

# Relation Network for Multilabel Aerial Image Classification

Yuansheng Hua, *Student Member, IEEE*, Lichao Mou, *Student Member, IEEE*,  
and Xiao Xiang Zhu<sup>IP</sup>, *Senior Member, IEEE*

**Abstract**—Multilabel classification plays a momentous role in perceiving intricate contents of an aerial image and triggers several related studies over the last years. However, most of them deploy few efforts in exploiting label relations, while such dependencies are crucial for making accurate predictions. Although an long short term memory (LSTM) layer can be introduced to modeling such label dependencies in a chain propagation manner, the efficiency might be questioned when certain labels are improperly inferred. To address this, we propose a novel aerial image multilabel classification network, attention-aware label relational reasoning network. Particularly, our network consists of three elemental modules: 1) a label-wise feature parcel learning module; 2) an attentional region extraction module; and 3) a label relational inference module. To be more specific, the label-wise feature parcel learning module is designed for extracting high-level label-specific features. The attentional region extraction module aims at localizing discriminative regions in these features without region proposal generation, yielding attentional label-specific features. The label relational inference module finally predicts label existences using label relations reasoned from outputs of the previous module. The proposed network is characterized by its capacities of extracting discriminative label-wise features and reasoning about label relations naturally and interpretably. In our experiments, we evaluate the proposed model on two multilabel aerial image data sets, of which one is newly produced. Quantitative and qualitative results on these two data sets demonstrate the effectiveness of our model. To facilitate progress in the multilabel aerial image classification, our produced data set will be made publicly available.

**Index Terms**—Attentional region extraction, convolutional neural network (CNN), high-resolution aerial image, label relational reasoning, multilabel classification.

Manuscript received July 15, 2019; revised November 24, 2019; accepted December 10, 2019. Date of publication February 7, 2020; date of current version June 24, 2020. This work was supported in part by the China Scholarship Council, in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Program, So2Sat, under Grant ERC-2016-StG-714087, in part by the Helmholtz Association through the Framework of the Young Investigators Group SiPEO under Grant VH-NG-1018, in part by the Helmholtz Artificial Intelligence Cooperation Unit (HAICU)—Local Unit Munich Unit at Aeronautics, Space and Transport (MASTr), and in part by the Helmholtz Excellent Professorship Data Science in Earth Observation—Big Data Fusion for Urban Research. (*Corresponding author: Xiao Xiang Zhu.*)

The authors are with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany, and also with the Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: yuansheng.hua@dlr.de; lichao.mou@dlr.de; xiaoxiang.zhu@dlr.de).

Color versions of one or more of the figures in this article are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TGRS.2019.2963364

## I. INTRODUCTION

RECENT advancements of remote sensing techniques have boosted the volume of attainable high-resolution aerial images, and massive amounts of applications, such as urban cartography [1]–[4], traffic monitoring [5]–[7], terrain surface analysis [8]–[11], and ecological scrutiny [12], [13], have benefited from these developments. For this reason, the aerial image classification has become one of the fundamental visual tasks in the remote sensing community and has drawn a plethora of research interests [14]–[21]. The classification of aerial images refers to assigning these images with specific labels according to their semantic contents, and a common hypothesis shared by many relevant studies is that an image should be labeled with only one semantic category, such as scene categories (see Fig. 1). Although such image-level labels [22], [23] are capable of delineating images from a macroscopic perspective, it is infeasible for them to provide a comprehensive view of objects in aerial images. To tackle this, huge quantities of algorithms have been proposed to identify each pixel in an image [24]–[26] or localize objects with bounding boxes [27]–[29]. However, the acquisition of requisite groundtruths (i.e., pixel-wise annotations and bounding boxes) demands enormous expertise and human labor, which makes relevant data sets expensive and difficult to access. With this intention, multilabel image classification now attracts increasing attention in the remote sensing community [30]–[34] owing to that 1) a comprehensive picture of aerial image contents can be drawn and 2) data sets required in this task are not expensive (only image-level labels are needed).

Fig. 1 illustrates the difference between image-level scene labels and object labels. As shown in Fig. 1, although these four images are assigned with the same scene label, their multiple object labels vary a lot. It is worth noting that the identification of some objects can actually offer important cues to understand a scene more deeply. For example, the existence of *building* and *pavement* indicates a high probability that rivers in Fig. 1(c) and (d) are very close to areas with frequent human activities, while rivers in Fig. 1(a) and (b) are more likely in the wild due to the absence of human activity cues. In contrast, simply recognizing scene labels can hardly provide such information. Therefore, in this article, we dedicate our efforts to explore an effective model for the multilabel classification of aerial images.

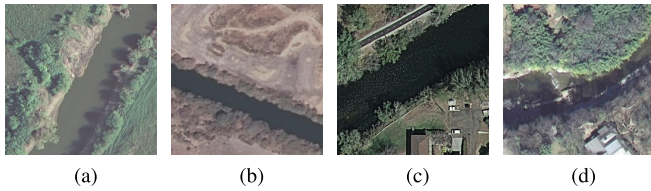


Fig. 1. Example aerial images of scene `river` and `objects` present in them. (a) `Bare soil`, `grass`, `tree`, and `water`. (b) `Water`, `bare soil`, and `tree`. (c) `Water`, `building`, `grass`, `car`, `tree`, `pavement`, and `bare soil`. (d) `Water`, `building`, `grass`, `bare soil`, `tree`, and `sand`.

### A. Challenges of Identifying Multiple Labels

In identifying multiple labels of an aerial image, two main challenges need to be faced with. One is how to extract semantic feature representations from raw images. This is crucial but difficult especially for high-resolution aerial images, as they always contain complicated spatial contextual information. Conventional approaches mainly resort to manually crafted features and semantic models [22], [35]–[38], while these methods cannot effectively extract high-level semantics and lead to a limited performance in classification [23]. Hence an efficient high-level feature extractor is desirable.

The other challenge is how to take full advantage of label correlations to infer multiple object labels of an aerial image. In contrast to single-label classification, which mainly focuses on modeling image-label relevance, exploring and modeling label-label correlations plays a supplementary yet essential role in identifying multiple objects in aerial images. For instance, the presence of ships confidently infers the co-occurrence of water or sea, while the existence of a car suggests a high probability of the appearance of pavements. Unfortunately, such label correlations are scarcely addressed in the literature. One solution is to use a recurrent neural network (RNN) to learn label dependencies. However, this is done with a chain propagation fashion, and its performance heavily depends on the learning effectiveness of its long-term memorization. Moreover, in this way, label relations are modeled implicitly, which leads to a lack of interpretability.

Overall, an efficient multilabel classification model is supposed to be capable of not only learning high-level feature representations but also modeling label correlations effectively.

### B. Related Work

Zegeye and Demir [39] propose a multilabel active learning framework using a multilabel support vector machine (SVM), relying on both the multilabel uncertainty and diversity. Koda *et al.* [32] introduced a spatial and structure SVM for multilabel classification by considering spatial relations between a given patch and its neighbors. Similarly, Zeggada *et al.* [33] employed a conditional random field (CRF) framework to model spatial contextual information among adjacent patches for improving the performance of classifying multiple object labels.

With the development of computational resources and deep learning, very recent approaches mainly resort to deep networks for multilabel classification. Zeggada *et al.* [31] make use of a standard convolutional neural network (CNN)

architecture to extract feature representations and then feed them into a multilabel classification layer, which is composed of customized thresholding operations, for predicting multiple labels. Stivaktakis *et al.* [40] demonstrated that training a CNN for multilabel classification with a limited amount of labeled data usually leads to an underwhelming-performance model and propose a dynamic data augmentation method for enlarging training sets. More recently, Sumbul and Demir [41] proposed a CNN-RNN method for identifying labels in multispectral images, where a bidirectional long short term memory (LSTM) is employed to model spatial relationships among image patches. In order to explore inherent correlations among object labels, Hua *et al.* [34] proposed a CNN-LSTM hybrid network architecture to learn label dependencies for classifying object labels of aerial images. Besides, we also notice that several zero short learning research studies focus on employing prior knowledge to model label relations. For instance, Sumbul *et al.* [42] apply an unsupervised word embedding model to encoding labels into word vectors, which are supposed to contain label semantics, and then model label relationships with these vectors. Lee *et al.* [43] proposed to learn label relations from structured knowledge graphs observed from the real world.

### C. Motivation of Our Work

In order to explicitly model label relations, we propose a label relational inference network for multilabel aerial image classification. This article is inspired by recent successes of relation networks in visual question answering [44], object detection [45], video classification [46], activity recognition in videos [47], and semantic segmentation [48]. A relation network is characterized by its inherent capability of inferring relations between an individual entity (e.g., a region in an image or a frame in a video) and all other entities (e.g., all regions in the image or all frames in the video). Besides, to increase the effectiveness of relational reasoning, we make use of a spatial transformer, which is often used to enhance the transformation invariance of deep neural networks [49], to reduce the impact of irrelevant semantic features.

More specifically, in this article, an innovative end-to-end multilabel aerial image classification network, termed as attention-aware label relational reasoning network, is proposed and characterized by its capabilities of localizing label-specific discriminative regions and explicitly modeling semantic label dependencies for the task. This article’s contributions are threefold, which are given as follows.

- 1) We propose a novel multilabel aerial image classification network, attention-AL-RN-CNN, which consists of three imperative components: a label-wise feature parcel learning module, an attentional region extraction module, and a label relational inference module. To our best knowledge, it is the first time that the idea of relation networks is employed to predict multiple object labels of aerial images, and experimental results demonstrate its effectiveness.
- 2) We extract attentional regions from the label-wise feature parcels in a proposal-free fashion. Particularly, a learnable spatial transformer is employed to localize

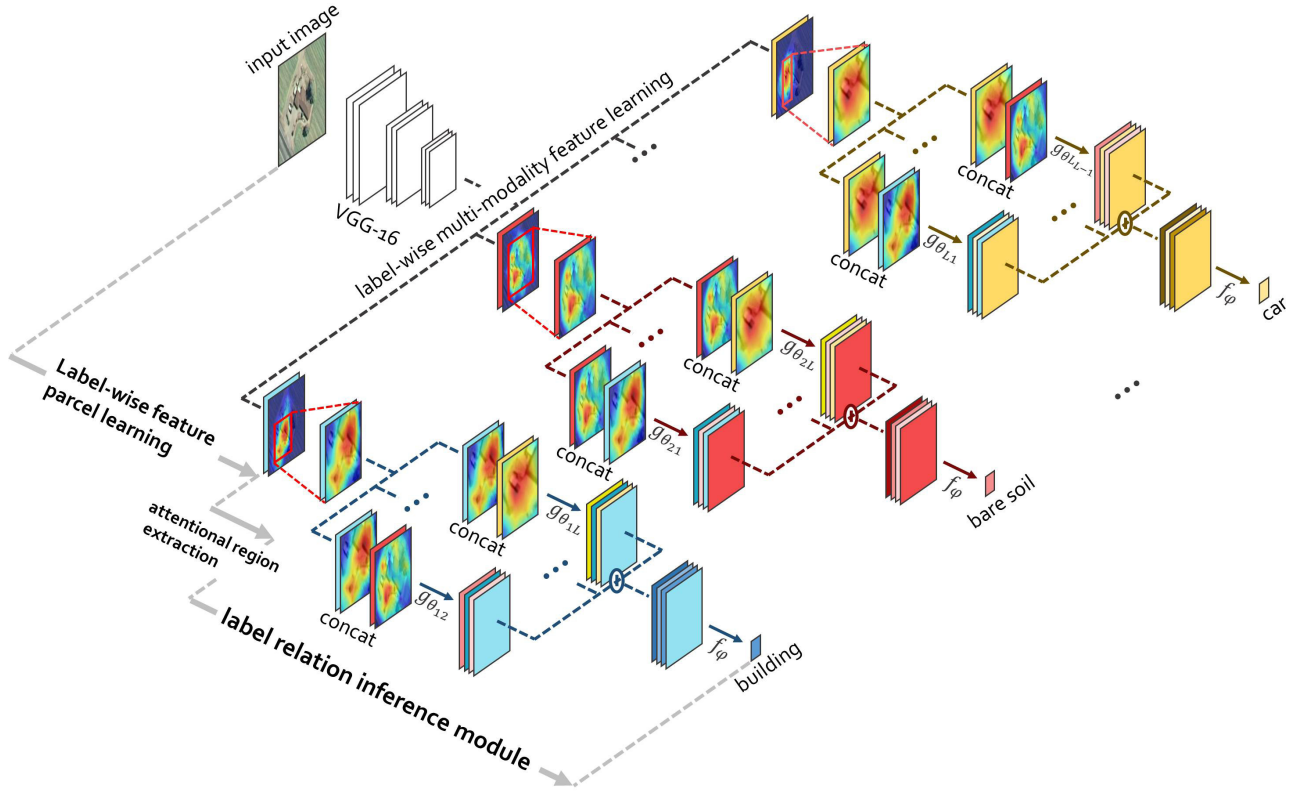


Fig. 2. Architecture of the proposed attention-aware label relational reasoning network.

attentional regions, which are assumed to contain discriminative information, and then re-coordinate them into a given size. By doing so, attentional feature parcels can be yielded.

- 3) To facilitate progress in the multilabel aerial image classification, we produce a new data set, AID multilabel data set, by relabeling images in the AID data set [23]. In comparison with the UCM multilabel data set [50], the proposed data set is more challenging due to diverse spatial resolutions of images, more scenes, and more samples.

Sections II–IV of this article are organized as follows. Section II delineates three elemental modules of our proposed network, and Section III introduces experiments, where experimental setups are given and results are analyzed and discussed. Eventually, Section IV draws a conclusion of this article.

## II. METHODOLOGY

### A. Network Architecture

As illustrated in Fig. 2, the proposed network comprises three components: a label-wise feature parcel learning module, an attentional region extraction module, and a label relational inference module. Let  $L$  be the number of object labels and  $l$  be the  $l$ th label. The label-wise feature parcel learning module is designed to extract high-level feature maps  $X_l$  with  $K$  channels, termed as feature parcel (for more details refer to Section II-B), for each label  $l$ . The attentional region extraction module is used to localize discriminative regions in each

$X_l$  and generate an attentional feature parcel  $A_l$ , which is supposed to contain the most relevant semantics with respect to the label  $l$ . Finally, relations among  $A_l$  and all other label-wise attentional feature parcels are reasoned about by the label relational inference module for predicting the presence of the object  $l$ .

Details of the proposed network are introduced in Sections II–IV.

### B. Label-Wise Feature Parcel Learning

The extraction of high-level features is crucial for visual recognition tasks, and many recent studies adopt CNNs owing to their remarkable performance in learning such features [15], [51]–[56]. Hence, we take a standard CNN as the backbone of the label-wise feature parcel learning module in our model. As shown in Fig. 2, an aerial image is first fed into a CNN (e.g., VGG-16), which consists of only convolutional and max-pooling layers, for generating high-level feature maps. Subsequently, these features are encoded into  $L$  feature parcels for each label  $l$  via a label-wise multimodality feature learning layer. To implement this layer, we first employ a convolutional layer with  $KL$  filters, whose size is  $1 \times 1$ , to extract  $KL$  feature maps. Afterward, we divide these features into  $L$  feature parcels, and each includes  $K$  feature maps. That is to say, for each label,  $K$  specific feature maps are learned, so-called feature parcel, to extract discriminative semantics after the end-to-end training of the whole network.

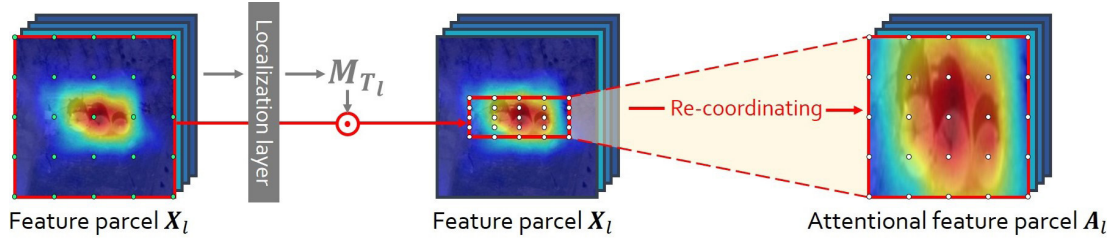


Fig. 3. Illustration of the attentional region extraction module. Green dots in the left image indicate the feature parcel grid  $G_{X_l}$ . White dots in the middle image represent the attentional feature parcel grid  $G_{X_l^{\text{attn}}}$ , while those in the right image indicate re-coordinated  $G_{X_l^{\text{attn}}}$ . Notably, the structure of re-coordinated  $G_{X_l^{\text{attn}}}$  is identical to that of  $G_{X_l}$ , and values of pixels located at grid points in re-coordinated  $G_{X_l^{\text{attn}}}$  are obtained from those in  $G_{X_l^{\text{attn}}}$ . For example, the pixel at the left top corner grid point in re-coordinated  $G_{X_l^{\text{attn}}}$  is assigned with the value of that at the left top corner of  $G_{X_l^{\text{attn}}}$ .

We denote the feature parcel for label  $l$  as  $X_l$  in the following statements.

In our experiment, we notice that  $X_l$  with a higher resolution is beneficial for the subsequent module to localize discriminative regions, as more spatial contextual cues are included. Accordingly, we discard the last max-pooling layer in VGG-16, leading to a spatial size of  $14 \times 14$  for outputs. Weights are initialized with pre-trained VGG-16 on ImageNet but updated during the training phase.

### C. Attentional Region Extraction Module

Although label-wise feature parcels can be directly applied to exploring label dependencies [34], less informative regions (see blue areas in Fig. 3) may bring noise and further reduce the effectiveness of these feature parcels. As shown in the left image of Fig. 3, weakly activated regions indicate a loose relevance to the corresponding label, while highlighted regions suggest a strong region-label relevance. To diminish the influence of unrelated regions, we employ an attentional region extraction module to automatically extract discriminative regions from label-wise feature parcels.

We localize and re-coordinate attentional regions from  $X_l$  with a learnable spatial transformer. Particularly, we sample a feature parcel  $X_l$  into a regular spatial grid  $G_{X_l}$  (see green dots in the left image of Fig. 3) according to the spatial resolution of  $X_l$  and regard pixels in  $X_l$  as points on the grid  $G_{X_l}$  with coordinates  $(x_l, y_l)$ . Similarly, we can define coordinates of a new grid, attentional region grid  $G_{X_l^{\text{attn}}}$  (see white dots in the middle image of Fig. 3), as  $(x_l^{\text{attn}}, y_l^{\text{attn}})$ , and the number of grid points along with the height and width is equivalent to that of  $G_{X_l}$ . As demonstrated in [49] that  $G_{X_l^{\text{attn}}}$  can be learned by performing spatial transformation on  $G_{X_l}$ ,  $(x_l^{\text{attn}}, y_l^{\text{attn}})$  can be calculated with the following equation:

$$\begin{bmatrix} x_l^{\text{attn}} \\ y_l^{\text{attn}} \end{bmatrix} = \mathbf{M}_{T_l} \begin{bmatrix} x_l \\ y_l \\ 1 \end{bmatrix} \quad (1)$$

where  $\mathbf{M}_{T_l}$  is a learnable transformation matrix, and grid coordinates,  $x_l$  and  $y_l$ , are normalized to  $[-1, 1]$ . Considering that this module is designed for localization, we only adopt scaling and translation in our case. Hence (1) can be

rewritten as

$$\begin{bmatrix} x_l^{\text{attn}} \\ y_l^{\text{attn}} \end{bmatrix} = \begin{bmatrix} s_{x_l} & 0 & t_{x_l} \\ 0 & s_{y_l} & t_{y_l} \end{bmatrix} \begin{bmatrix} x_l \\ y_l \\ 1 \end{bmatrix} \quad (2)$$

where  $s_{x_l}$  and  $s_{y_l}$  indicate scaling factors along the  $x$ - and  $y$ -axes, respectively, and  $t_{x_l}$  and  $t_{y_l}$  represent how feature maps should be translated along both axes. Notably, since different objects distribute differently in aerial images,  $\mathbf{M}_{T_l}$  is learned for each object label  $l$  individually. In other words, extracted attentional regions are label-specific and capable of improving the effectiveness of label-wise features.

As to the implementation of this module, we first vectorize  $X_l$  with a flatten function and then employ a localization layer (e.g., a fully connected layer) to estimate elements in  $\mathbf{M}_{T_l}$  from the vectorized  $X_l$ . Afterward, attentional region coordinates  $(x_l^{\text{attn}}, y_l^{\text{attn}})$  can be learned from  $(x_l, y_l)$  with (2), and values of pixels at  $(x_l^{\text{attn}}, y_l^{\text{attn}})$  are able to be obtained from neighboring pixels by bilinear interpolation. Finally, the attentional region grid  $G_{X_l^{\text{attn}}}$  is re-coordinated to a regular spatial grid, which shares an identical structure with  $G_{X_l}$ , for yielding the final attentional feature parcel  $A_l$ .

### D. Label Relational Inference Module

Being the core of our model, the label relational inference module is designed to fully exploit label interrelations for inferring existences of all labels. Before diving into this module, we define the pairwise label relation as a composite function with the following equation:

$$\text{LR}(A_l, A_m) = f_\phi(g_{\theta_{lm}}(A_l, A_m)) \quad (3)$$

where the input is a pair of attentional feature parcels,  $A_l$  and  $A_m$ , and  $l$  and  $m$  range from 1 to  $L$ . The functions  $g_{\theta_{lm}}$  and  $f_\phi$  are used to reason about the pairwise relation between label  $l$  and  $m$ . More specifically, the role of  $g_{\theta_{lm}}$  is to reason about whether there exist relations between the two objects and how they are related. In previous works [44], [47], a multilayer perceptron (MLP) is commonly employed as  $g_{\theta_{lm}}$  for its simplicity. However, spatial contextual semantics are not taken into account in this way. To address such an issue, here, we make use of  $1 \times 1$  convolution instead of an MLP to explore spatial information. Furthermore,  $f_\phi$  is applied to encode the output of  $g_{\theta_{lm}}$  into the final pairwise label relation

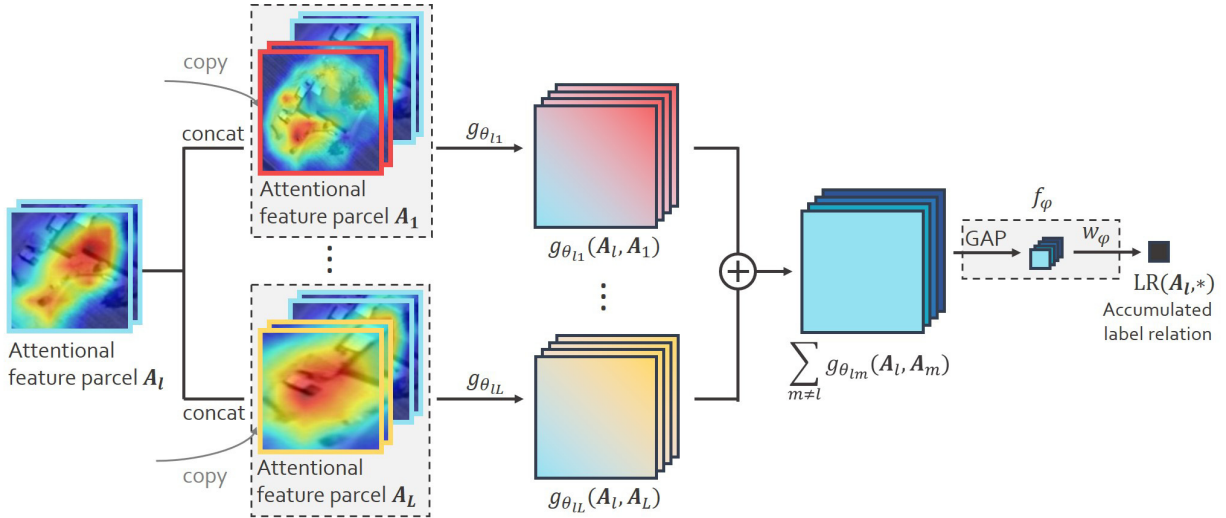


Fig. 4. Illustration of the label relation module.

$LR(A_l, A_m)$ . In our case,  $f_\phi$  consists of a global average pooling layer and an MLP, which finally yields the relation between label  $l$  and  $m$ .

Following the motivation of our work, we infer each label by accumulating all related pairwise label relations, and the accumulated label relation for object label  $l$  is defined as:

$$LR(A_l, *) = f_\phi \left( \sum_{m \neq l} g_{\theta_{lm}}(A_l, A_m) \right) \quad (4)$$

where  $*$  represents all attentional feature parcels except  $A_l$ . Based on this formula, we implement the label relational inference module with the following steps (taking the prediction of label  $l$  as an example): 1)  $A_l$  and every other attentional feature parcel are concatenated and fed into a  $1 \times 1$  convolutional layer, respectively; 2) afterward, a global average pooling layer is employed to transform  $g_{\theta_{lm}}(A_l, A_m)$  into vectors, which are then element-wise added; and 3) finally, the output is fed into an MLP layer with trainable parameters  $\phi$  to produce the accumulated label relation  $LR(A_l, *)$ . Note that  $g_{\theta_{lm}}$  is a learnable unit, which models pairwise relations using convolutions. Through the end-to-end training, it could be expected to learn data-driven label relations. Experiments in Sections III-D and III-E have verified that learned label relations are in line with prior knowledge. Since we expect the model to predict probabilities, an activation function  $\sigma$  is utilized to restrict each output digit to  $[0, 1]$ . For label  $l$ , a digit approaching 1 implies a high probability of its presence, while one closing 0 suggests the absence. Fig. 4 presents an visual illustration of the label relational inference module.

Compared to other multilabel classification methods, our model has three benefits.

- 1) The module can inherently reason about label relations as indicated by (3) and requires no particular prior knowledge about relations among all objects. That is to say, our network does not need to learn *how to compute label relations* and *which object relations should be considered*. All relations are automatically learned through

a data-driven way and proven to meet the reality in our experiments.

- 2) The learning effectiveness is independent of long short-term memory, leading to increased robustness. This is because, in (4), accumulated label relations are calculated with a summation function instead of chain architecture, e.g., an LSTM.
- 3) The function  $g_{\theta_{lm}}$  is learned for each object label pair  $l$  and  $m$  separately, which suggests that pairwise label relations are encoded in a specific way. Besides, our implementation of  $g_{\theta_{lm}}$  can extend the applicability of relational reasoning compared to using an MLP.

Since [34] shares the same design philosophy that modeling label relations is crucial, here, we emphasize two differences between our network and [34]: 1) the proposed network learns to extract discriminative regions as label-wise features for modeling label relations (see Section II-C) instead of directly using entire feature maps as in [34] and 2) the proposed label relation inference module encodes label relations explicitly with composite functions, while in [34], label relations are modeled implicitly via an RNN whose effectiveness depends heavily on the learning effect of long-term memorization. Quantitative comparisons between these two approaches are shown in Section III.

### III. EXPERIMENTS AND DISCUSSION

In this section, we conduct experiments on the UCM [50] and proposed the AID multilabel data set for evaluating our model. Specifically, Section III-A presents a description of these two data sets. Afterward, we introduce training strategies and thoroughly discuss experimental results in Sections III-B–III-F.

#### A. Data Set Introduction

1) *UCM Multilabel Data Set*: UCM multilabel data set [50] is reproduced by assigning all aerial images collected in the UCM data set [22] with newly defined object labels.

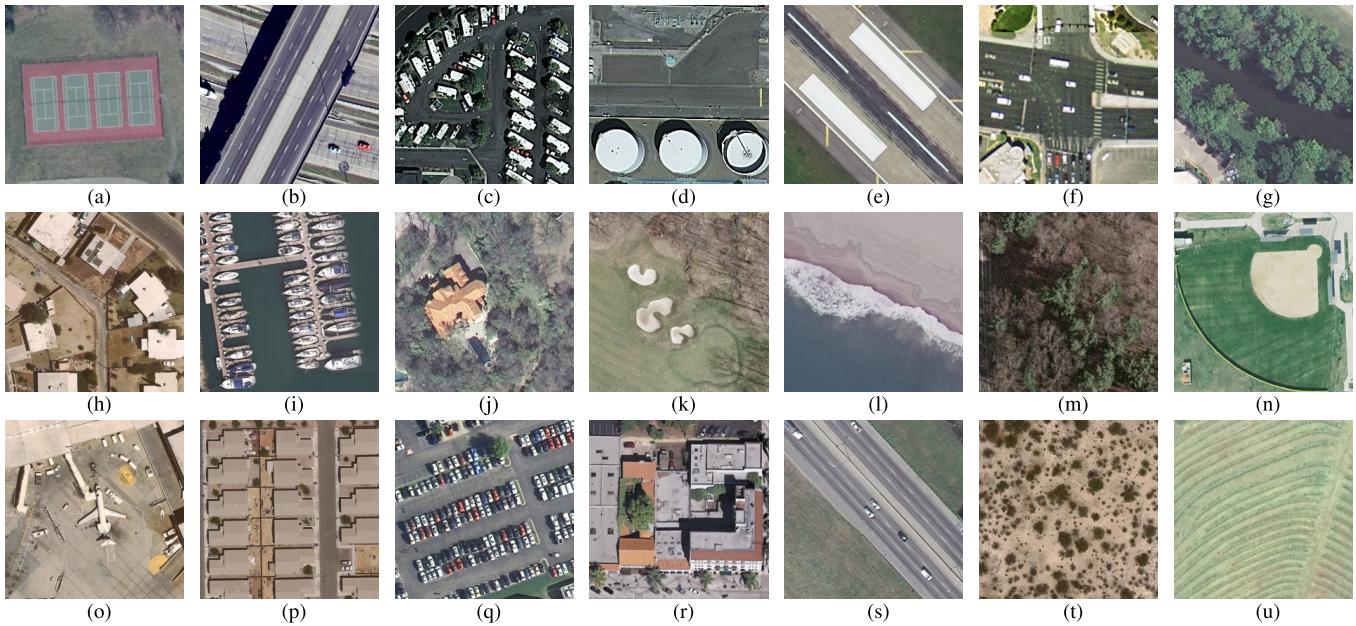


Fig. 5. Samples of various scene categories in the UCM multilabel data set as well as associated *object* labels. The spatial resolution of each image is one foot, and the size is  $256 \times 256$  pixels. Scene and *object* labels of each sample are as follows. (a) Tennis court: *tree, grass, court, and bare soil*. (b) Overpass: *pavement, bare soil, and car*. (c) Mobile home park: *pavement, grass, bare soil, tree, mobile home, and car*. (d) Storage tank: *tank, pavement, and bare soil*. (e) Runway: *pavement and grass*. (f) Intersection: *car, tree, pavement, grass, and building*. (g) River: *water, tree, and grass*. (h) Medium residential: *pavement, grass, car, tree, and building*. (i) Harbor: *ship, water, and dock*. (j) Sparse residential: *car, tree, grass, pavement, building, and bare soil*. (k) Golf course: *sand, pavement, tree, and grass*. (l) Beach: *sea and sand*. (m) Forest: *tree, grass, and building*. (n) Baseball diamond: *pavement, grass, building, and bare soil*. (o) Airplane: *airplane, car, bare soil, grass, and pavement*. (p) Dense residential: *tree, building, pavement, grass, and car*. (q) Parking lot: *pavement, grass, and car*. (r) Building: *pavement, car, and building*. (s) Free way: *tree, car, pavement, grass, and bare soil*. (t) Chaparral: *chaparral and bare soil*. (u) Agricultural: *tree and field*.

The number of all candidate object labels is 17: building, sand, dock, court, tree, sea, bare soil, mobile home, ship, field, tank, water, grass, pavement, chaparral, and car. It is worth noting that labels, such as tank, airplane, and building, exist in both [22] and [50] while at different levels. In [22], such terms are considered as scene-level labels due to the fact that related images can be characterized and depicted by them, while in [50], they mean objects that may present in aerial images.

As to the properties of images in this data set, the spatial resolution of each sample is one foot, and the size is  $256 \times 256$  pixels. All images are manually cropped from aerial imagery contributed by the National Map of the U.S. Geological Survey (USGS), and there are 2100 images in total. For each object category, the number of images is listed in Table I. Besides, 80% of image samples per scene class are selected to train our model, and the other 20% of images are used to test our model. The number of images assigned to training and test sets with respect to all object labels is available in Table I as well. Some visual examples are shown in Fig. 5.

2) *AID Multilabel Data Set*: In order to further evaluate our network and meanwhile promote progress in the area of multiclass classification of high-resolution aerial images, we produce a new data set, named AID multilabel data set, based on the widely used AID scene classification data set [23]. The AID data set consists of 10000 high-resolution aerial images collected from worldwide Google Earth imagery, including scenes from China, the United States, England, France, Italy, Japan, and Germany. In contrast to the UCM data set, spatial resolutions of images in the AID data set

TABLE I  
NUMBER OF IMAGES FOR DIFFERENT OBJECT CATEGORIES IN THE UCM MULTILABEL DATA SET

Category No.	Category Name	Training	Test	Total
1	bare soil	577	141	718
2	airplane	80	20	100
3	building	555	136	691
4	car	722	164	886
5	chaparral	82	33	115
6	court	84	21	105
7	dock	80	20	100
8	field	79	25	104
9	grass	804	171	975
10	mobile home	82	20	102
11	pavement	1047	253	1300
12	sand	218	76	294
13	sea	80	20	100
14	ship	80	22	102
15	tank	80	20	100
16	tree	801	208	1009
17	water	161	42	203
-	All	1680	420	2100

vary from 0.5 to 8 m/pixel, and the size of each aerial image is  $600 \times 600$  pixels. Besides, the number of images in each scene category ranges from 220 to 420. Overall, the AID data set is more challenging compared to the UCM data set.

Here, we manually relabel some images in the AID data set. With extensive human visual inspections, 3000 aerial images from 30 scenes in the AID data set are selected and assigned

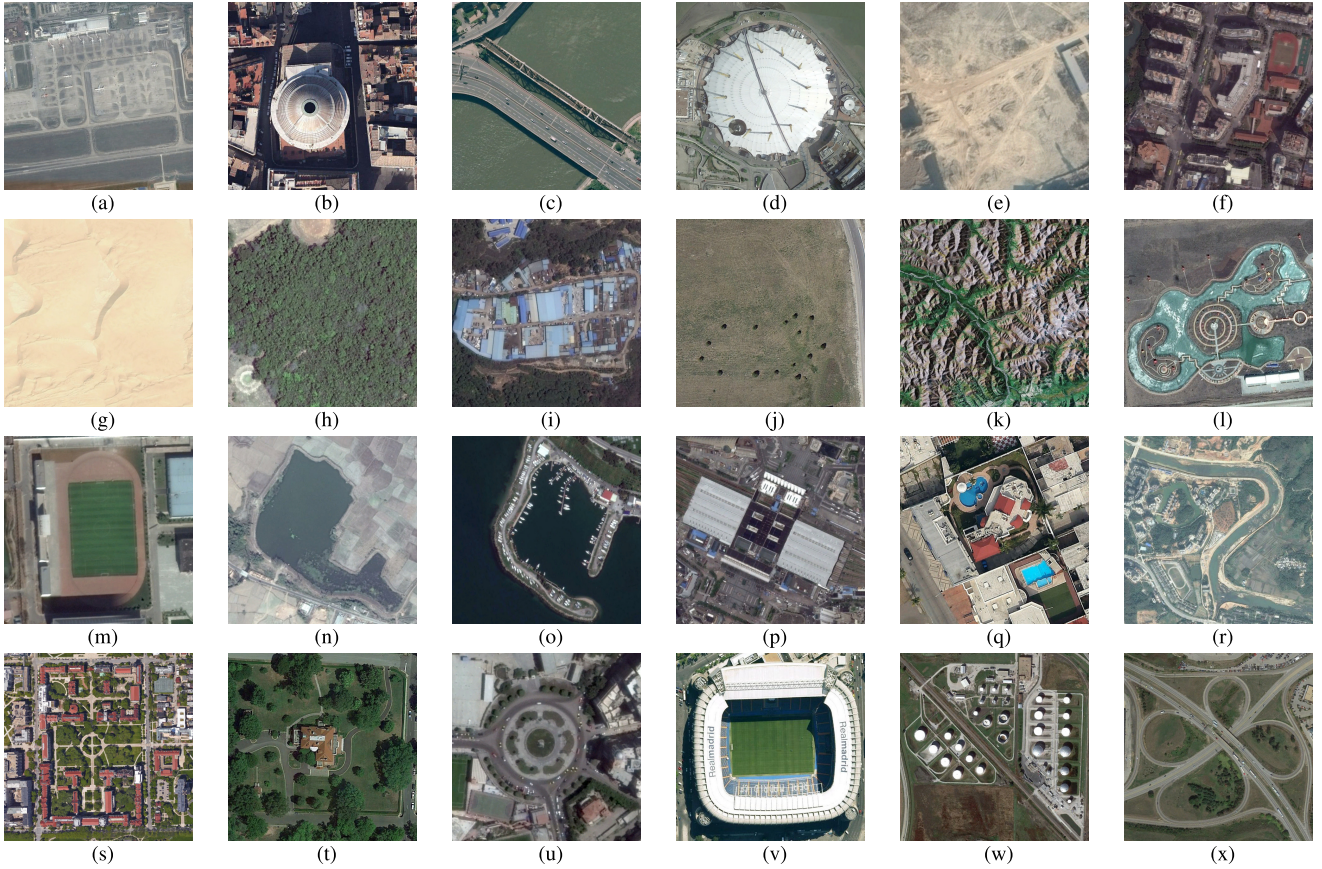


Fig. 6. Samples of various scene categories in the AID multilabel data set and their associated *object* labels. The spatial resolution of each image varies from 0.5 to 8 m/pixel, and the size is  $600 \times 600$  pixels. Here are scene and *object* labels of selected samples. (a) Airport: *car, building, tank, tree, airplane, grass, pavement, and bare soil*. (b) Church: *pavement, car, and building*. (c) Bridge: *building, car, grass, pavement, tree, and water*. (d) Center: *grass, building, tree, car, bare soil, and pavement*. (e) Bare land: *bare soil, building, pavement, and water*. (f) Commercial: *building, car, court, grass, pavement, tree, and water*. (g) Desert: *sand*. (h) Forest: *bare soil and tree*. (i) Industrial: *pavement, grass, car, bare soil, and building*. (j) Meadow: *pavement and grass*. (k) Mountain: *tree and grass*. (l) Park: *bare soil, building, court, grass, pavement, tree, and water*. (m) Playground: *car, grass, and pavement*. (n) Pond: *building, field, grass, pavement, tree, and water*. (o) Port: *ship, sea, car, grass, pavement, tree, building, and dock*. (p) Railway: *tree, car, pavement, building, and grass*. (q) Resort: *pavement, building, car, tree, field, bare soil, and water*. (r) River: *car, building, bare soil, dock, water, grass, pavement, tree, ship, and field*. (s) School: *pavement, tank, grass, court, building, and car*. (t) Sparse residential: *pavement, car, building, tree, and grass*. (u) Square: *tree, car, court, pavement, grass, and building*. (v) Stadium: *car, pavement, tree, court, grass, building, and bare soil*. (w) Storage tanks: *tank, tree, car, grass, pavement, building, and bare soil*. (x) Viaduct: *pavement, car, bare soil, tree, grass, and building*.

with multiple object labels, and the distribution of samples in each category is shown in Table II. Besides, 80% of all images is taken as training samples, while the rest is used for testing our model. Several example images are shown in Fig. 6.

### B. Training Details

As to the initialization of our network, different modules are done in different ways. For the label-wise feature parcel learning module, we initialize the backbone and weights in other convolutional layers with a pre-trained ImageNet [57] model and a Glorot uniform initializer, respectively. Regarding the attentional region extraction module, we initialize the transformation matrix in (1) as an identical transformation

$$\mathbf{M}_{T_i} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \quad (5)$$

In the label relational inference module, weights in both  $f_\phi$  and  $g_{\theta_m}$  are initialized with a Glorot uniform initializer and updated during the training phase. Notably, the entire

network is trained in an end-to-end manner, and weights in the backbone are fine-tuned as well.

In our case, multiple labels are encoded into multihot binary sequences instead of one-hot vectors widely used in single-label classification tasks. The length of such multihot binary sequence is identical to the number of total object categories, i.e., 17 in our case, and as to each digit, 0 suggests an absent object, while 1 indicates the presence of its corresponding object label. Accordingly, we define the network loss as the binary cross-entropy. Besides, Adam with Nesterov momentum [58], which shows faster convergence than stochastic gradient descent (SGD) for our task, are selected and its parameters are set as recommended [58]:  $\epsilon = 1e - 08$ ,  $\beta_1 = 0.9$ , and  $\beta_2 = 0.999$ . The learning rate is initially defined as  $1e - 04$  and decayed by a factor of 10 if the validation loss fails to decrease. Notably, we randomly select 10% of the training samples as the validation set. That is, during the training procedure, we use 90% of the training samples to learn network parameters.

TABLE II  
NUMBER OF IMAGES FOR DIFFERENT OBJECT CATEGORIES IN THE AID  
MULTILABEL DATA SET

Category No.	Category Name	Training	Test	Total
1	bare soil	1171	304	1475
2	airplane	79	20	99
3	building	1744	417	2161
4	car	1617	409	2026
5	chaparral	75	37	112
6	court	269	75	344
7	dock	221	50	271
8	field	175	39	214
9	grass	1829	466	2295
10	mobile home	1	1	2
11	pavement	1870	458	2328
12	sand	207	52	259
13	sea	177	44	221
14	ship	237	47	284
15	tank	87	21	108
16	tree	1923	483	2406
17	water	674	178	852
-	All	2400	600	3000

Our model is implemented on TensorFlow-1.12.0 and trained for 100 epochs. The computational resource is an NVIDIA Tesla P100 GPU with a 16-GB memory. As a compromise between the training speed and GPU memory capacities, we set the size of training batches as 32. To avoid overfitting, the training progress is terminated once the validation loss increases continuously in five epochs.

### C. Experimental Setup

To fully explore the capacity of our proposed network, we extend our researches by replacing the backbone with GoogLeNet (Inceptionv3) [59] and ResNet (ResNet-50 in our case) [60]. Specifically, we adapt GoogLeNet by removing global average pooling and fully connected layers as well as reducing the stride of convolutional and pooling layers in “mixed8” to 1 to improve the spatial resolution. Besides, in order to preserve receptive fields of subsequent convolutional layers, filters in “mixed9” are replaced with atrous convolutional filters, and the dilation rate is defined as 2. Regarding ResNet, we set the convolution stride and dilation rate of filters as 1 and 2, respectively, in the last residual block. Global average pooling and fully-connected layers are removed as well.

In our experiments, we compare the proposed attention-AL-RN-CNN with the following competitors: a standard CNN, CNN-RBFNN [31], and CA-CNN-BiLSTM [34]. Regarding the CNN, we replace its last softmax layer, designed for single-label classification, with a sigmoid layer to produce multihot sequences. For the CA-CNN-BiLSTM, we follow the experimental configurations in [34]. Specifically, we first initialize the feature extraction module of CA-CNN-BiLSTM and weights in the bidirectional LSTM layer with CNNs pre-trained on ImageNet data set and random values from  $-0.1$  to  $0.1$ , respectively. Afterward, we fine-tune the entire network in the training phase with Nesterov Adam optimizer, and the initial learning rate is set to  $1e-04$ . The loss is calculated with the

binary cross-entropy, and the size of training batches is 32. Notably, for all models, output sequences are binarized with a threshold of 0.5 to generate final predictions.

### D. Results on the UCM Multilabel Data Set

1) *Quantitative Analysis*: In our experiment, we employ  $F_1$  [61] and  $F_2$  [62] scores as evaluation metrics to quantitatively assess the performance of different models. Specifically, these two  $F$  scores are calculated with the following equation:

$$F_\beta = (1 + \beta^2) \frac{p_e r_e}{\beta^2 p_e + r_e}, \quad \beta = 1, 2 \quad (6)$$

where  $p_e$  indicates the example-based precision and recall [63] of predictions. Formulas for calculating  $p_e$  and  $r_e$  are

$$p_e = \frac{TP_e}{TP_e + FP_e}, \quad r_e = \frac{TP_e}{TP_e + FN_e} \quad (7)$$

where  $TP_e$  (example-based true-positive) indicates the number of correctly predicted positive labels in an example, while  $FP_e$  (example-based false-positive) denotes the number of those failed to be recognized. Besides,  $FN_e$  (example-based false-negative) represents the number of incorrectly predicted negative labels in an example. Here, an example stands for an aerial image and its associated multiple labels.

To evaluate our network comprehensively, we take mean  $F_1$  and  $F_2$  score as principal indexes. Moreover, we also report mean  $p_e$  and mean  $r_e$ . In addition to the example-based perspective, label-based precision and recall are also considered and calculated with

$$p_l = \frac{TP_l}{TP_l + FP_l}, \quad r_l = \frac{TP_l}{TP_l + FN_l} \quad (8)$$

to demonstrate the performance of networks from the perspective of each object label.

Table III exhibits experimental results on the UCM multilabel data set. We can observe that our model surpasses all competitors on the UCM multilabel data set with variant backbones. Specifically, AL-RN-VGGNet increases mean  $F_1$  and  $F_2$  scores by 7.16% and 5.64%, respectively, in comparison with VGGNet. Compared to CA-VGG-BiLSTM, which resorts to employing a bidirectional LSTM structure for exploring label dependencies, our network obtains an improvement of 5.92% in the mean  $F_1$  score. Besides, although CA-VGG-BiLSTM is superior to VGGNet in both mean  $F_1$  and  $F_2$  scores, it achieves decreased mean precisions and recalls. In contrast, AL-RN-VGGNet outperforms VGGNet not only in mean  $F_1$  and  $F_2$  scores but also in mean example- and label-based precisions and recalls. For another backbone, GoogLeNet, our network gains the best mean  $F_1$  and  $F_2$  scores. As shown in Table III, AL-RN-GoogLeNet increases the mean  $F_1$  score by 4.56% and 3.42% with respect to GoogLeNet and CA-GoogLeNet-BiLSTM, respectively. For the mean  $F_2$  score and precisions, our model also surpasses other competitors, which proves the effectiveness and robustness of our method. AL-RN-ResNet achieves the best mean  $F_1$  score, 0.8676, and  $F_2$  score, 0.8667, in comparison with all other models. Furthermore, it obtains the best mean example-based precision, 0.8881, and label-based precision, 0.9233,



TABLE III  
COMPARISONS OF THE CLASSIFICATION PERFORMANCE ON UCM MULTILABEL DATA SET (%)

Network	mean $F_1$	mean $F_2$	mean $p_e$	mean $r_e$	mean $p_l$	mean $r_l$
VGGNet [64]	78.54	80.17	79.06	82.30	86.02	80.21
VGG-RBFNN [31]	78.80	81.14	78.18	83.91	81.90	82.63
CA-VGG-BiLSTM [34]	79.78	81.69	79.33	83.99	85.28	76.52
AL-RN-VGGNet	85.70	85.81	87.62	86.41	91.04	81.71
GoogLeNet [59]	80.68	82.32	80.51	84.27	87.51	80.85
GoogLeNet-RBFNN [31]	81.54	84.05	79.95	86.75	86.19	84.92
CA-GoogLeNet-BiLSTM [34]	81.82	84.41	79.91	87.06	86.29	84.38
AL-RN-GoogLeNet	85.24	85.33	87.18	85.86	91.03	81.64
ResNet-50 [60]	79.68	80.58	80.86	81.95	88.78	78.98
ResNet-RBFNN [31]	80.58	82.47	79.92	84.59	86.21	83.72
CA-ResNet-BiLSTM [34]	81.47	85.27	77.94	89.02	86.12	84.26
AL-RN-ResNet	86.76	86.67	88.81	87.07	92.33	85.95

TABLE IV  
EXAMPLE IMAGES AND PREDICTED LABELS ON THE UCM AND AID MULTILABEL DATA SET

Samples from the UCM Multi-label Dataset					
Ground Truths	building, car, court, grass, tree, and pavement	building, bare soil, pavement, and grass	car, tree, building, grass, and bare soil	pavement, grass, tree, and bare soil	car, pavement, and building
Predictions	building, car, court, grass, tree, and pavement	building, bare soil, pavement, and grass	tree, car, building, grass, bare soil, and <b>pavement</b>	pavement, grass, tree, and bare soil	car, pavement, and <b>building</b>
Samples from the AID Multi-label Dataset					
Ground Truths	building, car, grass, tree, and pavement	car, bare soil, court, building, grass, tree, pavement, and water	building, car, tree, dock, grass, pavement, sea, and ship	bare soil, building, car, pavement, grass, tree, and water	court, building, car, bare soil, grass, tree, and pavement
Predictions	building, car, grass, tree, and pavement	car, bare soil, court, building, grass, tree, pavement, and <b>water</b>	building, car, tree, dock, grass, pavement, sea, <b>water</b> , and ship	bare soil, car, building, pavement, water, <b>sand</b> , tree, and grass	court, building, car, bare soil, grass, tree, and pavement

Red predictions indicate false positives, while blue predictions are false negatives.

and recall, 0.8595. To summarize, comparisons between AL-RN-CNN and other models demonstrate the effectiveness of our network. Moreover, comparisons between AL-RN-CNN and CA-CNN-BiLSTM illustrate that the composite function-based proposed model performs better than a BiLSTM framework in terms of both accuracy and robustness. Reasons could be that: 1) a chain-like BiLSTM architecture might suffer from

the error propagation [41] and thus is sensitive to the order of predictions, while in our network, all pair-wise label relations are encoded separately and the final summation function is order invariant [44] and 2) a BiLSTM-based structure models label relations implicitly, whereas our network encodes such relations in an explicit and direct way. Table IV presents several example predictions from the UCM multilabel data set.

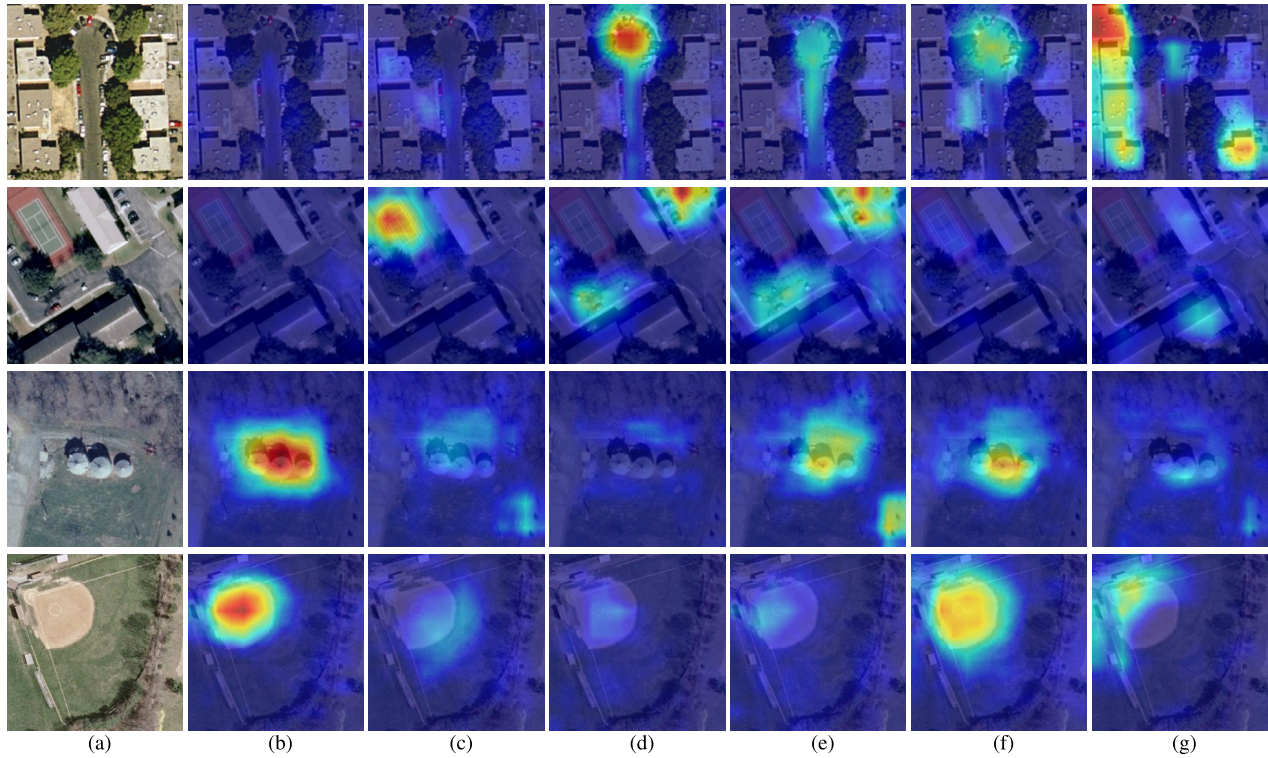


Fig. 7. Example label-specific features of (a) samples selected from the UCM multilabel data set regarding, (b) tank, (c) court, (d) pavement, (e) car, (f) bare soil, and (g) building. Red implies strong activations, whereas blue indicates weak activations.

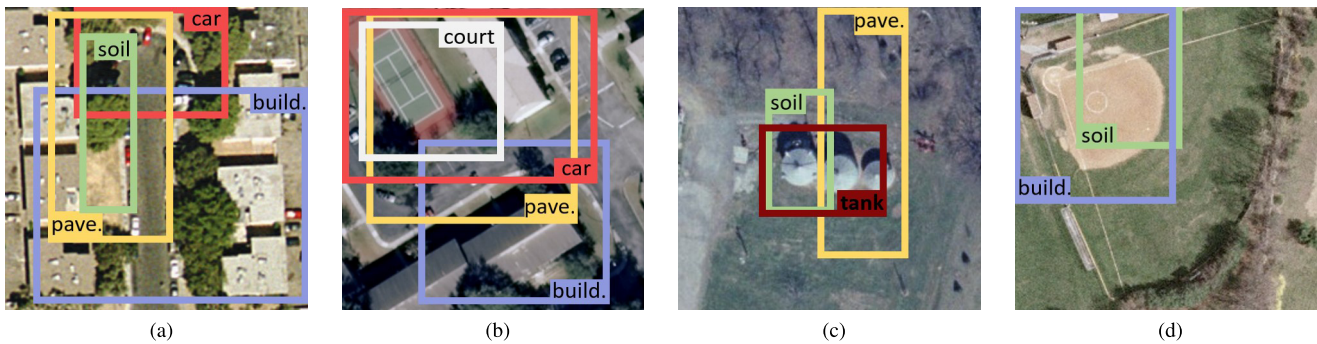


Fig. 8. (a)–(d) Example attentional regions for car, bare soil (soil), building (build.), pavement (pave.), court, and tank in various scenes in the UCM multilabel data set. For each scene, only positive labels mentioned in Fig. 7 are considered.

As a supplementary study, we evaluate the robustness of our proposed model by performing cross-validation in the training phase. More specifically, we randomly divide training samples into fivefold and train our best-performing model, i.e., AL-RN-ResNet, five times. For each training progress, we select one of fivefold as the validation set and train our model with the remaining fourfold. We observe that variances of mean  $F_1$  and  $F_2$  scores are 0.38% and 0.71%, respectively. Compared to improvements brought by our network, variances are limited, and this demonstrates the robustness of our proposed network.

2) *Qualitative Analysis*: In order to figure out what is going on inside our network, we further visualize features learned from each module and validate the effectiveness of the proposed network in a qualitative manner. In Fig. 7, a couple of feature parcels regarding bare soil, building, car, pavement, court, and tank is displayed for several example images. Note that for  $K$  feature maps in each feature parcel, we select

the most strongly activated one as the representative. We can observe that discriminative regions related to positive labels are highlighted in these feature maps, whereas less informative regions are weakly activated. As an exception, the feature map at the bottom left (BL) of Fig. 7 shows that the baseball field is misidentified as tanks, which may lead to incorrect predictions.

For evaluating the localization ability of the proposed network, we visualize attentional regions learned from the second module. Coordinates of BL and top right (TR) corners of attentional region grids are calculated with the following equation:

$$\begin{bmatrix} x_{BL}^{attn} & x_{TR}^{attn} \\ y_{BL}^{attn} & y_{TR}^{attn} \end{bmatrix} = M_{T_i} \begin{bmatrix} -1 & 1 \\ -1 & 1 \\ 1 & 1 \end{bmatrix}. \quad (9)$$

Fig. 8 shows some examples of learned attentional regions. As we can see, most attentional regions concentrate on areas

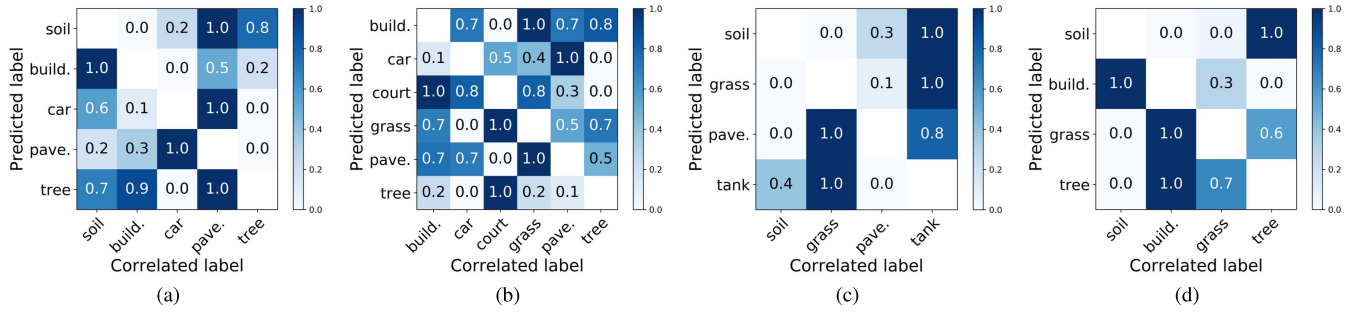


Fig. 9. Example pairwise relations among labels present in scene (a)–(d), which are shown in Fig. 8. Each label at Y-axis represents the predicted label  $l$ , and labels at X-axis are correlated labels. Normalization is performed according to each row, and white color represents null values.

TABLE V  
COMPARISONS OF THE CLASSIFICATION PERFORMANCE ON AID MULTILABEL DATA SET (%)

Network	mean $F_1$	mean $F_2$	mean $p_e$	mean $r_e$	mean $p_l$	mean $r_l$
VGGNet [64]	85.52	85.60	87.41	86.32	70.60	58.89
VGG-RBFNN [31]	84.58	85.99	84.56	87.85	62.90	69.15
CA-VGG-BiLSTM [34]	86.68	86.88	88.68	87.83	72.04	60.00
proposed AL-RN-VGGNet	88.09	88.31	89.96	89.27	76.94	68.31
GoogLeNet [59]	86.27	85.77	89.49	86.00	74.18	53.69
GoogLeNet-RBFNN [31]	84.85	86.80	84.68	89.14	65.41	72.26
CA-GoogLeNet-BiLSTM [34]	85.36	85.21	88.05	85.79	68.80	59.36
proposed AL-RN-GoogLeNet	88.17	88.25	90.03	88.77	77.92	69.50
ResNet-50 [60]	86.23	85.57	89.31	85.65	72.39	52.82
ResNet-RBFNN [31]	83.77	85.87	82.84	88.32	60.85	70.45
CA-ResNet-BiLSTM [34]	87.63	88.03	89.03	88.99	79.50	65.60
proposed AL-RN-ResNet	88.72	88.54	91.00	88.95	80.81	71.12

covering objects of interest. Besides, it is noteworthy that even objects are distributed dispersedly, the learned attentional regions can still cover most of them, e.g., buildings in Fig. 8(a) and cars in Fig. 8(b).

Furthermore, learned pairwise label relations are visualized in the format of matrix, where an element at  $(l, m)$  indicates  $LR(A_l, A_m)$ . Fig. 9 exhibits some examples for the four scenes in Fig. 8. In these examples, we take only positive object labels into consideration and perform normalization alongside each row to yield a distinct visualization of “label relations.” Since  $m$  differs from  $l$ , we assign null values to diagonal elements and mark them as white color in Fig. 9. It can be seen that in Fig. 9(a) and (b), relations between car and pavement contribute significantly to predicting the presence of both car and pavement. Besides, Fig. 9(d) shows that the existence of a tree highly suggests the presence of bare soil, but not vice versa. These observations illustrate that even without prior knowledge, the proposed network can reason about relations that are in line with reality.

### E. Results on the AID Multilabel Data Set

1) *Quantitative Analysis:* To further evaluate the proposed network, we report experimental results on the AID multilabel data set. Evaluation metrics here are the same as those in previous experiments, and results are presented in Table V. As we can observe, the proposed AL-RN-CNN behaves

superior to all competitors in most of the metrics. To be more specific, AL-RN-VGGNet improves the mean  $F_1$  and  $F_2$  score by 2.57% and 2.71%, respectively, compared to the baseline model. In comparison with CA-VGG-BiLSTM, our network gains an improvement of 1.41% in the mean  $F_1$  score and 1.43% in the mean  $F_2$  score. Regarding the other two backbones, similar phenomena can be observed as well. AL-RN-GoogLeNet achieves the highest mean  $F_1$  and  $F_2$  score, 0.8817 and 0.8825, compared to GoogLeNet and CA-GoogLeNet-BiLSTM, while AL-RN-ResNet surpasses the second-best model by 1.09% and 0.51% in the mean  $F_1$  and  $F_2$  score, respectively. Besides, it is noteworthy that although CA-GoogLeNet-BiLSTM shows a decreased performance compared to the baseline model, our network still achieves higher scores in all metrics. Moreover, we notice that the proposed AL-RN-CNNs outperform baseline CNNs by a large margin in the mean label-based recall, and the maximum improvement can reach 18.30%. In conclusion, these comparisons suggest that explicitly modeling label relations can improve the robustness and retrieval ability of a network. Several example predictions on the AID multilabel data set are presented in Table IV.

2) *Qualitative Analysis:* To dive deep into the model, we visualize label-specific features and attentional regions in Figs. 10 and 11, respectively. In Fig. 10, representative feature maps in various feature parcels for bare soil, building,

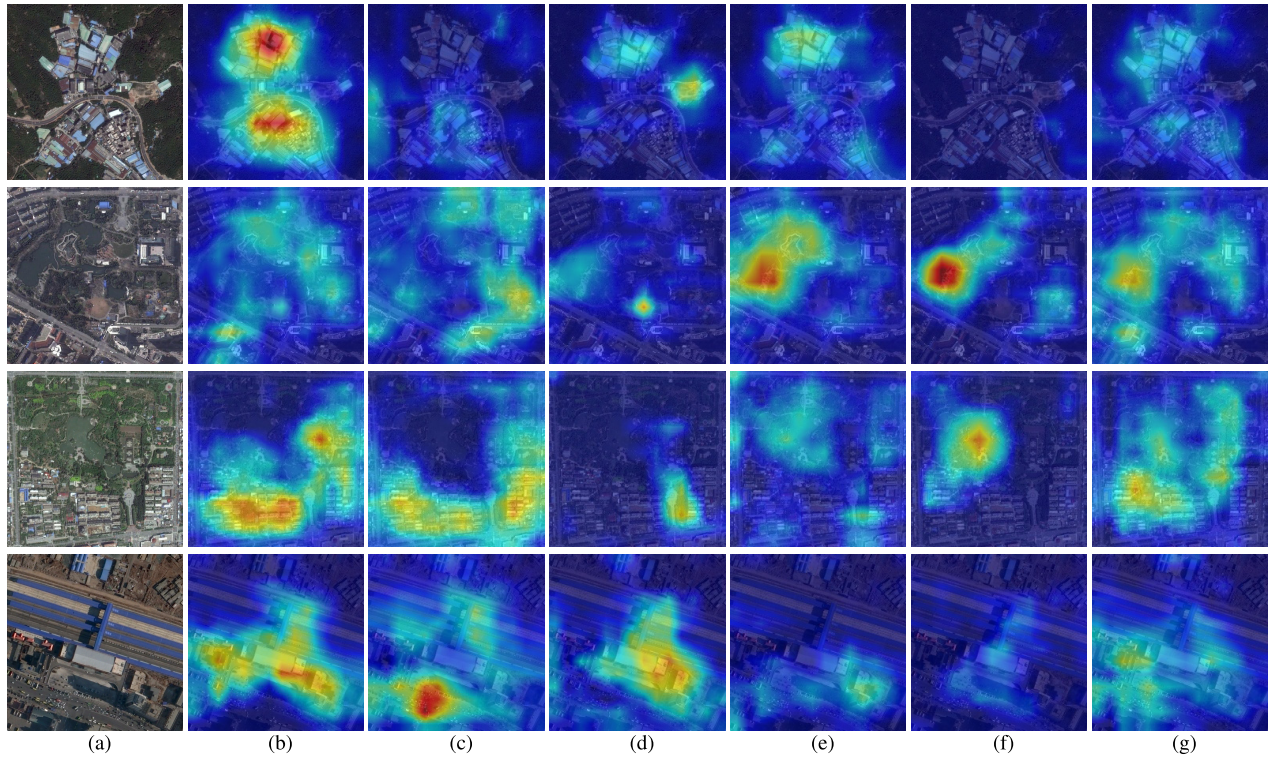


Fig. 10. Example label-specific features of (a) samples selected from the AID multilabel data set regarding, (b) building, (c) car, (d) bare soil, (e) tree, (f) water, and (g) pavement. Red implies strong activations, whereas blue indicates weak activations.

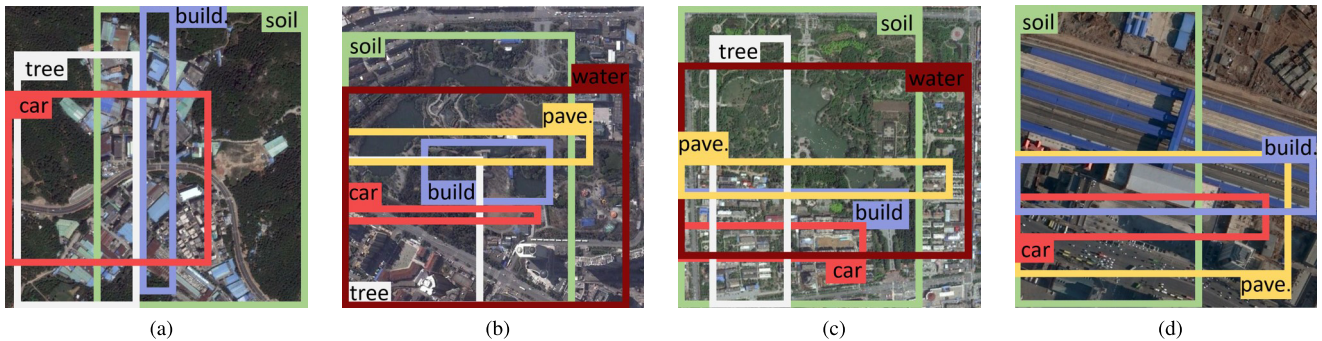


Fig. 11. Example attentional regions for car, bare soil (soil), building (build.), pavement (pave.), court, and tank in various scenes (a)–(d) in the AID multilabel data set. For each scene, only positive labels mentioned in Fig. 10 are considered.

car, pavement, tree, and water are displayed. As shown here, regions with label-related semantics are highlighted, while less informative regions present weak activations. For instance, regions of ponds are considered as discriminative regions for identifying *water*. Residential and industrial areas are strongly activated in feature maps for recognizing *building*. In Fig. 11, it can be observed that attentional regions learned from our network are able to capture areas of semantic objects, such as cars and trees. We also note that some attentional regions in Fig. 11 are coarser than those in Fig. 8, which is because the AID multilabel data set has a lower spatial resolution.

Furthermore, pairwise relations among positive labels are visualized in Fig. 12. As shown in Fig. 12(b)–(d), existences of both tree and pavement contribute significantly to the identification of car, while the occurrence of the car only

suggests a high probability that pavement presents. Strong pairwise relations between the building and other labels, e.g., car, pavement, and tree, indicate that the presence of building can heavily assist in predicting those labels.

#### F. Discussion on the Relational Inference Module

Regarding the relational inference module, the function  $g_{\theta_{lm}}$  is an important component, which reasons about relations between two objects. Hence, in this section, we discuss about different implementations of  $g_{\theta_{lm}}$ . Specifically, we compare our AL-RN-CNN with LR-CNN [65], which employs a global average pooling layer and an MLP as  $g_{\theta_{lm}}$ , on both the UCM and AID multilabel data sets. Experimental results are reported in Table VI. As shown in Table VI, our network

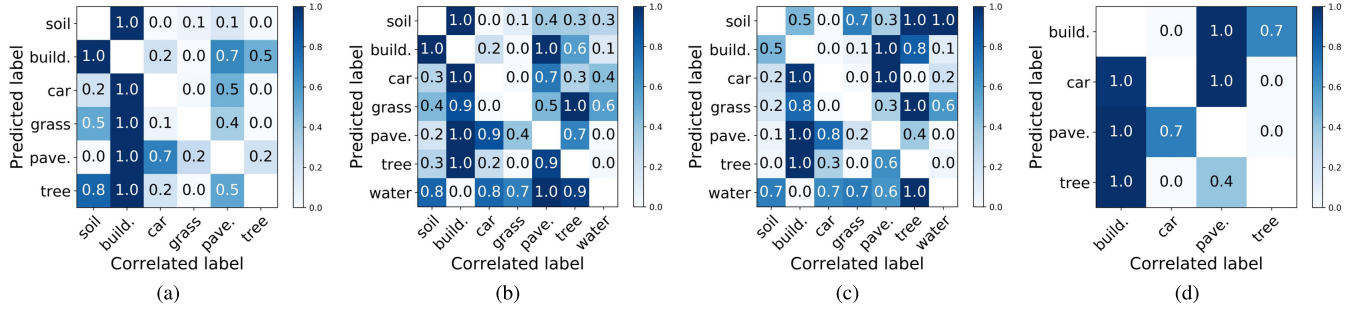


Fig. 12. Example pairwise relations among labels present in scene (a)–(d), which are shown in Fig. 11. Each label at Y-axis represents the predicted label  $l$ , and labels at X-axis are correlated labels. Normalization is performed according to each row, and white color represents null values.

TABLE VI  
COMPARISON BETWEEN DIFFERENT  $g\theta_{lm}$  (%)

Dataset	$g\theta_{lm}$	$V^*_{F_1}$	$G^*_{F_1}$	$R^*_{F_1}$	$V^*_{F_2}$	$G^*_{F_2}$	$R^*_{F_2}$
UCM mul.	MLP	82.11	83.02	85.36	81.99	84.02	86.09
	Conv.	85.70	85.24	86.76	85.81	85.33	86.67
AID mul.	MLP	87.79	84.92	87.10	87.74	86.97	86.83
	Conv.	88.09	88.17	88.72	88.31	88.25	88.54

$V^*_{F_1}$ ,  $G^*_{F_1}$ , and  $R^*_{F_1}$  indicate the mean  $F_1$  score achieved by VGGNet-, GoogLeNet-, and ResNet-based networks.

$V^*_{F_2}$ ,  $G^*_{F_2}$ , and  $R^*_{F_2}$  indicate the mean  $F_2$  score achieved by VGGNet-, GoogLeNet-, and ResNet-based networks.

gains the best mean  $F_1$  and  $F_2$  score on both data sets with variant backbones. AL-RN-VGGNet achieves the highest improvements of 3.59% and 3.82% for the mean  $F_1$  and  $F_2$  score, respectively, compared to LR-VGGNet on the UCM multilabel data set. AL-RN-GoogLeNet increases the mean  $F_1$  and  $F_2$  score by 3.25% and 1.28%, respectively, in comparison with LR-ResNet on the AID multilabel data set. Moreover, AL-RN-CNN can encode label relations through various fields of view by simply changing the size of convolutional filters in  $g\theta_{lm}$ .

#### IV. CONCLUSION

In this article, we propose a novel aerial image multilabel classification network, namely attention-aware label relational reasoning network. This network comprises three components: a label-wise feature parcel learning module, an attentional region extraction module, and a label relational inference module. To be more specific, the label-wise feature parcel learning module is designed to learn high-level feature parcels, which are proven to encompass label-relevant semantics, and the attentional region extraction module further generates finer attentional feature parcels by preserving only features located in discriminative regions. Afterward, the label relational inference module reasons about pairwise relations among all labels and exploit these relations for the final prediction. In order to assess the performance of our network, experiments are conducted on the UCM multilabel data set and a newly proposed AID multilabel data set. In comparison with other deep learning methods, our network can offer better classification results. In addition, we visualize extracted feature parcels, attentional regions, and relation matrices for demonstrating

the effectiveness of each module in a qualitative way. Looking into the future, such network architecture has several potentials, e.g., weakly supervised object detection and semantic segmentation.

#### ACKNOWLEDGMENT

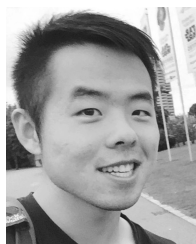
The authors would like to thank X. Liu for supporting this work with data annotation.

#### REFERENCES

- [1] D. Marmanis, K. Schindler, J. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogram. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.
- [2] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogram. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.
- [3] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogram. Remote Sens.*, vol. 145, pp. 96–107, Nov. 2018, doi: 10.1016/j.isprsjprs.2018.01.021.
- [4] L. Mou and X. X. Zhu, "RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images," May 2018, *arXiv:1805.02091*. [Online]. Available: <https://arxiv.org/abs/1805.02091>
- [5] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6699–6711, Nov. 2018.
- [6] L. Mou and X. X. Zhu, "Spatiotemporal scene interpretation of space videos via deep neural network and tracklet analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1823–1826.
- [7] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7147–7161, Dec. 2018.
- [8] S. Lucchesi, M. Giardino, and L. Perotti, "Applications of high-resolution images and DTMs for detailed geomorphological analysis of mountain and plain areas of NW Italy," *Eur. J. Remote Sens.*, vol. 46, no. 1, pp. 216–233, Jan. 2013.
- [9] L. Mou and X. X. Zhu, "IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," Feb. 2018, *arXiv:1802.10249*. [Online]. Available: <https://arxiv.org/abs/1802.10249>
- [10] Q. Weng, Z. Mao, J. Lin, and X. Liao, "Land-use scene classification based on a CNN using a constrained extreme learning machine," *Int. J. Remote Sens.*, vol. 39, no. 19, pp. 6281–6299, Oct. 2018.
- [11] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
- [12] P. Zarco-Tejada, R. Diaz-Varela, V. Angileri, and P. Loudjani, "Tree height quantification using very high resolution imagery acquired from an unmanned aerial vehicle (UAV) and automatic 3D photo-reconstruction methods," *Eur. J. Agronomy*, vol. 55, pp. 89–99, Apr. 2014.

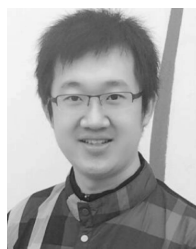
- [13] D. Wen, X. Huang, H. Liu, W. Liao, and L. Zhang, "Semantic Classification of Urban Trees Using Very High Resolution Satellite Imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1413–1424, Apr. 2017.
- [14] K. Nogueira, O. A. Penatti, and J. A. Dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, Jan. 2017.
- [15] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.
- [16] B. Demir and L. Bruzzone, "Histogram-based attribute profiles for classification of very high resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2096–2107, Apr. 2016.
- [17] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, Nov. 2015.
- [18] F. Hu, G.-S. Xia, W. Yang, and L. Zhang, "Recent advances and opportunities in scene classification of aerial images with deep models," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 4371–4374.
- [19] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.
- [20] X. Huang, H. Chen, and J. Gong, "Angular difference feature extraction for urban scene classification using ZY-3 multi-angle high-resolution satellite imagery," *ISPRS J. Photogram. Remote Sens.*, vol. 135, pp. 127–141, Jan. 2018.
- [21] L. Mou *et al.*, "Multitemporal very high resolution from space: Outcome of the 2016 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3435–3447, Aug. 2017.
- [22] 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. (GIS)*, 2010, pp. 270–279.
- [23] G.-S. 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.
- [24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
- [26] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," Nov. 2015, *arXiv:1511.00561*. [Online]. Available: <https://arxiv.org/abs/1511.00561>
- [27] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Aug. 2005, p. 1.
- [28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
- [29] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun, "Object detection networks on convolutional feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1476–1481, Jul. 2017.
- [30] K. Karalás, 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.
- [31] 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.
- [32] S. Koda, A. Zeggada, F. Melgani, and R. Nishii, "Spatial and structured SVM for multilabel image classification," *IEEE Trans. Geosci. Remote Sens.*, to be published.
- [33] 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.
- [34] 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. Photogram. Remote Sens.*, vol. 149, pp. 188–199, Mar. 2019.
- [35] W. Shao, W. Yang, G. Xia, and G. Liu, "A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization," in *Proc. Int. Conf. Comput. Vis. Syst.*, Jul. 2013, pp. 324–333.
- [36] V. Risojevic and Z. Babic, "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.
- [37] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [38] Q. Zhu, Y. Zhong, B. Zhao, G.-S. 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.
- [39] B. Teshome Zegeye and B. Demir, "A novel active learning technique for multi-label remote sensing image scene classification," *Proc. SPIE Photo-Opt. Instrum. Eng.*, vol. 10789, Oct. 2018, Art. no. 107890B.
- [40] 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.
- [41] 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. (IGARSS)*, 2019, pp. 5726–5729.
- [42] G. Sumbul, R. G. Cinbis, and S. Aksoy, "Fine-grained object recognition and zero-shot learning in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 770–779, Feb. 2018.
- [43] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C.-F. Wang, "Multi-label zero-shot learning with structured knowledge graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1576–1585.
- [44] A. Santoro *et al.*, "A simple neural network module for relational reasoning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4967–4976.
- [45] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3588–3597.
- [46] L. Wang, W. Li, W. Li, and L. Van Gool, "Appearance-and-relation networks for video classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1430–1439.
- [47] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 831–846.
- [48] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," Apr. 2019, *arXiv:1904.05730*. [Online]. Available: <https://arxiv.org/abs/1904.05730>
- [49] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 2017–2025.
- [50] 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.
- [51] F. Hu, G.-S. Xia, W. Yang, and L. Zhang, "Mining deep semantic representations for scene classification of high-resolution remote sensing imagery," *IEEE Trans. Big Data*, to be published.
- [52] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.
- [53] S. Srivastava, J. E. Vargas-Muñoz, and D. Tuia, "Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution," *Remote Sens. Environ.*, vol. 228, pp. 129–143, Jul. 2019.
- [54] W. Huang, Q. Wang, and X. Li, "Feature sparsity in convolutional neural networks for scene classification of remote sensing image," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 3017–3020.
- [55] C. Qiu, L. Mou, M. Schmitt, and X. X. Zhu, "Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network," *ISPRS J. Photogram. Remote Sens.*, vol. 154, pp. 151–162, Aug. 2019.
- [56] C. Qiu, L. Mou, M. Schmitt, and X. X. Zhu, "Fusing multi-seasonal Sentinel-2 imagery for urban land cover classification with residual convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, to be published.
- [57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. 2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [58] T. Dozat, *Incorporating Nesterov Momentum into Adam*. Accessed: Sep. 22, 2017. [Online]. Available: [http://cs229.stanford.edu/proj2015/054\\_report.pdf](http://cs229.stanford.edu/proj2015/054_report.pdf)

- [59] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [61] X. Wu and Z. Zhou, "A unified view of multi-label performance measures," Sep. 2016, *arXiv:1609.00288*. [Online]. Available: <https://arxiv.org/abs/1609.00288>
- [62] *Planet: Understanding the Amazon From Space*. [Online]. Available: <https://www.kaggle.com/c/planet-understanding-the-amazon-from-space#eva>
- [63] G. Tsoumakas and I. Vlahavas, "Random klabelsets: An ensemble method for multilabel classification," in *Proc. Eur. Conf. Mach. Learn.*, 2007, pp. 406–417.
- [64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv:1409.1556*. [Online]. Available: <https://arxiv.org/abs/1409.1556>
- [65] 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. (IGARSS)*, Jul. 2019, pp. 5244–5247.



**Yuansheng Hua** (Student Member, IEEE) received the bachelor's degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2014, and the master's degree in earth oriented space science and technology (ESPACE) from the Technical University of Munich (TUM), Munich, Germany, in 2018. He is currently pursuing the Ph.D. degree with the German Aerospace Center (DLR), Weßling, Germany, and the Technical University of Munich (TUM), Munich, Germany.

In 2019, he was a Visiting Researcher with Wageningen University & Research, Wageningen, The Netherlands. His research interests include remote sensing, computer vision, and deep learning, especially their applications in remote sensing.



**Lichao Mou** (Student Member, IEEE) received the bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, and the master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015. He is currently pursuing the Ph.D. degree with the German Aerospace Center (DLR), Weßling, Germany, and the Technical University of Munich (TUM), Munich, Germany.

In 2015, he spent six months at the Computer Vision Group, University of Freiburg, Freiburg im Breisgau, Germany. In 2019, he was a Visiting Researcher with the University of Cambridge, Cambridge, U.K. His research interests include remote sensing, computer vision, and machine learning, especially deep networks and their applications in remote sensing.

Mr. Mou was a recipient of the First Place in the 2016 IEEE GRSS Data Fusion Contest and Finalists for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and the 2019 Joint Urban Remote Sensing Event.



**Xiao Xiang Zhu** (Senior Member, IEEE) received the M.Sc., Dr.Ing., and the Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is currently the Professor for signal processing in earth observation (SiPEO) with the Technical University of Munich (TUM) and the German Aerospace Center (DLR), Weßling, Germany, the Head of the Department of EO Data Science with the DLR's Earth Observation Center, and the Head of the Helmholtz Young Investigator Group SiPEO with DLR and TUM. She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, The University of Tokyo, Tokyo, Japan, and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Since 2019, she has been co-coordinating the Munich Data Science Research School. She is also leading the Helmholtz Artificial Intelligence Cooperation Unit (HAICU)—Research Field "Aeronautics, Space and Transport." Her main research interests are remote sensing and Earth observation, signal processing, machine learning and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of the Young Academy (Junge Akademie/Junges Kolleg) with the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.