Gated Recurrent Multiattention Network for VHR Remote Sensing Image Classification

Boyang Li, Yulan Guo, Jungang Yang, Longguang Wang, Yingqian Wang, and Wei An

***Abstract*— With the advances of deep learning, many recent CNN-based methods have yielded promising results for image classification. In very high-resolution (VHR) remote sensing images, the contributions of different regions to image classification can vary significantly, because informative areas are generally limited and scattered throughout the whole image. Therefore, how to pay more attention to these informative areas and better incorporate them over long distances are two main challenges to be addressed. In this article, we propose a gated recurrent multiattention neural network (GRMA-Net) to address these problems. Because informative features generally occur at multiple stages in a network (i.e., local texture features at shallow layers and global profile features at deep layers), we use multilevel attention modules to focus on informative regions to extract more discriminative features. Then, these features are arranged as spatial sequences and fed into a deep-gated recurrent unit (GRU) to capture long-range dependency and contextual relationship. We evaluate our method on the UC Merced (UCM), Aerial Image dataset (AID), NWPU-RESISC (NWPU), and Optimal-31 (Optimal) datasets. Experimental results have demonstrated the superior performance of our method as compared to other state-of-the-art methods.**

***Index Terms*— Gated recurrent unit (GRU), multilevel attention mechanism, scene classification, very high-resolution (VHR) remote sensing.**

I. Introduction

With the development of satellite imaging sensors, very high-resolution (VHR) satellite images have become available for remote sensing (RS) scene classification [1]–[3] and thus promoted the prosperity of geospatial object detection [4], [5] land cover/land use classification [6], [7], and natural hazard detection [8]. Nevertheless, diverse semantic categories, complex spatial information, and high intraclass and low interclass variations in VHR RS images introduce great challenges to accurate classification. Consequently, it is necessary to develop a discriminative method for VHR RS image classification.

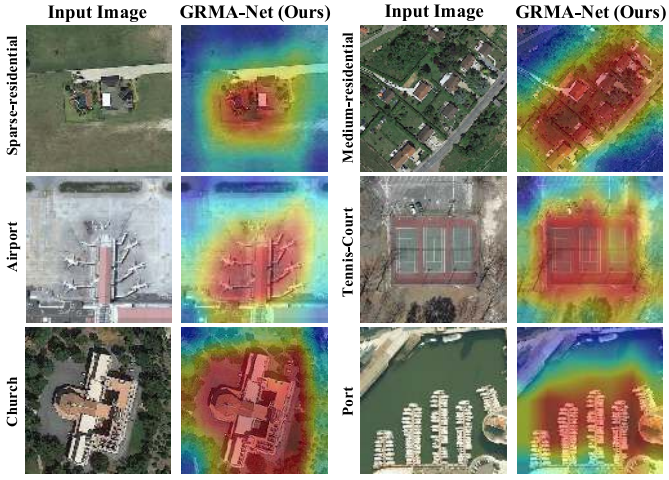


Fig. 1. Visualization of the attention maps produced by GRMA-Net for different VHR RS images. The informative and irrelevant areas are highlighted in red and blue. GRMA-Net can assign discriminative weights for informative areas and suppress the irrelevant ones.

As shown in Fig. 1, RS images generally have complex spatial structures. They usually cover a large-scale area with many types of objects. The informative areas usually occupy a small part of the image. Although the classic CNN (i.e., ResNets [9]) can generate the global representation by cascaded convolutions, they fail to assign discriminative weights to the informative local areas. The irrelevant areas cannot be well suppressed. This problem easily leads to misclassification of the network. Moreover, because of the long imaging distance, informative areas generally scatter around the whole image and exhibit complex spatial distribution. How to effectively aggregate these widely distributed features is the other problem to be solved.

Attention mechanism is widely used to address the allocation of available processing resources toward the most informative components of an input signal [11]. It has achieved promising results in the area of neural language processing (NLP) [12], [13] and image recognition [11], [14]–[17]. However, existing attention methods in RS field [18], [19] mainly concentrate on enhancing the global features description ability. It has been shown that multiscale local features are also important for RS image classification [20]–[24]. Intuitively, different layers have different

Manuscript received January 4, 2021; revised May 25, 2021; accepted June 27, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61972435, Grant 61401474, and Grant 61921001; and in part by the Tianjin Natural Science Foundation of China under Grant 18JCZDJC40300. *(Corresponding author: Jungang Yang.)*

Boyang Li, Jungang Yang, Longguang Wang, Yingqian Wang, and Wei An are with the College of Electronic Science and Technology, National University of Defense Technology (NUDT), Changsha 410073, China (e-mail: liboyang20@nudt.edu.cn; yangjungang@nudt.edu.cn; wanglongguang15@nudt.edu.cn; wangyingqian16@nudt.edu.cn; anwei@nudt.edu.cn).

Yulan Guo is with the College of Electronic Science and Technology, National University of Defense Technology (NUDT), Changsha 410073, China, and also with the School of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou 510275, China (e-mail: yulan.guo@nudt.edu.cn).

Digital Object Identifier 10.1109/TGRS.2021.3093914

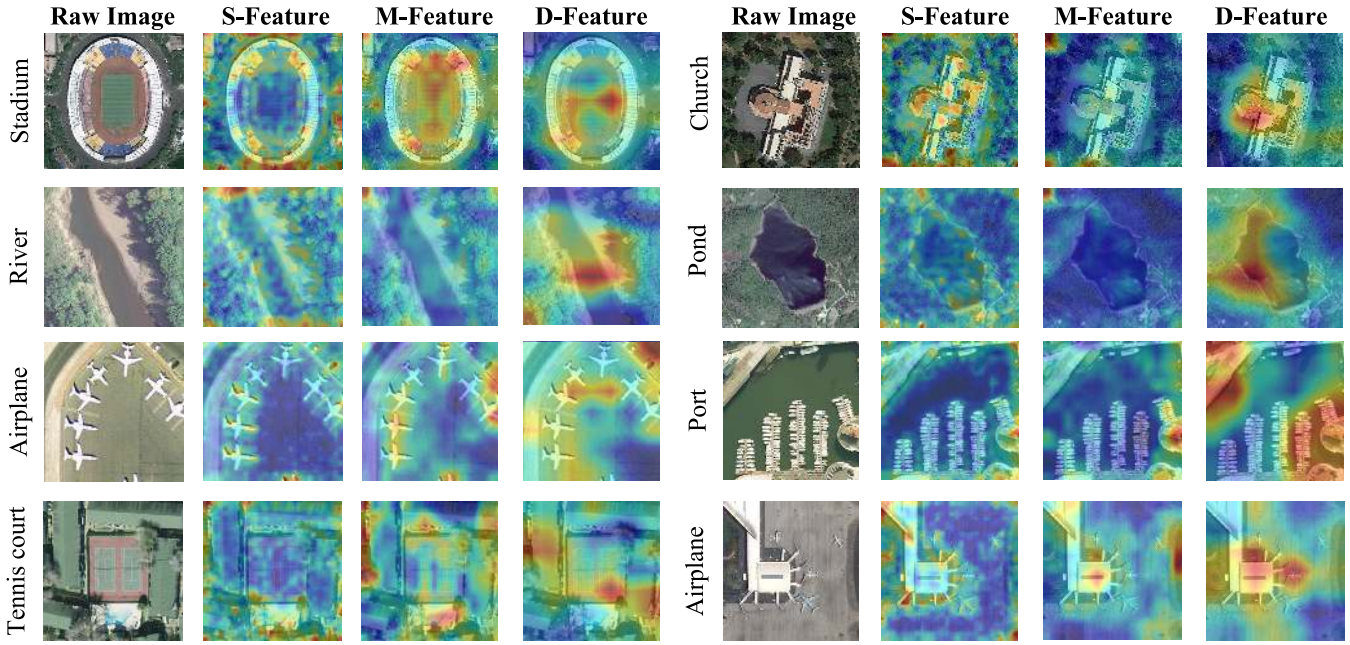


Fig. 2. Visualization of the change of interest regions using 10 randomly selected images from the AID dataset [10]. S-Feature, M-Feature, and D-Feature represent the features on the shallow, middle, and deep layers, respectively. These features are from the last convolutional layer of the conv2-x, conv3-x, and conv4-x blocks in the ResNet50 networks [9]. As the neural network goes deeper, the interest regions change from local texture to global profile.

regions of interest, as shown in Fig. 2. As the network goes deeper, the regions of interest grow from local textures to global profiles. These multiscale features are all essential to RS image classification. Therefore, it is nontrivial to incorporate attention mechanism in multiscale feature extraction for more powerful representations. To achieve effective aggregation of informative areas, pioneering works either directly concatenate multiscale features sequentially [25] or impose an adaptive factor on these features [26] to perform weighted summation. These methods do not fully exploit the spatial relationship and contextual dependency of these features. Actually, these widely distributed areas generally have rich spatial relationship and contextual dependency, which is essential for accurate classification.

To address the first problem, we design a multilevel attention module to focus on regions of interest at multiple scales, as shown in Fig. 3. High-level semantic information extracted by global features can be used to guide local features to focus on informative cues. If we directly add the multiscale local features and global features to generate attention map, the huge magnitude difference among multiscale features and global features will weaken the guidance of global features. Therefore, we introduce an adaptive convolution to adjust local features during feature aggregation. Inspired by the effectiveness of recurrent neural network (RNN) in modeling long-range dependency [12], [13], we introduce RNN to exploit the relationship among different locations. We re-arrange multiscale features as spatial sequences and then sequentially process them using a deep RNN.

In summary, the contribution of this article can be summarized as follows.

1) We propose a gated recurrent multiattention neural network (GRMA-Net) to address the problem of weak representation for local informative areas and weak dependency among widely distributed informative features.
2) A multilevel attention module and a gated recurrent unit (GRU)-based feature aggregation module are proposed to assign discriminative weights for multiscale local features and exploit the spatial dependency of features at different locations, respectively. As shown in Fig. 1, our method can increase the response of informative areas and meanwhile suppress other areas.
3) It is demonstrated that our GRMA-Net has achieved the state-of-the-art performance on the UC Merced (UCM), AID, NWPU and Optimal.

The remainder of this article is organized as follows. Section II discusses the related work on VHR RS image classification and attention mechanism. Section III introduces the details of our GRMA-Net. Section IV presents the experimental results. Section V gives the conclusion.

II. Related Work

In this section, we briefly review the related work for VHR remote sensing scene classification and attention mechanism.

A. Scene Classification for VHR Remote Sensing Images

1) Hand-Crafted Feature-Based Methods: Hand-crafted feature-based methods have been extensively investigated before the wide application of deep learning. These methods mainly focus on human-designed feature extractors. Typical features include histogram of oriented gradient (HOG) [27], scale invariant feature transformation (SIFT) [28], local binary pattern (LBP) [29] and median robust extended local binary pattern (MRELBP) [30]. Then, post-encoding methods have been proposed to improve the discriminativeness of low-level semantic descriptors, including hierarchical coding

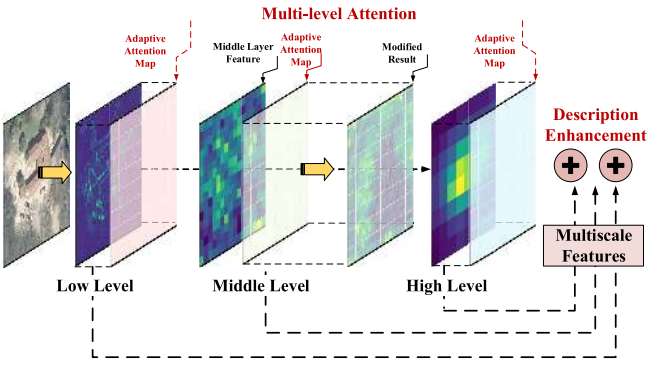


Fig. 3. Illustration of the multilevel attention module. The multiattention module imposes multiple attention maps on corresponding layer to enhance their feature representation ability and generate comprehensive representation.

vector (HCV) [31], spatial pyramid match kernel (SPMK) [32], and randomized spatial partition (RSP) [33].

Although these methods have achieved good performance, they are essentially low-level descriptors. Compared to deep features extracted by pretrained CNNs, these features are lack of high-level semantic information and suffer from limited performance

2) Deep Learning-Based Methods: Hu *et al.* [34] first used pretrained networks such as (e.g., VGG [35], AlexNet [36]) to extract high-level semantic features. Cheng *et al.* [37] and Li *et al.* [38] proposed multiple post-encoding methods (e.g., bag of visual word, fisher vector) to optimize extracted features. Afterward, Castelluccio *et al.* [39] adopted a pretrained GoogLeNet [40] and then fine-tuned it on the target RS dataset. Similarly, Li *et al.* [41] activated baseline CNNs layer by layer to search for the optimal activation strategy. These methods [34], [37]–[39], [42] transfer existing baseline CNNs without any modification for RS target dataset. Hence, they are inferior in high-level semantic representation as compared to recent deep-learning-based methods [43]–[45].

Subsequently, complex networks have been developed in deep-learning-based method in RS. Zhao *et al.* [43] proposed a multilayer perception structure to reduce the over-fitting problem. Liu *et al.* [44] adopted adaptive deep pyramid matching to enhance the multiscale representation ability. In [45], the cross entropy loss was replaced by the metric learning regularization to make baseline CNNs more discriminative. Because of the limited data of RS datasets, it is hard to train very deep networks with only thousands of images. Many deep networks (e.g., DenseNet [46], InceptionNet [47]), which perform well on the ImageNet dataset, cannot be well transferred into RS image classification.

Apart from traditional VHR RS image classification, some new subfields have drawn increasing attention recently, e.g., ship species classification [48], tree species classification [49] in fine-grained image classification, and high-dimension RS images retrieval [50] in multilabel image classification. These methods further explore rich details in RS images, which may ultimately contribute to RS image coarse classification.

B. Attention Mechanism in CNNs

The pioneering work of attention mechanism was developed for natural language processing (NLP). Later, attention mechanisms were introduced to solve different computer vision tasks such as image classification [16], [51], [52], fine-grained visual categorization [53], and image super-resolution [54]–[58]. Generally, attention mechanism in computer vision can be divided into three main categories: spatial, channel, and hybrid attention. Jaderberg *et al.* [15] proposed the first spatial attention-based learning method, named spatial transformer network (STN). Although STN is simple and shallow, it performs patch-level attention to achieve significant improvements over traditional classification methods [59], [60]. Wang *et al.* [61] proposed a refined pixel-level spatial attention network, in which nonlocal operations are employed to capture long-range dependencies to achieve further improvements over STN. Afterward, Hu *et al.* [11] proposed the first channel attention-based method, i.e., squeeze and excitation networks (SENet), to adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies among different channels. Subsequently, several attention mechanisms are developed to fuse both spatial and channel information. Wang *et al.* [51] proposed the first hybrid attention-based method (i.e., residual attention network). Specifically, residual attention learning was used in both spatial and channel domains to achieve further improvements over SENet. Similarly, Woo *et al.* [14] proposed a more general hybrid attention module, i.e., convolutional block attention module (CBAM), which can be integrated into any CNN architectures. CBAM consists of a channel and a spatial attention module. It helps the CNN to learn what and where to emphasize or suppress in images. Therefore, CBAM achieves further improvements over SENet. Although sophisticated attention modules have achieved better performance, they inevitably increase model complexity. Recent works [62], [63] pay more attention to lightweight design. Wang *et al.* [62] proposed a local cross-channel interaction-based method, i.e., efficient channel attention (ECANet), to generate channel attention through a fast 1-D convolution. In this way, the trade-off between network performance and complexity is achieved. Then, Hou *et al.* [63] proposed coordinate attention (CA) to factorize channel attention into two fast 1-D feature encoding processes, which aggregate features along the two spatial directions. In this way, CA achieves significant improvements with nearly no computational overhead.

Attention mechanism also achieves excellent performance in RS image classification. Wang *et al.* [18] imposed a spatial attention map on the last feature map of backbone CNNs to improve their global representation ability and thus obtained significant improvements over traditional classification methods [34], [38]. Afterward, Tong *et al.* [64] proposed a channel attention-based learning method, i.e., channel attention-based densenet (CAD). They used DenseNet121 as the backbone and adopted a channel attention module to strengthen the important channels. Following CAD, Zhao *et al.* [65] proposed a hybrid attention-based method, i.e., enhanced attention module (EAM). They use ResNet101 as the backbone and adopt spatial and channel attention modules to enhance the features in both domains. Therefore, EAM achieves further improvements over [18]. Different from these works that plug attention modules into the backbone networks, some works

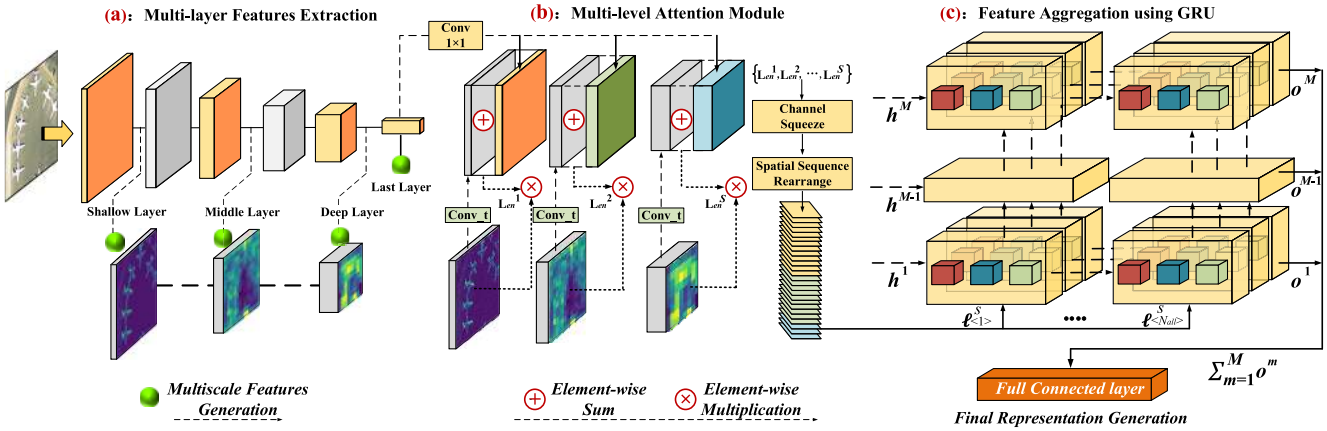


Fig. 4. Illustration of the proposed gated recurrent multiattention network. (a) Multiscale feature extraction. Input images are fed into the backbone CNN to extract multiscale local and global features. (b) Multilevel attention module. After feature extraction, the resultant features are fed into a softmax layer to generate the corresponding attention map. After element-wise multiplication, the enhanced multiscale features are obtained. (c) Optimization by GRU. Multiscale features are arranged as spatial location sequences. These sequences are fed into deep GRUs to fully exploit spatial relationship and contextual dependency.

try to use attention as a post-processing module at the end of networks. Li *et al.* [66] proposed multiinstance learning (MIL) by adding a spatial attention pooling module into the end of the network. Chen *et al.* [19] proposed an attention-guided sparse filter (SGSF) by embedding a spatial attention module into deep sparse filter networks. These methods achieved substantial performance improvements.

Although the performance is continuously improved by recent attention-based methods, the weak representation of local informative areas and weak dependency among widely distributed informative features have not been well addressed in literature. Therefore, our GRMA-Net first combines multilevel attention module and deep GRUs to both selectively enhance informative local features and capture contextual relationship of these widely distributed features. In this way, the informative areas can be given more attention and meanwhile the long-range dependency of these widely distributed features can be captured.

III. Methodology

In this work, we develop a multilevel attention module to enable the network to pay more attention to informative areas and suppress irrelevant areas. Besides, we propose a recurrent module to exploit the spatial relationship and contextual dependency among informative areas of an RS image. The overall architecture of the proposed method is shown in Fig. 4.

A. Overall Architecture

Section III-B introduces our multilayer feature extraction approach. Input images are first preprocessed and then fed into the backbone CNN to extract multiscale local features $\mathbf{L}^s \in \mathbb{R}^{C_s \times H_s \times W_s}$ and global feature $\mathbf{G} \in \mathbb{R}^{C_g \times 1 \times 1}$. Section III-C presents the multilevel attention module. Features $\mathbf{L}^s \in \mathbb{R}^{C_s \times H_s \times W_s}$ ($s \in \{1, 2, 3, \ldots, S\}$) at single scale are fed into a transition convolution to generate $\mathbf{L}_0^s$. The global feature $\mathbf{G} \in \mathbb{R}^{C_g \times 1 \times 1}$ is fed into a 1×1 convolution to generate $\mathbf{G}_0 \in \mathbb{R}^{C_s \times 1 \times 1}$ and then is stretched to the size of $\mathbf{G}_1 \in \mathbb{R}^{C_s \times H_s \times W_s}$. After element-wise sum between $\mathbf{L}_0^s$ and $\mathbf{G}_1$, the obtained score map $\boldsymbol{F}^s$ is fed into softmax operation to generate corresponding attention map $\boldsymbol{\alpha}^s$ at scale s. After element-wise multiplication $\mathbf{L}_{\text{en}}^s = \boldsymbol{\alpha}^s \otimes \mathbf{L}^s$, the enhanced multiscale features $\mathbf{L}_{\text{en}} = \{\mathbf{L}_{\text{en}}^1, \mathbf{L}_{\text{en}}^2, \ldots, \mathbf{L}_{\text{en}}^S\}$ are obtained. Section III-D shows the GRU optimization. Multiscale features are arranged as spatial location sequences $\mathbf{L}_{\text{en}} = \{\ell_1, \ell_2, \ldots, \ell_{N_{\text{all}}}\}$. These sequences are fed into deep GRUs to search for the optimal spatial relationship and contextual dependency. The image label is obtained by $\mathbf{Y} = \mathbf{GRU}(\mathbf{L}_{\text{en}})$.

B. Multiscale Feature Extraction

The multiscale feature extraction module consists of several cascaded layers. As shown in Fig. 2, as the neural network goes deeper, the interest region of the network changes from local textures to global profiles. Because these features are all important to RS image classification, we design a multilevel attention module to improve the multiscale representation ability of backbone networks.

In our module, we first extract multiscale local features as the input of the attention operation. Here, the local feature at scale s is given as

$$\mathbf{L}^s = \left\{\mathbf{l}_1^s, \mathbf{l}_2^s, \mathbf{l}_3^s, \ldots, \mathbf{l}_{N_s}^s\right\} \tag{1}$$

where C_s, H_s, W_s denote the number of channels, height, and width of $\mathbf{L}^s$, respectively. $\mathbf{l}_n^s$ represents the value of local feature $\mathbf{L}^s$ at spatial location $n \in \{1, 2, 3, \ldots, N_s\}$, at a given convolutional layer $s \in \{1, 2, 3, \ldots, S\}$. Then, global feature $\mathbf{G} \in \mathbb{R}^{C_g \times 1 \times 1}$ is also generated by the first nonconvolutional layer before the softmax layer. C_g denotes the channels of $\mathbf{G}$.

C. Multilevel Attention Module

Assume $\mathbf{L}$ denotes the local coarse feature, $\mathbf{G}$ is the global discriminative feature. High-level semantic information extracted by global features can be used to guide local features to focus on informative cues. If we directly add multiscale

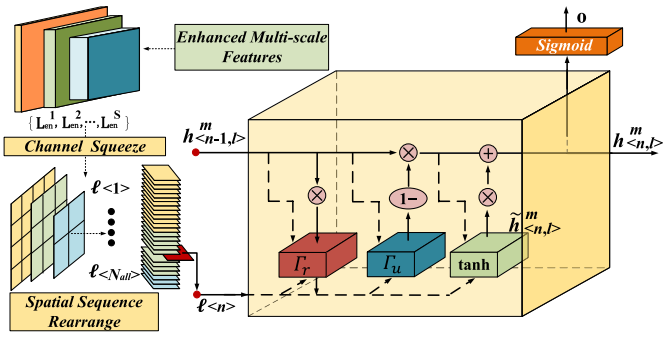


Fig. 5. Architecture of GRU. GRU integrates multiscale features to enhance their mutual long spatial relationship.

local features and global features to generate attention map, the huge magnitude difference among multiscale features and global features will weaken the guidance of global features.

Therefore, we first feed the local features $\mathbf{L}^s \in \mathbb{R}^{C_s \times H_s \times W_s}$ into transition convolution *Conv_t* to adaptively adjust their magnitudes at scale s, resulting in $\mathbf{L}_0^s$

$$\mathbf{L}_0^s = \text{Conv\_}t(\mathbf{L}^s), \quad \mathbf{L}_0^s \in \mathbb{R}^{C_s \times H_s \times W_s}. \tag{2}$$

The global feature $\mathbf{G}$ is fed to a 1×1 convolution to generate $\mathbf{G}_0 \in \mathbb{R}^{C_s \times 1 \times 1}$. Then, $\mathbf{G}_0$ is stretched to the size of $\mathbf{G}_1 \in \mathbb{R}^{C_s \times H_s \times W_s}$. After element-wise sum between $\mathbf{L}_0^s$ and $\mathbf{G}_1$. The score map $\boldsymbol{F}^s$ at scale s can be generated according to

$$\boldsymbol{F}^s = \sigma\left(\mathbf{L}_0^s + \mathbf{G}_1\right) \tag{3}$$

where σ is ReLU activation function.

Once $\boldsymbol{F} = \{\boldsymbol{F}^1, \boldsymbol{F}^2, \ldots, \boldsymbol{F}^S\}$ is generated, a softmax layer is used to obtain the normalized attention map

$$\boldsymbol{\alpha}_n^s = \frac{\exp\left(\boldsymbol{f}_n^s\right)}{\sum_{n=1}^{N_s} \exp\left(\boldsymbol{f}_n^s\right)}, \quad n \in \{1, 2, 3, \ldots, N_s\} \tag{4}$$

where $\boldsymbol{f}_n^s$ denotes the score map $\mathbf{F}_n^s$ at location n, at a given scale $s \in \{1, 2, 3, \ldots, S\}$.

Finally, we perform element-wise multiplication between the normalized attention weight value $\boldsymbol{\alpha}_n^s$ and corresponding local features $\mathbf{l}_n^s$. That is, $\mathbf{L}_{\text{en}}^s = \{\ell_1, \ell_2, \ldots, \ell_{N_s}\}$ is generated as the final descriptor for the image at each scale s.

D. Feature Aggregation Using GRU

In the multilevel attention module, we have extracted sufficient multiscale features, which are scattered throughout the images with long spatial ranges. How to better fuse these widely distributed features is a problem to be solved. RNN can naturally capture the mutual dependencies of information. As a special kind of RNN, as shown in Fig. 5, GRU can memorize long-range information to achieve better performance than normal RNN structures. To fully exploit long-range dependency among these local and global information, we use GRU in our network to sequentially process these multiscale features and automatically find the optimal combination through continuous iteration.

Similar to the application of GRU in NLP, which arranges features in time series, feature extracted by multiattention module can be considered as spatial series. As shown in Fig. 5, we first used an 1×1 convolution operation to squeeze the channel of multiscale features $\mathbf{L}_{\text{en}} = \{\mathbf{L}_{\text{en}}^1, \mathbf{L}_{\text{en}}^2, \ldots, \mathbf{L}_{\text{en}}^s\} \in \mathbb{R}^{C_{\text{en}} \times H_{\text{en}} \times W_{\text{en}}}$ into a single channel and generated $\mathbf{L}_{\text{en}} \in \mathbb{R}^{1 \times H_{\text{en}} \times W_{\text{en}}}$. Then, the single-channel features are stretched into an one-dimension sequence $\mathbf{L}_{\text{en}} = \{\ell_1, \ell_2, \ldots, \ell_{N_1}, \ell_1, \ell_2, \ldots, \ell_{N_s}, \ell_1, \ell_2, \ldots, \ell_{N_{\text{all}}}\} \in \mathbb{R}^{1 \times (H_{\text{en}} W_{\text{en}})}$. For feature $\boldsymbol{\ell}_n$ at the nth spatial location, mth recurrence and lth layer, the operation of GRU can be formulated as

$$\widetilde{\mathbf{h}}_{\langle n,l\rangle}^m = \tanh\left(\mathbf{W}_c\left[\boldsymbol{\Gamma}_r * \mathbf{h}_{\langle n-1,l\rangle}^m, \boldsymbol{\ell}_{\langle n,l\rangle}^m\right] + \mathbf{b}_c\right) \tag{5}$$

$$\boldsymbol{\Gamma}_u = \sigma\left(\mathbf{W}_u\left[\mathbf{h}_{\langle n-1,l\rangle}^m, \boldsymbol{\ell}_{\langle n,l\rangle}^m\right] + \mathbf{b}_u\right) \tag{6}$$

$$\boldsymbol{\Gamma}_r = \sigma\left(\mathbf{W}_r\left[\mathbf{h}_{\langle n-1,l\rangle}^m, \boldsymbol{\ell}_{\langle n,l\rangle}^m\right] + \mathbf{b}_r\right) \tag{7}$$

$$\mathbf{h}_{\langle n,l\rangle}^m = \boldsymbol{\Gamma}_{\mathbf{u}} * \widetilde{\mathbf{h}}_{\langle n,l\rangle}^m + (1 - \boldsymbol{\Gamma}_u) * \mathbf{h}_{\langle n-1,l\rangle}^m \tag{8}$$

$$\mathbf{o}_{\langle n,l\rangle}^m = \text{sigmoid}\left(\mathbf{W}_o * \mathbf{h}_{\langle n,l\rangle}^m + \mathbf{b}_c\right). \tag{9}$$

Note that, $\mathbf{h}_{\langle n,l\rangle}^m$, $\boldsymbol{\ell}_{\langle n,l\rangle}^m$, $\mathbf{o}_{\langle n,l\rangle}^m$ are the hidden state, input feature, and output feature at the nth spatial location, the mth recurrence, and the lth layer, respectively. $\boldsymbol{\Gamma}_{\mathbf{u}}$ and $\boldsymbol{\Gamma}_{\mathbf{r}}$ represent the update gate and reset gate, respectively. In each spatial step, these parameters determine whether the hidden state $\mathbf{h}_{\langle n,l\rangle}^m$ should be memorized or forgotten.

Then, as shown in Fig. 6, the hidden state $\mathbf{h}_{\langle n,l\rangle}^m$ is passed through all the layers and spatial locations to generate last-layer hidden state $\mathbf{h}^m$ and output $\mathbf{o}^m$ at the mth recurrence

$$\mathbf{h}^m = \left\{\mathbf{h}_{\langle N_{\text{all}},1\rangle}^m, \mathbf{h}_{\langle N_{\text{all}},2\rangle}^m, \ldots, \mathbf{h}_{\langle N_{\text{all}},L\rangle}^m\right\} \tag{10}$$

$$\mathbf{o}^m = \left\{\mathbf{o}_{\langle 1,L\rangle}^m, \mathbf{o}_{\langle 2,L\rangle}^m, \ldots, \mathbf{o}_{\langle N_{\text{all}},L\rangle}^m\right\} \tag{11}$$

where the last-layer hidden state $\mathbf{h}^m$ at the mth recurrence is treated as the initial hidden state at the $(m+1)$th recurrence. After M iterations, the output $\mathbf{o}^M$ at the Mth recurrence is generated as

$$\mathbf{o}^M = \left\{\mathbf{o}_{\langle 1,L\rangle}^M, \mathbf{o}_{\langle 2,L\rangle}^M, \ldots, \mathbf{o}_{\langle N_{\text{all}},L\rangle}^M\right\}. \tag{12}$$

Finally, $\mathbf{o}^m$ from all M iterations are summed and passed through a fully connected layer to generate the final output

$$\mathbf{Y} = \text{FC}\left(\sum_{m=1}^{M} \mathbf{o}^m\right). \tag{13}$$

IV. Experiment

The performance of our GRMA-Net is comprehensively evaluated in this section. We perform VHR remote sensing scene classification and attention map visualization experiments on the UCM [32], AID [10], NWPU [67], and Optimal [18] datasets. Our method is compared to several state-of-the-art methods.

A. Datasets

1) UC Merced Land-Use Dataset: The UCM dataset [32] is the most popular dataset in the area of VHR remote sensing scene classification. This dataset consists of 21 land-use classes. Each class contains 100 images of 256×256 pixels with an aerial-to-ground spatial resolution of 0.3 m per pixel. The challenge of the UCM dataset lies in its high intraclass, low interclass variations and highly overlapping land-use classes.

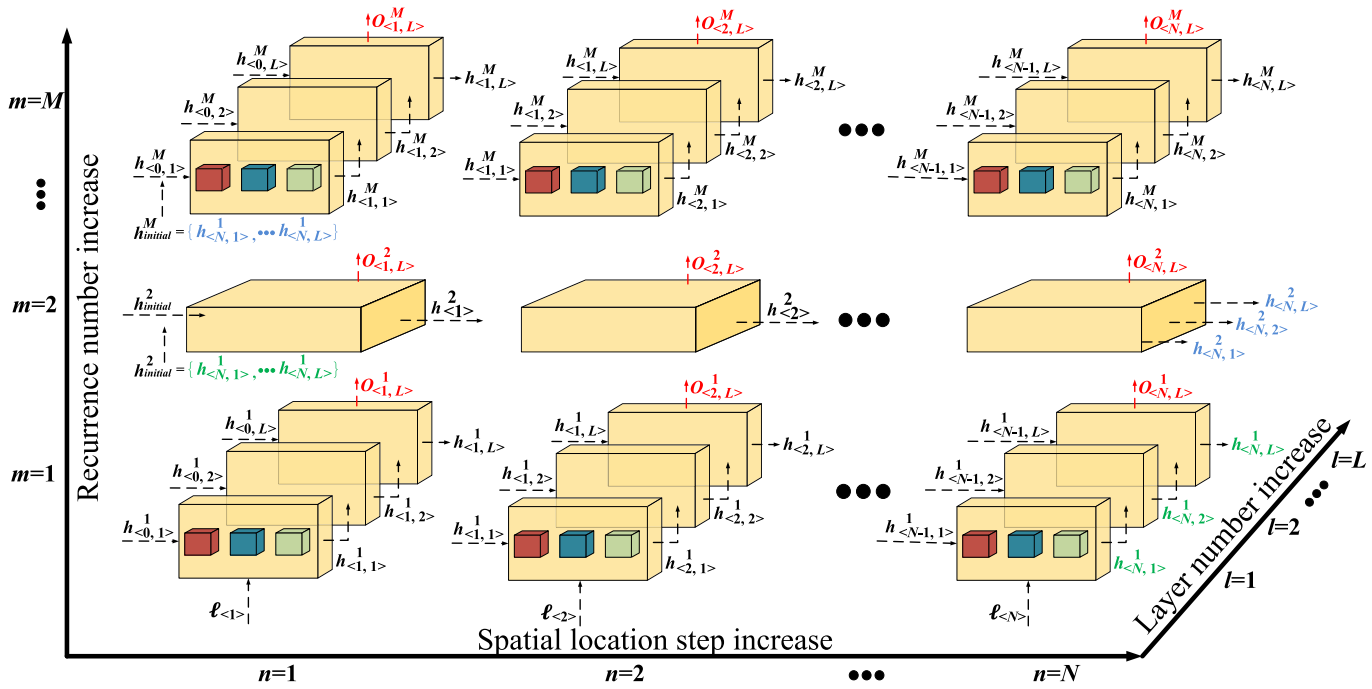


Fig. 6. Architecture of deep GRUs. Salient areas are split into multiple spatial steps. These spatial steps are fed into deep GRUs to capture long-range dependency.

2) Aerial Image Dataset: The AID [10] dataset is a large dataset for aerial scene image classification. It contains 30 common scene classes, while each class contains different number of images ranging from 220 to 420. The size of each image is 600 × 600 pixels with different aerial-to-ground spatial resolutions ranging approximately from 0.5 to 8 m. Variation of multiscale images and multicategory images are the two main challenges of this dataset.

3) NWPU-RESISC Dataset: The NWPU dataset [67] is the largest RS dataset. It contains 45 scene classes. Each class contains 700 images with a resolution of 256 × 256. The aerial-to-ground spatial resolution ranges from 0.2 to 30 m. Large image scale, rich spatial resolution variations, high intraclass diversity, and interclass similarity make this dataset really challenging.

4) OPTIMAL-31 Dataset: The OPTIMAL [18] is a small dataset with 31 classes. Each class contains only 60 images with a resolution of 256 × 256. Small size and multiple classes make it difficult for end-to-end training.

B. Evaluation Metrics

1) Overall Accuracy: Overall accuracy represents the ratio of correctly predicted images to overall images. In this article, we use the K-fold cross validation as the final classification result.

2) Inference Time: Inference time measures the computational efficiency of different algorithms. In this article, we use the inference time per image as the evaluation metrics.

C. Training Protocol

1) Data Augmentation: All input images with different initial sizes were first resized to a resolution of 256 × 256. Then, we randomly cropped these images into patches of size 224 × 224, performed randomly horizontal and vertical flipping, and randomly scaling for data augmentation. Afterward, we used color jitter to enrich image contrast. Finally, to accelerate the network convergence, these images were normalized by Z-score to ensure that their values are centered at zero.

2) Parameter Setting: We used ResNets (i.e., ResNet18, ResNet50, ResNet101) as backbone networks, which was pretrained on the ImageNet [69] dataset. The parameters of our designed modules were all initialized using the Xavier method [70]. We set the batch size to 64 and the learning rate to 0.001. Our model was trained using the stochastic gradient descent (SGD) optimization algorithm. The L2 weight decay regularization coefficient was set to 0.01, and the momentum was set to 0.9. The learning rate was decayed by a factor of 0.1 if the training loss does not decrease within 30 epochs.

3) Implementation Details: We modified the ResNets by adding three attention branches into corresponding convolutional layers. Given that ResNets was composed of four convolution blocks, we chose the final layer of conv2-x, conv3-x, and conv4-x block as shallow, middle, and deep layers, respectively. Moreover, the training process has two phases. We first trained the backbone network by 100 epochs on the RS dataset and then performed end-to-end training (including backbone network, multiple attention models, and deep GRUs) until convergence. Experiment results show that the network achieves promising performance with this training strategy.

4) Hardware and Software Platforms: All models were implemented in PyTorch [71] on a computer with an Intel i7 7700H @ 2.80 GHz CPU and an Nvidia GeForce1080Ti GPU.

TABLE I

OA VALUES ACHIEVED BY DIFFERENT SOTA METHODS ON THE UCM, AID, NWPU, AND OPTIMAL-31 DATASETS. MEAN ± STANDARD ERROR IS REPORTED

Method Description	UCM (OA%)		AID (OA%)		NWPU (OA%)		OPTIMAL-31
	Tr=50%	Tr=80%	Tr=20%	Tr=50%	Tr=10%	Tr=20%	Tr=80%
Transferring Deep CNN with IFK [34]	96.81 ± 0.29	98.49	89.88 ± 0.61	93.60 ± 0.39	87.66 ± 0.89	90.31 ± 0.65	94.66 ± 0.48
Retraining Deep CNN-ResNet101 [41]	97.39 ± 0.42	99.00 ± 0.21	94.27 ± 0.58	95.85 ± 0.22	91.54 ± 0.66	93.88 ± 0.52	95.68 ± 0.37
Two-stream fusion [24]	97.79 ± 0.56	98.90 ± 0.95	94.09 ± 0.34	95.99 ± 0.35	85.02 ± 0.25	87.01 ± 0.19	95.37 ± 0.43
Binary pattern encoded CNNs [68]	96.91 ± 0.36	97.72 ± 0.54	93.81 ± 0.12	95.73 ± 0.16	-	-	-
Discriminative CNN [45]	-	98.93 ± 0.10	90.82 ± 0.16	96.89 ± 0.10	89.22 ± 0.50	91.89 ± 0.22	-
Aggregated Deep Fisher Feature [38]	97.22 ± 0.45	98.81 ± 0.51	89.22 ± 0.56	93.22 ± 0.63	86.01 ± 0.15	88.79 ± 0.17	94.27 ± 0.29
Recurrent Attention Network [18]	96.81 ± 0.14	99.12 ± 0.40	88.75 ± 0.41	93.10 ± 0.55	91.69 ± 0.39	93.34 ± 0.88	92.70 ± 0.35
Multiple Attention Network [26]	98.34 ± 0.26	99.28 ± 0.21	94.75 ± 0.23	96.93 ± 0.16	91.08 ± 0.24	93.49 ± 0.17	96.22 ± 0.28
Channel Attention Based DenseNet [64]	98.57 ± 0.33	**99.66 ± 0.27**	95.73 ± 0.22	97.16 ± 0.26	92.70 ± 0.32	94.58 ± 0.26	-
Enhanced Attention Module [65]	98.81 ± 0.26	99.21 ± 0.26	94.26 ± 0.11	97.06 ± 0.19	91.91 ± 0.22	94.29 ± 0.09	96.45 ± 0.28
Multiple Instance Learning [66]	98.57 ± 0.44	99.45 ± 0.12	94.93 ± 0.18	96.60 ± 0.21	92.25 ± 0.21	93.92 ± 0.13	-
GRMA-Net-ResNet18 (ours)	98.81 ± 0.52	99.52 ± 0.13	94.58 ± 0.25	97.05 ± 0.37	92.84 ± 0.36	94.26 ± 0.27	96.23 ± 1.56
GRMA-Net-ResNet50 (ours)	98.90 ± 0.88	99.19 ± 0.10	95.43 ± 0.32	97.39 ± 0.24	93.19 ± 0.42	94.72 ± 0.25	96.71 ± 0.68
GRMA-Net-ResNet101 (ours)	**99.29 ± 0.29**	99.38 ± 0.18	**96.19 ± 0.48**	**97.84 ± 0.39**	**93.67 ± 0.21**	**95.32 ± 0.28**	**97.42 ± 0.46**

TABLE II

CONFIGURATION AND MAIN PARAMETERS (I.E., BACKBONE, ATTENTION METHOD, BATCHSIZE, LEARNING RATE, AND OPTIMIZER) FOR FIVE RECENT ATTENTION-BASED COMPARISON METHODS

Configuration & Parameters	Methods				
	ARCNet	MAN	CAD	EAM	MIL
Backbone	VGG16	VGG16	DenseNet121	ResNet101	VGG16
Pretrain Model	√	√	×	√	√
Attention	Channel	Spatial	Channel	Hybrid	Spatial
BatchSize	32	32	16	-	32
Learning Rate	0.0001	0.005	-	0.001	0.0001
Optimizer	Adam	SGD	SGD	NAG	Adam

D. Comparison to the State-of-the-Art Methods

To demonstrate the superiority of our methods, we compare our GRMA-Net to several state-of-the-art (SOTA) methods on the UCM [32], AID [10], NWPU [67], and OPTIMAL [18] datasets. As summarized in Table I, our GRMA-Net outperforms state-of-the-art methods on four benchmark datasets except for the UCM dataset (under a training ratio of 80%).

The parameter settings of five main attention-based compared methods are summarized in Table II. The introduction of these compared methods are listed as follows:

1) *ARCNet [18]:* It is the first work to combine attention mechanism and RNN. It used VGG-16 as backbone to extract global features and then optimized these features by LSTM.
2) *MAN [26]:* This article used VGG-16 as backbone to extract multilayer features. Then, this model aggregated these features and enhanced them by a channel attention module.
3) *CAD [64]:* This article used DenseNet121 as backbone and inserted SENet to adaptively strengthen the weights of the important feature channels.
4) *EAM [65]:* This article used ResNet101 as backbone and added CBAM to achieve hybrid attention. In this way, both informative spatial and channel features are enhanced.
5) *MIL [66]:* This article used VGG16 as backbone and replaced the max pooling with an attention mechanism, which considered the contribution of each instance to the bag label and achieved better performance.

1) Quantitative Results: Quantitative results are presented in Table I. Our GRMA-Net achieves the highest OA scores on four datasets (i.e., UCM [32], AID [10], NWPU [67], and OPTIMAL [18]). It is also worth noting that the improvements of OA scores achieved by our GRAM-Net on the AID and NWPU datasets are significant. That is because the spatial resolution of the AID and NWPU datasets vary significantly. Previous methods can generate the global representation by cascaded convolutions, they fail to assign discriminative weights to the informative local areas. Our GRMA-Net can capture long-range dependency to better exploit spatial cues over long distances by using the multilevel attention module and deep GRUs. Moreover, our method achieves much better results than existing RNN-based methods [18]. GRMA-Net-ResNet101 achieves an improvement of 7.44%. Our method achieves consistent improvements (1.44%, 0.46%, 1.93%,

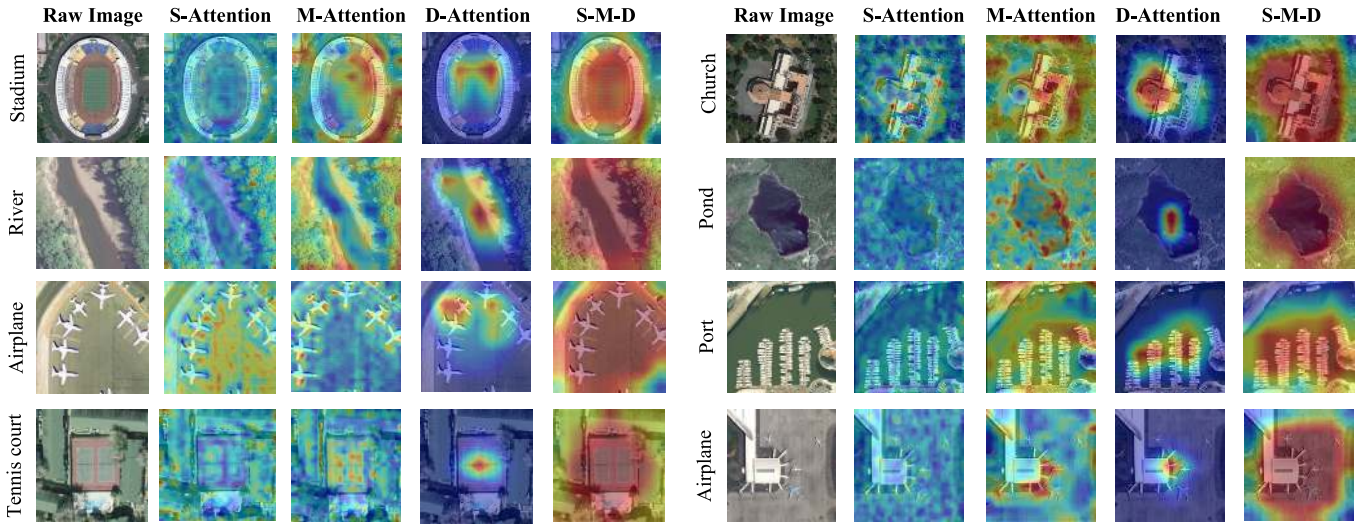


Fig. 7. Visualization of attention maps. We randomly selected ten images from the AID dataset [10]. S-Attention, M-Attention, and D-Attention denote the attention maps from shallow, middle, and deep layers in our GRMA-Net, respectively. S-M-D is the weighted average of all multiscale attention maps.

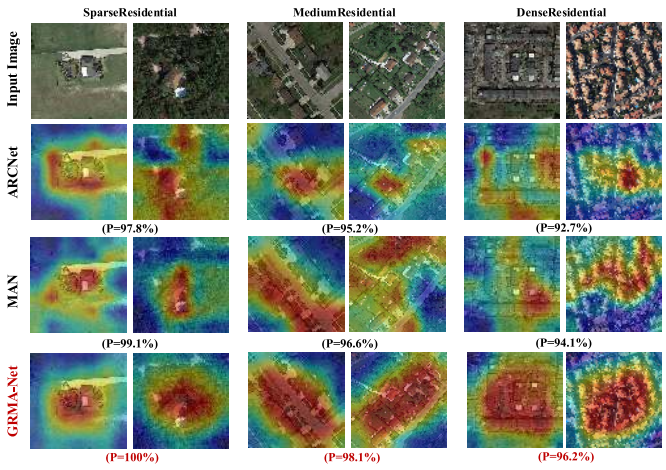


Fig. 8. Visualization of the attention maps produced by ARCNet [18], MAN [26], and our network. Our GRMA-Net can capture more informative areas and thus achieve higher confidence scores than previous attention-based methods. P means the classification accuracy of this subclass.

1.26% higher than [26], [64], [65], and [66], respectively) compared to the attention-based method on AID under a training ratio of 20%. Similar results are observed with the other datasets and training ratios. This demonstrates that the combination of multilevel attention and deep GRUs is effective.

2) Qualitative Results: We visualized the attention maps of 10 randomly selected images from the AID dataset in Fig. 7. It shows that shallow, middle, and deep attention maps have different interest regions. Specifically, the shallow, medium, and deep layers focus on local textures, key parts of objects, and central objects, respectively. It is also worth noting that, by comparing Fig. 7 and 2, the GRMA-Net captures more informative areas than the baseline method [9].

As shown in Fig. 8, when we compared GRMA-Net with previous attention-based methods [18], [26], our method can produce visualization maps containing more informative areas under higher confidence values. The irrelevant areas are suppressed, while the informative areas are assigned discriminative weights. That is because, our designed GRMA can effectively fuse multiscale informative features and fully exploit the spatial dependency of informative features at different locations. In this way, our GRMA-Net can achieve better performance. Comparative results are shown in Fig. 9. It can be observed that the statistical significance difference between GRMA-Net and recent attention-based methods is significant.

3) Computational Efficiency: We compared our GRMA-Net to several competitive methods (i.e., ADFF [38], ARC-Net [18], MIL [66], BAM [41]) in terms of the number of parameters (i.e., #Params) and FLOPs. Our GRMA-Net-ResNet18 achieves the best OA score with a small number of parameters and lower FLOPs. Because the deep GRU module is hard to converge, it takes more time to train the network. The time cost of both the first and second training phases are summarized in Table III. Although the training time of our network is longer than previous methods, the test time of our GRMA-Net-ResNet18 is the shortest. That is because, we adopt a lightweight RNN structure to capture long-range dependency. Compared to BAM, our network (GRMA-NetResNet18) achieves much better performance with a comparable model size.

E. Ablation Study

In this section, we compare our GRMA-Net with several variants to investigate the potential benefits introduced by our network modules and design choices.

1) Different Backbones: Because of the promising performance of ResNets in classification, we adopt three ResNet variants (i.e., ResNet18, ResNet50, ResNet101) as backbone networks in our GRMA-Net. As deeper networks generally achieve better classification accuracy, but introduce high computational burden, we evaluate the performance of different backbone networks to achieve a good trade-off between computational efficiency and classification accuracy. In this part,

TABLE III

COMPARISON TO SOTA METHODS IN TERMS OF PARAMETERS, FLOPS, TEST TIME, AND TRAINING TIME ON THE AID DATASET UNDER TRAINING RATIOS OF 50%. + MEANS TWO-PHASE TRAINING METHOD

Method	Evaluation Metircs			
	Params (M)	FLOPS (G)	Test (ms)	Train (h)
ADFF [38] (VGG16)	27.2	6.7	270.0	12.5
BAM [41] (ResNet101)	44.6	7.8	50.0	6.88
ARCNet [18] (VGG16)	**16.5**	15.4	1.9	**0.79**
MIL [66] (VGG16)	19.6	-	1.2	-
GRMA-Net (ResNet18)	17.1	**1.93**	**0.7**	0.73+3.60
GRMA-Net (ResNet50)	35.1	4.4	1.3	1.24+6.79
GRMA-Net (ResNet101)	54.1	8.2	2.1	1.68+9.37

TABLE IV

OA VALUES ACHIEVED BY GRMA-NET AND ITS VARIATIONS ON THE AID DATASET UNDER TRAINING RATIOS OF 20% AND 50%

Bakbone	Split Method	w/o MAM&DGM	w/o DGM	GRMA
ResNet18	AID (20% Train)	92.71 ± 0.46	94.08 ± 0.33	**94.58 ± 0.25**
	AID (50% Train)	94.46 ± 0.27	96.46 ± 0.24	**97.05 ± 0.37**
ResNet50	AID (20% Train)	93.29 ± 0.32	94.74 ± 0.41	**95.43 ± 0.32**
	AID (50% Train)	95.19 ± 0.22	96.68 ± 0.17	**97.39 ± 0.24**
ResNet101	AID (20% Train)	93.74 ± 0.22	95.25 ± 0.20	**96.19 ± 0.48**
	AID (50% Train)	95.55 ± 0.13	96.89 ± 0.13	**97.84 ± 0.39**

we gradually removed the multilevel attention module (MAM) and the deep GRU module (DGM) to evaluate the performance improvements introduced by the above modules for three backbone networks.

Experimental results on the AID dataset are summarized in Table IV. GRMA-Net-ResNet101 achieves the best performance. It introduces an improvement of 1.61%/0.79% in terms of OA scores than GRMA-Net-ResNet18 under training ratios of 20%/50% and introduces 0.76%/0.45% improvements than GRMA-Net-ResNet50 under training ratios of 20%/50%, respectively. It demonstrates that deeper backbones introduce larger classification improvements to GRMA-Net. Moreover, our MAM and DGM also introduce significant improvements on all backbone networks, resulting in an improvement of 2.45% and 2.29% in terms of OA scores on GRMA-Net-ResNet101 under training ratios of 20% and 50%, respectively.

Although deeper networks introduce larger classification performance improvements, they also cause a higher computational burden. We can see from Fig. 10 that as the network goes deeper, the improvements brought by two modules tend to be saturated, but the network parameters and computational cost increase significantly. For example, the improvements

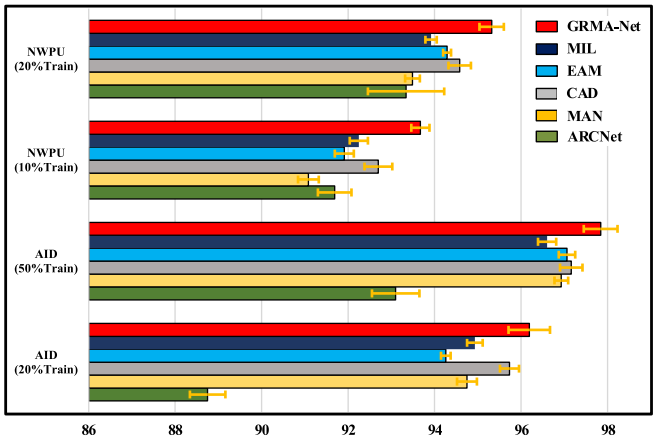


Fig. 9. OA values achieved by different recent attention-based methods on the AID and NWPU datasets. All experiments were tested by ten times. Mean ± standard error is reported.

of GRMA-Net-ResNet101 over GRMA-Net-ResNet18 are about 1.61% and 0.79% in terms of OA scores under training ratios of 20% and 50%, respectively. But the network parameters and computational cost increase 2.6 times and 3.7 times, respectively. It demonstrates that excessively increasing the depth of the network is not a good choice. GRMA-Net-ResNet18 achieves a better trade-off between classification accuracy and computational efficiency. Therefore, we use it as our basic model in the subsequent ablation study.

2) Multilevel Attention Module (MAM): As the core module of our GRMA-Net, MAM makes our network to pay more attention to informative areas at multiple levels. Here, we use attn S, attn M, and attn D to represent the attention modules at different stages and evaluate the effectiveness of MAM by introducing the following five variants:

1) *GRMA-Net w/o MAM:* We removed the multilevel attention module in this variant to investigate their contributions. Specially, we gradually replaced the attention modules with simple channel squeeze operation to keep the dimension identical as before.
2) *GRMA-Net w/o Score F:* We mainly investigate the benefit of score map $\boldsymbol{F}$. Specially, we replace the fused score map with simple self-scale score map, which means we do not introduce the global features $\mathbf{G}$ to instruct the distribution of multiscale local features $\mathbf{L}$.
3) *GRMA-Net w/o Conv_t:* To investigate the benefit introduced by the transition convolution *Conv_t*, we replaced the transition convolution a constant value (value = 1). It means the huge magnitude difference between local and global features cannot be adaptively adjusted by *Conv_t*.
4) *GRMA-Net With Channel Attention:* We used the channel attention operation of [14] to replace the spatial attention operation in this variant to investigate the effectiveness of channel attention.
5) *GRMA-Net With Hybrid Attention:* We replaced the spatial attention operation with hybrid attention [14] in this variant to investigate the effectiveness of hybrid attention.

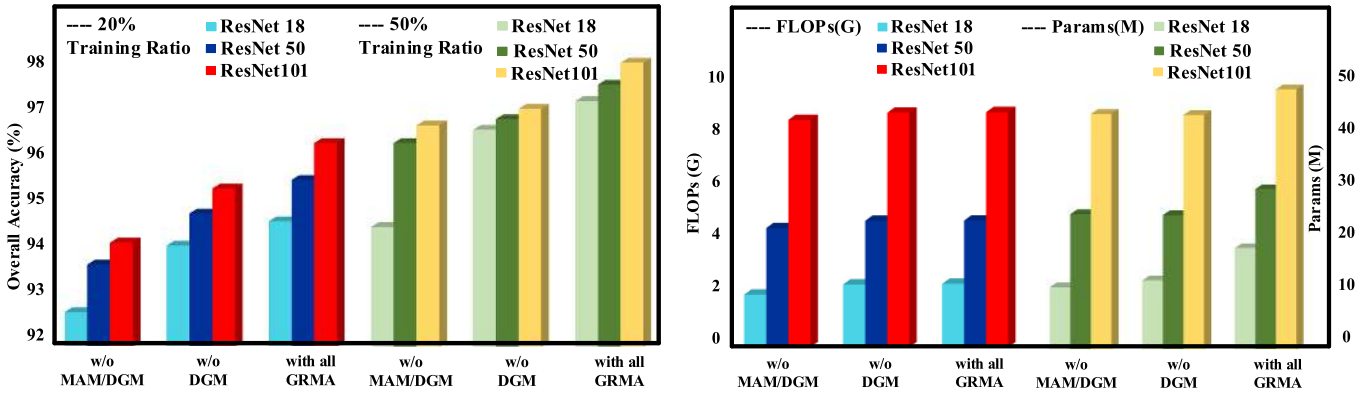


Fig. 10. Performance on different backbone networks. (a) OA performance on different backbone networks on the AID dataset under training ratios of 20% and 50%. (b) FLOPs and #Params of different backbone networks on the AID dataset under training ratios of 20% and 50%.

TABLE V
PERFORMANCE OF OUR NETWORK WITH DIFFERENT SETTINGS OF MAM ON THE AID DATASET UNDER TRAINING RATIOS OF 20% AND 50%

Model	#Params (M)	Datasets	
		AID (20%)	AID (50%)
GRMA-Net *w/o attn_S*	16.82	93.89 ± 0.49	95.88 ± 0.62
GRMA-Net *w/o attn_M*	16.51	93.74 ± 0.33	96.17 ± 0.32
GRMA-Net *w/o attn_S&M*	16.24	93.24 ± 0.36	95.67 ± 0.29
GRMA-Net *w/o score F*	16.29	93.07 ± 0.48	95.23 ± 0.29
GRMA-Net *w/o Conv_t*	16.75	94.19 ± 0.31	96.36 ± 0.19
GRMA-Net *with Channel-A*	17.13	94.32 ± 0.29	96.57 ± 0.33
GRMA-Net *with Hybrid-A*	17.14	**94.97 ± 0.26**	**97.28 ± 0.15**
GRMA-Net-ResNet18	17.10	94.58 ± 0.25	97.05 ± 0.37

TABLE VI
PERFORMANCE OF OUR NETWORK WITH DIFFERENT SETTINGS OF MAM ON THE AID DATASET UNDER TRAINING RATIOS 50%

Model	#Train Time (M)	#Test Time (ms)	AID (50%)
GRMA-Net *w/o DGM*	2.47	0.63	96.46 ± 0.22
GRMA-Net *with GCN*	7.76	1.79	96.83 ± 0.39
GRMA-Net-ResNet18	0.73+3.60	0.70	**97.05 ± 0.37**

Table V summarizes comparative results achieved by GRMA-Net and its variants. It can be observed that the OA value of GRMA-Net w/o attn_S&M suffers a decrease of 1.34% and 1.38% compared to GRMA-Net on the AID dataset under training ratios of 20% and 50%, respectively. That is because multilayer features contain rich local informative cues. Multilevel attention module helps to enhance the representation of these local features and thus achieve better performance. Moreover, the performance degradation is also significant for GRMA-Net w/o score F. It results in about 1.51% and 1.82% decrease. That is because the global feature **G** can help local features **L** to generate better distribution, which is important for the fusion of multiscale features.

It is worth noting that GRMA-Net w/o Conv_t suffers decreases of 0.39% and 0.69% on AID compared to GRMA-Net. Without conv_t, the huge magnitude gap between local and global features hinders our GRMA-Net to exploit mutual information. In contrast, conv_t can effectively alleviates this gap to facilitate our network to achieve better performance.

As summarized in Table V, GRMA-Net with channel attention suffers a decrease of 0.26% and 0.48% on AID as compared to GRMA-Net. That is because complex spatial distribution of RS images requires powerful spatial representation ability. Although channel attention help to capture informative feature channel, it cannot replace spatial attention.

When we replaced the spatial attention with hybrid attention, this new variant introduces minor improvements, which is 0.39% and 0.23% on the AID dataset compared to GRMA-Net. That is because both informative spatial areas and representative feature channels are enhanced by hybrid attention. In this way, GRMA-Net with hybrid attention achieves better performance. Because the objective of this article is to demonstrate the effectiveness of the proposed combination of multilevel attention module and deep GRU-based feature aggregation, we try to make our network architecture simple and did not use the delicately designed hybrid attention module for this minor performance improvement.

3) Deep GRU Module (DGM): Deep GRU module is used in our GRMA-Net to capture long spatial range dependency. Here, we validate the effectiveness of DGM by introducing the following three variants:

1) *GRMA-Net With GCN:* In this variant, we replaced DGM with a graph convolutional network (GCN) [72] to capture the spatial dependency of features at different locations.
2) *GRMA-Net w/o DGM:* We removed the DGM in this variant to investigate its contribution to GRMA-Net. Specially, we replaced the DGM with a fully connected layer to generate the predicted labels.

TABLE VII

PERFORMANCE ACHIEVED BY OUR NETWORK WITH DIFFERENT SETTINGS OF DGM ON AID UNDER A TRAINING RATIO OF 50%

Recurrence Number	Overall Accuracy (%)				
	#Layer=0	#Layer=1	#Layer=2	#Layer=3	#Layer=4
0	96.46	-	-	-	-
5	-	95.46	96.04	96.13	96.39
10	-	96.17	96.04	96.39	96.56
15	-	96.58	96.86	**97.05**	96.88
20	-	96.35	96.22	96.46	96.08

3) *Depth vs Width in DGM:* We investigate the two main components (i.e., recurrence number and layer number) in the experiments, where recurrence number represents the depth of DGM and layer number represents the width of DGM. The hidden size is fixed to 500.

As summarized in Table VI, both GRMA-Net with GCN and GRMA-Net-ResNet18 achieve obvious improvements in terms of OA scores over GRMA-Net w/o DGM. These spatial re-arrangement operations result in improvements of 0.59% and 0.37% for GRMA-Net-ResNet18 and GRMA-Net with GCN in term of OA values under AID dataset with 50% training ratio. That is because, the spatial re-arrangement operation can help to capture long-range dependency among multilevel features. Then, when we compare GRMA-Net with GCN with GRMA-Net-ResNet18, GRMA-Net with GCN suffers a decrease of 0.22% in terms of OA scores and increases of 3.43 h, 1.09 ms in terms of training time and test time over GRMA-Net-ResNet18. Although the GRMA-Net with GCN is hard to converge and needs longer test time, the comparable OA scores also demonstrate the effectiveness of GCN. The potential of GCN is worthy of further exploring.

As summarized in Table VII, GRMA-Net achieves an improvement of 0.59% (97.05% vs 96.46%) in terms of OA scores over GRMA-Net w/o DGM. This is because our DGM can better capture long-range dependency to achieve better performance. Moreover, we test the performance of our network with different numbers of GRU layers and recurrence. It can be observed that our network achieves the best performance with three GRU layers and ten iterations. It demonstrates that excessive recurrence and layer number can increase the difficulty of network fitting, leading to degraded performance.

V. CONCLUSION

In this article, we propose a GRMA-Net for VHR remote sensing scene classification. By incorporating multiscale attention module, our GRMA-Net can focus on informative regions at multiple scales to extract discriminative features. Moreover, our GRMA-Net uses GRUs to better exploit the spatial dependency and contextual relationship of features at different locations. Experimental results demonstrate the superiority of our GRMA-Net over state-of-the-art methods on four benchmark datasets.

REFERENCES

[1] 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.

[2] X. Zhang, S. Du, and Y. Zhang, "Semantic and spatial co-occurrence analysis on object pairs for urban scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2630–2643, Aug. 2018.

[3] W. Zhao and S. Du, "Learning multiscale and deep representations for classifying remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 113, pp. 155–165, Mar. 2016.

[4] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[5] S. Bhagavathy and B. S. Manjunath, "Modeling and detection of geospatial objects using texture motifs," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 12, pp. 3706–3715, Dec. 2006.

[6] J. Han, X. Yao, G. Cheng, X. Feng, and D. Xu, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[7] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[8] S. Cui and M. Datcu, "Comparison of approximation methods to Kullback–Leibler divergence between Gaussian mixture models for satellite image retrieval," *Remote Sens. Lett.*, vol. 7, no. 7, pp. 651–660, 2016.

[9] 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.

[10] 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.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[12] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1422–1432.

[13] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1480–1489.

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

[15] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 2017–2025.

[16] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 256–270.

[17] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," in *Medical Imaging With Deep Learning*. Amsterdam, The Netherlands: PMLR, 2018.

[18] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[19] J. Chen, C. Wang, Z. Ma, J. Chen, D. He, and S. Ackland, "Remote sensing scene classification based on convolutional neural networks pretrained using attention-guided sparse filters," *Remote Sens.*, vol. 10, no. 2, p. 290, Feb. 2018.

[20] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.

[21] L. Huang, C. Chen, W. Li, and Q. Du, "Remote sensing image scene classification using multi-scale completed local binary patterns and Fisher vectors," *Remote Sens.*, vol. 8, no. 6, p. 483, Jun. 2016.

[22] J. Zou, W. Li, C. Chen, and Q. Du, "Scene classification using local and global features with collaborative representation fusion," *Inf. Sci.*, vol. 348, pp. 209–226, Jun. 2016.

[23] P. Du, E. Li, J. Xia, A. Samat, and X. Bai, "Feature and model level fusion of pretrained CNN for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2600–2611, Aug. 2019.

[24] Y. Liu, Y. Liu, and L. Ding, "Scene classification based on two-stage deep feature fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 183–186, Feb. 2018.

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

[26] J. Ji, T. Zhang, L. Jiang, W. Zhong, and H. Xiong, "Combining multilevel features for remote sensing image scene classification with attention model," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1647–1651, Sep. 2020.

[27] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[28] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[29] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, Jan. 1996.

[30] L. Liu, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Local binary features for texture classification: Taxonomy and experimental study," *Pattern Recognit.*, vol. 62, pp. 135–160, Feb. 2017.

[31] H. Wu, B. Liu, W. Su, W. Zhang, and J. Sun, "Hierarchical coding vectors for scene level land-use classification," *Remote Sens.*, vol. 8, no. 5, p. 436, May 2016.

[32] 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.

[33] Y. Jiang, J. Yuan, and G. Yu, "Randomized spatial partition for scene recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Basel, Switzerland: Springer, 2012, pp. 730–743.

[34] 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.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 128–142.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2012, pp. 1097–1105.

[37] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.

[38] B. Li *et al.*, "Aggregated deep Fisher feature for VHR remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3508–3523, Sep. 2019.

[39] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," 2015, *arXiv:1508.00092*. [Online]. Available: http://arxiv.org/abs/1508.00092

[40] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[41] B. Li *et al.*, "Further exploring convolutional neural networks' potential for land-use scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1687–1691, Oct. 2020.

[42] K. Nogueira, O. A. B. 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.

[43] B. Zhao, B. Huang, and Y. Zhong, "Transfer learning with fully pretrained deep convolution networks for land-use classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1436–1440, Sep. 2017.

[44] Q. Liu, R. Hang, H. Song, F. Zhu, J. Plaza, and A. Plaza, "Adaptive deep pyramid matching for remote sensing scene classification," 2016, *arXiv:1611.03589*. [Online]. Available: http://arxiv.org/abs/1611.03589

[45] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[46] G. Huang, Z. Liu, V. D. M. Laurens, and K. Q. Weinberger, "Densely connected convolutional networks," ACM, New York, NY, USA, Tech. Rep., 2016.

[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 38–49.

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

[49] Z. He and D. He, "Bilinear squeeze-and-excitation network for fine-grained classification of tree species," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 7, pp. 1139–1143, Jul. 2020.

[50] 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.

[51] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[52] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," *Int. J. Autom. Comput.*, vol. 14, no. 2, pp. 119–135, Apr. 2017.

[53] J. Han, X. Yao, G. Cheng, X. Feng, and D. Xu, "P-CNN: Part-based convolutional neural networks for fine-grained visual categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 6, 2020, doi: 10.1109/TPAMI.2019.2933510.

[54] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[55] L. Wang *et al.*, "Learning parallax attention for stereo image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12250–12259.

[56] X. Dong, L. Wang, X. Sun, X. Jia, L. Gao, and B. Zhang, "Remote sensing image super-resolution using second-order multi-scale networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3473–3485, Apr. 2021.

[57] X. Ying, Y. Wang, L. Wang, W. Sheng, W. An, and Y. Guo, "A stereo attention module for stereo image super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 496–500, 2020.

[58] L. Wang *et al.*, "Parallax attention for unsupervised stereo correspondence learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 25, 2020, doi: 10.1109/TPAMI.2020.3026899.

[59] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.

[60] X. Zhang, J. Zou, X. Ming, K. He, and J. Sun, "Efficient and accurate approximations of nonlinear convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1984–1992.

[61] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.

[62] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.

[63] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," 2021, *arXiv:2103.02907*. [Online]. Available: http://arxiv.org/abs/2103.02907

[64] W. Tong, W. Chen, W. Han, X. Li, and L. Wang, "Channel-attention-based DenseNet network for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4121–4132, 2020.

[65] Z. Zhao, J. Li, Z. Luo, J. Li, and C. Chen, "Remote sensing image scene classification based on an enhanced attention module," *IEEE Geosci. Remote Sens. Lett.*, early access, Aug. 4, 2020, doi: 10.1109/LGRS.2020.3011405.

[66] Z. Li, K. Xu, J. Xie, Q. Bi, and K. Qin, "Deep multiple instance convolutional neural networks for learning robust scene representations," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3685–3702, May 2020.

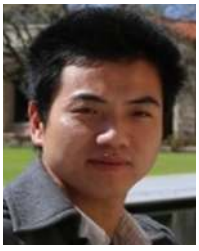
[67] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[68] R. M. Anwer, F. S. Khan, J. van de Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 74–85, Apr. 2018.

[69] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[70] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[71] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 8026–8037.
[72] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: http://arxiv.org/abs/1609.02907

Boyang Li received the B.E. degree in mechanical design manufacture and automation from Tianjin University, Tianjin, China, in 2017, and the M.S. degree in biomedical engineering from the National Innovation Institute of Defense Technology, Academy of Military Sciences, Beijing, China, in 2020. He is pursuing the Ph.D. degree in information and communication engineering with the National University of Defense Technology (NUDT), Changsha, China.

His research interests focus on VHR remote sensing image classification, infrared small target detection, and deep learning.



Yulan Guo received the B.Eng. and Ph.D. degrees from the National University of Defense Technology (NUDT), Changsha, China, in 2008 and 2015, respectively.

He was a Visiting Ph.D. Student with The University of Western Australia, Perth, Australia, from 2011 to 2014. He worked as a Post-Doctoral Research Fellow with the Institute of Computing Technology, Chinese Academy of Sciences, Guangzhou, China, from 2016 to 2018. He is an associate professor. He has authored over 100 journals articles and conferences papers, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and IJCV. His current research interests focus on 3-D vision, particularly on 3-D feature learning, 3-D modeling, 3-D object recognition, and scene understanding.

Dr. Guo received the ACM China SIGAI Rising Star Award in 2019, the Wu-Wenjun Outstanding AI Youth Award in 2019, and the CAAI Outstanding Doctoral Dissertation Award in 2016. He served as the Area Chair for CVPR 2021 and ICPR 2020, a Guest Editor for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, and an Associate Editor for *IET Computer Vision* and *IET Image Processing*.

Jungang Yang received the B.E. and Ph.D. degrees from the National University of Defense Technology (NUDT), Changsha, China, in 2007 and 2013, respectively.

He was a Visiting Ph.D. Student with The University of Edinburgh, Edinburgh, U.K., from 2011 to 2012. He is an Associate Professor with the College of Electronic Science, NUDT. His research interests include computational imaging, image processing, compressive sensing, and sparse representation.

Dr. Yang received the New Scholar Award of Chinese Ministry of Education in 2012, the Youth Innovation Award, and the Youth Outstanding Talent of NUDT in 2016.

Longguang Wang received the B.E. degree in electrical engineering from Shandong University (SDU), Jinan, China, in 2015, and the M.E. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2017, where he is pursuing the Ph.D. degree with the College of Electronic Science and Technology.

His research interests include low-level vision and deep learning.

Yingqian Wang received the B.E. degree in electrical engineering from Shandong University (SDU), Jinan, China, in 2016, and the M.E. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2018, where he is pursuing the Ph.D. degree with the College of Electronic Science and Technology.

His research interest focuses on low-level vision, particularly on light field imaging and image super -resolution.

Wei An received the Ph.D. degree from the National University of Defense Technology (NUDT), Changsha, China, in 1999.

She was a Senior Visiting Scholar with the University of Southampton, Southampton, U.K., in 2016. She is a Professor with the College of Electronic Science and Technology, NUDT. She has authored or coauthored over 100 journal and conference publications. Her current research interests include signal processing and image processing.