Transformer-Driven Semantic Relation Inference for Multilabel Classification of High-Resolution Remote Sensing Images

Xiaowei Tan, Zhifeng Xiao, *Member, IEEE*, Jianjun Zhu, Qiao Wan, Kai Wang, and Deren Li, *Member, IEEE*

***Abstract*—It is hard to use a single label to describe an image for the complexity of remote sensing scenes. Thus, it is a more general and practical choice to use multilabel image classification for high-resolution remote sensing (HRS) images. How to construct the relation between categories is a vital problem for multilabel classification. Some researchers use the recurrent neural network (RNN) or long short-term memory (LSTM) to exploit label relations over the last years. However, the RNN or LSTM could model such category dependence in a chain propagation manner. The performance of the RNN/LSTM might be questioned when a specific category is improperly inferred. To address this, we propose a novel HRS image multilabel classification network, transformer-driven semantic relation inference network. The network comprises two modules: semantic sensitive module (SSM) and semantic relation-building module (SRBM). The SSM locates the semantic attentional regions in the features extracted by a deep convolutional neural network and generates a discriminative content-aware category representation (CACR). The SRBM uses label relation inference from outputs of the SSM to predict final results. The characteristic of the proposed method is that it can extract semantic attentional regions relevant to the category and generate a discriminative CACR and natural and interpretable reasoning about label relations. Experiments were performed on the public UCM multilabel and MLRSNet datasets. Quantitative and qualitative analyses on state-of-the-art multilabel benchmarks proved that the proposed method could effectively locate semantic regions and build relationships between categories with better robustness.**

***Index Terms*—Deep convolutional neural network (DCNN), label dependence, multilabel scene classification, remote sensing, semantic relation learning.**

I. Introduction

OWING to the complexity of remote sensing scenes, using a single label to describe an image for the complexity of remote sensing scenes is hard. Thus, multilabel image classification for high-resolution remote sensing (HRS) images is more general and practical than single-label image classification. HRS scenes comprise various categories with correlations and differences between them. For example, for correlations between categories, as shown in Fig. 1(a) and (b), “road” and “car” often appear simultaneously in the remote sensing image, “grass” and “water” accompany “golf course,” and “airport” usually contains “aircraft.” For differences between categories, as shown in Fig. 1(c) and (d), they both have two categories of road and water; however, the relationship between road and water is different, and the road shapes in the two pictures differ too, so their categories are different. In picture (c), both roads are close to the water body, and one of them is in a spiral and overlapping position, so its category is an overpass. In picture (d), the road is above the water body, so this type of road is a bridge. From Fig. 1, there are not only semantic associations but also spatial location associations between categories. Even if they have the same feature type, their spatial relationships are different, and they may belong to different categories. How to construct the relationship between categories is a complex problem for multilabel image classification.

With the rapid progress of artificial intelligence and machine learning (ML), many deep convolutional neural networks (DCNNs) have been proposed to extract high-level semantic features, e.g., VGG-Net [1], GoogLeNet [2], ResNet [3], and DenseNet [4]. These networks have been successfully used in many computer vision tasks such as image classification [5], object detection [6], semantic segmentation [7], and video tracking [8], and achieved satisfactory performances. Nevertheless, most applications are for single-label remote sensing image classification [9]–[13]. Although these networks achieve some improvements for remote sensing image classification, they only consider the case that each image contains only one label; they do not consider that an image may be associated with multiple semantic labels.

In ML and other fields, the multilabel issue has attracted much attention. Many researchers have shown that it is helpful to use label correlation to classify multilabel images [14]–[19]. However, how to efficiently make use of label correlations is still an open issue. Most scholars use recurrent neural network (RNN) or long short-term memory (LSTM), to find the dependence between categories. However, due to its chain propagation fashion, RNN/LSTM’s performance heavily depends on its long-term memorization learning effectiveness. In addition, the categories relationship is implicitly modeled in this way, resulting in a

Manuscript received August 22, 2021; revised October 25, 2021 and December 10, 2021; accepted January 12, 2022. Date of publication January 25, 2022; date of current version February 28, 2022. This work was supported in part by the Science and Technology Project of State Grid Corporation of China under Grant 52789921001M. *(Corresponding author: Zhifeng Xiao.)*

Xiaowei Tan, Zhifeng Xiao, Qiao Wan, and Deren Li are with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: cug_txw@163.com; xzf@whu.edu.cn; wanqiao@whu.edu.cn; drli@whu.edu.cn).

Jianjun Zhu is with the National Bio Energy Company Ltd., Beijing 100052, China (e-mail: zhujianjun@sgecs.sgcc.com.cn).

Kai Wang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: whu-wk@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3145042

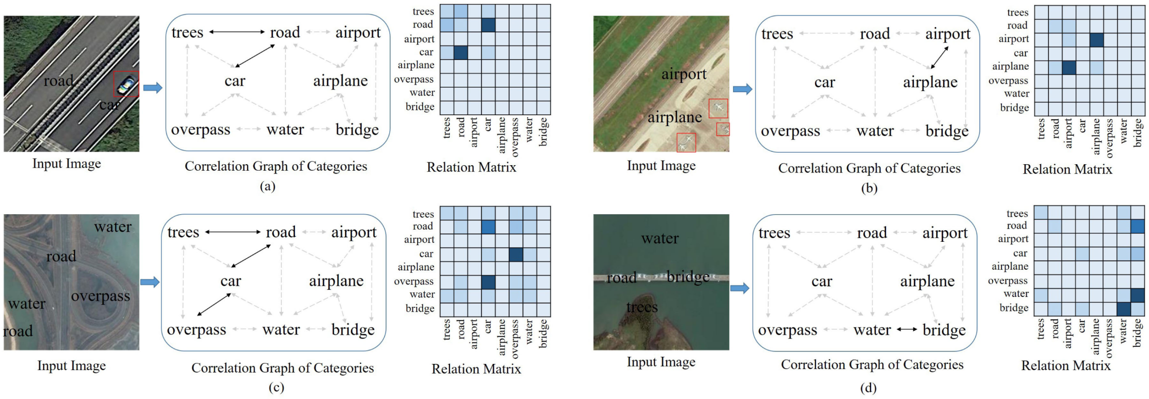


Fig. 1. Examples for semantic relation of categories in remote sensing image scenes. For an input image, we can get a correlation graph between categories. The solid line indicates higher relation of the categories, and the dashed line indicates lower relation of the categories. We can obtain a relation matrix of categories from their correlation. The darker color in the matrix indicates the stronger correlation of the categories, and the lighter color indicates the weaker correlation of the categories.

lack of interpretability. Moreover, the RNN only considers the correlation between adjacent labels, while nonadjacent labels are ignored.

The transformer solves the problems of the RNN and LSTM. Transformer processes a sentence as a whole and does not rely on the hidden state of the past to capture the dependence on previous words, so there is no risk of losing (or "forgetting") past information. In addition, both multihead attention and position embedding provide information about the relationship between different words. Some scholars have currently tried to apply the transformer in image classification tasks [20], [21]. One of the main problems in applying the transformer to image classification is how to convert the image into the sequence fashion in a semantic manner. In [20], the author proposes iGPT and directly converts the image into a sequence as input. Notably, the general input for image classification is $384 \times 384 \times tex3$ or $224 \times 224 \times 3$, which is too large for iGPT to directly reshape into sequence length, so the image is reshaped to $32 \times 32 \times 3$, $48 \times 48 \times 3$, or $64 \times 64 \times 3$. In [21], images have been divided into 16×16 image patches, and then, reshaped into a sequence. The aforementioned methods are not good enough to process pictures and cannot model the local and semantic information. Some objects are relatively small in remote sensing scenes and are ignored when reducing the image resolution, which affects the final classification result. In addition, it is more suitable to understand images from the global scene for multilabel remote sensing image classification. Cutting an image into several small patches is unconducive to understanding the image as a whole, let alone constructing the relationship between categories. Moreover, the transformer does not have a local receptive field. How to transform the image into a sequence in a semantic manner and make the transformer have a local receptive field remains an open issue.

To address the aforementioned problems, we propose a novel end-to-end transformer-driven semantic relation inference network (SR-Net) for the multilabel classification of HRS images. Its characteristic is locating the category-specific semantic regions without bounding box and segmentation and modeling the semantic category relation autonomously for the task. The network comprises two modules: semantic sensitive module (SSM) and semantic relation-building module (SRBM). We used the DCNN as the feature extractor to extract high-level semantic features. Then, used the SSM to locate the category-specific semantic region and obtain a discriminative content-aware category representation (CACR). Finally, the SRBM uses label relation inference from outputs of the SSM to predict final results. The main contributions of our work can be summarized as follows.

1) The significant contribution of this article is that we introduce a novel transformer-driven relation reasoning from CACRs for multilabel HRS image classification. The transformer-driven relation reasoning is competent to model category relations for a specific HRS image in an adaptive way, further enhancing its representative and discriminative ability. For all we know, this is the first method of using the transformer to build the relationship between categories for multilabel classification of HRS images.
2) The features extracted by the DCNN contain rich contexts and semantic information. We design the SSM to further utilize this information to locate the semantic attentional regions in the feature and generate a discriminative CACR. It makes up for the lack of local receptive field in the transformer while enhancing features' representative and discriminative ability.
3) We propose a novel end-to-end HRS image multilabel classification network, transformer-driven SR-Net. The network comprises two modules: sSSM and SRBM). We design the SSM to locate semantic attentional regions from the features extracted by the DCNN without bounding box and segmentation and generate a discriminative CACR. We design the SRBM to use label relation inference from outputs of the SSM, that is, CACR, to predict final results.

The rest of this article is organized as follows. Section II introduces related multilabel classification and transformer methods. Section III introduces an overview of the proposed method. Section IV presents the employed dataset, model training parameters, evaluation indicators, experimental results, and analyses. Finally, Section V concludes this article.

II. Related Work

A. Traditional ML Methods for Multilabel Classification

The traditional ML methods for solving the multilabel classification problem include mainly two solutions: problem transformation and algorithm adaptation.

The main idea of the problem transformation method is converting a multilabel dataset into a single-label dataset in some way and proceeding with the single-label classification method. This method can include transforming to binary classification [22]–[24], transforming to label ranking [25], and transforming to multiclass classification [26].

The adaptive algorithm is based on the particularity of multilabel classification, improving the existing single-label classification algorithm, mainly including multilabel k-nearest neighbor (ML-KNN) [27], ranking support vector machine (Rank-SVM) [28], multilabel decision tree (ML-DT) [29], and collective multilabel classifier (CML) [30].

B. Region-Based Methods for Multilabel Classification

Region-based methods aim to first roughly locate multiple regions, and then, use the DCNN to recognize each region. Wei *et al.* [15] propose a hypotheses–convolutional-neural-network (CNN)–pooling (HCP) framework that generates many proposals through an object detection method [31], [32] and considers each proposal as a single-label image classification issue. Yang *et al.* [33] regarded the task as a multiclass and multi-instance learning problem. Specifically, they generated an instance package for each image and used label information to enhance discriminant features to incorporate local information. These methods are complex and computationally expensive because they lead to many categories of unknown regions. In addition, these methods generally ignore label dependence and regional relations, which are crucial for the multilabel image classification.

C. Relation-Based Methods for Multilabel Classification

Relation-based methods are designed to exploit dependencies or region relationships between objects [34]–[39]. Wang *et al.* [34] proposed the CNN–RNN framework, which used RNN to predict final scores and formulate label relations. Wang *et al.* [35] employed LSTM and a spatial transformer to locate the attention area iteratively to find the relation. Hua *et al.* [40] used the class attention learning layer to capture discriminative class-specific features, and used the bidirectional LSTM-based subnetwork to classify class dependence in both directions and produce structured multiple object labels. Sumbul *et al.* [41] used the DCNN to develop the spatial and spectral features of local areas in images, used LSTM to characterize the importance scores of different local areas in each image, and then, defined a global descriptor for each image based on those scores. Ji *et al.* [42] introduced the attention module to separate the features extracted from the DCNN by channel, and then, sent the separated features to the LSTM network for a prediction label. These relation-based methods explore the relationship between categories or semantic regions using the RNN or LSTM; they cannot fully explore the direct relationship between categories or semantic regions. Different from those methods, some researchers have tried to solve this problem through graphical architectures. Before the age of neural networks, people used to use graphical models [36], [43]–[45]. Guo *et al.* [43] used the circular directed graph model to construct the dependence relationship between labels. Micusik *et al.* [44] constructed the Markov random field on super superpixels. Li *et al.* [36] handle such relations by image-dependent conditional label structures with a graphical Lasso framework. Li *et al.* [46] use a maximum spanning tree algorithm to create a tree structure graph in the label space. Recently, with the concept of graph convolution and the outstanding performance of the graph convolution network (GCN) in multiple visual tasks, people used GCNs to model the correlation between nodes [37], [38], [47]–[49]. Particularly, Chen *et al.* [37] use the GCN to propagate prior label representations (e.g., word embeddings) and generate a classifier that replaces the last linear layer in a standard deep convolutional neural network such as ResNet [3]. Chen *et al.* [38] compute a probabilistic matrix as the relation edge between each label in a graph with the help of label annotations. Khan *et al.* [50] proposed a new multilabel deep GCN by modeling the subsequent supervised learning problem. Li *et al.* [49] extracted image features through the DCNN and inferred spatiotopological relationships between images and features using the graph neural network (GNN).

D. Transformer-Based Methods

Concerning transformers, the attention mechanism is paramount. The attention mechanism was first proposed in the visual field. Mnih *et al.* [51] incorporated the attention mechanism into the RNN to classify images, thereby making the attention mechanism popular. Bahdanau *et al.* [52] applied the attention mechanism to the natural language processing (NLP) field, using the Seq2Seq and attention model for machine translation, thereby improving the performance. Vaswani *et al.* [53] proposed the transformer structure, entirely abandoning network structures such as RNN and CNN. They only used the attention mechanism for machine translation tasks and achieved good results. The transformer is a deep learning model entirely based on the self-attention mechanism. Self-attention was first applied to machine translation, with a focus on the contextual relationship between words. Self-attention uses the attention mechanism to calculate the relationship between each word and all other words. In other words, the encoder reads the input data, using the self-attention mechanism of superimposed layers to obtain a new representation of each word that considers the context information. In the past two years, several scholars have applied transformer to visual tasks, such as target detection [54], image classification [20], [21], and semantic segmentation [55]. Specifically, in the field of remote sensing scene classification, Deng *et al.* [56] proposed a joint framework of the CNN and transformer, which has two branches (CNN and Transformer) and combines the features extracted from the two branches to make predictions.

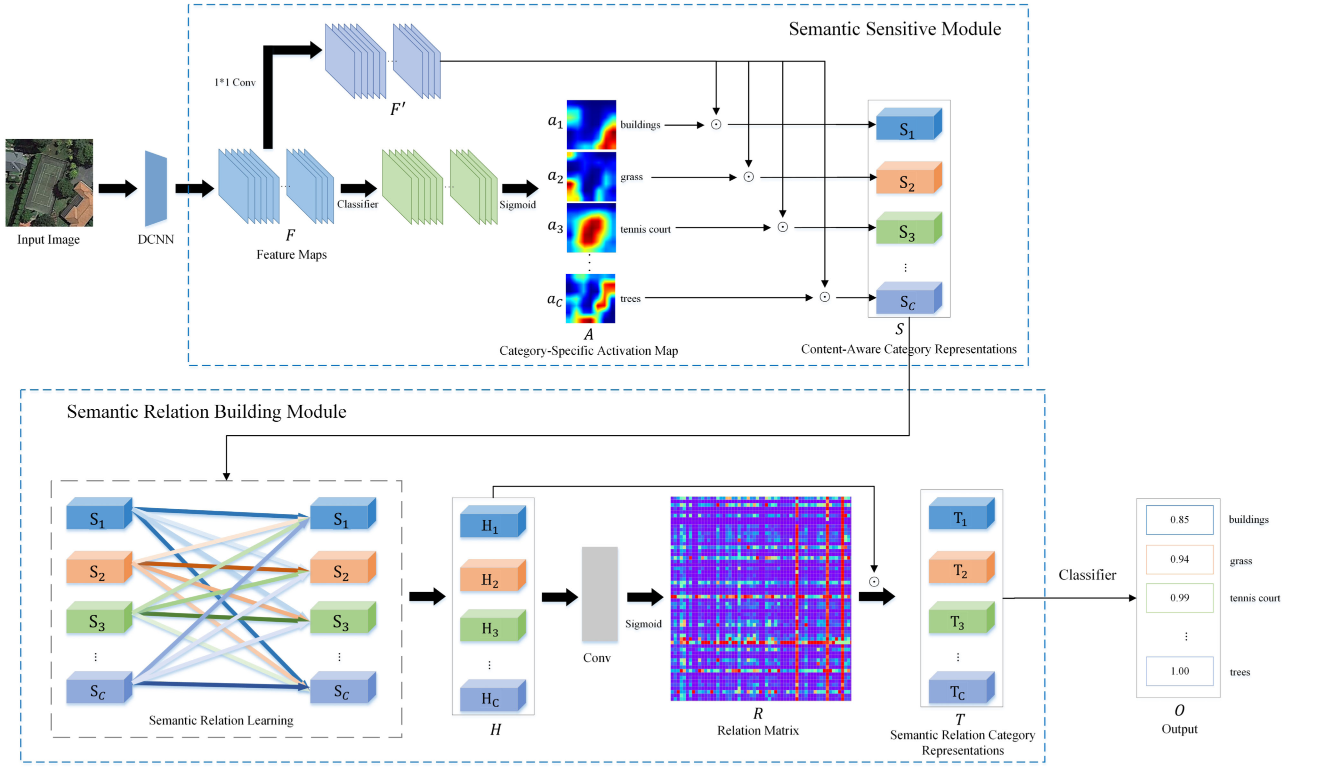


Fig. 2. Overall framework of our approach.

III. Methodology

Our method focuses on the semantic attentional regions location and semantic relationship modeling between categories. It comprises two modules: SSM and SRBM, as shown in Fig. 2.

High-level semantic features extraction is vital for visual classification tasks. Many recent studies adopt DCNNs for visual classification tasks owing to their remarkable performance in learning such features. Hence, we take a standard DCNN, such as VGG, ResNet, and DenseNet, as the backbone of our network. In Fig. 2, given an image, we use a DCNN as the feature extractor to obtain the feature map F, which contains rich semantic and location information. We designed the SSM to locate the semantic attentional regions in F and obtain the category-specific activation map (CSAM), thereby obtaining the CACR. We use the SRBM to ratiocinate the relation matrix between categories to generate the final robust semantic relationship category representation (SRCR), which contains rich relationship information with other categories. Finally, we use a binary classifier to classify SRCR and obtain the final result.

This section introduces our proposed network framework. The implementation details of the SSM and SRBM will be described in detail in Sections III-A and III-B, respectively. The classification and loss function will be introduced in Section III-C.

A. Semantic Sensitive Module (SSM)

Semantic relations are essential for multilabel classification, but obtaining the semantic information of DCNN extracted features is complex. Although the features extracted by the DCNN can be directly applied to explore label dependencies, some regions of features that are less relevant to the category (such as the blue area in Fig. 3) may bring noise and further reduce the effectiveness of feature representations. Fig. 3 is a visual example of images of CSAMs. Fig. 3(a) and (c) are the original images, and Fig. 3(b) and (d) are the activation maps of

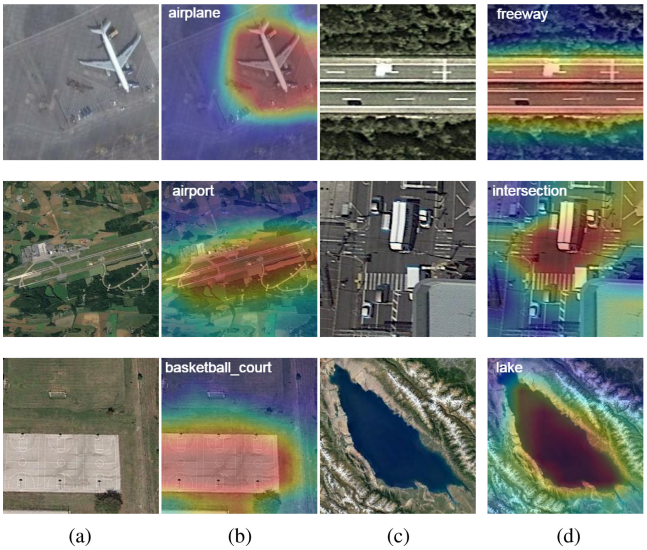


Fig. 3. Example images of CSAMs. (a) and (c) Original images, and (b) and (d) activation maps (AM) for a certain category. (a) Images. (b) AM. (c) Images. (d) AM.

a specific category. In Fig. 3, the weakly activated area indicates that the correlation with the corresponding category is weak, and the highlighted area indicates that the correlation between the region and the category is vital. In order to strengthen the feature representations ability related to the category, we designed the SSM to locate the regions related to the category of the features and enhance the semantic feature representation ability.

Each convolution in the DCNN plays the role of an object detector, and it can locate objects. However, this ability is lost when using a fully connected (FC) layer for classification. We can avoid using an FC layer and instead use a global average pooling (GAP) to establish the relationship between the feature map and the categories [57]. We generate a CSAM A based on class activation mapping (CAM). The CAM technology is to extract implicit attention region on the image in a proposal-free fashion. Specifically, we can perform GAP or global max pooling (GMP) on the feature map F, and use the FC classifier to classify these pooled features, and then, perform the process by combining the weight of the FC classifier with the feature map F convolution, using these classifiers to identify the CSAM. Owing to GMP only recognizes one discriminative part, much information will be lost, and it is disadvantageous for multilabel region extraction. Thus, we use a classifier before GMP, which will make up for GMP's disadvantages. Moreover, unlike the FC classifier, we use a convolution layer and sigmoid activation function as a classifier.

We use $A = [a_1, a_2, \ldots, a_C] \in \mathbb{R}^{H\times W\times C}$ to represent the CSAM, where C is the number of categories. a_C can be calculated using the following formula.

$$a_C = \text{Sigmoid}(\text{Conv}(f_C)) \tag{1}$$

where Sigmoid$(\cdot)$ is sigmoid activation function; Conv$(\cdot)$ is a convolution with a 1*1 convolution kernel; and f_C is the cth feature vector of the feature map F.

Then, multiply A and $F' \in \mathbb{R}^{H\times W\times D'}$ to obtain the CACR, F' is obtained by reshaping the feature map F after convolution, and we use $1 * 1$ convolution kernel for convolution. We use $S = [s_1, s_2, \ldots, s_C] \in \mathbb{R}^{C\times D}$ for the CACR. Specifically, each class representation s_C as a weighted sum on F' so that the generated s_C can selectively highlight features related to its specific category c. s_C can be calculated using the following formula.

$$s_C = a_C^T F' = \sum_{i=1}^{H}\sum_{j=1}^{W} a_{i,j}^C f'_{i,j} \tag{2}$$

where $a^C_{i,j}$ is the weight of cth activation map at (i,j); and $f'_{i,j} \in \mathbb{R}^{D'}$ is the feature vector of the feature map F' at (i, j).

B. Semantic Relation Building Module (SRBM)

In the SSM, we obtain the CACR, which was represented by S. As mentioned in the previous section, each dimension in S contains the feature code of its corresponding category, and the code contains weighted location and semantic information. Using the information to construct the relationship between categories is a problem that the SRBM needs to solve. We introduced the transformer encoder to model the relationship between the categories further and get the relation matrix.

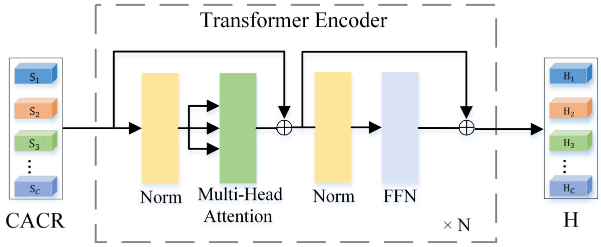


Fig. 4. Semantic relation learning. In the transformer encoder, MSA is followed by an FFN, this FFN contains two FC layers, and the nonlinear activation function in the middle uses GeLU. Both MSA and FFN contain the same skip connection as ResNet, and both MSA and FFN contain the layer norm layer.

One of the main problems in applying the transformer to image classification is how to transform the image into a sequence in a semantic manner and make the transformer have a local receptive field. In this work, we use the SSM's output, CACR, as the transformer encoder's input. The CACR is a C-dimensional sequence (C is the number of categories analogous to the sentence length) whose each dimension represents the semantically activated feature encoding of its corresponding category. The length of each dimension is X, and X is the feature code length of each category analogous to the word code length. Therefore, we can use the self-attention mechanism to explore the relationship between categories.

We use the transformer encoder architecture to learn relations in the CACR. The framework is shown in Fig. 4. In the transformer encoder, multihead self-attention (MSA) is followed by a feed-forward network (FFN), containing two FC layers, and the nonlinear activation function in the middle uses GeLU. Both MSA and FFN contain the same skip connection and number of layer as ResNet, and norm layer, respectively.

MSA is to define h attention heads, i.e., h self-attention is applied to the input sequence; the sequence can be split into h sequences Xss of size $N \times d$, where $D = hd$. Then, the outputs obtained from h different heads are concatenated. For the ith h sequence, the attention head will learn three weight matrices, namely W_i^Q, W_i^K, and W_i^V, and through (5), three vectors Q'_i, K'_i, and V'_i can be obtained. The formulas are as follows:

$$\text{MSA}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{3}$$

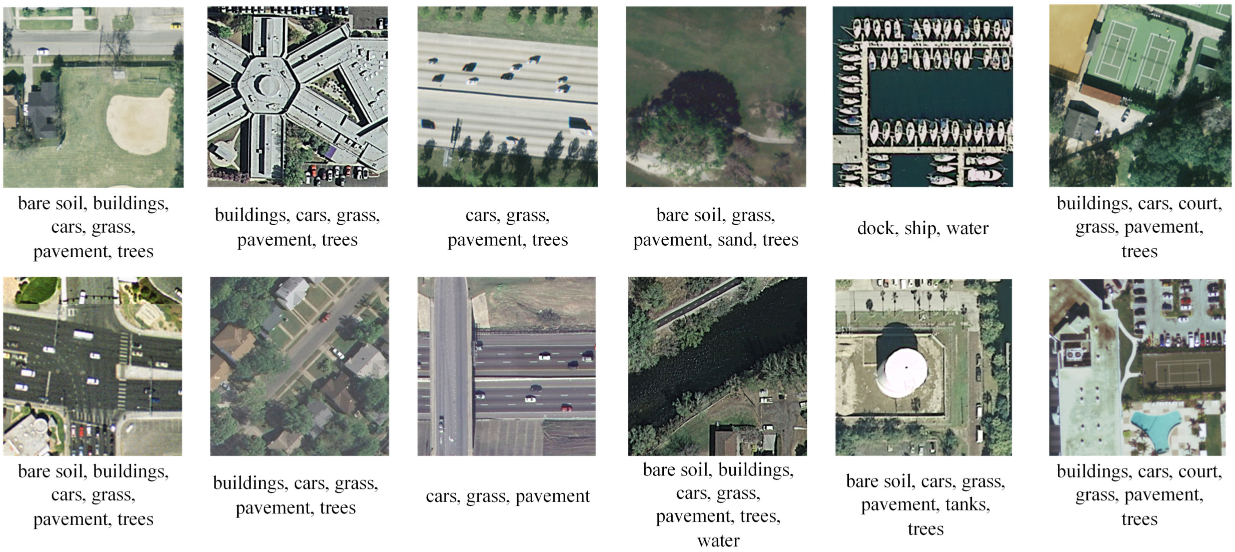
$$\text{head}_i = \text{Attention}(Q'_i, K'_i, V'_i) \tag{4}$$

$$Q'_i = XW_i^Q, K'_i = XW_i^K, V'_i = XW_i^V \tag{5}$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where $\sqrt{d_k}$ is the scaling factor to avoid the variance effect caused by the dot product.

As shown in Figs. 2 and 4, the SRBM takes the $S \in \mathbb{R}^{C\times D}$ as input node features and sequentially feeds them into the transformer encoder. Specifically, the transformer encoder's output is defined as H, where $H = [h_1, h_2, \ldots, h_C] \in \mathbb{R}^{C\times D'}$. Next, we introduce a relation matrix $R \in \mathbb{R}^{C\times C}$, adaptively estimated



bare soil, buildings, cars, grass, pavement, trees

buildings, cars, grass, pavement, trees

cars, grass, pavement, trees

bare soil, grass, pavement, sand, trees

dock, ship, water

buildings, cars, court, grass, pavement, trees

bare soil, buildings, cars, grass, pavement, trees

buildings, cars, grass, pavement, trees

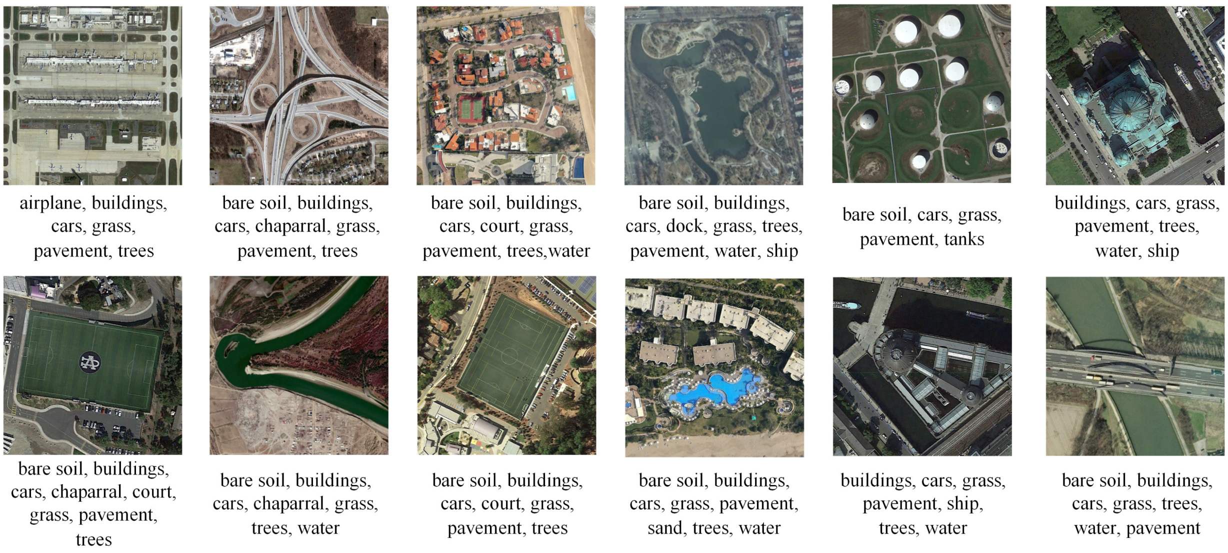
cars, grass, pavement

bare soil, buildings, cars, grass, pavement, trees, water

bare soil, cars, grass, pavement, tanks, trees

buildings, cars, court, grass, pavement, trees

Fig. 5. Example images of the UCM dataset are shown, with the corresponding categories.



airplane, buildings, cars, grass, pavement, trees

bare soil, buildings, cars, chaparral, grass, pavement, trees

bare soil, buildings, cars, court, grass, pavement, trees,water

bare soil, buildings, cars, dock, grass, trees, pavement, water, ship

bare soil, cars, grass, pavement, tanks

buildings, cars, grass, pavement, trees, water, ship

bare soil, buildings, cars, chaparral, court, grass, pavement, trees

bare soil, buildings, cars, chaparral, grass, trees, water

bare soil, buildings, cars, court, grass, pavement, trees

bare soil, buildings, cars, grass, pavement, sand, trees, water

buildings, cars, grass, pavement, ship, trees, water

bare soil, buildings, cars, grass, trees, water, pavement

Fig. 6. Example images of the AID dataset are shown, with the corresponding categories.

from the feature H. Since the category in each picture has a different R, the model is highly representative. Formally, the output $T \in \mathbb{R}^{C\times D''}$ of the SRBM can be defined as follows:

$$T = f(RH) \tag{7}$$

where $R = S$ (Conv(H)); $f(\cdot)$ LeakyReLU activation function; $S(\cdot)$ sigmoid activation function.

Notably, the relation matrix R is specific for each image, which may capture content-dependent category dependencies. Overall, the module can autonomously reason about the category relationship and does not require specific prior knowledge about the relationship between all categories. All relationships are automatically learned in a data-driven way and proved to be realistic in our experiments.

C. *Classification and Loss Function*

As shown in Fig. 2, the final SRCR is represented by T, $T = [t_1; t_2; \ldots; t_C]$ is used for the final classification. Since each vector is aligned with its specific class and contains rich relationship information with other vectors, we put each class vector into a binary classifier to predict its class score. We supervise the final score $O = [O^1; O^2; \ldots; O^C]$ and use the multilabel classification loss function to train the entire network. The loss function comprises two parts: sigmoid function and binary cross-entropy loss (BCEloss).

In multilabel image classification, we assume that there are five categories in total, and the label form is [1, 0, 0, 1, 1], i.e., images have the 0th, 3rd, and 4th categories. We judge whether an image has each category, which is a two-class classification problem. We can use BCEloss for the classification. BCELoss is

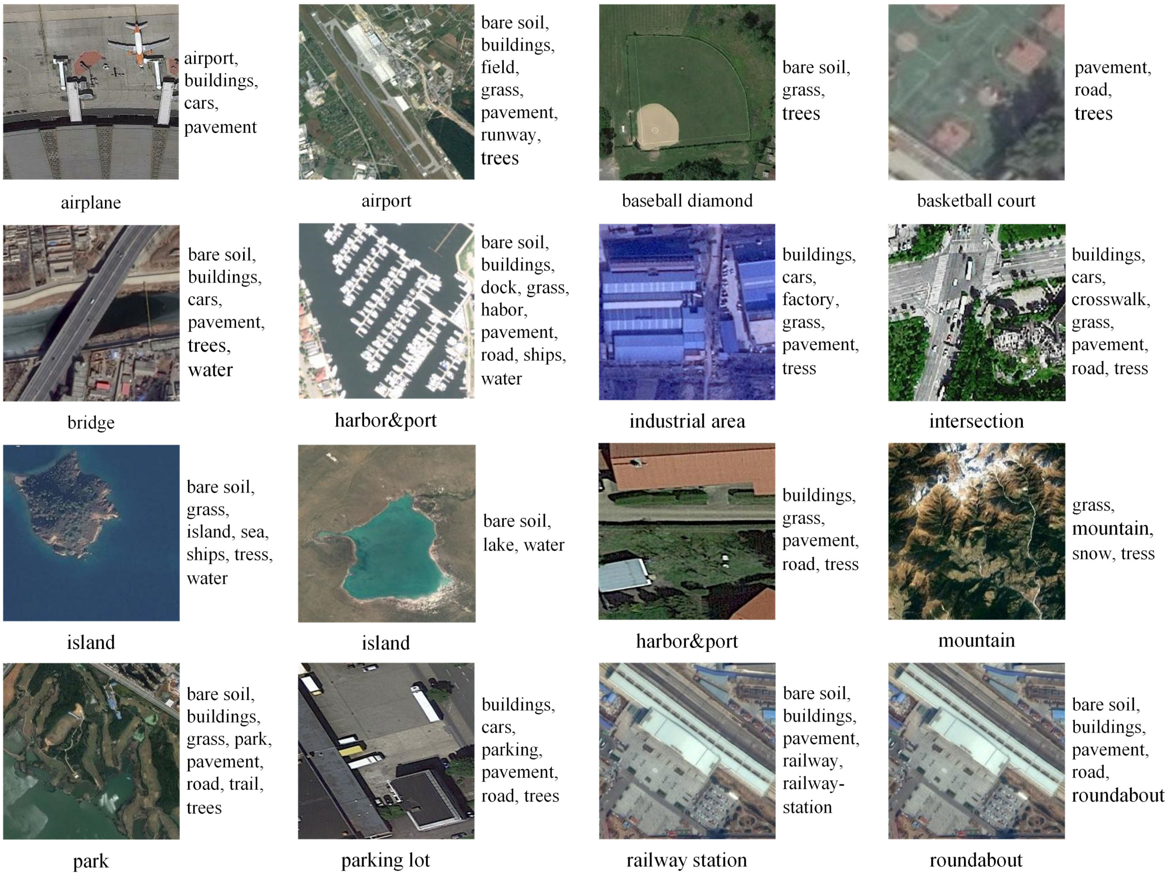

airport, buildings, cars, pavement
airplane
bare soil, buildings, field, grass, pavement, runway, trees
airport
bare soil, grass, trees
baseball diamond
pavement, road, trees
basketball court
bare soil, buildings, cars, pavement, trees, water
bridge
bare soil, buildings, dock, grass, habor, pavement, road, ships, water
harbor&port
buildings, cars, factory, grass, pavement, tress
industrial area
buildings, cars, crosswalk, grass, pavement, road, tress
intersection
bare soil, grass, island, sea, ships, tress, water
island
bare soil, lake, water
island
buildings, grass, pavement, road, tress
harbor&port
grass, mountain, snow, tress
mountain
bare soil, buildings, grass, park, pavement, road, trail, trees
park
buildings, cars, parking, pavement, road, trees
parking lot
bare soil, buildings, pavement, railway, railway-station
railway station
bare soil, buildings, pavement, road, roundabout
roundabout

Fig. 7. Example images of the MLRSNet dataset are shown, with the corresponding multilabel image on the right side of the image.

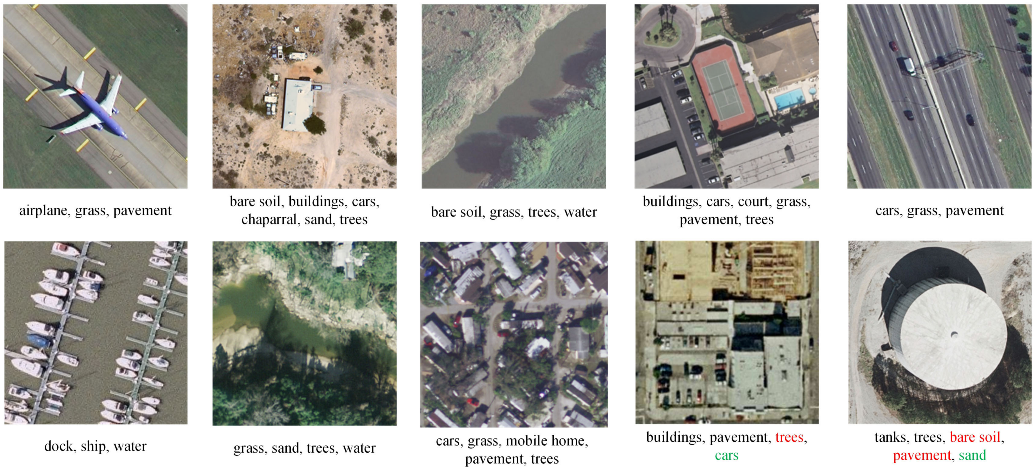

airplane, grass, pavement
bare soil, buildings, cars, chaparral, sand, trees
bare soil, grass, trees, water
buildings, cars, court, grass, pavement, trees
cars, grass, pavement
dock, ship, water
grass, sand, trees, water
cars, grass, mobile home, pavement, trees
buildings, pavement, trees, cars
tanks, trees, bare soil, pavement, sand

Fig. 8. Example images and predicted labels by the ResNet50-SR-Net on the UCM multilabel dataset. Multilabel of each image are reported below the related image, and the red font indicates the incorrect classification result, while the green font indicates the correct labels, but the model is not tagged.

to create a standard to measure the binary cross-entropy between label and output; it is expressed as follows:

$$L(O, y) = -\frac{1}{C} * \sum_{i=1}^{C} y_i * \log(O_i) + (1 - y_i) * \log(1 - O_i) \quad (8)$$

where C the number of labels; O the label predicted by the model, the shape of O is (N, C), N is the batch size; y the real label.

BCEloss requires the input value between $(0, 1)$, so we need to use the sigmoid function to convert O. The sigmoid function is a differentiable and bounded function, often used in binary classification problems; it is expressed as follows:

$$S(x) = -\frac{1}{1 + e^{-x}}. \quad (9)$$

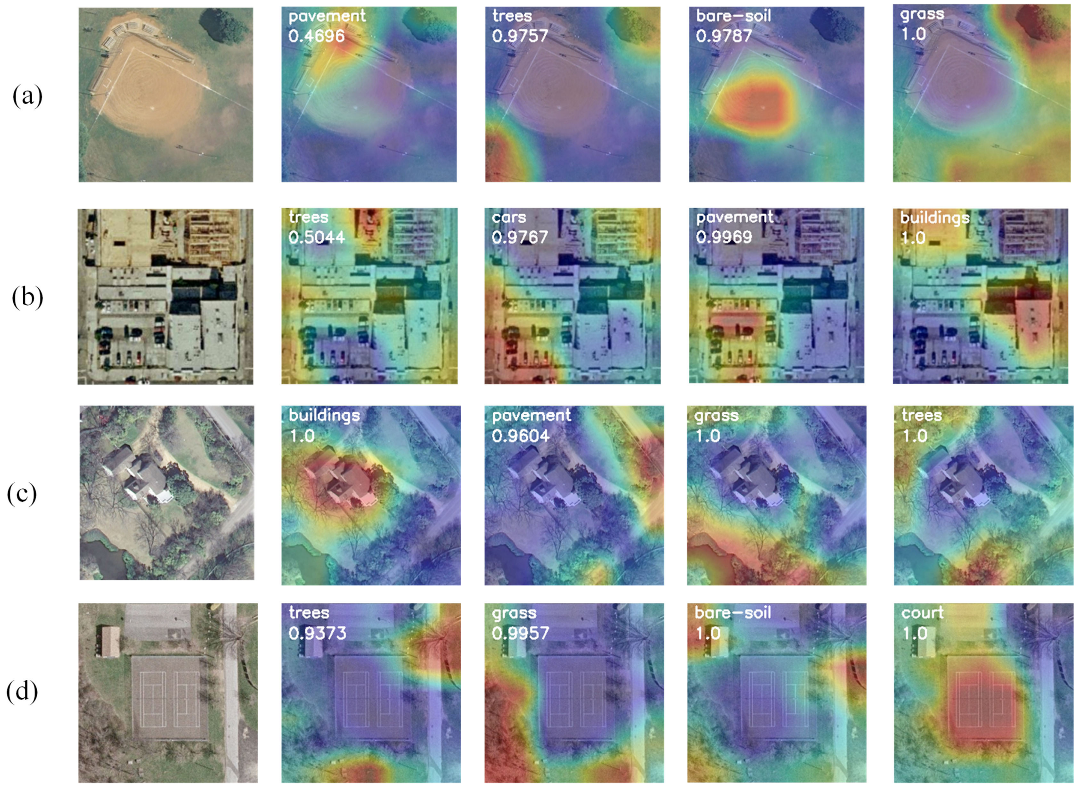


Fig. 9. Visualization results of CSAMs on UCM multilabel dataset.

Therefore, the final loss function is represented by the following formula:

$$L(O, y) = -\frac{1}{C} * \sum_{i=1}^{C} y_i * \log(1 + \exp(-O_i)^{-1}) + (1 - y_i) * \log\left(\frac{\exp(-O_i)}{1 + \exp(-O_i)}\right). \quad (10)$$

IV. Experimental Results and Discussion

Extensive experiments and analyses are presented in this section. There are six subsections. The first subsection demonstrates experimental setups, including the datasets and training details. The second subsection describes the evaluation metrics. The third section introduces the DCNN feature extraction algorithm. The fourth to sixth subsections present the experimental results and qualitative and quantitative analyses of the tested methods on the UCM multilabel, AID multilabel, and MLRSNet datasets.

A. Datasets and Training Details

To verify the effectiveness of the proposed method, we selected three challenging datasets, UCM multi-label [58], AID multi-label [17], and MLRSNet [59].

1) UCM Multilabel Dataset: The UCM multilabel dataset reproduces all the aerial images collected in the UCM dataset by assigning them to newly defined object tags. The UCM dataset is extracted from aerial imagery provided by the National Map of the U.S. Geological Survey. It contains 2100 images and is divided into 21 categories. These categories correspond to different land cover and land use types. Each category has 100 images with a size of 256 × 256 × 3 and a spatial resolution of 0.3 m. In the UCM multilabel dataset, there are a total of 17 object-level labels, including airplane, bare soil, buildings, cars, chaparral, court, dock, field, grass, mobile home, pavement, sand, sea, ship, tanks, trees, and water. Each image is labeled with one or more (up to seven) tags. Fig. 5 shows some UCM multilabel dataset examples, and Table I shows the details of the datasets. In order to compare with other methods, we follow the principle of dataset division in [17] and [40]. We select 80% of the image samples in each scene category to train our model, and the other 20% of the images are used to test our model.

TABLE I
Number of Images Present in the UCM Dataset for Each Class Label

Class label	Number	Class label	Number
Airplane	100	Mobile home	102
Bare soil	718	Pavement	1300
Buildings	691	Sand	294
Cars	886	Sea	100
Chaparral	115	Ship	102
Court	105	Tanks	100
Dock	100	Trees	1009
Field	104	Water	203
Grass	975		

There are 17 predefined class labels in total.

2) AID Multilabel Dataset: AID multilabel dataset selects 3000 aerial images from 30 scenes in the AID dataset and assigns

TABLE II
NUMBER OF IMAGES ARE PRESENT IN THE AID DATASET FOR EACH CLASS LABEL

Class label	Number	Class label	Number
Airplane	99	Mobile home	2
Bare soil	1475	Pavement	2328
Buildings	2161	Sand	259
Cars	2026	Sea	221
Chaparral	112	Ship	284
Court	344	Tanks	108
Dock	271	Trees	2406
Field	214	Water	852
Grass	2295		

There are 17 predefined class labels in total.

TABLE III
NUMBER OF IMAGES PRESENT IN THE MLRSNET DATASET FOR EACH CLASS LABEL

Class label	Number	Class label	Number
Airplane	2306	Mountain	5468
Airport	2480	Overpass	2652
Bare soil	39345	Park	1682
Baseball diamond	1996	Parking lot	7061
Basketball court	3726	Parkway	2537
Beach	2485	Pavement	56383
Bridge	2772	Railway	4399
Buildings	51305	Railway station	2187
Cars	34013	River	2493
Chaparral	5903	Road	37783
Cloud	1798	Roundabout	2039
Containers	2500	Runway	2259
Crosswalk	2673	Sand	11014
Dense residential area	2774	Sea	4980
Desert	2537	Ships	4092
Dock	2492	Snow	3565
Factory	2667	Snowberg	2555
Field	15142	Sparse residential area	1829
Football feld	1057	Stadium	2462
Forest	3562	Swimming pool	5078
Freeway	2500	Tanks	2500
Golf course	2515	Tennis court	2499
Grass	49391	Terrace	2345
Greenhouse	2601	Track	3693
Gully	2413	Trail	12376
Habor	2492	Transmission tower	2500
Intersection	2497	Trees	70728
Island	2493	Water	27834
Lake	2499	Wetland	3417
Mobile home	2499	Wind turbine	2049

There are 60 predefined class labels in total.

multiple object labels. The AID dataset contains 10 000 high-resolution aerial images collected from Google Earth around the world, including scenes from China, the United States, the United Kingdom, France, Italy, Japan, and Germany, with spatial resolutions ranging from 0.5 to 8 m and sizes of 600 × 600 × 3. In the AID multilabel dataset, there are a total of 17 object-level labels, consistent with the UCM multilabel dataset. Fig. 6 shows some examples of AID multilabel datasets, and Table II shows details of the datasets. In order to compare with other methods, we follow the principle of dataset division in [18] and [40]. We select 80% of the image samples in each scene category to train our model, and the other 20% of the images are used to test our model.

3) MLRSNet Dataset: MLRSNet [59] comprises 109 161 labeled RGB images globally, labeled into 46 broad categories. The number of sample images varies in the range of 1500–3000 for each category. In addition, this dataset contains 60 predefined class categories, and the number of categories associated with each image ranges from 1 to 13. Table III lists the number of images present in the dataset associated with each predefined label, and Fig. 7 shows some examples of images with corresponding multilabel. Besides, MLRSNet has various resolutions, approximately 10–0.1 m. Each multilabel image has a size of 256×256 pixels to cover a scene with various resolutions. In the DCNN training, training data are often related to experimental results. The more training data, the better the testing effect of the model. In order to verify that our model has better generalization ability and robustness, we randomly select 10% of the dataset for training and 90% for testing.

4) Training Details: For the entire framework, we, respectively, use VGG-Net [1], ResNet [3], and DenseNet [4] as our backbone. During training, we adopt the data augmentation suggested in [37] to avoid overfitting: for the UCM multilabel and MLRSNet datasets, the input image is randomly cropped and resized to 224×224 with random horizontal flips for data augmentation; for the AID multilabel dataset, the input image is randomly cropped and resized to 512×512 with random horizontal flips for data augmentation. To make the proposed model converge quickly, we follow [38] to choose the model trained on VOC2012 as the pretrain model. We chose SGD as the optimizer, with a 0.9 momentum and 10^{-4} weight decay. The initial learning rate was set to 0.01 for the SSM and SRBM and 0.001 for the backbone DCNN. We trained our model for 100 epochs on UCM multilabel and AID multilabel datasets and 50 epochs on the MLRSNet dataset, and the learning rate was reduced by a factor of 0.1 at 25, 50, and 75 epochs, respectively. The batch size was 32 for the UCM multilabel dataset and 16 for the AID multilabel and MLRSNet datasets. All experiments are implemented based on PyTorch.

B. Evaluation Metrics

To fairly compare with existing methods, we follow previous works [18], [19], [40] to adopt the average of example-based precision and recall (P_e/R_e), label-based precision and recall (P_l/R_l), and the mean F1 (mF1) as evaluation metrics. Suppose that there is a test set $D = (x_i, y_i) \mid 1 \leq i \leq N$, where the binary vector $y_i = [y_i 1, y_i 2, \ldots, y_C]^T \in 0, 1^C$ is the ground-truth label of the ith test sample. Let $\hat{y_i} = [\hat{y_i} 1, \hat{y_i} 2, \ldots, \hat{y_C}^T \in 0, 1^C]$ denote the predicted label vector for the sample x_i, $Y = [y_1, y_2, \ldots, y_N]^T = [I_1, I_2, \ldots, I_C] \in R^{N \times C}$ denote the ground truth label matrix, and $\hat{Y} = [\hat{y_1}, \hat{y_2}, \ldots, \hat{y_N}]^T = [\hat{I_1}, \hat{I_2}, \ldots, \hat{I_C}] \in R^{N \times C}$ denote the predicted label matrix. The

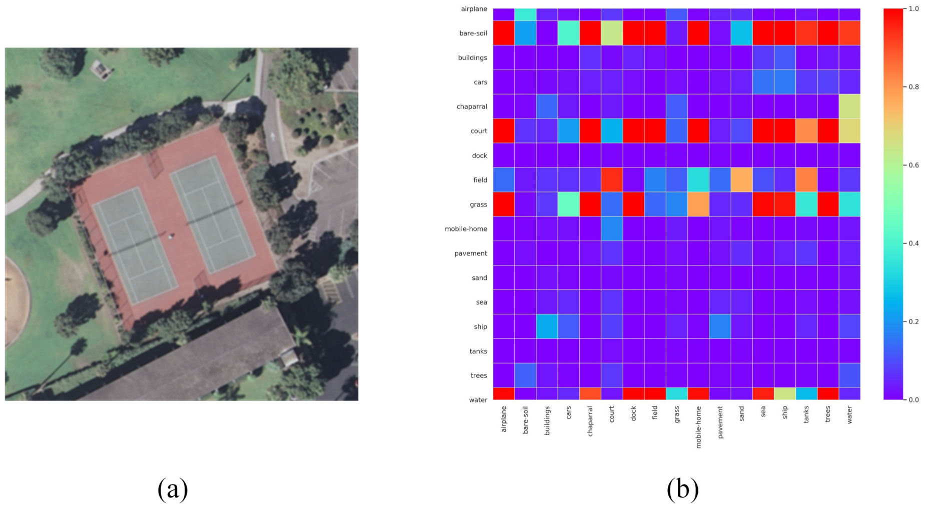


Fig. 10. Visualization of an example and what its relation matrix R looks like on the UCM multilabel dataset. (a) Input Image (b) Relation Matrix

formulas are defined as follows:

$$P_e = \frac{1}{N}\sum_{i}^{N}\frac{\mid \hat{y}_i \bigcap y_i \mid}{\mid \hat{y}_i \mid}, P_l = \frac{1}{C}\sum_{i}^{C}\frac{\mid \hat{I}_j \bigcap I_j \mid}{\mid \hat{I}_j \mid} \quad (11)$$

$$R_e = \frac{1}{N}\sum_{i}^{N}\frac{\mid \hat{y}_i \bigcap y_i \mid}{\mid y_i \mid}, R_l = \frac{1}{C}\sum_{i}^{C}\frac{\mid \hat{I}_j \bigcap I_j \mid}{\mid I_j \mid} \quad (12)$$

$$mF1 = \frac{2P_eR_e}{P_e + R_e}. \quad (13)$$

When measuring P_e/R_e/P_l/R_l/mF1, the label is positive if its confident score is greater than 0.5. To fairly compare with state-of-the-arts, generally speaking, the mF1 is more vital than other indicators.

C. Compare Methods

Five popular CNN architectures, VGGNet16 [1], VGGNet19 [1], ResNet50 [3], ResNet101 [3], and DenseNet201 [4] are chosen as our backbone to verify the effectiveness of the proposed method.

1) VGGNet16 and VGGNet19 [1]: The Visual Geometry Group of Oxford proposed VGG at ILSVRC 2014, which mainly proved that increasing the depth of a network can affect the final performance of the network to a certain extent. Simonyan and Zisserman used a continuous 3 × 3 convolution kernel instead of a larger convolution kernel to increase the depth of networks, which showed a significant improvement in accuracies. We use two models that show corresponding performance in scene classification, namely VGGNet16 and VGGNet19.

2) ResNet50 and ResNet101 [3]: The proposal of the deep residual network (ResNet) is a milestone event in the history of CNN imaging. This model won first place in the ILSVRC 2015 classification task. It proposes residual learning, which solves the problem of decreasing accuracy as the network structure deepens. ResNet50 and ResNet101 are 50- and 101-layer ResNet, respectively.

3) DenseNet201 [4]: Dense convolutional network (Dense Net) uses a feed-forward method to connect each layer to other layers. In the traditional DCNN, if the network has L layers, then there will be L connections, but in DenseNet, there will be $\frac{L(L+1)}{2}$ connections. Simply, the input of each layer comes from the output of all previous layers. DenseNets are widely used because they have several compelling advantages, such as alleviating the problem of vanishing gradients, enhancing feature propagation, promoting feature reuse, and significantly reducing the number of parameters. DensesNet201 is the 201-layer DenseNet.

4) DCNN + SSM: This method uses the DCNN (such as VGG16, ResNet101, and DenseNet201) as a feature extractor and feeds features into the SSM. Use a binary classifier upon the output of the SSM directly.

5) DCNN + SRBM: This method uses the DCNN (such as VGG16, ResNet101, and DenseNet201) as a feature extractor and feeds features into the SRBM. Use a binary classifier upon the output of the SRBM directly.

6) DCNN-RBFNN [60]: This method uses the DCNN for feature extraction and the RBFNN [60] for classification.

7) CA-DCNN-BiLSTM [40]: This method uses the DCNN for feature extraction, uses an attention learning layer to capture class-specific features, and uses the bidirectional LSTM network to model the class dependence for final classification.

8) AL-RN-DCNN [17]: This method uses the DCNN for feature extraction. Then, localize discriminative regions in these features and uses a 1 × 1 convolution to explore spatial information. Finally, use an MLP layer to produce the label relation for final classification.

9) MLRSSC-CNN-GNN [49]: This method combines a CNN and a GNN to generate high-level appearance features using CNN's perception of visual elements in learning scenes. Based on the trained CNN, a scene graph of each scene was further constructed, and the superpixel region of the scene represented the nodes in the graph.

TABLE IV
COMPARISON OF THE PROPOSED METHOD AND THE STATE-OF-THE-ART METHOD ON THE UCM MULTILABEL DATASET

	mF1	mPe	mRe	mPl	mRl
VGG16	80.18	77.86	82.63	79.92	74.33
VGG16-RBFNN	78.80	78.18	83.91	81.90	82.63
CA-VGG16-BiLSTM	79.78	79.33	83.99	85.28	76.52
AL-RN-VGG16Net	**85.70**	**87.62**	**86.41**	**91.04**	**81.71**
VGG16-SR-Net	83.43	82.33	84.55	86.90	80.68
ResNet50	76.68	80.86	81.95	88.78	78.98
ResNet50-RBFNN	80.58	79.92	84.59	86.21	83.72
CA-ResNet50-BiLSTM	81.47	77.94	89.02	86.12	84.26
AL-RN-ResNet50	86.76	**88.81**	87.07	86.12	84.26
MLRSSC-CNN-GNN	86.39	87.11	88.41	-	-
ResNet50-SR-Net	**88.67**	87.96	**89.40**	**93.52**	**91.51**

TABLE V
COMPARISON OF THE PROPOSED METHOD AND THE STATE-OF-THE-ART METHOD ON THE AID MULTILABEL DATASET

	mF1	mPe	mRe	mPl	mRl
VGG16	85.52	87.41	86.32	70.60	58.89
VGG16-RBFNN	84.58	84.56	87.85	62.90	69.15
CA-VGG16-BiLSTM	86.68	88.68	87.83	72.04	60.00
AL-RN-VGG16Net	**88.09**	**89.96**	**89.27**	**76.94**	**68.31**
VGG16-SR-Net	87.15	86.84	87.46	75.79	65.38
ResNet50	86.23	89.31	85.65	72.39	52.82
ResNet50+RBFNN	83.77	82.84	88.32	60.85	70.45
CA-ResNet50-BiLSTM	87.63	89.03	88.99	79.50	65.60
AL-RN-ResNet50	88.72	**91.00**	88.95	80.81	71.12
MLRSSC-CNN-GNN	86.39	87.11	88.41	-	-
ResNet50-SR-Net	**89.97**	89.42	**90.52**	**87.24**	**82.25**

D. Experiments on UCM Multilabel Dataset

1) Quantitative Analysis: Generally, mF1 is more vital than other indicators. Table IV shows the experimental results of the UCM multilabel dataset. Specifically, compared with VGG16, VGG16-SR-Net improves the mF1 score by 3.25%. Compared with CA-VGG16-BiLSTM, our method improves the mF1 score by 3.65%. Compared with VGG16-RBFNN, our method improves the mF1 score by 4.63%.

In contrast, VGG16-SR-Net is superior to VGG16, VGG16-RBFNN, and CA-VGG16-BiLSTM in mF1 scores and mean example- and label-based precisions and recalls. For another backbone, ResNet50, our network got the best mF1 score. As shown in Table IV, ResNet50-SR-Net increases the mF1 scores of ResNet50, ResNet50-RBFNN, CA-ResNet50-BILSTM, AL-RN-ResNet50, and MLRSSC-CNN-RNN by 11.99%, 8.09%, 7.2%, 1.91%, and 2.28%, respectively. Our method also surpasses other competitors for other indicators, which proves the effectiveness and robustness of our method. Compared with all other methods, the mF1 score of ResNet50-SR-Net is 0.8867. In addition, it achieved the best mean example-based recall rate of 0.8940, mean label-based accuracy of 0.9352, and recall rate of 0.9151. Fig. 8 shows an example of classification results on ResNet50-SR-Net. Although our mean example-based precision is slightly lower than AL-RN-ResNet50, we have achieved a significant improvement in label-based precision and recall. The classification results are shown below the picture. Here, the black font is the correctly classified category, the red font is the wrong category, and the green is the correct label, but the model does not recognize its correct label.

All in all, the comparison between DCNN-SR-Net and other models proves the effectiveness of our method. In addition, when using VGG16 as our backbone network, our method is lower than AL-RN-VGG16Net in various accuracy, indicating that our method is more suitable for using the DCNN with a deeper network structure as a feature extractor. At the same time, VGG16 is a shallower DCNN network, and the experimental results are easily affected by the division of the dataset, while the UCM multilabel dataset is relatively small, so we added a more extensive dataset for supplementary experiments.

2) Qualitative Analysis: In order to figure out what is learned inside our method, we further visualize each module to verify the effectiveness of the proposed method in a qualitative way. In Fig. 9, we visualized the original image and its corresponding CSAM to illustrate the SSM's ability to capture the semantic regions of each category appearing in the image. We used the standard ResNet50 as a backbone. Each row shows the original image and activation map of a specific category. The category and score are displayed in the upper left corner of the corresponding category activation map. The proposed model could locate discriminative semantic regions related to positive categories are highlighted in these feature maps, whereas less informative regions are weakly activated. We can observe that in these activation maps, the discriminative regions related to the positive categories are highlighted, while the regions with less information are weakly activated. As an exception, there is an activation map in Fig. 9(a) that is incorrectly identified as a pavement, which may lead to mispredictions.

Furthermore, in Fig. 10, we visualized an original image with its corresponding relation matrix R to illustrate what relations the SRBM had learned. For the input image in Fig. 10(a), its labels are "buildings," "bridge," "court," "grass," "pavement," and "trees." Fig. 10(b) is the visualization of the R of the input image. $R^{\text{court,trees}}$ ranked top (about top 10%) in the row of "court." It means that "trees" were more relevant for "court" in the image. From the relation matrix visualization, the SRBM can build such semantic relations for a specific input image.

E. Experiments on AID Multilabel Dataset

1) Quantitative Analysis: Generally, mF1 is more vital than other indicators. Table V shows the experimental results of the AID multilabel dataset. Specifically, compared with VGG16, VGG16-SR-Net improves the mF1 score by 1.63%. Compared with CA-VGG16-BiLSTM, our method improves the mF1 score by 0.47%. Compared with VGG16-RBFNN, our method improves the mF1 score by 2.57%.

In contrast, VGG16-SR-Net is superior to VGG16, VGG16-RBFNN, and CA-VGG16-BiLSTM in mF1 scores and mean example- and label-based precisions and recalls. For another backbone, ResNet50, our network got the best mF1 score. As shown in Table V, ResNet50-SR-Net increases the mF1 scores of ResNet50, ResNet50-RBFNN, CA-ResNet50-BILSTM AL-RN-ResNet50, and MLRSSC-CNN-RNN by 3.74%, 6.20%, 2.34%, 1.25% and 3.58%, respectively. Our method also

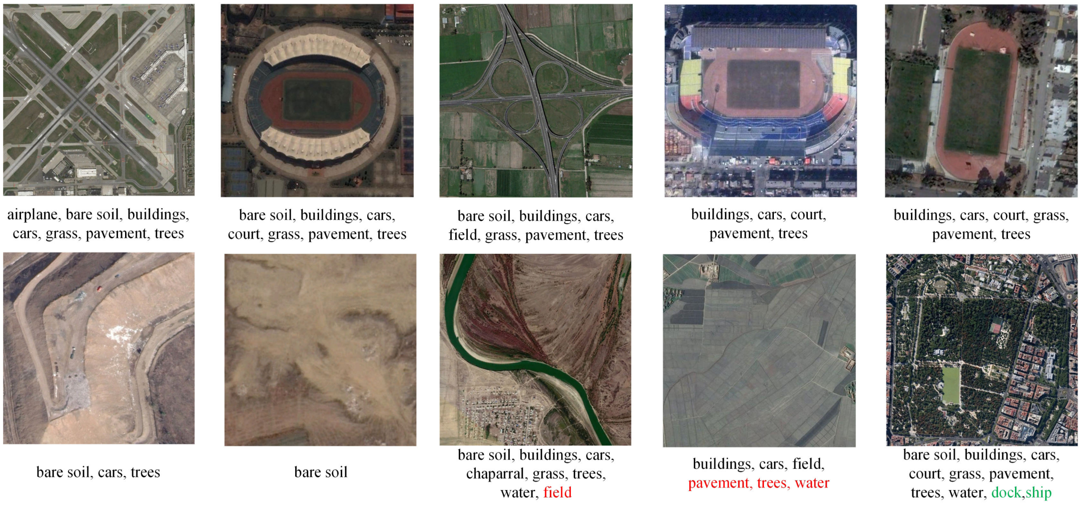

airplane, bare soil, buildings, cars, grass, pavement, trees
bare soil, buildings, cars, court, grass, pavement, trees
bare soil, buildings, cars, field, grass, pavement, trees
buildings, cars, court, pavement, trees
buildings, cars, court, grass, pavement, trees
bare soil, cars, trees
bare soil
bare soil, buildings, cars, chaparral, grass, trees, water, field
buildings, cars, field, pavement, trees, water
bare soil, buildings, cars, court, grass, pavement, trees, water, dock,ship

Fig. 11. Example images and predicted labels by the ResNet50-SR-Net on the AID multilabel dataset. Multilabel of each image are reported below the related image, and the red font indicates the incorrect classification result, while the green font indicates the correct labels, but the model is not tagged.

surpasses other competitors for other indicators, which proves the effectiveness and robustness of our method. Compared with all other methods, the mF1 score of ResNet50-SR-Net is 0.8997. In addition, it achieved the best mean example-based recall rate of 0.9052, mean label-based accuracy of 0.8724, and recall rate of 0.8225. Fig. 11 shows an example of classification results on ResNet50-SR-Net. Although our mean example-based precision is slightly lower than AL-RN-ResNet50, we have achieved a significant improvement in label-based precision and recall. The classification results are shown below the picture. Here, the black font is the correctly classified category, the red font is the wrong category, and the green is the correct label, but the model does not recognize its correct label.

All in all, the comparison between DCNN-SR-Net and other models proves the effectiveness of our method. In addition, when using VGG16 as our backbone network, our method is lower than AL-RN-VGG16Net in various accuracy, indicating that our method is more suitable for using the DCNN with a deeper network structure as a feature extractor.

2) Qualitative Analysis: In order to figure out what is learned inside our method, we further visualize each module to verify the effectiveness of the proposed method in a qualitative way. In Fig. 12, we visualized the original image and its corresponding CSAM to illustrate the SSM's ability to capture the semantic regions of each category appearing in the image. We used the standard ResNet50 as a backbone. Each row shows the original image and activation map of a specific category. The category and score are displayed in the upper left corner of the corresponding category activation map. The proposed model could locate discriminative semantic regions related to positive categories are highlighted in these feature maps, whereas less informative regions are weakly activated. We can observe that in these activation maps, the discriminative regions related to the positive categories are highlighted, while the regions with less information are weakly activated.

Furthermore, in Fig. 13, we visualized an original image with its corresponding relation matrix R to illustrate what relations the SRBM had learned. For the input image in Fig. 13(a), its labels are "buildings," "cars," "dock," "grass," "pavement," "sea," "ship," and "trees." Fig. 10(b) is the visualization of the R of the input image. $R^{\text{ship,dock}}$ ranked top (about top 10%) in the row of "ship." It means that "dock" were more relevant for "ship" in the image. From the relation matrix visualization, the SRBM can build such semantic relations for a specific input image.

TABLE VI
COMPONENT EFFECTIVENESS EVALUATION OF THE PROPOSED METHOD ON MLRSNET DATASET

	mF1	mPe	mRe	mPl	mRl
VGG16	68.01	70.84	65.41	64.71	43.56
VGG16+SSM	72.44	75.71	69.43	75.37	50.40
VGG16+SRBM	71.26	75.77	67.25	73.86	49.47
VGG16-SR-Net	**73.80**	**78.25**	**69.82**	**76.07**	**54.13**
VGG19	67.10	71.79	62.98	63.10	43.25
VGG19+SSM	72.91	**78.00**	68.43	**76.78**	51.86
VGG19+SRBM	70.12	73.61	66.95	70.52	46.95
VGG19-SR-Net	**73.33**	77.43	**69.63**	75.33	**51.86**
ResNet50	85.68	85.65	85.71	86.64	83.81
ResNet50+SSM	86.58	87.15	86.02	**88.98**	85.72
ResNet50+SRBM	86.07	**87.94**	84.28	88.68	85.50
ResNet50-SR-Net	**87.21**	87.08	**87.34**	88.79	**86.73**
ResNet101	86.05	86.75	85.23	87.41	84.85
ResNet101+SSM	86.92	87.65	86.21	89.90	85.84
ResNet101+SRBM	86.71	**88.90**	84.62	**90.55**	85.03
ResNet101-SR-Net	**87.55**	87.84	**87.26**	89.41	**87.48**
DenseNet201	86.17	85.97	86.37	86.41	85.99
DenseNet201+SSM	86.56	87.12	86.00	88.14	87.06
DenseNet201+SRBM	86.26	86.65	85.87	88.57	86.40
DenseNet201-SR-Net	**87.36**	**87.23**	**87.49**	**88.80**	**87.87**

F. Experiments on MLRSNet Dataset

1) Quantitative Analysis: In order to explore whether our method can achieve a better performance with less training data on a large dataset and to verify the influence of different modules in our method, we conducted the following experiments with the MLRSNet dataset. Table VI shows in detail the evaluation of different modules in SR-Net methods. Generally, mF1 is more

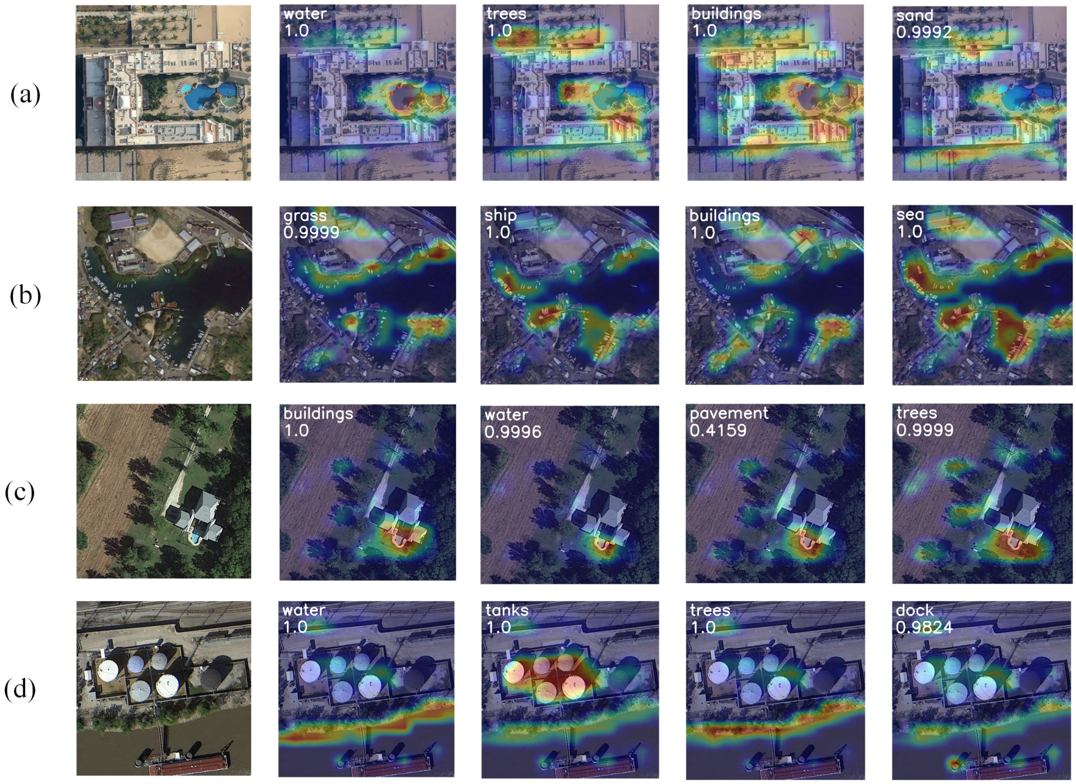
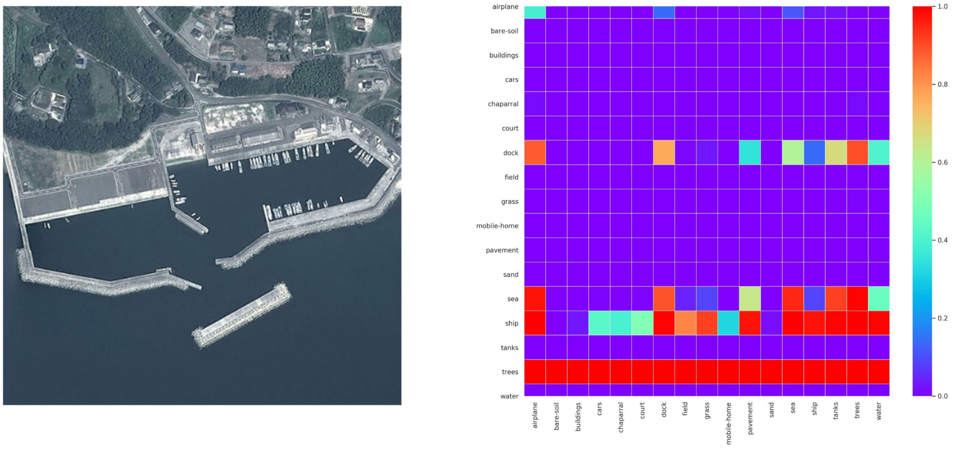


Fig. 12. Visualization results of CSAMs on AID multilabel dataset.

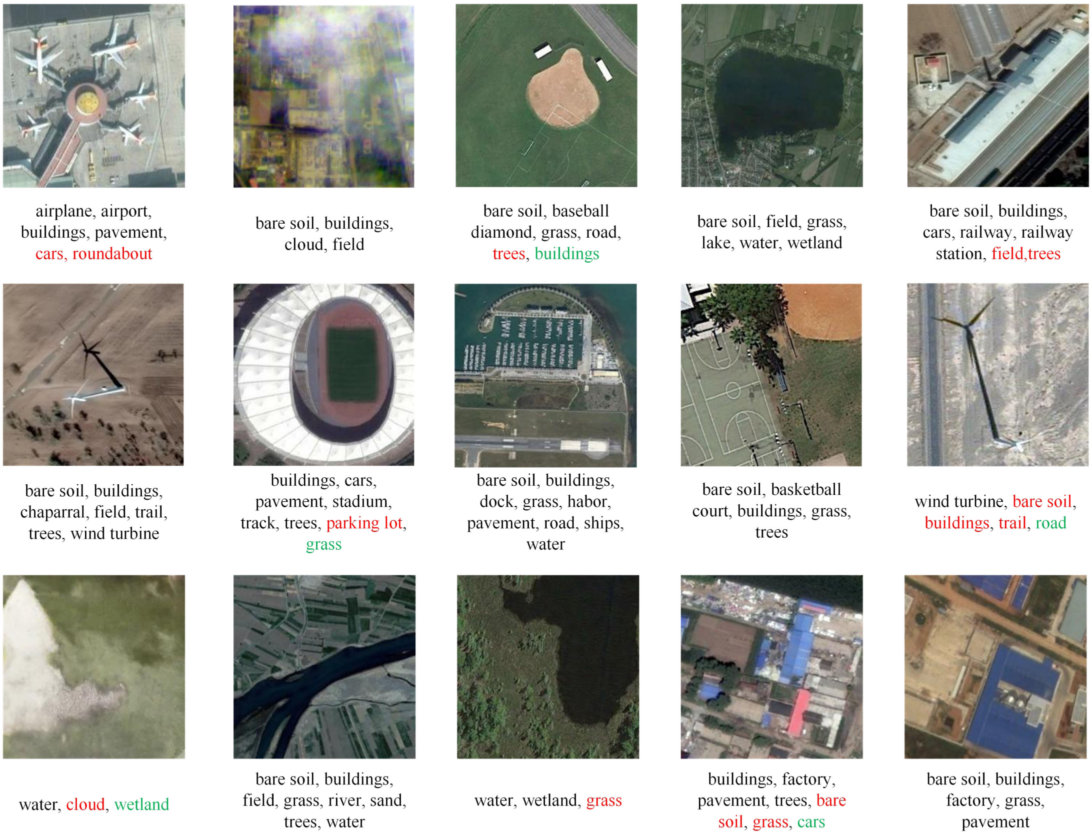


(a) (b)

Fig. 13. Visualization of an example and what its relation matrix R looks like on the AID multilabel dataset.

vital than other indicators. In Table VI, the performance of each DCNN model improves after our method is added. Specifically, compared with VGG16, VGG19, ResNet50, ResNet101, and DenseNet201, our method improves the mF1 score by 5.79%, 6.23%, 1.53%, 1.5%, and 1.19%, respectively.

The improvement of the SSM showed that the decomposed representation had a more vital discriminative ability. We also found that the SRBM could improve the feature recognition ability of the results compared with the baseline. However, if the features learned from the DCNN were directly input to the SRBM, the result would not be as good as combining with the SSM and using the CACR as the input. That means increasing the semantic information in the feature can further improve the ability of the SRBM to learn relationships, and enhancing the semantic information in the features can further enhance the ability of the SRBM to learn relationships. Although for VGG19, ResNet50, and ResNet101 models, our method does not perform the best on mean example- and label-based precisions, but our method sacrifices a little bit of the accuracy of mean example- and label-based precisions, but the mean



airplane, airport, buildings, pavement, cars, roundabout

bare soil, buildings, cloud, field

bare soil, baseball diamond, grass, road, trees, buildings

bare soil, field, grass, lake, water, wetland

bare soil, buildings, cars, railway, railway station, field,trees

bare soil, buildings, chaparral, field, trail, trees, wind turbine

buildings, cars, pavement, stadium, track, trees, parking lot, grass

bare soil, buildings, dock, grass, habor, pavement, road, ships, water

bare soil, basketball court, buildings, grass, trees

wind turbine, bare soil, buildings, trail, road

water, cloud, wetland

bare soil, buildings, field, grass, river, sand, trees, water

water, wetland, grass

buildings, factory, pavement, trees, bare soil, grass, cars

bare soil, buildings, factory, grass, pavement

Fig. 14. Example images and predicted labels by the ResNet101-SR-Net on MLRSNet dataset. Multilabel of each image are reported below the related image, and the red font indicates the incorrect classification result, while the green font indicates the correct labels, but the model is not tagged.

example- and label-based recalls have been significantly improved. Overall, combining SSM and SRBM would further improve performance, as expected, because it focuses on different aspects.

Notably, VGG16 and VGG19 had the most significant improvement. The proposed method significantly improved on the VGG model because the VGG model is not as deep as the network structure of ResNet and DenseNet, so the high-level semantic information in the features extracted by VGG is less than those extracted by ResNet and DenseNet. The proposed SSM could capture the semantic regions in the DCNN extracted features and strengthen the representation ability of DCNN features. In addition, the proposed SRBM could construct the relationship between category representations and further enhance the representation ability of features. Experiments have proven the effectiveness of the proposed method. The results of the proposed model were significantly superior to those of models with shallower network structures, such as VGG. There were also specific improvements on more complex models with deeper network structures, such as ResNet and DendeNet.

Fig. 14 shows an example of classification results using ResNet101 as the backbone. The classification results are shown below the picture. Here, the black font is the correctly classified category, the red font is the wrong category, and the green is the correct label, but the model does not recognize its correct label. Table VII shows the models' number of parameters, GFLOPs, and inference speed.

TABLE VII
MODELS' NUMBER OF PARAMETERS(PARAMS), GFLOPS, AND INFERENCE SPEED

	params	GFLOPs	FPS
VGG16	134M	15.5	154.14
VGG16-SR-Net	31M	15.4	103.83
VGG19	140M	19.6	139.79
VGG19-SR-Net	36M	19.5	102.53
ResNet50	24M	4.1	114.84
ResNet50-SR-Net	42M	4.2	74.07
ResNet101	43M	7.8	84.15
ResNet101-SR-Net	61M	7.9	54.88
DenseNet201	18M	4.3	40.26
DenseNet201-SR-Net	39M	4.5	33.90

All FLOPs are measured with a size of 224 over the first 20 000 images of the MLRSNet dataset. Moreover, the FPS in the table is calculated with batch size one on 1080Ti from the total inference pure compute time reported in the Pytorch.

2) Qualitative Analysis: In order to figure out what is learned inside our method, we further visualize each module to verify the effectiveness of the proposed method in a qualitative way. In Fig. 15, we visualized the original image and its corresponding CSAM to illustrate the SSM's ability to capture the semantic regions of each category appearing in the image. We used the standard ResNet101 as a backbone, and the training–testing ratio was $10-90\%$. Each row shows the original image and activation map of a specific category. The category and score are displayed in the upper left corner of the corresponding category

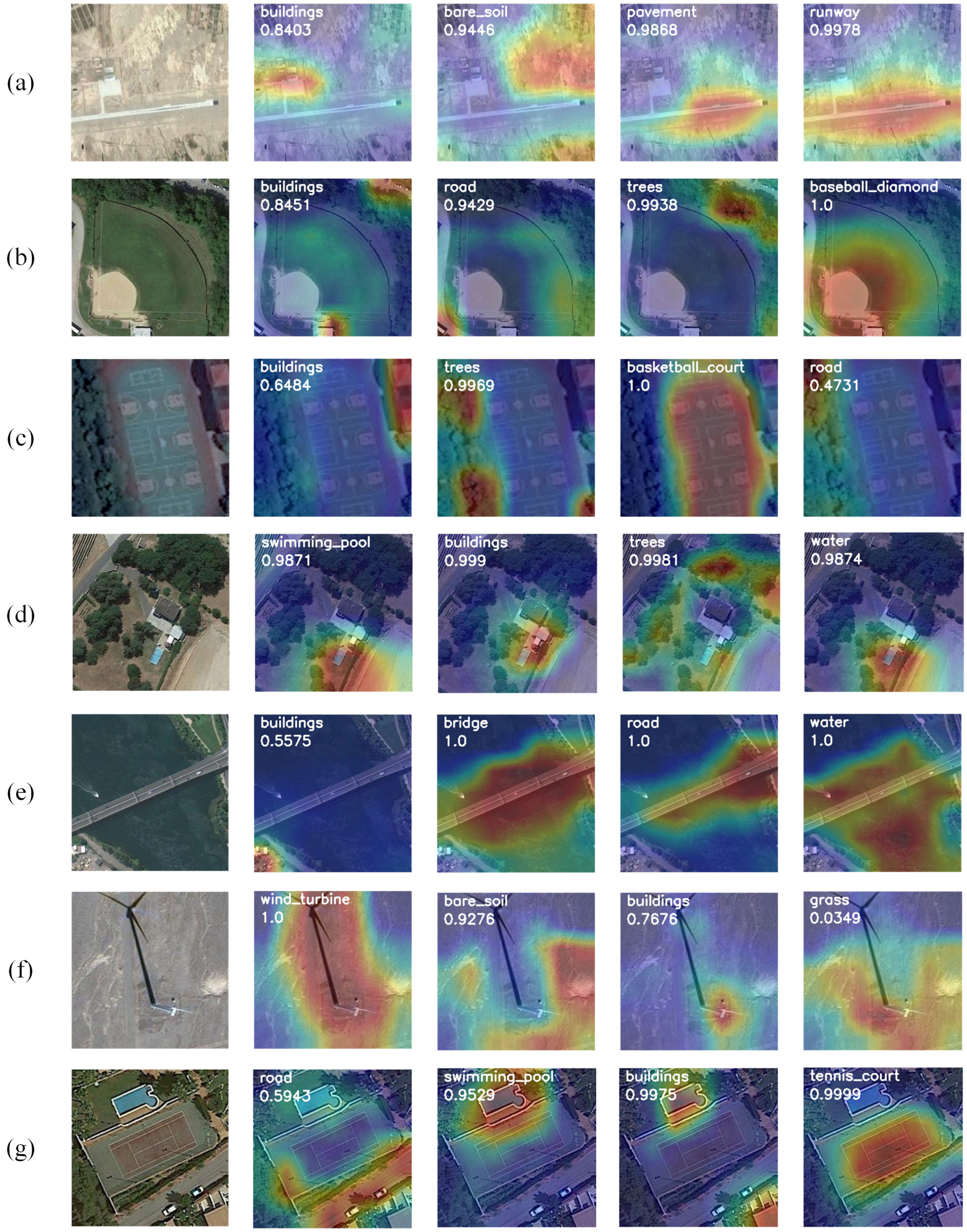
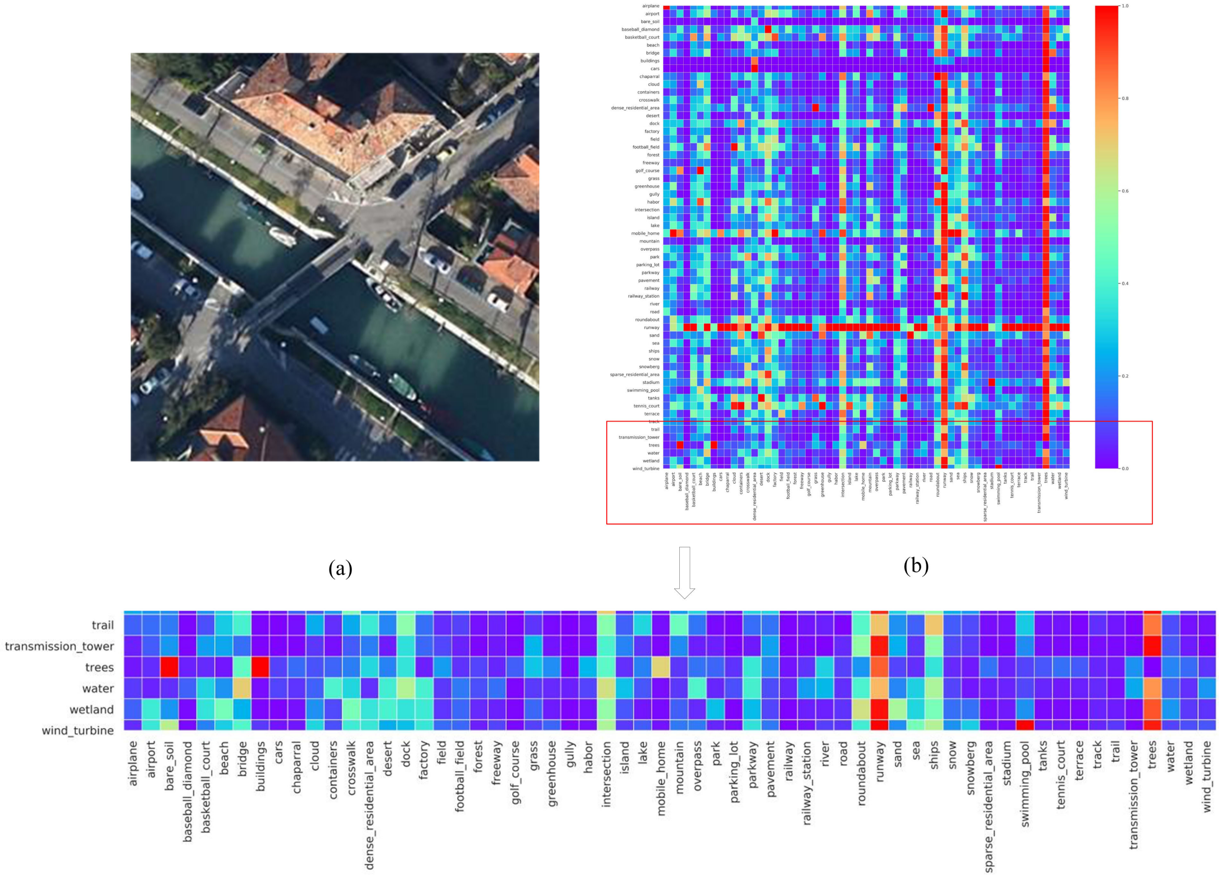


Fig. 15. Visualization results of CSAMs on MLRSNet dataset.

activation map. The proposed model could locate discriminative semantic regions related to positive categories are highlighted in these feature maps, whereas less informative regions are weakly activated. For example, in Fig. 15(e), the activated area of the road is the road itself, whereas the activated area of the bridge includes the road and water around the bridge, which considers the semantic information of the bridge and water. The SR-Net could accurately highlight relevant semantic regions. In addition, the final score indicated that the CACR was sufficiently discriminative and could be accurately identified using the proposed method.

Furthermore, in Fig. 16, we visualized an original image with its corresponding relation matrix R to illustrate what relations the SRBM had learned. For the input image in Fig. 16(a), its labels are "buildings," "bridge," "cars," "grass," "pavement," "road," "ships," "trees," and "water." Fig. 14(b) is the visualization of the R of the input image. $R^{\text{water,bridge}}$ and $R^{\text{water,ships}}$ ranked top (about top 10%) in the row of "water." It means that "bridge" and "ships" were more relevant for "water" in the image. From the relation matrix visualization, the SRBM can build such semantic relations for a specific input image.



(a)

(b)

(c)

Fig. 16. Visualization of an example and what its relation matrix R looks like on MLRSNet dataset. (a) Input Image (b) Relation Matrix (c) Partial Graph of Relation Matrix

V. Conclusion

In this study, we propose a novel HRS image multilabel classification network, transformer-driven SR-Net. The proposed network contains two modules: SSM and SRBM. The SSM captures the semantic regions of features extracted by the DCNN and generates a discriminative CACR. The SRBM further uses label relation inference from outputs of the SSM, that is, CACR, to obtain the relation matrix of categories for final classification. We conduct extensive experiments on the public UCM multilabel, AID multilabel, and MLRSNet datasets to evaluate the proposed method. The experimental results showed that our network using the deep network (e.g., ResNet) could offer better classification results. In addition, we visualize extracted semantic attentional regions and relation matrix for qualitatively demonstrating each module's effectiveness.

Acknowledgment

The numerical calculations in this article have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 19–36.

[2] C. Szegedy, L. Wei, Y. Jia, P. Sermanet, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

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

[4] G. Huang, Z. Liu, V. Laurens, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[5] Y. Yuan, J. Fang, X. Lu, and Y. Feng, "Remote sensing image scene classification using rearranged local features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1779–1792, Mar. 2019.

[6] Z. Xiao, K. Wang, Q. Wan, X. Tan, and F. Xia, "A2S-Det: Efficiency anchor matching in aerial image oriented object detection," *Remote Sens.*, vol. 13, no. 1, p. 73, 2020.

[7] X. Tan, Z. Xiao, Q. Wan, and W. Shao, "Scale sensitive neural network for road segmentation in high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 533–537, Mar. 2021.

[8] J. Shao, B. Du, C. Wu, M. Gong, and T. Liu, "HRSiam: High-resolution siamese network, towards space-borne satellite video tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 3056–3068, 2021.

[9] L. Wan, N. Liu, Y. Guo, H. Huo, and T. Fang, "Local feature representation based on linear filtering with feature pooling and divisive normalization for remote sensing image classification," *J. Appl. Remote Sens.*, vol. 11, no. 1, 2017, Art. no. 0 16017.

[10] H. Li-jun, H. Bin, and Z. Da-biao, "A destriping method with multi-scale variational model for remote sensing images," *Opt. Precis. Eng.*, vol. 25, pp. 198–207, Jan. 2017.

[11] K. Xu, H. Huang, and P. Deng, "Remote sensing image scene classification based on global-local dual-branch structure model," *IEEE Geosci. Remote Sens. Lett.*, vol. 30, pp. 3056–3068, 2021.

[12] K. Xu, H. Huang, P. Deng, and Y. Li, "Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2021.3071369.

[13] K. Xu, H. Huang, Y. Li, and G. Shi, "Multilayer feature fusion network for scene classification in remote sensing," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1894–1898, Nov. 2020.

[14] Z. Feng, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2027–2036.

[15] Y. Wei *et al.*, "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Softw. Eng.*, vol. 38, no. 9, pp. 1901–1907, Sep. 2016.

[16] P. F. Zhang, H. Y. Wu, and X. S. Xu, *A Dual-CNN Model for Multi-Label Classification by Leveraging Co-Occurrence Dependencies Between Labels*. Cham, Switzerland: Springer, 2017.

[17] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4558–4572, Jul. 2020.

[18] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020.

[19] R. Huang, F. Zheng, and W. Huang, "Multilabel remote sensing image annotation with multiscale attention and label correlation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6951–6961, 2021.

[20] M. Chen *et al.*, "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2020, vol. 119, pp. 1691–1703.

[21] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," *Int. Conf. Learn. Representations*, 2021.

[22] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, 2004.

[23] E. Tanaka *et al.*, "A multi-label approach using binary relevance and decision trees applied to functional genomics," *J. Biomed. Informat.*, vol. 54, pp. 85–95, 2015.

[24] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discov. Databases, II*, 2009, pp. 254–269.

[25] J. Fürnkranz, E. Hüllermeier, E. L. Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Mach. Learn.*, vol. 73, no. 2, pp. 133–153, 2008.

[26] G. Tsoumakas and I. Vlahavas, "Random K-labelsets: An ensemble method for multilabel classification," in *Proc. Eur. Conf. Mach. Learn.*, 2007, pp. 406–417.

[27] M. L. Zhang and Z. H. Zhou, "Ml-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007.

[28] A. E. Elisseeff and J. Weston, "A Kernel method for multi-labelled classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 681–687.

[29] A. Clare and R. D. King, "Knowledge discovery in multi-label phenotype data" in *Proc. Eur. Conf. Principles Data Mining Knowl. Discov.*, 2001, pp. 42–53.

[30] N. Ghamrawi and A. McCallum, "Collective multi-label classification," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, 2005, pp. 195–200.

[31] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300FPS," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3286–3293.

[32] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[33] H. Yang, J. T. Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai, "Exploit bounding box annotations for multi-label object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 280–288.

[34] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2285–2294.

[35] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 464-472, doi: 10.1109/ICCV.2017.58.

[36] Q. Li, M. Qiao, W. Bian, and D. Tao, "Conditional graphical lasso for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2977–2986.

[37] Z. Chen, X. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 5172–5181, doi: 10.1109/CVPR.2019.00532.

[38] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 522–531.

[39] Y. Wang *et al.*, "Multi-label classification with label graph superimposing," 2019, *arXiv:1911.09243*.

[40] Y. Hua, L. Mou, and X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 188–199, Feb. 2019.

[41] G. Sumbul and B. Demİr, "A deep multi-attention driven approach for multi-label remote sensing image classification," *IEEE Access*, vol. 8, pp. 95934–95946, 2020.

[42] J. Ji, W. Jing, G. Chen, J. Lin, and H. Song, "Multi-label remote sensing image classification with latent semantic dependencies," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1110. [Online]. Available: https://www.mdpi.com/2072-4292/12/7/1110

[43] Y. Guo and S. Gu, "Multi-label classification using conditional dependency networks," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1300–1305.

[44] B. Micusik and T. Pajdla, "Multi-label image segmentation via max-sum solver," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–6.

[45] M. Tan *et al.*, "Learning graph structure for multi-label image classification via clique generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4100–4109.

[46] X. Li, F. Zhao, and Y. Guo, "Multi-label image classification with a probabilistic label enhancement model," in *Proc. 30th Conf. Uncertainty Artif. Intell.*, Jan. 2014, pp. 430–439.

[47] A. Pal, M. Selvakumar, and M. Sankarasubbu, "Magnet: Multi-label text classification using attention-based graph neural network," in *Proc. 12th Int. Conf. Agents Artifi. Intell.*, vol. 2, 2020, pp. 494–505.

[48] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5172–5181.

[49] Y. Li, R. Chen, Y. Zhang, M. Zhang, and L. Chen, "Multi-label remote sensing image scene classification by combining a convolutional neural network and a graph neural network," *Remote Sens.*, vol. 12, no. 23, 2020, Art. no. 4003.

[50] 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, May 2019.

[51] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[52] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations*, Jan. 2015.

[53] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[54] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[55] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[56] P. Deng, K. Xu, and H. Huang, "When CNNs meet vision transformer: A joint framework for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, pp. 1–5, vol. 19, 2022.

[57] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[58] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.

[59] A. Xq *et al.*, "MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 337–350, 2020.

[60] A. Zeggada, F. Melgani, and Y. Bazi, "A deep learning approach to UAV image multilabeling," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 694–698, May 2017.

Xiaowei Tan received the B.S. degree in remote sensing science and technology in 2017 from the China University of Geosciences, Wuhan, China, where she is currently working toward the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Surveying, Mapping and Remote Sensing Information Engineering.

Her research interests include the application of classification and semantic segmentation in remote sensing.

Zhifeng Xiao (Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2008.

From 2014 to 2015, he was a Visiting Scholar with the Computational Biomedicine Imaging and Modeling Center, Rutgers University, New Brunswick, NJ, USA. He is currently an Associate Professor with the State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University. His work consists of object detection in remote sensing images, large-scale content-based remote sensing image retrieval, and scene analysis on remote sensing images. His research interests include remote sensing image processing, computer vision, and machine learning.

Jianjun Zhu received the M.S. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2005. He is a Deputy General Manager and a Senior Engineer with National Bio Energy Company, Ltd., Beijing, China.

He has in-depth research on biomass power generation and integrated energy development based on biomass energy utilization at the county level.

Qiao Wan received the B.S. degree in geographic information system and M.S. degree in cartography and geographic information systems in 2012 and 2015, respectively, from the Wuhan University, Wuhan, China, where she is currently working toward the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Surveying, Mapping and Remote Sensing Information Engineering.

She is currently a Teacher with the College of Computer Science and Technology, Guizhou University, Guiyang, China. Her research interests include domain generalization in remote sensing images.

Kai Wang received the B.S. degree in geographic information science and the M.S. degree in cartography and geographic information systems in 2018 and 2021, respectively, from Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree in remote sensing science and technology with School of Remote Sensing and Information Engineering.

His research interests include object detection in remote sensing.

Deren Li (Member, IEEE) received the Ph.D. degree in photogrammetry from the University of Stuttgart, Stuttgart, Germany, in 1986 and the Honorary Doctorate degree from ETH Zürich, Zürich, Switzerland.

He is a Scientist in surveying, mapping and remote sensing with Wuhan University, Wuhan, China.

Dr. Li is a member of the Chinese Academy of Sciences and the Chinese Academy of Engineering. He is also a member of International Eurasia Academy of Sciences and International Academy of Astronautics. He was the recipient of the Honorary Member and the Brock Gold Medal in recognition of outstanding contributions to photogrammetry from the International Society for Photogrammetry and Remote Sensing.