Multilabel Remote Sensing Image Annotation With Multiscale Attention and Label Correlation

Rui Huang, *Member, IEEE*, Fengcai Zheng, *Student Member, IEEE*, and Wei Huang

***Abstract*—Deep-learning-based multilabel image annotation is receiving increasing attention in the field of remote sensing due to the great success of deep networks in single-label remote sensing image classification. Compared with those low-level features, the features extracted by the convolutional neural network (CNN) are more informative and can alleviate the problem of semantic gap. However, the CNN model tends to ignore the smaller objects when objects of different sizes exist in an image. In addition, how to efficiently leverage the correlation among multiple labels to enhance annotation performance remains an open issue. In this article, we propose an end-to-end deep learning framework for multilabel remote sensing image annotation. The framework is composed of a multiscale feature fusion module, a channel-spatial attention learning module, and a label correlation extraction module. The multiscale features from different layers of a CNN model are first fused and refined by using a channel-spatial attention mechanism. Then, the label correlation information is extracted from a label co-occurrence matrix and embedded into the multiscale attentive features to increase the discriminative ability of the resulting image features. The experiments on two benchmark datasets demonstrate the superiority of the proposed method in comparison with the state-of-the-art methods.**

***Index Terms*—Attention mechanism, convolutional neural network (CNN), label correlation, multilabel image annotation, multiscale features, remote sensing image.**

I. Introduction

With the rapid progress of sensor technology, a large number of remote sensing (RS) images have represented a major resource for land-cover monitoring, urban planning, disaster forecasting, and many more. RS image annotation, which is to associate one or several semantic labels with an image, can provide a comprehensive understanding of the image content. Compared with the pixelwise classification of RS images, image annotation just tells us whether the interesting objects exist in the image and thus helps to quickly acquire the images of interest. In recent years, the image scene classification technique [1] has been an active topic in the context of RS analysis. In particular, as an RS image usually involves multiple object classes and can be simultaneously assigned to different land-cover class labels, multilabel RS image annotation has received increasing attention and has been applied to many applications, such as atmospheric conditions classifications [2], [3] and land cover/land use investigation [4].

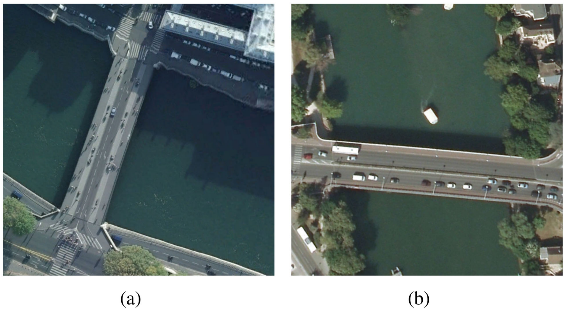


Fig. 1. Example images of the scene label and object labels. (a) Scene label: *bridge*. object labels: *cars, buildings, pavement, trees, water*. (b) Scene label: *bridge*. object labels: *buildings, cars, dock, grass, pavement, ship, tree, water*.

Fig. 1 illustrates the difference between single- and multilabel image annotation. Scene labels from single-label image annotation usually describe the whole picture with only one relatively general label, such as *river, forest,* and *bridge*. However, multilabel image annotation uses more specific object-level labels to represent the image, The object-level labels provide important cues for understanding a scene more deeply. For example, *pavement* has a large probability to appear when *cars* and *buildings* exist. In contrast, we cannot infer such clues from the scene label.

Although in multilabel RS image annotation, much effort has been spent to develop better scene understanding, there are many challenging problems. Two main issues are semantic feature representation and label correlation exploitation [5].

High-resolution RS images contain complex spatial and geometric information of objects with varying scale properties. The most used low-level features (color, texture, shape, etc.) cannot represent high-level semantics, and there is a well-known semantic gap between low-level features and high-level semantic concepts. Recently, since convolutional neural networks (CNNs) are able to extract informative features through adaptive image learning, deep learning models have become generic image descriptors. However, the representational power of CNNs needs to be improved especially in situations where objects of different sizes exist. In those cases, smaller objects tend to be misidentified.

In multilabel learning, label correlations are helpful to infer multiple object labels and, thus, improve the classification

Manuscript received March 10, 2021; revised June 3, 2021; accepted June 10, 2021. Date of publication June 22, 2021; date of current version July 22, 2021. This work was supported by grant from the National Natural Science Foundation of China under Grant 61772513. *(Corresponding author: Rui Huang.)*

The authors are with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: huangr@shu.edu.cn; 15858701908@163.com; lyxhw@shu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3091134

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

performance. Some works have been done to exploit label correlations, but how to efficiently make use of label correlations is still an open issue. In CNN-based image annotation systems, label correlations are often modeled by recurrent neural networks (RNNs). However, the chain propagation fashion of the RNN affects the efficiency of label prediction. Moreover, the RNN only considers the correlation between adjacent labels, while the correlation between nonadjacent labels is ignored [6].

To address the above problems, we propose an end-to-end deep learning framework for multilabel RS image annotation in this article. The framework consists of three modules involving multiscale feature fusion, channel-spatial attention learning, and label correlation extraction modules. To identify objects of various sizes, multiscale features from different layers of a CNN are fused, and the resulting feature maps are adaptively refined through the channel-spatial attention module. In the label correlation extraction module, a label co-occurrence matrix is constructed and concatenated with the refined image features from the channel-spatial attention module. Then, the joint embedding features are generated by several full connection (FC) layers. The main contributions of our work can be summarized as follows.

1) We propose an end-to-end deep learning framework for multilabel RS image annotation. The framework that consists of a multiscale feature fusion module, a channel-spatial attention learning module, and a label correlation extraction module has better performance than some state-of-the-art methods.
2) The framework uses a cascade fusion strategy to integrate the multiscale feature maps and a channel-spatial attention mechanism to refine the fused feature. The multiscale attentive feature has better representation ability for classification and can help to detect objects of different sizes.
3) The label correlation information is extracted from a label co-occurrence matrix and integrated with the refined image features through a two-step fusion, which can help improve generalization performance.

II. Related Works

In multilabel annotation, there are two key steps, namely, feature extraction and multilabel classification. In the feature extraction stage, researchers have developed different handcrafted features to describe object properties of color, texture, shape, and so on [7]–[11]. These global or local features are usually integrated by bag of visual words, which is an intermediate feature representation and can help to bridge the semantic gap [12]–[14]. However, the discriminative capability of low-level features is limited. In the classification stage, class labels are predicted for each instance based on the extracted features. Classical algorithms of ML-KNN [15] and Rank-SVM [16] do not perform well in the presence of high-dimensional image features. Recently, graph theory and sparse representation have been introduced to multilabel RS image annotation and retrieval and have shown promising performance [17]–[19].

With the important advances in deep learning theory, deep-learning-based algorithms have provided an attractive solution to the problem of multilabel image annotation. In these deep learning frameworks, CNN models are usually applied as image descriptors, and classification is done by using different networks. Wang *et al.* [20] proposed a CNN-RNN framework for multilabel image classification where the RNN model learns a joint image-label embedding from the CNN features for label prediction. Differently, Zeggada *et al.* [21] fed these features into a radial basis function neural network (RBFNN) with a multilabeling layer. In [22] and [23], autoencoder neural networks instead of CNNs acted as feature extractors, and canonical correlation analysis and a multilabel conditional random field were adopted for classification. In addition, some works introduce an attention mechanism into the CNN-RNN framework to improve network performance. Hua *et al.* [24] added a class attention learning layer to explore features with respect to each category. Sumbul and Demir [25] proposed a novel multiattention mechanism to learn scores for each local descriptor. In [26], a self-attention process was used to learn semantic dependencies and spatial relevance of features simultaneously. In the field of RS, Stivaktakis *et al.* [27] adopted a dynamic data augmentation to improve the performance of a CNN framework when a limited number of training samples are available. In [28], a two-branch neural network, which consists of an image branch and a label branch, was proposed to deal with RS image classification. In [29], dual-level semantic concepts were applied for multilabel RS image annotation, and an attention mechanism was introduced for salient object detection. Besides, semisupervised learning [30] and zero-shot learning [31] are adopted for solving the problem of high cost of data labeling. The influence of loss function [32], [33] on multilabel classification is also worth studying. In [32], label occurrence is calculated and introduced into the loss function to deal with the imbalance between positive and negative training samples. In [33], the binary cross-entropy function is improved by introducing scalable neighbor discriminative loss to embed a graph structure into the network.

Label correlation is an important cue for multilabel classification, but only a few works have considered exploiting the correlations in the CNN-based image annotation frameworks. In the CNN-RNN framework presented in [20], the memory mechanism of the RNN was used to predict labels in an ordered prediction path. To model label correlations, label co-occurrence matrices were used in [33]–[35]. Ji *et al.* [34] calculated a parameter based on the co-occurrence matrix and introduced it into the loss function. Zhang *et al.* [35] fed the label co-occurrence matrix into two convolutional layers and two fully connected layers to learn label correlation. With the emergence of graph convolution network, a label graph structure is explored in [36]–[39].

III. Methodology

As illustrated in Fig. 2, the overall network architecture of our proposed approach is composed of three components: a multiscale feature fusion module, a channel-spatial attention learning module, and a label correlation extraction module. We will describe the three modules in detail.

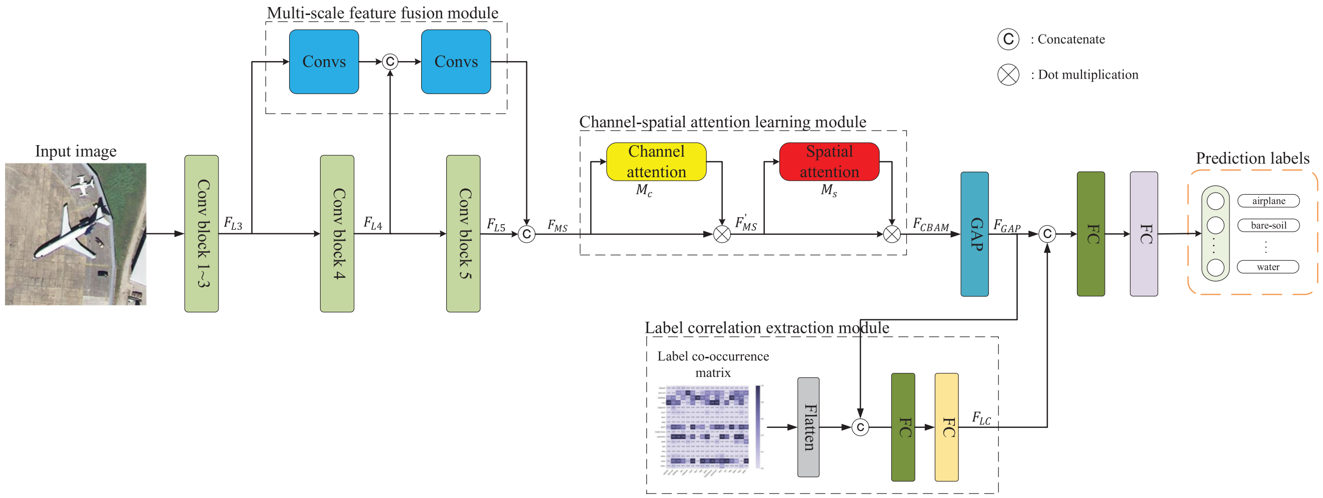


Fig. 2. Framework of our method for multilabel RS image annotation. It consists of a multiscale feature fusion module, a channel-spatial attention learning module, and a label correlation extraction module.

A. Multiscale Feature Fusion

There are different-sized objects in RS images. Furthermore, the size of the same object class may change owing to the distinct spatial resolution of images and their inherent appearance. The top-level features of CNNs with rich semantic information and larger receptive fields are useful to identify larger objects. But the feature extraction mechanism tends to ignore the information of small-sized objects. Since the features of the first few layers have rich spatial information and smaller receptive fields, which help to find smaller objects, the combination of outputs of different layers is a natural solution to recognize objects of varying size [40]–[42]. The multilevel features present an effective image representation with different scale information.

In the proposed method, we choose VGG16 to extract image features. The feature maps of *Conv block 3*, *Conv block 4*, and *Conv block 5* of VGG16 can act as three-level image descriptors. First, the low-level features are convolved by a standard convolution layer and concatenated with the mid-level features. Then, the obtained features are convolved and concatenated with the high-level features. Through the cascade strategy of fusing multiscale features, the geometric and semantic properties of objects are kept simultaneously. Supposing that the input image is denoted as X, the three-level features are formulated as follows:

$$\begin{aligned} F_{\mathrm{L3}} &= VGG_{\mathrm{conv3}}(X) \\ F_{\mathrm{L4}} &= VGG_{\mathrm{conv4}}(X) \\ F_{\mathrm{L5}} &= VGG_{\mathrm{conv5}}(X). \end{aligned} \tag{1}$$

These features at three scales will be concatenated in channel dimension to obtain the multiscale feature F_{MS}. The process can be summarized as follows:

$$F_{\mathrm{MS}} = [F_{\mathrm{L5}}; g([F_{\mathrm{L4}}; g(F_{\mathrm{L3}})])] \tag{2}$$

where $g(\cdot)$ denotes a composite function of three consecutive operations including a 2 × 2 convolution with a stride of 2, batch normalization (BN), and a rectified linear unit (ReLU). The number of filters of the convolution layer is 64 and 128 for F_{L3} and F_{L4}, respectively. In our case, the sizes of F_{L3}, F_{L4}, F_{L5}, and F_{MS} are 32 × 32 × 256, 16 × 16 × 512, 8 × 8 × 512, and 8 × 8 × 640, respectively.

B. Channel-Spatial Attention Learning

To exploit both spatial and channelwise attention, we adopt an attention learning mechanism named convolutional block attention module (CBAM) [43]. The CBAM is a simple and efficient attention module and has shown superior performances on some benchmark image classification datasets. It consists of a channel attention module and a spatial attention module. The channel attention module learns to find "what" to focus on and the spatial attention module concentrates on discovering "where" is attractive. First, the channel attention map is learned from the input feature map by using max-pooling and average-pooling with a shared multilayer perceptron (MLP) network. The channel attention values are broadcasted along the spatial dimension of the input feature map through elementwise multiplication. Second, the spatial attention map is learned from the channel-refined feature map by utilizing max-pooling and average-pooling with a convolution layer. Finally, the refined feature map is obtained by broadcasting the spatial attention information along the channel dimension of the channel-refined feature map.

In our case, the multiscale feature map F_{MS} is fed into the CBAM, and the channel attention map is first obtained as follows:

$$\begin{aligned} M_c(F_{\mathrm{MS}}) = \sigma\,&(\mathrm{Mlp}\,(\mathrm{AvgPool}\,(F_{\mathrm{MS}})) \\ &+ \mathrm{Mlp}\,(\mathrm{MaxPool}\,(F_{\mathrm{MS}}))) \end{aligned} \tag{3}$$

where σ denotes the sigmoid function, Mlp refers to the shared MLP with one hidden layer, and $AvgPool$ and $MaxPool$ denote average pooling and max-pooling operations, respectively.

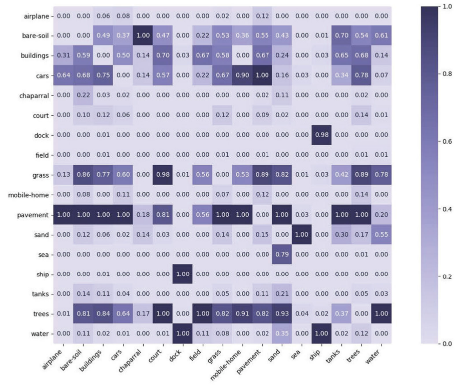


Fig. 3. Label co-occurrence matrix of the UCM multilabel dataset.

Then, the multiscale feature map F_{MS} is refined by the channel attention map $M_c(F_{\mathrm{MS}})$ along the spatial dimension. The process can be formulated as

$$F'_{\mathrm{MS}} = M_c(F_{\mathrm{MS}}) \otimes F_{\mathrm{MS}} \tag{4}$$

where $\otimes$ denotes elementwise multiplication. In the multiplication process, the attention value of the channel attention is broadcasted. Subsequently, the spatial attention map is learned from the above channel-refined feature map as

$$\begin{aligned} M_s(F'_{\mathrm{MS}}) = \sigma\,(&\mathrm{Conv}\,([\mathrm{AvgPool}(F'_{\mathrm{MS}});\\ &\mathrm{MaxPool}\,(F'_{\mathrm{MS}})])) \end{aligned} \tag{5}$$

where σ denotes the sigmoid function and Conv represents a convolution operation with the filter size of 7 × 7. The final refined feature map is obtained as follows:

$$F_{\mathrm{CBAM}} = M_s(F'_{\mathrm{MS}}) \otimes F'_{\mathrm{MS}}. \tag{6}$$

Through the CBAM, important features are focused and unnecessary features are suppressed. Furthermore, to exploit the global contextual information, the refined feature map F_{CBAM} is squeezed into a channelwise descriptor F_{GAP} by using the global average pooling (GAP) [44].

C. Label Correlation Extraction

We use the label co-occurrence matrix to model the label correlation. Let $C \in R^{q\times q}$ be the matrix, where q is the number of class labels. Each entry $C(i, j)$ in the matrix is the number of the ith label and the jth label appearing together ($1 \le i, j \le q$, $i \ne j$). The co-occurrence matrix is further normalized to the range [0, 1] as follows:

$$C'(i,j) = \begin{cases} 0, & i = j \\ \frac{C(i,j)-\mathrm{Min}(C(\cdot,j))}{\mathrm{Max}(C(\cdot,j))-\mathrm{Min}(C(\cdot,j)),} & i \ne j \end{cases}. \tag{7}$$

Fig. 3 shows the normalized label co-occurrence matrix of the UCM multilabel dataset. The depth of the color indicates the possibility of two labels appearing at the same time. The darker the color is, the greater the possibility of the co-occurrence is, and *vice versa*. It is noted that the matrix shows global label correlation without considering local label correlation, which is helpful to boost the classification.

Different from [33]–[35], the information in the label co-occurrence matrix is integrated into the refined image features through two fusion steps. In the first step, we flatten the normalized label co-occurrence matrix L'_{corr} into an $N \times 1$ vector and concatenate the vector with F_{GAP} obtained in the previous subsection. After *BN*, the concatenated vector is fed into two FC layers followed by ReLU, and a new feature vector F_{LC} is obtained. F_{LC} is an adaptive label correlation feature, which can represent the global label correlation to some extent and change with different image samples. F_{LC} contains the label correlation information, which is enhanced by the image features. The number of units in the FC layers is 64 and 128, respectively. In the second step, the image descriptor vector F_{GAP} is concatenated with F_{LC} and fed into two FC layers. The activation functions of the two layers are ReLU and sigmoid, respectively. The final image feature vector enhanced with label correlation is obtained, and label prediction is then estimated based on it. The number of units of the first FC layer is set to 128. Dropout with a probability equal to 0.5 is added before each FC layer to prevent overfitting. This process is summarized as follows:

$$\begin{aligned} F_{\mathrm{LC}} &= \mathrm{ReLU}\,(\mathrm{Fc}\,(\mathrm{ReLU}\,(\mathrm{Fc}\,([\mathrm{Flatten}\,(C')\,; F_{\mathrm{GAP}}])))) \\ \hat{y} &= \sigma\,(\mathrm{Fc}\,(\mathrm{ReLU}\,(\mathrm{Fc}\,([L_{\mathrm{GAP}}; F_{\mathrm{LC}}])))) \end{aligned} \tag{8}$$

where $\hat{y}$ is the predicted label.

IV. Experiments

In this section, we conduct experiments to investigate the performance of our proposed method on two multilabel RS image datasets. The experiments consist of two parts: evaluating different modules of the proposed method and comparing the proposed method with some other state-of-the-art multilabel image annotation methods.

A. Datasets

1) UCM Multilabel Dataset: The UCM dataset [12] is extracted from the aerial imagery contributed by the U.S. Geological Survey National Map. It contains 2100 images, which are divided into 21 categories at the scene level. These categories are corresponding 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 [17], 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 assigned with one or more (up to seven) labels. Fig. 4 shows some examples of the UCM multilabel dataset, and Fig. 6(a) gives the details of the dataset at the object level.

2) AID Multilabel Dataset: The AID dataset [45] was published in 2017 by Wuhan University. It consists of 30 categories at the scene level. Each category has 220–420 images with a size of 600 × 600 × 3 and a spatial resolution varying from 0.5 to 8 m. The multilabel dataset [46] generated from the AID dataset contains 3000 images with multiple object labels. There are a total of 17 object-level labels: bare-soil, airplane, building, car, chaparral, court, dock, field, grass, mobile home, pavement,

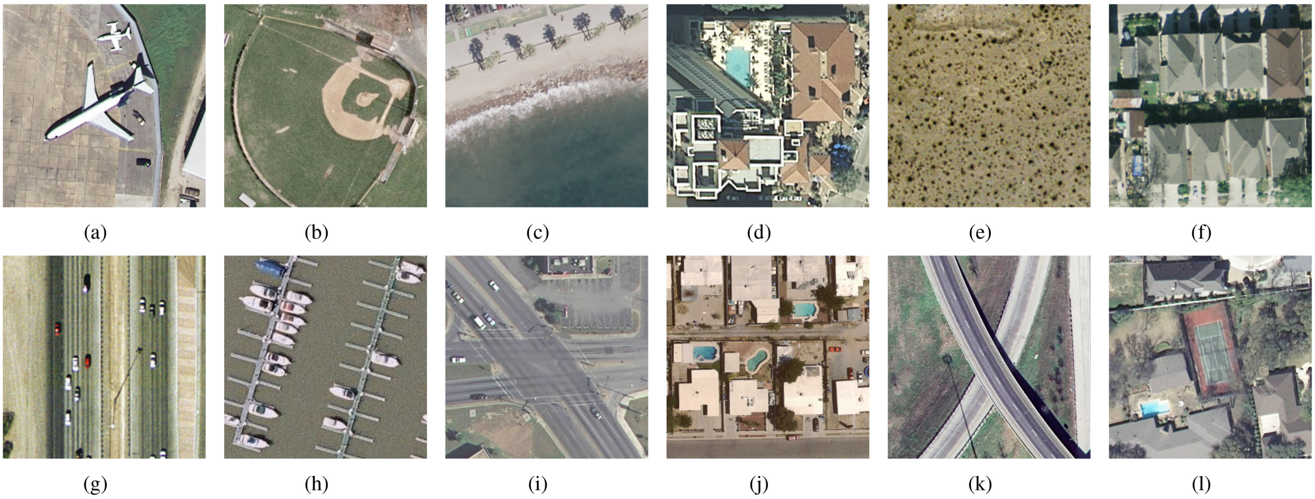


Fig. 4. Samples of the UCM multilabel dataset from different scenes. The labels at the scene level and object level are as follows. (a) Airplane: *airplane, buildings, cars, grass*, and *pavement*. (b) Baseball diamond: *bare-soil, buildings, grass*, and *pavement*. (c) Beach: *cars, pavement, sand, sea*, and *trees*. (d) Buildings: *bare-soil, buildings*, and *trees*. (e) Chaparral: *bare-soil* and *chaparral*. (f) Dense residential: *buildings, grass, pavement*, and *trees*. (g) Freeway: *bare-soil, cars*, and *pavement*. (h) Harbor: *dock, ship*, and *water*. (i) Intersection: *bare-soil, buildings, cars, grass, pavement*, and *trees*. (j) Medium residential: *bare-soil, buildings, cars, grass, pavement*, and *trees*. (k) Overpass: *bare-soil, pavement*, and *grass*. (l) Tennis court: *bare-soil, buildings, cars, court, grass*, and *trees*.

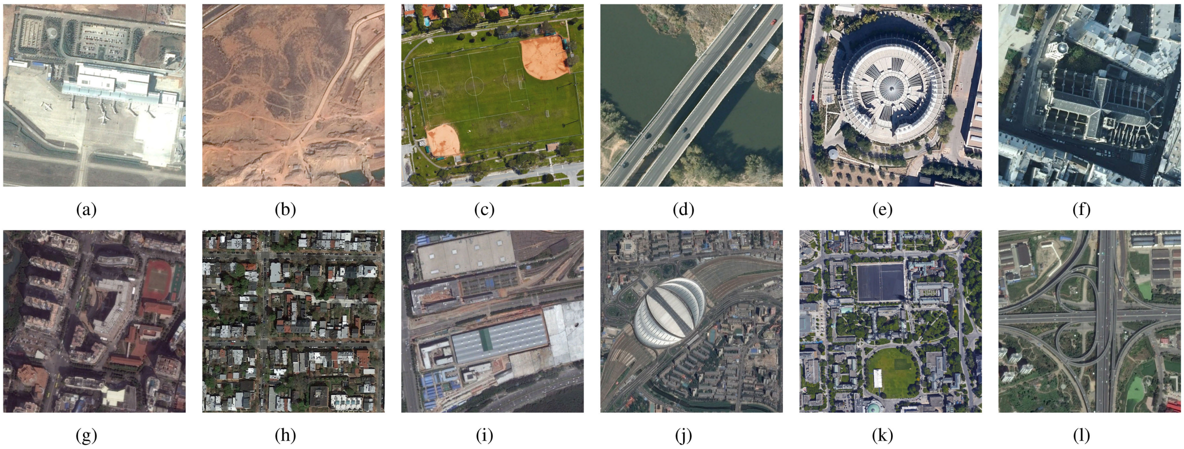


Fig. 5. Samples of AID multilabel dataset from different scenes. The labels at the scene level and object level are as follows. (a) Airplane: *airplane, buildings, cars, bare-soil, grass, pavement*, and *trees*. (b) Bare land: *bare-soil, buildings, cars, pavement, trees*, and *water*. (c) Baseball field: *bare-soil, buildings, cars, court, grass, pavement*, and *trees*. (d) Bridge: *bare-soil, cars, grass, pavement, trees*, and *water*. (e) Center: *bare-soil, buildings, cars, grass, pavement*, and *trees*. (f) Church: *buildings, cars, grass, pavement*, and *trees*. (g) Commercial: *buildings, cars, court, grass, pavement*, and *trees*. (h) Dense residential: *buildings, cars, grass, pavement*, and *trees*. (i) Industrial: *bare-soil, buildings, cars, grass, pavement*, and *trees*. (j) Railway station: *bare-soil, buildings, cars, grass, pavement, trees*, and *water*. (k) School: *bare-soil, buildings, cars, grass, pavement*, and *trees*. (l) Viaduct: *bare-soil, buildings, cars, grass, pavement*, and *trees*.

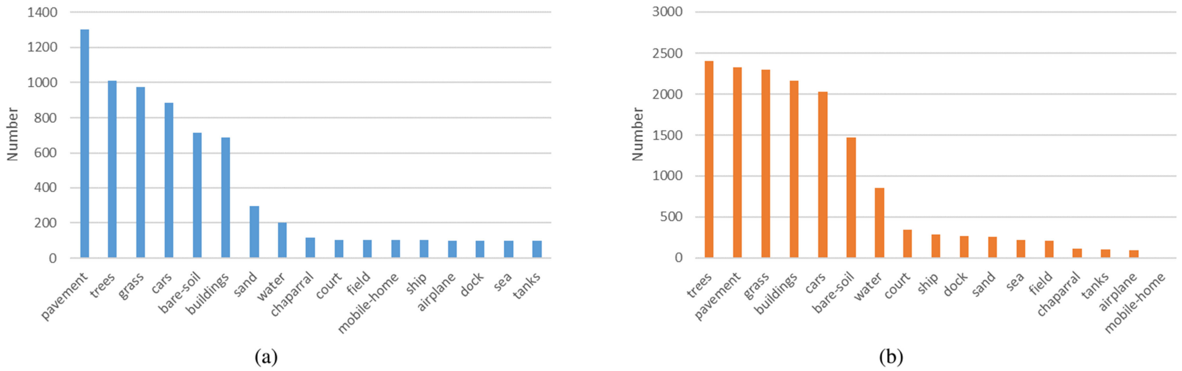


Fig. 6. Number of images per object-level label in the two datasets. (a) UCM multilabel dataset. (b) AID multilabel dataset.

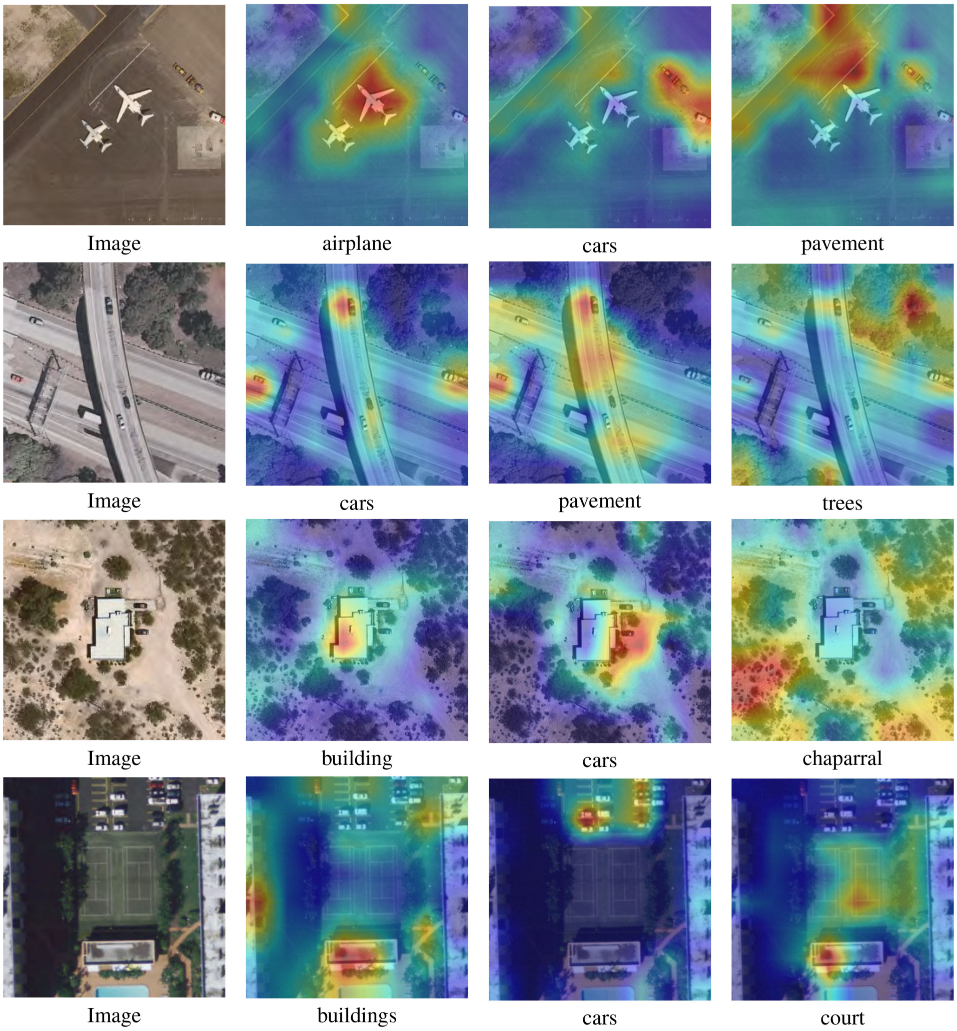


Fig. 7. Visualization results of four images from the UCM dataset. Red indicates the areas that our model focuses on when predicting the corresponding labels.

TABLE I
SOME STATISTICS OF EXPERIMENTAL DATASETS

Data set	Samples	Object labels	Scene labels	LC	LD
UCM	2100	17	21	3.334	0.196
AID	3000	17	30	5.152	0.303

sand, sea, ship, tank, tree, and water. Fig. 5 shows some visual examples of the AID multilabel dataset. Fig. 6(b) gives the number of images associated with each object-level label.

Some statistics for the two dataset are listed in Table I. Here, LC denotes the label cardinality, which calculates the average number of class labels associated with each sample, and LD represents the label density, which is the cardinality normalized by the number of labels.

B. Evaluation Metrics

To investigate the performance of our proposed method from multiple perspectives, we choose three example-based metrics and three label-based metrics, including precision, recall, and F1 score for evaluation. Suppose that there is a test set $\mathbf{D} = \{(\mathbf{x}_i, \mathbf{y}_i) | 1 \le i \le n\}$, where the binary vector $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iq}]^T \in \{0,1\}^q$ is the ground-truth label of the ith test sample. $y_{ij} = 1$ and $y_{ij} = 0$ correspond to the presence and absence of the jth label for sample $\mathbf{x}_i$, respectively. Let $\hat{\mathbf{y}}_i = [\hat{y}_{i1}, \hat{y}_{i2}, \ldots, \hat{y}_{iq}]^T \in \{0,1\}^q$ denote the predicted label vector for sample $\mathbf{x}_i$, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]^T = [\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_q] \in R^{n \times q}$ denote the ground truth label matrix, and $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_n]^T = [\hat{\mathbf{l}}_1, \hat{\mathbf{l}}_2, \ldots, \hat{\mathbf{l}}_q] \in R^{n \times q}$ denote the predicted label matrix. Then, the six metrics can be computed as follows:

$$P_E = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{\mathbf{y}}_i \cap \mathbf{y}_i|}{|\hat{\mathbf{y}}_i|}, \quad P_L = \frac{1}{q} \sum_{j=1}^{q} \frac{\left|\hat{\mathbf{l}}_j \cap \mathbf{l}_j\right|}{\left|\hat{\mathbf{l}}_j\right|}$$

$$R_E = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{\mathbf{y}}_i \cap \mathbf{y}_i|}{|\mathbf{y}_i|}, \quad R_L = \frac{1}{q} \sum_{j=1}^{q} \frac{\left|\hat{\mathbf{l}}_j \cap \mathbf{l}_j\right|}{|\mathbf{l}_j|}$$

$$F1_E = \frac{2P_E R_E}{P_E + R_E}, \quad F1_L = \frac{2P_L R_L}{P_L + R_L} \tag{9}$$

TABLE II
COMPONENT EFFECTIVENESS EVALUATION OF THE PROPOSED METHOD ON THE UCM DATASET (MEAN% ± STD%)

Models	$P_E(\%)$	$R_E(\%)$	$F1_E(\%)$	$P_L(\%)$	$R_L(\%)$	$F1_L(\%)$
VGG16	82.27 ± 1.77	82.86 ± 1.94	82.56 ± 1.75	81.73 ± 2.38	73.61 ± 2.36	77.45 ± 2.16
VGG16+attention	86.24 ± 1.05	87.65 ± 1.26	86.93 ± 1.07	89.07 ± 2.46	83.29 ± 1.96	86.07 ± 2.01
VGG16+attention+multi-scale fusion	90.34 ± 1.27	92.63 ± 0.71	91.46 ± 0.78	**93.83 ± 1.01**	91.85 ± 2.04	92.82 ± 1.24
Proposed Method	**90.54 ± 1.11**	**92.98 ± 1.00**	**91.74 ± 0.77**	93.73 ± 1.04	**92.75 ± 1.34**	**93.23 ± 0.82**

where the subscripts E and L indicate that the corresponding metric is obtained from the perspective of examples or labels.

In addition, the hamming loss is calculated to evaluate the number of misclassified instance-label pairs, i.e., a relevant label is not predicted or an irrelevant label is predicted

$$\mathrm{HL} = \frac{1}{nq}\sum_{i=1}^{n} |\hat{\mathbf{y}}_i \Delta \mathbf{y}_i| \tag{10}$$

where Δ denotes the symmetric difference between two sets.

C. Implementation Details

In our proposed method, VGG16 is used to extract features from images. The weights of VGG16 are pretrained on ImageNet [47]. The other components are initialized by the Xavier uniform initializer [48]. We train the deep neural network with the Adam optimizer [49], and the learning rate is initially set to 0.001 with a decay factor of 0.9 every ten epochs. Other parameters of the optimizer are set as recommended: $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Binary cross-entropy loss (BCE) function is adopted as the loss function and is calculated as follows:

$$L_{\mathrm{BCE}} = -\sum_{i=1}^{n}\sum_{j=1}^{q} y_{ij}\log \hat{y}_{ij} + (1 - y_{ij})\log(1 - \hat{y}_{ij}). \tag{11}$$

The model is implemented on Tensorflow and is trained on NVIDIA GeForce GTX 1080 Ti GPU. We train the model for 100 epochs with a batch size of 32. The original images are resized to 256 × 256 × 3. To reduce the risk of overfitting, we adopt various data augmentation technique, where images can be flipped, rotated, and shifted.

D. Evaluation for Model Components

Our proposed multilabel annotation framework is composed of three modules, including multiscale feature fusion, channel-spatial attention learning, and label correlation extraction modules. To evaluate the effectiveness of different modules, we compare our method with the following approaches.

1) *VGG16:* The original CNN is used as the image descriptor and followed by two FC layers for classification.
2) *VGG16 + attention:* The image features generated from VGG16 are refined by the attention mechanism of CBAM and then fed into the FC layers.
3) *VGG16 + attention + multiscale fusion:* The multiscale features from different layers of VGG16 are fused before they are refined by the attention module.

We run all these methods on the UCM multilabel dataset and the AID multilabel dataset. In each dataset, we randomly select 80%, 10%, and 10% of images for training, validation, and test, respectively. The final results are averaged over ten realizations and listed in Tables II and III . The best results are shown in bold. For all metrics, the higher the value, the better the evaluation is.

From the comparison of VGG16 and VGG16 + attention, we can conclude that the CBAM significantly improves the multilabel RS image annotation performances. The six metric values rise 6.46% and 7.97% on average on the UCM and AID datasets, respectively. Particularly, R_L values increase by 9.68% on the UCM dataset and 13.7% on the AID dataset.

The effectiveness of multiscale feature fusion can be evaluated through the comparison between VGG16 + attention and VGG16 + attention + multiscale fusion. By introducing multiscale feature information, the latter method achieves better performances in terms of all the six metrics. Specifically, the score of R_L obtains increments of 8.56% on the UCM dataset and 8.14% on the AID dataset.

The proposed method generally outperforms the other three methods, which indicates the effectiveness of integrating the three modules. Compared with the three methods, the average gains are 12.42%, 5.95%, and 0.34% on the UCM dataset, 13.26%, 5.30%, and 0.41% on the AID dataset, respectively.

E. Comparisons With State-of-the-Art Methods

For a comprehensive evaluation, we compare the proposed method with the following state-of-the-art multilabel annotation methods.

1) *ML-KNN:* This method uses VGG16 for feature extraction and ML-KNN [15] for classification.
2) *Gardner [3]:* This method uses VGG16 for feature extraction and three FC layers with dropout for classification.
3) *CNN-RNN [20]:* This method combines CNN and RNN networks. The CNN is used to extract visual features and the RNN is used to model label correlation.
4) *RBFNN [21]:* This method uses the pretrained VGG16 for feature extraction and the RBFNN for classification.
5) *Stivaktakis [27]:* This method employs a dynamic data augmentation technique for CNN architecture to solve the problem of a small amount of data.
6) *Zhu [29]:* This is a deep learning framework using dual-level semantic concepts, where scene labels are used to guide multilabel classification.
7) *ML-GCN [36]:* This method uses the pretrained Resnet-101 for feature extraction and graph convolution neural network for label correlation extraction.

TABLE III
COMPONENT EFFECTIVENESS EVALUATION OF THE PROPOSED METHOD ON THE AID DATASET (MEAN% ± STD%)

Models	$P_E(\%)$	$R_E(\%)$	$F1_E(\%)$	$P_L(\%)$	$R_L(\%)$	$F1_L(\%)$
VGG16	80.44 ± 1.84	80.38 ± 1.91	80.41 ± 1.81	71.98 ± 3.02	52.38 ± 2.27	60.60 ± 2.01
VGG16+attention	86.73 ± 1.29	86.04 ± 1.53	86.38 ± 1.36	77.51 ± 4.20	66.08 ± 1.62	71.26 ± 1.50
VGG16+attention+multi-scale fusion	90.63 ± 0.89	90.44 ± 0.79	90.53 ± 0.67	80.37 ± 2.15	74.22 ± 1.67	77.15 ± 1.40
Proposed Method	**91.03 ± 1.31**	**91.88 ± 1.06**	**91.44 ± 0.35**	**81.37 ± 2.56**	**74.60 ± 1.48**	**77.79 ± 0.69**

TABLE IV
COMPARISONS OF THE PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON THE UCM DATASET (MEAN% ± STD%)

Models	$P_E(\%)$	$R_E(\%)$	$F1_E(\%)$	$P_L(\%)$	$R_L(\%)$	$F1_L(\%)$	HL
ML-KNN	86.82 ± 1.09	88.16 ± 0.96	87.49 ± 0.99	87.60 ± 1.94	87.68 ± 2.27	87.63 ± 1.84	0.06 ± 0.00
Gardner [3]	88.29 ± 0.57	84.51 ± 1.29	86.35 ± 0.74	88.66 ± 2.97	79.22 ± 1.99	83.64 ± 1.83	0.06 ± 0.00
CNN-RNN [20]	74.79 ± 2.50	79.88 ± 2.60	75.12 ± 2.22	65.70 ± 4.23	62.47 ± 4.15	61.22 ± 3.80	0.11 ± 0.01
Stivaktakis [27]	85.16 ± 0.93	87.45 ± 1.21	86.29 ± 0.68	87.96 ± 1.02	84.48 ± 0.92	86.18 ± 0.81	0.06 ± 0.00
RBFNN [21]	88.37 ± 1.27	87.75 ± 1.55	88.05 ± 1.10	92.40 ± 1.43	87.79 ± 1.57	90.02 ± 1.28	0.05 ± 0.00
Zhu [29]	**91.75 ± 0.83**	91.65 ± 0.76	90.62 ± 0.62	92.96 ± 0.98	92.60 ± 0.52	92.66 ± 0.47	0.04 ± 0.00
ML-GCN [36]	90.03 ± 1.47	90.70 ± 3.05	90.25 ± 2.00	92.46 ± 1.50	90.17 ± 5.17	91.21 ± 3.00	0.04 ± 0.00
Proposed Method	90.54 ± 1.11	**92.98 ± 1.00**	**91.74 ± 0.77**	**93.73 ± 1.04**	**92.75 ± 1.34**	**93.23 ± 0.82**	**0.04 ± 0.00**

TABLE V
COMPARISONS OF THE PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON THE AID DATASET (MEAN% ± STD%)

Models	$P_E(\%)$	$R_E(\%)$	$F1_E(\%)$	$P_L(\%)$	$R_L(\%)$	$F1_L(\%)$	HL
ML-KNN	83.82 ± 0.90	84.65 ± 0.86	84.23 ± 0.84	64.96 ± 2.90	66.47 ± 2.26	65.69 ± 2.39	0.09 ± 0.00
Gardner [3]	86.15 ± 1.12	82.46 ± 1.36	84.26 ± 0.83	76.30 ± 1.31	57.87 ± 1.16	65.80 ± 0.75	0.08 ± 0.00
CNN-RNN [20]	84.06 ± 2.73	85.01 ± 1.44	82.34 ± 1.60	56.90 ± 5.63	55.72 ± 3.64	54.11 ± 3.81	0.10 ± 0.01
Stivaktakis [27]	87.69 ± 0.56	87.92 ± 0.61	87.79 ± 0.28	73.40 ± 2.76	66.45 ± 1.80	69.74 ± 1.96	0.07 ± 0.00
RBFNN [21]	88.52 ± 1.13	86.56 ± 1.31	87.52 ± 0.98	78.78 ± 3.59	64.16 ± 1.82	70.70 ± 2.29	0.07 ± 0.00
Zhu [29]	89.72 ± 0.44	88.41 ± 0.65	87.49 ± 0.18	80.89 ± 1.84	74.08 ± 3.11	76.50 ± 2.39	0.07 ± 0.00
ML-GCN [36]	89.69 ± 1.99	89.48 ± 1.67	89.58 ± 1.57	78.91 ± 2.55	**75.06 ± 4.29**	76.90 ± 3.18	0.06 ± 0.00
Proposed Method	**91.03 ± 1.31**	**91.88 ± 1.06**	**91.44 ± 0.35**	**81.37 ± 2.56**	74.60 ± 1.48	**77.79 ± 0.69**	**0.06 ± 0.00**

In the experiments, we randomly divide the UCM dataset and the AID dataset such that 80% images are training samples, 10% images are validation samples, and 10% images are test samples, respectively. On each dataset, each method is run ten times to get the average performance. All seven metrics are computed, and the average metric values are listed in Tables IV and V. The best results are shown in bold. For the hamming loss, the lower the value, the better the evaluation.

From the tables, we can see that the proposed method achieves the best performance on the whole. On the UCM multilabel dataset, shown in Table IV, the proposed method ranks in first place among the seven comparing methods except that the Zhu method [29] outperforms it in the P_E metric. Among the methods, CNN-RNN performs poorly because the chain propagation fashion of the RNN is not good at utilizing the label correlation. The methods ML-KNN, Gardne [3], and Stivaktakis [27] show comparable performances, which are inferior to the methods of RBFNN [21], ML-GCN [36], and Zhu [29]. On the AID multilabel dataset, shown in Table V, a decrease in the performance of each method can be observed, especially in the three label-based metrics. It is because the AID dataset is more challenging for classification than the UCM dataset. However, the proposed method also competes with the other methods in all metrics except the R_L metric. The above experimental results suggest that our proposed method can significantly improve the multilabel RS image annotation performances.

F. Per-Class Case Studies

Tables VI and VII show the class-specific annotation results (F1 score) of the UCM dataset and the AID dataset, respectively. The best results in each category are shown in bold. As shown in Table VI, the results of the proposed method outperform the compared methods in most classes (11/17 in terms of F1 scores). For all the 17 categories except for the bare-soil class, the proposed method obtains more than 83% on the F1 score. Especially, the F1 score of the proposed method reaches 100% on the classes of airplane, dock, and ship. As shown in Table VII, the results of the proposed method outperform the compared methods in half of classes (9/17 in terms of F1 scores). In particular, the F1 score of the proposed method in the class airplane is significantly ahead of the second-best method by

TABLE VI
PER-CLASS F1 SCORES FROM EIGHT METHODS ON UCM DATASET (MEAN% ± STD%)

Labels	ML-KNN	Gardner [3]	CNN-RNN [20]	Stivaktakis [27]	RBFNN [21]	Zhu [29]	ML-GCN [36]	Proposed Method
airplane	91.25 ± 5.26	91.81 ± 2.92	8.16 ± 10.44	96.09 ±1.30	99.70 ± 0.91	100.00 ± 0.00	95.38 ± 2.73	**100.00** ± **0.00**
baresoil	73.33 ± 4.45	65.88 ± 3.95	51.07 ± 6.17	74.17 ± 0.92	70.83 ± 2.45	**78.23** ± **3.6**4	75.30 ± 4.59	76.95 ± 1.56
buildings	85.42 ± 2.06	82.56 ± 1.81	71.95 ± 3.22	79.12 ± 1.60	86.65 ± 1.96	87.98 ± 2.12	83.56 ± 2.84	**89.79** ± **1.03**
cars	81.30 ± 3.20	83.02 ± 1.40	80.81 ± 2.33	83.24 ± 0.52	84.74 ± 3.20	85.52 ± 1.47	88.74 ± 1.75	**89.46** ± **1.39**
chaparral	99.05 ± 1.84	92.53 ± 1.73	90.17 ± 6.02	91.88 ± 3.87	98.58 ± 1.77	93.59 ± 1.57	95.02 ± 5.20	**99.57** ± **1.24**
court	65.51 ± 10.55	19.17 ± 10.09	0.00 ± 0.00	56.03 ± 4.82	81.79 ± 9.91	87.05 ± 3.99	86.23 ± 16.03	**89.28** ± **4.09**
dock	99.05 ± 1.84	99.60 ± 1.20	90.42 ± 9.14	100.00 ± 0.00	99.26 ± 1.49	99.52 ± 1.43	99.63 ± 1.06	**100.00** ± **0.00**
field	96.66 ± 3.39	**99.23** ± **2.31**	75.87 ± 6.70	96.17 ± 5.04	95.05 ± 6.99	94.37 ± 1.73	98.75 ± 3.58	87.53 ± 1.33
grass	83.82 ± 2.40	81.74 ± 1.74	75.87 ± 4.20	84.87 ± 1.41	82.27 ± 3.59	**89.13** ± **1.26**	86.64 ± 2.73	88.33 ± 1.45
mobilehome	87.24 ± 7.00	82.09 ± 7.61	24.85 ± 21.27	87.95 ± 6.54	94.86 ± 5.70	**100.00** ± **0.00**	84.78 ± 8.26	95.36 ± 3.98
pavement	90.78 ± 1.55	**94.23** ± **0.43**	89.18 ± 1.64	89.93 ± 0.85	91.91 ± 1.57	93.29 ± 0.81	91.76 ± 0.95	92.77 ± 0.86
sand	80.65 ± 4.16	73.01 ± 2.29	55.23 ± 11.81	74.50 ± 3.62	84.14 ± 3.75	88.86 ± 2.52	90.20 ± 5.93	**93.63** ± **1.98**
sea	89.00 ± 7.76	96.80 ± 1.60	80.35 ± 10.30	94.12 ± 0.00	97.96 ± 2.67	98.70 ± 3.91	97.45 ± 3.23	**99.52** ± **1.36**
ship	99.05 ± 1.84	96.80 ± 1.60	87.17 ± 6.18	100.00 ± 0.00	98.78 ± 1.89	98.66 ± 2.89	99.63 ± 1.06	**100.00** ± **0.00**
tanks	77.65 ± 12.84	81.69 ± 12.76	8.26 ± 11.18	74.34 ± 7.39	87.15 ± 9.62	**95.32** ± **2.88**	87.55 ± 13.47	83.40 ± 9.98
trees	84.21 ± 2.90	87.60 ± 1.08	77.70 ± 2.90	83.96 ± 1.56	85.73 ± 1.74	87.86 ± 1.87	88.13 ± 1.26	**89.59** ± **1.74**
water	91.23 ± 7.09	83.79 ± 3.87	66.53 ± 10.29	89.13 ± 2.98	93.10 ± 2.32	97.06 ± 2.36	94.45 ± 3.99	**94.60** ± **2.46**

TABLE VII
PER-CLASS F1 SCORES FROM SIX METHODS ON AID DATASET (MEAN% ± STD%)

Labels	ML-KNN	Gardner [3]	Stivaktakis [27]	RBFNN [21]	ML-GCN [36]	Proposed Method
airplane	65.05 ± 11.27	75.65 ± 8.18	72.75 ± 8.37	76.97 ± 14.62	80.34 ± 16.50	**92.36** ± **4.53**
baresoil	71.91 ± 1.84	71.44 ± 1.33	74.27 ± 1.87	73.74 ± 2.02	75.36 ± 4.13	**77.98** ± **1.59**
buildings	94.05 ± 1.35	94.53 ± 0.39	93.40 ± 0.46	95.41 ± 0.88	94.17 ± 0.41	**95.65** ± **0.48**
cars	90.23 ± 1.05	93.10 ± 0.74	92.77 ± 0.47	92.73 ± 1.05	92.85 ± 1.38	**94.43** ± **0.59**
chaparral	28.03 ± 10.36	0.00 ± 0.00	9.45 ± 9.48	22.31 ± 16.71	34.37 ± 19.95	**41.03** ± **6.28**
court	49.31 ± 6.17	22.52 ± 8.63	57.98 ± 3.14	54.85 ± 6.48	**67.37** ± **8.39**	58.93 ± 4.56
dock	57.54 ± 10.77	61.75 ± 5.38	62.60 ± 2.80	68.42 ± 10.63	**70.52** ± **6.25**	68.19 ± 3.17
field	59.78 ± 7.67	58.73 ± 3.64	48.06 ± 4.29	67.52 ± 8.07	73.38 ± 5.99	**76.96** ± **2.50**
grass	91.98 ± 0.95	92.75 ± 0.9 5	92.45 ± 0.72	93.46 ± 1.23	94.00 ± 1.10	**95.43** ± **0.40**
mobilehome	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
pavement	95.82 ± 1.38	96.23 ± 0.70	95.41 ± 0.30	96.76 ± 0.41	97.18 ± 0.94	**98.07** ± **0.21**
sand	76.97 ± 6.42	55.85 ± 4.15	78.40 ± 3.35	83.05 ± 6.12	89.78 ± 3.54	**90.56** ± **0.99**
sea	72.57 ± 9.67	58.83 ± 5.47	88.72 ± 4.32	81.30 ± 6.23	**89.20** ± **6.43**	89.00 ± 3.62
ship	56.51 ± 11.20	67.78 ± 4.18	68.75 ± 4.35	69.29 ± 5.67	**71.50** ± **5.56**	69.54 ± 4.98
tanks	64.07 ± 8.66	72.43 ± 9.96	73.37 ± 4.62	**92.14** ± **4.99**	91.67 ± 7.08	79.93 ± 4.37
trees	93.53 ± 1.15	94.89 ± 0.53	**94.91** ± **0.55**	94.36 ± 1.51	94.31 ± 0.82	94.50 ± 0.28
water	58.93 ± 4.41	94.89 ± 3.37	63.77 ± 2.75	62.48 ± 4.21	**75.62** ± **2.70**	71.94 ± 1.13

12.02%. However, the F1 score of each method on the AID dataset is generally lower than on the UCM dataset, especially for the classes of chaparral and mobile-home. The possible reason lies in the fact that the samples with the two labels in the AID dataset are relatively few [as illustrated in Fig. 6(b)]. The networks cannot be fully trained due to the lack of training samples. In addition, the images collected from different sources with varying spatial resolutions make the classification more difficult.

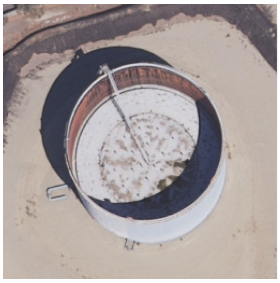
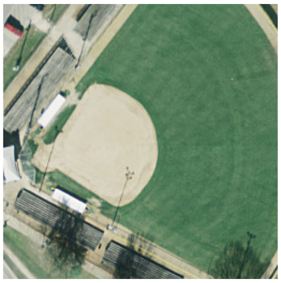
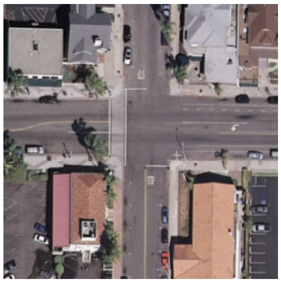
G. Annotation Case Studies

Table VIII displays the annotation results of three images from the UCM dataset. The comparison is made among all the eight methods except for CNN-RNN. In the table, the ground-truth labels and the labels predicted by each method are listed. The correct predictions are shown in green, the false positive predictions are shown in red, and the false negative predictions are shown in blue. Compared with other methods, the proposed method is the only one that can correctly predict the corresponding labels of three images.

H. Visualization

In order to verify the effectiveness of our model for detecting objects of different sizes in an image, we conduct a visualization analysis on four images of the UCM dataset. The feature maps are shown in Fig. 7. It can be observed that the proposed model can highlight the discriminative regions of objects and identify them accurately. For example, there are airplanes, cars, and pavement in the first image. The scale of these objects changes largely, but the model can locate them well.

TABLE VIII
ANNOTATION RESULTS OF THREE IMAGES FROM THE UCM DATASET

Samples from the UCM multi-label data set			
Ground truth	sand, tanks, bare-soil	bare-soil, trees, grass, pavement, buildings	bare-soil, trees, cars, pavement, buildings
ML-KNN	sand, tanks, bare-soil, pavement	bare-soil, trees, grass, pavement, buildings	bare-soil, trees, cars, pavement, buildings, grass
Gardner [3]	sand, tank, pavement	bare-soil, trees, grass, pavement, buildings	bare-soil, trees, cars, pavement, buildings
Stivaktakis [27]	sand, tanks, bare-soil, pavement	bare-soil, trees, grass, pavement, buildings	bare-soil, trees, cars, pavement, buildings, grass
RBFNN [21]	sand, tanks, bare-soil	bare-soil, trees, grass, pavement, buildings	bare-soil, trees, cars, pavement, buildings, grass
Zhu [29]	sand, tanks, pavement	bare-soil, trees, grass, pavement, buildings	bare-soil, trees, cars, grass, pavement, buildings
ML-GCN [36]	sand, tanks, bare-soil	bare-soil, trees, grass, pavement, buildings	bare-soil, trees, cars, pavement, buildings
Proposed Method	sand, tanks, bare-soil	bare-soil, trees, grass, pavement, buildings	bare-soil, trees, cars, pavement, buildings

V. CONCLUSION

In this article, we propose an end-to-end deep learning framework for multilabel RS image annotation. The framework consists of a multiscale feature fusion module, a channel-spatial attention learning module, and a label correlation extraction module. The multiscale features are first extracted from different layers of a VGG16 model and fused by a cascade fusion strategy with a series of operations, involving convolution, BN, ReLU, and concatenation. Then, the fused features are refined by the channel-spatial attention module for salient object detection. Finally, the multiscale attentive features are further enhanced by the label correlation extraction module, where the label correlation information from a label co-occurrence matrix is embedded into the features through a two-step fusion. Experimental results on the UCM and AID multilabel datasets show that our proposed method achieves the best performance compared with the state-of-the-art methods.

In the label correlation extraction model, the pairwise label correlation is used for the label co-occurrence matrix. In future work, we plan to leverage high-order label correlations in the model.

ACKNOWLEDGMENT

The authors would like to thank Prof. Mingyi He for his valuable comments.

REFERENCES

[1] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[2] I. Shendryk, Y. Rist, R. Lucas, P. Thorburn, and C. Ticehurst, "Deep learning—A new approach for multi-label scene classification in planetscope and Sentinel-2 imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 1116–1119.

[3] D. Gardner and D. Nichols, "Multi-label classification of satellite images with deep learning," 2017, [Online]. Available: http://vision.stanford.edu/teaching/cs231n/reports/2017/pdfs/908.pdf

[4] K. Karalas, G. Tsagkatakis, M. Zervakis, and P. Tsakalides, "Land classification using remotely sensed data: Going multilabel," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3548–3563, Jun. 2016.

[5] 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.

[6] Y. Hua, L. Mou, and X. X. Zhu, "Label relation inference for multi-label aerial image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5244–5247.

[7] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1465–1472.

[8] W. Shao, W. Yang, G.-S. Xia, and G. Liu, "A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization," in *Proc. Int. Conf. Comput. Vis. Syst.*, 2013, pp. 324–333.

[9] E. Aptoula, "Remote sensing image retrieval with global morphological texture descriptors," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 3023–3034, May 2014.

[10] V. Risojević and Z. Babić, "Fusion of global and local descriptors for remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 836–840, Jul. 2013.

[11] G. J. Scott, M. N. Klaric, C. H. Davis, and C. Shyu, "Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 5, pp. 1603–1616, May 2011.

[12] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.

[13] Q. Zhu, Y. Zhong, B. Zhao, G. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, Jun. 2016.

[14] L. Zhao, P. Tang, and L. Huo, "Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 12, pp. 4620–4631, Dec. 2014.

[15] M.-L. Zhang and Z.-H. Zhou, "Ml-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, pp. 2038–2048, Jul. 2007.

[16] A. Jiang, C. Wang, and Y. Zhu, "Calibrated rank-SVM for multi-label image categorization," in *Proc. IEEE Int. Joint Conf. Neural Netw./IEEE World Congr. Comput. Intell.*, 2008, pp. 1450–1455.

[17] 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.

[18] Q. Tan, Y. Liu, X. Chen, and G. Yu, "Multi-label classification based on low rank representation for image annotation," *Remote Sens.*, vol. 9, no. 2, 2017, Art. no. 109.

[19] O. E. Dai, B. Demir, B. Sankur, and L. Bruzzone, "A novel system for content-based retrieval of single and multi-label high-dimensional remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2473–2490, Jul. 2018.

[20] 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.

[21] 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.

[22] A. Zeggada, S. Benbraika, F. Melgani, and Z. Mokhtari, "Multilabel conditional random field classification for UAV images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 399–403, Mar. 2018.

[23] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang, "Learning deep latent spaces for multi-label classification," *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 2838–2844.

[24] 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 J. Photogrammetry Remote Sens.*, vol. 149, pp. 188–199, 2019.

[25] G. Sumbul and B. Demir, "A novel multi-attention driven system for multi-label remote sensing image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5726–5729.

[26] T. Chen, Z. Wang, G. Li, and L. Lin, "Recurrent attentional reinforcement learning for multi-label image recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 6730–6737.

[27] R. Stivaktakis, G. Tsagkatakis, and P. Tsakalides, "Deep learning for multilabel land cover scene categorization using data augmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 7, pp. 1031–1035, Jul. 2019.

[28] Y. Bazi, "Two-branch neural network for learning multi-label classification in UAV imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 2443–2446.

[29] P. Zhu *et al.*, "Deep learning for multilabel remote sensing image annotation with dual-level semantic concepts," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4047–4060, Jun. 2020.

[30] L. Wang, Y. Liu, C. Qin, G. Sun, and Y. Fu, "Dual relation semi-supervised multi-label learning," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 6227–6234.

[31] D. Huynh and E. Elhamifar, "A shared multi-attention framework for multi-label zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8776–8786.

[32] T. Wu, Q. Huang, Z. Liu, Y. Wang, and D. Lin, "Distribution-balanced loss for multi-label classification in long-tailed datasets," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 162–178.

[33] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, "Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4355–4369, May 2021.

[34] 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.

[35] 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," in *Proc. Pacific Rim Conf. Multimedia*, 2018, pp. 315–324.

[36] 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. 5177–5186.

[37] B. Chen, J. Li, G. Lu, H. Yu, and D. Zhang, "Label co-occurrence learning with graph convolutional networks for multi-label chest X-ray image classification," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 8, pp. 2292–2302, Aug. 2020.

[38] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, and S. Wen, "Cross-modality attention with semantic graph embedding for multi-label classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 12709–12716.

[39] Y. Wang *et al.*, "Multi-label classification with label graph superimposing," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 12265–12272.

[40] X. Zhao, J. Zhang, J. Tian, L. Zhuo, and J. Zhang, "Residual dense network based on channel-spatial attention for the scene classification of a high-resolution remote sensing image," *Remote Sens.*, vol. 12, no. 11, 2020, Art. no. 1887.

[41] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12416–12425.

[42] Z. Yuan, X. Li, and Q. Wang, "Exploring multi-level attention and semantic relationship for remote sensing image captioning," *IEEE Access*, vol. 8, pp. 2608–2620, 2019.

[43] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[44] M. Lin, Q. Chen, and S. Yan, "Network in network," *Int. Conf. Learn. Representations*, 2014, *arXiv:1312.4400*.

[45] G. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[46] 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.

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

[48] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, May 13–15, 2010, pp. 249–256.

[49] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. Learn. Representations*, 2015, *arXiv:1412.6980*.

Rui Huang (Member, IEEE) received the B.E. degree in electrical engineering and the M.E. and D.E. degrees in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 1999, 2002, and 2006, respectively.

She is currently an Associate Professor with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. Her current research interests include remote sensing information processing, pattern recognition, and machine learning.

Fengcai Zheng (Student Member, IEEE) received the B.S. degree in communication engineering from the School of Electrical and Electronic Engineering, Wenzhou University, Wenzhou, China, in 2015. He is currently working toward the M.S. degree with the School of Communication and Information Engineering, Shanghai University, Shanghai, China.

His research interests include computer vision and multilabel learning.

Wei Huang received the B.E. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2008, respectively.

She is currently a Lecturer with the School of Communication and Information Engineering, Shanghai University Shanghai, China. Her research interests include remote sensing, images denoising and analysis, and image enhancement and restoration for display processing.