and 93.08% on EF_1 and EF_2, respectively. Compared with CA-ResNet-BiLSTM which is the state-of-the-art on this dataset, MSGM improves EP by 2.68% and ER by 13.59%. Furthermore, MSGM increases EF_1 and EF_2 scores by 8.6% and 11.69% over CA-ResNet-BiLSTM. In addition, the proposed MSGM outperforms state-of-the-art methods not only in example-based indexes but also in label-based scores. MSGM reaches 91.42% on LP and 90.70% on LR. In comparison with CA-ResNet-BiLSTM, MSGM is 30.65% higher on LR, proving the robustness of the MSGM model. The improvements by MSGM on the DFC15 dataset further demonstrate the robustness of the proposed model on the more challenging DFC15 dataset. By extracting the fine-grained image feature and learning the self-adapted semantic information, MSGM can provide a more effective solution for the current MAIC task.

3.2.3. Annotation Case Study. To further evaluate the effectiveness of MSGM, we conducted a case study with several images. The results are listed in Table 4. We note that the proposed MSGM generally works well for images with

Table 5: Attention maps for label-specific features of several samples selected from UCM.

 (a)	 (*dock*)	 (*ship*)	 (*water*)	 (*buildings*)
 (b)	 (*buildings*)	 (*cars*)	 (*pavement*)	 (*water*)
 (c)	 (*buildings*)	 (*cars*)	 (*grass*)	 (*tree*)

Note: regions marked as red imply strongly activated, and blue indicates weakly activated.

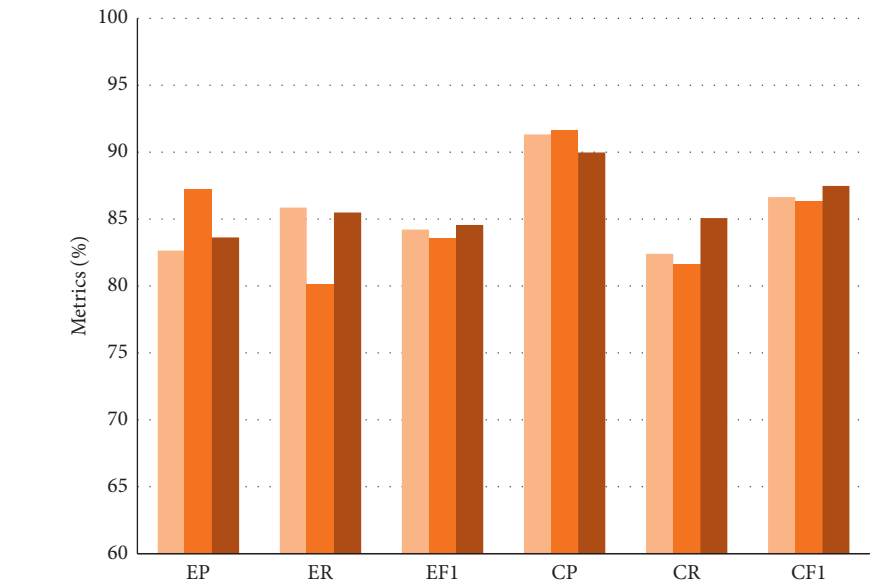


Figure 4: Results of MSGM with three grouping mechanisms.

dense labels, as shown in the example images (a) and (b) in Table 4. For instance, MSGM predicts all candidate labels for image (a) *bare-soil*, *cars*, *buildings*, *court*, *grass*, *pavement*, and *trees*.

In addition, MSGM can classify images accurately, even those with sparse labels. For image (c), the annotation includes three labels (*grass*, *trees*, and *water*). While in the prediction results, MSGM not only predicts all ground truths but also marks *sand* as positive. This provides more fine-grained semantic information of images for the follow-up computer vision tasks.

3.2.4. Fine-Grained Feature Visualization. Table 5 shows the predicted labels and the corresponding semantic feature maps of MSGM on the UCM dataset. For images (a), (b), and (c), the positive labels (ground truth) are in black and the negative labels are in red. Moreover, the activation areas of each label are concentrated on semantic-aware areas. It is intuitive that the label-correlated areas are activated. For image (a), the image patch corresponding to the label *ship* is annotated in red, while the whole image is in blue for the label *buildings*. It reveals that MSGM can learn task-specific features and explore label-region interaction.

3.2.5. Results on Three Grouping Mechanisms. As introduced previously, we designed three grouping mechanisms, FAG, FSG, and AFSG, for aerial image feature extraction. So in this part, we discuss the results on different mechanisms. For reading convenience, we name the proposed MSGM based on three modules as MSGM-FAG, MSGM-FSG, and MSGM-AFSG, respectively. The qualitative (the histogram) and quantitative (scores on EP, ER, EF_1, and CP, CR, CF_1) results are illustrated in Figure 4. It is intuitive that MSGM-AFSG surpasses the other two candidates on almost all metrics. Particularly on F-scores, MSGM-AFSG achieves 84.54% on EF_1, and 87.46% on LF_1, improving MSGM-FAG (the second place) by 0.34% and 0.83%. In addition, for results on precision, MSGM-FSG achieves the best on both EP and LP, indicating the effectiveness of concatenating the feature vectors with different dimensions directly. Another interesting observation is that all three modules outperform the existing MAIC methods, verifying the robustness and feasibility of our MSGM model.

4. Conclusions

This paper provides a new solution for fine-grained semantic understandings of aerial panoramic images. Focusing on the crucial challenges of this research task, we designed a comprehensive multilabel aerial image classification model, named MSGM. To tackle the problem of how to learn more task-specific features from aerial panoramic images, MSGM designs pyramid convolutional layers to extract multiple instances by multiscale feature encoders. And then, three grouping mechanisms are designed to integrate the instances into the final aerial panoramic image features. Furthermore, MSGM learns semantic features from label dependencies during the multilabel predicting phase. Inspired by the recently proposed GCN-based models, which can deal with graph structure directly, MSGM builds a concept graph to represent the label correlations and then feeds the graph into a designed GCN based on the attention mechanism. To this end, with the multigrained semantic features, a novel end-to-end multilabel aerial image classification considering label correlations is built. Three components constitute the whole framework of the proposed MSGM: the image feature extractor, the label correlation extractor, and the multilabel classifier. Experimental results verify the effectiveness of the proposed method both quantitatively and qualitatively on two benchmark aerial panoramic image datasets, UCM and DFC15. In the future, we will further explore the dimensions of SPC layers to provide more adaptive approaches in applications.

Data Availability

The data are available from the following: https://bigearth.eu/datasets; "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification" by Hua, Yuansheng et al.; and Isprs Journal of Photogrammetry and Remote Sensing 149 (2019): 188–199.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work is jointly supported by the National Natural Science Foundation of China (62076047), the Dalian Science and Technology Innovation Fund (2021JJ12SN44), and the Fundamental Research Funds for the Central Universities (DUT20LAB136).

References

[1] D. Lin, J. Lin, L. Zhao, Z. Jane Wang, and Z. Chen, "Multi-label aerial image classification with unsupervised domain adaptation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2021 [online], Article ID 5609613.

[2] Y. Li, Y. Chang, Y. Ye, X. Zou, S. Zhong, and L. Yan, "Category-aware aircraft landmark detection," *IEEE Signal Processing Letters*, vol. 28, pp. 61–65, 2021.

[3] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object detection with deep learning: a review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[4] Y. Su, D. Hong, Y. Li, and P. Jing, "Low-rank regularized deep collaborative matrix factorization for micro-video multi-label classification," *IEEE Signal Processing Letters*, vol. 27, pp. 740–744, 2020.

[5] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *Applied Soft Computing*, vol. 70, pp. 42–65, 2017, https://arXiv/abs/1704.06857.

[6] S. Rapinel and L. Hubert-Moy, "One-class classification of natural vegetation using remote sensing: a review," *Remote Sensing*, vol. 13, no. 10, p. 1892, 2021.

[7] M. Wang, D. Xiao, and Y. Xiang, "Low-cost and confidentiality-preserving multi-image compressed acquisition and separate reconstruction for internet of multimedia things," *IEEE Internet of Things Journal*, vol. 8, no. 3, pp. 1662–1673, 2021.

[8] S. Lucchesi, M. Giardino, and L. Perotti, "Applications of high-resolution images and DTMs for detailed geomorphological analysis of mountain and plain areas of NW Italy," *European Journal of Remote Sensing*, vol. 46, no. 1, pp. 216–233, 2013.

[9] D. Lin, J. Lin, L. Zhao, Z. J. Wang, and Z. Chen, "Multi-label aerial image classification with a concept attention graph neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–12, 2021.

[10] J. Gao, Z. Hu, K. Bian, X. Mao, and L. Song, "Aq360: uav-aided air quality monitoring by 360-degree aerial panoramic images in urban areas," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 428–442, 2021.

[11] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond rgb: very high resolution urban remote sensing with multimodal deep networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, pp. 20–32, 2018.

[12] A. Hanyu, Y. Kawamoto, and N. Kato, "Adaptive channel selection and transmission timing control for simultaneous receiving and sending in relay-based UAV network," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 4, pp. 2840–2849, 2020.

[13] X. Wang, X. Wang, and S. Mao, "Deep convolutional neural networks for indoor localization with csi images," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 316–327, 2020.

[14] J. Zhang, Q. Wu, C. Shen, J. Zhang, and J. Lu, "Multilabel image classification with regional latent semantic dependencies," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2801–2813, 2018.

[15] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional lstm network for multi-label aerial image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 149, pp. 188–199, 2019.

[16] W.-J. Yu, Z.-D. Chen, X. Luo, W. Liu, and X.-S. Xu, "Delta: a deep dual-stream network for multi-label image classification," *Pattern Recognition*, vol. 91, pp. 322–331, 2019.

[17] D. Parashar and D. K. Agrawal, "Automatic classification of glaucoma stages using two-dimensional tensor empirical wavelet transform," *IEEE Signal Processing Letters*, vol. 28, pp. 66–70, 2021.

[18] L. Mou and X. X. Zhu, "Im2height: height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," 2018, http://arXiv.org/abs/:1802.10249.

[19] B. Mei, Y. Xiao, R. Li, H. Li, X. Cheng, and Y. Sun, "Image and attribute based convolutional neural network inference attacks in social networks," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 869–879, 2020.

[20] I. Shendryk, Y. Rist, R. Lucas, P. Thorburn, and C. Ticehurst, "Deep learning-a new approach for multi-label scene classification in planet scope and sentinel-2 imagery," in *Proceedings of the . IEEE International. Geoscienceand Remote Sensing. Symposium. (IGARSS)*, pp. 1116–1119, IEEE, Valencia, Spain, July 2018.

[21] A. Melo and H. Paulheim, "Local and global feature selection for multilabel classification with binary relevance," *Artificial Intelligence Review*, vol. 51, no. 1, pp. 33–60, 2019.

[22] X. Wang, X. Xiong, and C. Ning, "Multi-label remote sensing scene classification using multi-bag integration," *IEEE Access*, vol. 7, Article ID 120410, 2019.

[23] M. Majzoubi and A. Choromanska, "Ldsm: logarithm-depth streaming multi-label decision trees," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 4247–4257, August 2020.

[24] H. Cevikalp, B. Benligiray, and O. N. Gerek, "Semi-supervised robust deep neural networks for multi-label image classification," *Pattern Recognition*, vol. 100, Article ID 107164, 2020.

[25] Y. Yan and S.-J. Huang, "Cost-effective active learning for hierarchical multi-label classification," in *Proceedings of the 27th International Joint Conference on Artificial IntelligenceJuly 2018 IJCAI*, pp. 2962–2968, AAAI Press, Stockholm, Sweden, July 2018.

[26] Y. Wang, D. He, F. Li et al., "Multi-label classification with label graph super imposing," 2019, http://arXiv.org/abs/:1911.09243.

[27] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3486–3495, IEEE, Venice, Italy, October 2017.

[28] I. B. Rejeb, S. Ouni, W. Barhoumi, and E. Zagrouba, "Fuzzy va-files for multi-label image annotation based on visual content of regions," *Signal, Image and Video Processing*, vol. 12, no. 5, pp. 877–884, 2018.

[29] Q. Li, M. Qiao, W. Bian, and D. Tao, "Conditional graphical lasso for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2977–2986, IEEE, Las Vegas, NV, USA, June 2016.

[30] J. Wu, A. Guo, V. S. Sheng, P. Zhao, and Z. Cui, "An active learning approach for multi-label image classification with sample noise," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 32, no. 03, Article ID 1850005, 2018.

[31] X. Yang, Y. Zhou, Q. Zhu, and Z. Wu, "Joint graph regularized extreme learning machine for multi-label image classification," *Journal of Computational Methods in Science and Engineering*, vol. 18, no. 1, pp. 213–219, 2018.

[32] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5513–5522, HI, USA, July 2017.

[33] G. Sumbul and B. Demir, "A novel multi-attention driven system for multi-label remote sensing image classification," in *Proceedings of the IEEE International. Geoscience. Remote Sensing. Symposium. (IGARSS)*, pp. 5726–5729, Yokohama, Japan, July 2019.

[34] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the. IEEE Conference. Computer. Vis. Pattern Recognition. (CVPR)*, pp. 5177–5186, Long Beach, CA, USA, June2019.

[35] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the 30th International Conference on International Conference on Machine Learning(ICML)*, p. 3, Atlanta GA USA, June 2013.

[36] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Proceedings of theAdvances in Neural Inf. Process. Syst*, pp. 5998–6008, CA, USA, December 2017.

[37] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multi-label remote sensing image retrieval using a semi-

supervised graph-theoretic method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1144–1158, 2017.

[38] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International. Conference on Advance Geographic Information System*, pp. 270–279, California, San Jose, November 2010.

[39] M. Campos-Taberner, A. Romero-Soriano, C. Gatta et al., "Processing of extremely high-resolution LiDAR and RGB data: outcome of the 2015 IEEE GRSS data fusion contest-Part A: 2-D contest," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 12, pp. 5547–5559, 2016.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, pp. 770–778, HI, USA, July 2016.

[41] A. Zeggada, F. Melgani, and Y. Bazi, "A deep learning approach to UAV image multilabeling," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 694–698, 2017.

[42] N. Khan, U. Chaudhuri, B. Banerjee, and S. Chaudhuri, "Graph convolutional network for multi-label vhr remote sensing scene recognition," *Neurocomputing*, vol. 357, pp. 36–46, 2019.