Attention Consistent Network for Remote Sensing Scene Classification

Xu Tang, *Member, IEEE*, Qiushuo Ma, Xiangrong Zhang, *Senior Member, IEEE*, Fang Liu, *Member, IEEE*, Jingjing Ma, *Member, IEEE*, and Licheng Jiao, *Fellow, IEEE*

***Abstract*—Remote sensing (RS) image scene classification is an important research topic in the RS community, which aims to assign the semantics to the land covers. Recently, due to the strong behavior of convolutional neural network (CNN) in feature representation, the growing number of CNN-based classification methods has been proposed for RS images. Although they achieve cracking performance, there is still some room for improvement. First, apart from the global information, the local features are crucial to distinguish the RS images. The existing networks are good at capturing the global features since the CNNs' hierarchical structure and the nonlinear fitting capacity. However, the local features are not always emphasized. Second, to obtain satisfactory classification results, the distances of RS images from the same/different classes should be minimized/maximized. Nevertheless, these key points in pattern classification do not get the attention they deserve. To overcome the limitation mentioned above, we propose a new CNN named attention consistent network (ACNet) based on the Siamese network in this article. First, due to the dual-branch structure of ACNet, the input data are the image pairs that are obtained by the spatial rotation. This helps our model to fully explore the global features from RS images. Second, we introduce different attention techniques to mine the objects' information from RS images comprehensively. Third, considering the influence of the spatial rotation and the similarities between RS images, we develop an attention consistent model to unify the salient regions and impact/separate the RS images from the same/different semantic categories. Finally, the classification results can be obtained using the learned features. Three popular RS scene datasets are selected to validate our ACNet. Compared with some existing networks, the proposed method can achieve better performance. The encouraging results illustrate that ACNet is effective for the RS image scene classification. The source codes of this method can be found in https://github.com/TangXu-Group/Remote-Sensing-Images-Classification/tree/main/GLCnet.**

***Index Terms*—Convolutional neural network (CNN), remote sensing (RS), scene classification.**

Manuscript received June 22, 2020; revised September 6, 2020 and October 7, 2020; accepted January 10, 2021. Date of publication January 14, 2021; date of current version February 1, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61801351, Grant 61802190, and Grant 61772400, in part by the China Postdoctoral Science Foundation Funded Project under Grant 2017M620441, in part by the Hong Kong Scholars Program under Grant XJ2019037, in part by the Aeronautical Science Foundation under Grant 20185181012, in part by the Xidian University Artificial Intelligence School Innovation Fund Project under Grant YJS2028, and in part by the Open Fund of Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education under Grant IPIU2019001. *(Corresponding author: Xu Tang.)*

Xu Tang is with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China, with the Science and Technology on Electro-optic Control Laboratory, and also with the Department of Computer Science, Hong Kong Baptist University, Hong Kong (e-mail: tangxu128@gmail.com).

Qiushuo Ma, Xiangrong Zhang, Jingjing Ma, and Licheng Jiao are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: 934627374@qq.com; xrzhang@mail.xidian.edu.cn; jjma@xidian.edu.cn; lchjiao@mail.xidian.edu.cn).

Fang Liu is with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: liufang_cs@njust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3051569

I. Introduction

With the development of remote sensing (RS) technology, an increasing number of RS images can be collected every day by the diverse earth observation satellites. Abundant information is provided by these images to scholars for understanding our planet. How to organize these huge volumes of RS images becomes an urgent and necessary task. As a fundamental and useful technology, RS image scene classification plays an important role in the RS community. Through assigning the semantic tags (e.g., "airport" and "beach") to the RS images, the large number of RS images could be categorized in different classes. Then, researchers can select the specific RS images according to the diverse semantics to accomplish their tasks. Due to this characteristic, RS image scene classification is popular in many practical applications, such as agriculture [1], hydrology [2], and forestry [3].

During the last decades, many successful RS image scene classification methods have been proposed [4]–[17]. At first, the two-stage framework dominates the scene classification community. In other words, researchers develop the methods to extract or learn the RS images' visual features first, and then some machine learning algorithms are adopted or designed to complete the categorization. For example, Sheng *et al.* [5] proposed a two-stage classification scheme in which the support vector machine (SVM) [18] is used to generate probability images with different handcrafted features in the first stage and the generated probability images with different features are fused in the second stage to obtain the final classification results. Another scene classification method was presented in the literature [10]. It extracts several handcrafted visual features from the RS images first. Then, the fully sparse semantic topic model is developed to fuse the contributions of diverse features. Finally, the fused features are classified by the SVM classifier. In this period, for the feature extraction/learning, the low-/mid-level visual features (e.g., Gabor feature [19] and bag of words feature [20]) are popular since they are easy in accomplishment and stable in performance. For the classification, some traditional machine

learning classifiers based on the statistical and Bayesian theories (e.g., random forest [21] and SVM) are favored by researchers.

Recently, due to the strong feature learning ability and the end-to-end classification framework, the convolutional neural network (CNN) attracts more and more scholars' attention. An ocean of CNN-based RS scene classification methods is proposed. Lu *et al.* [22] introduced an unsupervised deep feature learning method for scene classification. Inspired by the deconvolution network [23], [24], the weighted deconvolution network and the spatial pyramid model are combined to extract the effective features from RS images. Then, the scene classification results can be obtained by these features. Liu *et al.* [25] proposed a random-scale stretched CNN named SRSCNN, where the patches with a random scale are cropped from the image and stretched to the specified scale as input to train a CNN for solving the scale variation of the same object in a scene. To explore the semantic label information, Lu *et al.* [26] proposed an end-to-end feature aggregation CNN (FACNN). In FACNN, a supervised convolutional features' encoding module and a progressive aggregation strategy are proposed to leverage the semantic label information to aggregate the intermediate features. The good performance of the above methods mainly depends on CNN's hierarchical feature learning structure and a large number of labeled samples. However, they do not fully consider a key issue in the pattern classification [27], i.e., the interclass differences should be maximized and the intraclass variations should be minimized.

To overcome the limitation mentioned above, the Siamese network [28] is introduced to the RS images classification task. On the one hand, based on the basic CNN framework, the Siamese network can explore the high-level semantic features from the RS images. On the other hand, due to the dual-channel structure and specific loss function, the Siamese network can also mine the similarity relationships between the RS image pairs. For example, Liu *et al.* [29] imposed the metric learning regularization term on the original Siamese network, which enforces the Siamese networks to be more robust. Although the mentioned Siamese network based classification methods perform well for RS scenes, there is still some space to improve. First, to capture the complex contents within the RS images, not only the global information but also the local objects should be taken into account during the feature learning. In general, the global information of an RS image can be explored by the common CNN. Nevertheless, it is hard for a usual CNN to mine the objects from the RS images since they are diverse in type and huge in volume. Second, to compact the interclass samples and separate the intraclass images, the resemblance relationships between RS images should be measured from different aspects.

Based on the above discussion, we design a new method based on the Siamese network to accomplish RS image scene classification, which is named attention consistent network (ACNet). First, we adopt a popular CNN (VGG16Net [30]) to learn the intermediate feature maps (global information) from the RS image pairs. Second, with the help of the visual attention mechanism, a parallel-attention model is designed to mine the detailed information (objects' information) from the obtained feature maps. Third, to reduce the influence of differences between attention maps corresponding to the input pairs, we develop an attention consistent model here. Also, this model can narrow down the intraclass variations and enlarge interclass differences of RS images from the specific region aspect, which is beneficial to deeply explore the similarity relationships between RS images for the classification task. Finally, the RS scene classification results can be obtained using the learned deep features.

The major contributions of this article are as follows.

1) Based on the Siamese network, an end-to-end RS image scene classification model is proposed. The input RS image pairs for our model are constructed by the spatial rotation, which can not only augment the training data but also highlight the intraclass similarities.
2) Taking the characteristics of RS images into account, the parallel-attention model is developed to capture the local information from the spatial and spectral aspects. Accompanying with the global knowledge obtained by a successful CNN, the discrimination of the final features can be improved a lot.
3) To unify different kinds of attention maps and consider the resemblance between RS images for the scene classification, we design the attention consistent model. On the one hand, it can avoid the negative effects caused by the differences between attention maps of image pairs. On the other hand, this model is able to reduce the interclass differences and increase intraclass variations of RS images.
4) Extensive experiments are conducted on three benchmark datasets, and the encouraging results prove that our ACNet is effective for the RS image scene classification task.

The remainder of this article is organized as follows. Related work is reviewed in Section II. Then, the proposed ACNet is introduced in Section III. The evaluations of the proposed method are given in Section IV. Finally, the conclusions are summarized in Section V.

II. Related Work

RS scene classification is one of the challenging content understanding tasks in the RS community. In recent years, with the help of CNN, the performance of RS scene classification is enhanced dramatically. Here, we roughly divide the existing CNN-based classification methods into two groups according to their architecture.

In the first group, the structure of the classification networks is the *single-branch*. In other words, these networks have only one entrance. When the RS images are input the networks, they would be mapped into the feature vectors by some operations (e.g., convolution, pooling, fully connection, and other advanced techniques) for completing the classification task. The positive results of this kind of classification methods mainly depend on constructing the relationships between RS images and semantic labels by a large number of training data [31], [32]. At the very beginning, some classical CNNs (such as AlexNet [33], Overfeat [34], and VGG16 [30]) were applied to the RS scene classification directly [35]. Due to the pretrained weights (using ImageNet dataset [36]) and the strong feature learning capacity of CNN, the classification results are improved dramatically.

Then, considering the characteristics of RS images, a series of methods have been proposed based on the basic CNNs. In 2017, Han *et al.* [37] proposed a pretrained AlexNet-spatial pyramid pooling-side supervision model for RS scene classification, in which the spatial pyramid pooling and the side supervision models are embedded into a pretrained AlexNet to solve the problem of nonconvergence caused by the small quantities of RS images. Since the feature maps corresponding to different convolutional layers are fused, the multiscale information within RS images can be explored, so that this model enhances the RS image classification performance of AlexNet effectively. A multisource compensation network was proposed in the literature [38]. It combined a pretrained CNN, a cross-domain alignment model, and a classifier complement module to deal with the cross-source scene classification task. By finding a common space for RS images from different sensors, the homogeneous features of diverse RS images can be captured, which is beneficial to the multisource RS image scene classification. The mentioned two networks focus on mining the global information, however, the local information that is also important to scene classification is ignored. Fan *et al.* [39] proposed an attention-based residual network to fully explore the complex contents from RS images, where the common CNN with residual blocks is selected to mine the global information from RS images. Meanwhile, the visual attention mechanism is utilized to capture the local information from RS images by assigning the larger weights to key areas of RS images. Through combining the global and local information, the RS scene classification results were enhanced obviously. In 2019, Guo *et al.* [40] proposed an end-to-end global-local attention network (GLANet) in which the global attention blocks are designed to capture the global semantic information from RS images. Then, the local attention blocks, as a kind of attention mechanisms, are proposed to explicitly distinguish between key information and redundant information of RS images. Finally, to enhance the learning ability of the network, the two supplementary loss functions are applied to the GLANet. Although the mentioned two networks consider the global and local information simultaneously, and they achieve good performance. The similarities between RS images are not taken into account, which could further enhance the classification accuracy.

In the second group, the structure of the CNNs is the *multibranch*. Besides extracting the discriminative features from RS images, the methods in this group can deeply explore the intra/interclass relationships between RS images, which are important to the RS scene classification task. Among the diverse multibranch CNNs, the Siamese network [41] is a typical one. It combines two weight-shared CNNs and develops some specific objective functions to accomplish RS the image content understanding. Zhan *et al.* [42] proposed a deep Siamese convolution network for RS images. Different from the methods based on hand-crafted features, this model is developed to capture the visual features from the image pairs. Furthermore, the weighted contrastive loss function is imposed on the Siamese convolution network to ensure the discrimination of the features. Its effectiveness has been proved by the positive performance of the change detection task. In the literature [43], an RS scene classification network was introduced. It integrates the Siamese network and structural metric learning to accomplish the feature learning and develops the diversity-promoting scheme to enhance the representational ability of the network. In 2018, Ma *et al.* [44] proposed the Siamese hierarchical attention network (SHAN). To obtain the more discriminative semantic features from RS images, SHAN is designed based on the hierarchical recurrent structure. Duo to the characteristics of the Siamese network, it can effectively minimize the distance between the same class of samples and separate the distance between the different classes of samples, thereby improving the learning ability of the CNN. In 2019, Liu *et al.* [29] proposed a scene classification method based on the Siamese network, which consists of identification and verification models. The identification model is used to predict the input images' identity labels and the verification model is designed to measure the similarities between image pairs. Integrating those two models, the interclass samples could be compacted, whereas the intraclass images could be separated. Finally, the regularization term is imposed on the features, which effectively improve the performance of Siamese networks. The advantage of the above methods is developing the specific strategies for capturing the similarities between RS images, which could improve the classification performance through compacting/separating the RS images from the same/different classes. Nevertheless, they overlook the effectiveness of visual features. In other words, the characteristics of RS images are not fully taken into account during the feature learning.

III. Methodology

A. Overall Framework

Different from the ordinary multibranch network, which samples two images from the same/different classes to compose the positive/negative input data, our ACNet takes the image $\mathbf{I}$ and the image $\mathrm{T}(\mathbf{I})$ obtained by the spatial rotation as the input. The reasons for this operation are twofold. First, spatial rotation is a common data augmentation strategy. After the rotation, the volume of training data can be increased two times with visual perceptual consistency, which can not only reduce the overfitting risk but also help ACNet to fully understand RS images from different aspects. Second, since the RS images $\mathbf{I}$ and $\mathrm{T}(\mathbf{I})$ have the same semantics, this kind of input pairs can push ACNet to pay its attention to learn the rules for compacting the RS images from the same classes. In other words, the RS images from different semantic classes can be dispersed as well.

The framework of the proposed ACNet is shown in Fig. 1, which contains four parts, i.e., the intermediate feature extraction model, the parallel-attention model, the attention consistent model, and the classification model. First, the input image pairs $\mathbf{I}$ and $\mathrm{T}(\mathbf{I})$ are passed through the intermediate feature extraction model for the feature maps $\mathbf{X}$ and $\mathbf{X}'$, which contain the basic semantic information. Second, the parallel-attention model is developed to explore the complex contents within the RS images from the global and local aspects. Then, the convolutional representation is obtained by global average pooling (GAP) on the concatenated feature maps, which come from parallel attention models. Third, taking the influence of spatial rotation

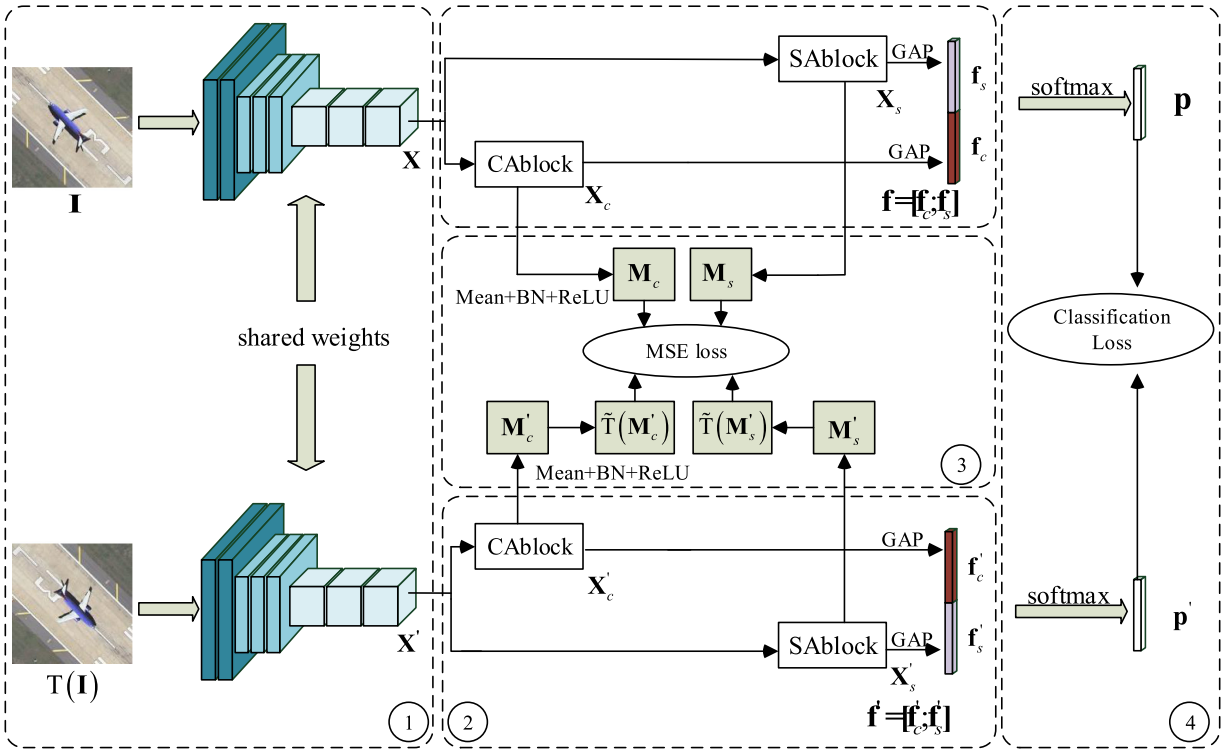

Fig. 1. Framework of our ACNet.

into account, we develop an attention consistent model to unify the attention areas of the image pairs with different angles. Also, this specific operation can compact the interclass samples and separate the intraclass images. Finally, the softmax function is combined to complete the classification. Note that, we assume that the spatial rotation $\mathrm{T}(\cdot)$ would not change the size of the images. Now, we discuss them in detail.

B. Intermediate Feature Extraction

In this article, rather than developing some new CNNs, we adopt VGG16 [30] to extract the intermediate features from the RS images. The whole VGG16 network consists of 13 convolution layers with the kernel size of 3×3, five max pooling layers with the size of 2×2, five nonlinear activation layers, and three fully connected (FC) layers. Here, our intermediate feature extraction network remains convolution layers, max pooling layers, and nonlinear activation layers. We use the pretrained weights (obtained by ImageNet [35]) to initialize the feature extraction model for speeding up the convergence. Note that, for the input RS image pairs $\mathbf{I}$ and $\mathrm{T}(\mathbf{I})$, the feature extraction networks are parameter shared. After the feature extracting mode, the feature maps $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ and $\mathbf{X}' \in \mathbb{R}^{H \times W \times C}$ can be obtained, where C, H, and W indicate the number of channels, height, and width of the feature map.

The reasons that we adopt VGG16 are summarized as follows. Considering the complex contents of RS images, we need a CNN that has strong nonlinear feature learning capacity. Also, taking the time complexity of network training into account, the select intermediate feature extraction network should be as light as possible. VGG16 just meets the demands mentioned above. Compared with some shallow CNNs (e.g., LeNet [45] and AlexNet [33]), VGG16 has more nonlinear layers that could improve the capacity of feature learning. Also, the sizes of convolutional kernels of VGG16 are 3×3, which are beneficial to capture the objects with diverse resolutions. Compared with some heavy models (e.g., GoogleNet [46] and ResNet101 [47]), optimizing VGG16 is not a tough issue. Therefore, we choose VGG16 as our intermediate feature extraction network in this article.

C. Parallel-Attention Models

The intermediate feature maps extracted by VGG16 can represent the contents of RS images from the global aspect. However, they are limited to explore the targets' information from RS images. To address this problem, we introduce the visual attention mechanism into our model. Furthermore, considering the characteristics of the complex targets within RS images, a parallel-attention model is designed to fully capture the local-level information for the scene classification task. Note that, since the operations on $\mathbf{X}$ and $\mathbf{X}'$ are the same, we use $\mathbf{X}$ as an example to explain our parallel-attention model clearly.

The parallel-attention model contains two blocks: channel-wise attention block (CAblock) and spatial-wise attention block (SAblock). CAblock focuses on emphasizing the significant channels and suppressing the insignificant channels of $\mathbf{X}$ to capture the salient targets from the RS images. To this end, we adopt the squeeze-and-excitation (SE) model [48] as our CAblock, and its framework is shown in Fig. 2. The SE model mines the

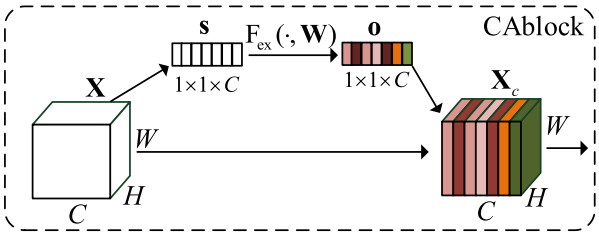

Fig. 2. Flowchart of CAblock.

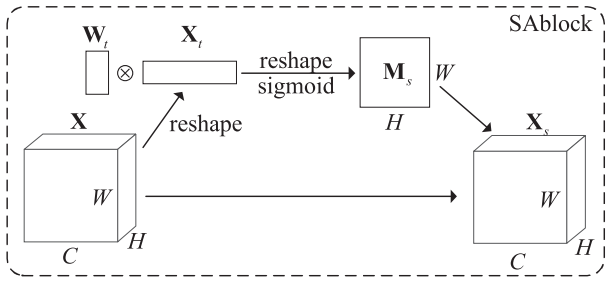

Fig. 3. Flowchart of SAblock.

local information from the input feature maps through adaptively assigning the weights to different channels within $\mathbf{X}$. First, to quantize the contributions of different feature responses, SE applies GAP on the input $\mathbf{X}$. The generated channel descriptor $\mathbf{s} \in \mathbb{R}^{1\times C}$ can be regarded as the compressed local descriptors. The squeeze operation mentioned above can be formulated as

$$s^C = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{x}^C(i,j) \tag{1}$$

where s^C indicates the Cth element of $\mathbf{s}$, and $\mathbf{x}^C$ means the Cth feature map in $\mathbf{X}$. Second, to construct correlations between channels and capture channel-wise dependencies, the channel descriptor $\mathbf{s}$ is passed through two FC layers with different activation functions. The output $\mathbf{o}$ is the channel weight that represents the importance of each channel. This excitation operation can be formulated as

$$\mathbf{o} = F_{\text{ex}}(\cdot, \mathbf{W}) = \eta\left(\mathbf{W}_2 \sigma\left(\mathbf{W}_1 \mathbf{s}\right)\right) \tag{2}$$

where $\mathbf{W_1}$ and $\mathbf{W_2}$ represent the parameters of two FC layers, $\sigma(\cdot)$ is the ReLU function, and $\eta(\cdot)$ is the sigmoid function. Finally, the output of SE $\mathbf{X}_c \in \mathbb{R}^{H\times \mathbb{W}\times C}$ can be obtained by the following function:

$$\mathbf{x}_c^l = o^l \cdot \mathbf{x}^l, l = 1, \ldots, C \tag{3}$$

where $\mathbf{x}_c^l$ means the lth channel of $\mathbf{X}_l$, o^l indicates the lth element of $\mathbf{o}$, and $\mathbf{x}^l$ denotes the lth feature map of $\mathbf{X}$.

SAblock aims to transform the input data $\mathbf{X}$ into a special space with the consideration of spatial information, in which the targets within RS images can be highlighted. For the output of the intermediate feature extraction network $\mathbf{X}$, each feature vector $\mathbf{f}_{\mathbf{X}}^{(i,j)} \in \mathbb{R}^{1\times 1\times C}$ corresponds to a $32 \times 32 \times 3$ spatial region of the RS image. To make clear the contributions of different spatial regions for exploring the local information, SAblock should learn a group of weights in the spatial domain for feature vectors within $\mathbf{X}$. To achieve this goal, we select a spatial attention (SA) model proposed in the literature [49] to be our SAblock. The structure of the SA model is exhibited in Fig. 3. First, to transform the input $\mathbf{X}$ into the spatial domain, the SA model reshapes $\mathbf{X}$ into $\mathbf{X}_t \in \mathbb{R}^{C\times WH}$. Second, to emphasize the significant feature vectors within $\mathbf{X}$ (i.e., highlight the attention regions), and analysis the spatial-wise dependencies between feature vectors, SA applies multiple nonlinear layers and reshape operation on $\mathbf{X}_t$. In particular, this step can be formulated as

$$\mathbf{M}_s = f_m\left(\sigma\left(\eta\left(\text{Re}\left(\mathbf{W}_t \mathbf{X}_t\right)\right)\right)\right) \tag{4}$$

where $\mathbf{W}_t \in \mathbb{R}^{1\times C}$ is the transformation matrix, $\text{Re}(\cdot)$ represents the operation of reshape, $\sigma(\cdot)$ is the ReLU function, $\eta(\cdot)$ is the sigmoid function, $f_m(\cdot)$ indicates the normalization function that maps the elements of $\mathbf{M}_s$ into the values between 0 and 1, and $\mathbf{M}_s \in \mathbb{R}^{W\times H}$ is the weight coefficient of $\mathbf{X}$ that reflects the importance of the different RS image spatial regions. Finally, the output of SA $\mathbf{X}_s \in \mathbb{R}^{H\times W\times C}$ can be obtained by

$$\mathbf{x}_s^{(i,j)} = \mathbf{x}^{(i,j)} \cdot \mathbf{M}_s^{(i,j)} \tag{5}$$

where $\mathbf{x}^{(i,j)}$ indicates the $1 \times 1 \times C$ feature vector located in the spatial position (i, j) of $\mathbf{X}$, and $\mathbf{M}_s^{(i,j)}$ denotes the (i, j) of $\mathbf{M}_s$.

To fuse the contributions of different attention feature maps and get complete local information, we first use the batch normalization (BN) and GAP operations to map them into the channel attention feature $\mathbf{f}_c \in \mathbb{R}^{1\times 1\times C}$ and SA feature $\mathbf{f}_s \in \mathbb{R}^{1\times 1\times C}$ Then, $\mathbf{f}_c$ and $\mathbf{f}_s$ are contacted together for the final feature representation $\mathbf{f} \in \mathbb{R}^{1\times 1\times 2C}$.

D. Attention Consistent Model

So far, we get the deep features $\mathbf{f}$ and $\mathbf{f}'$ for the image pairs $\mathbf{I}$ and $\mathrm{T}(\mathbf{I})$. They can represent the images' contents from both global and local aspects. Ideally speaking, we hope these two deep features could help us to group the image pairs into the same semantic class since they are constructed by the simple spatial rotation. Nevertheless, these two features would influence each other negatively due to the issue of visual attention inconsistency. In detail, as mentioned in Section III-A, the image $\mathrm{T}(\mathbf{I})$ is obtained by rotating $\mathbf{I}$ spatially. Thus, the attention areas of the two images may be different. Taking an RS image as an example (shown in Fig. 4), when we rotate the original image by 90°, the attention regions are changed as well for focusing on the planes to reflect the semantic of "Plane." For the original RS image [see Fig. 4(a)], the channel-wise attention and spatial-wise attention regions are concentrated in the right part. For the rotated RS image [see Fig. 4(b)], the channel-wise attention and spatial-wise attention regions are concentrated in the bottom part. Also, the objects within the attention regions are different slightly. For instance, the lounge bridges can be highlighted in the rotated scenario. In the original scenario, however, the lounge bridges cannot be extracted by the attention mechanism. To overcome the limitation discussed above, we develop an attention consistent model here. On the one hand, the proposed model could remain the consistency of the visual

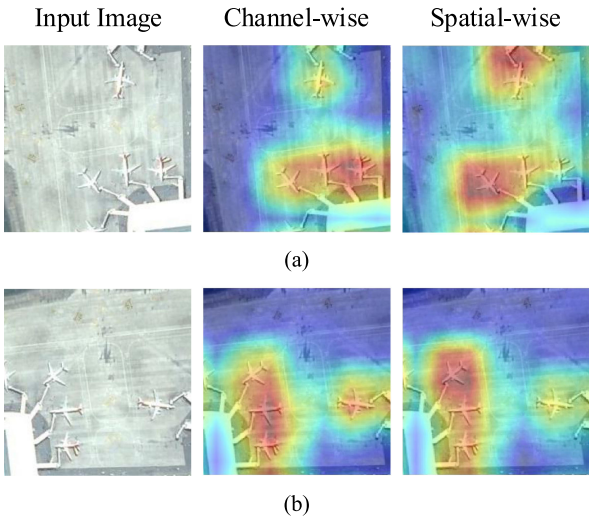

Fig. 4. Illustration of the visual attention inconsistency. The images within the first column are input images, the heat maps in the second column are the channel-wise attention maps, and the heat maps in the third column are the spatial-wise attention maps. (a) Original image and its attention maps. (b) Rotated image (90°) and its attention maps.

attention areas of a pair of RS images. On the other hand, the attention consistent model is beneficial to enlarging the differences between interclass RS images and narrowing down the distances between intraclass RS images.

For CAblock, to reduce the gap between two attention maps $\mathbf{X}_c$ and $\mathbf{X}'_c$ corresponding to image pairs $\mathbf{I}$ and $\mathrm{T}(\mathbf{I})$, we first extract the attention regions $\mathbf{M}_c \in \mathbb{R}^{W\times H}$ and $\mathbf{M}'_c \in \mathbb{R}^{W\times H}$ from $\mathbf{X}_c$ and $\mathbf{X}'_c$ by the operations of average, BN, and ReLU activation. In detail, the average (at the channel level) is used to extract the salient features from attention maps, and BN and ReLU operations are adopted to emphasize the attention regions. Second, we reversely rotate $\mathbf{M}'_c \in \mathbb{R}^{W\times H}$ according to the degrees of the input data rotation to get $\tilde{\mathrm{T}}(\mathbf{M}'_c)$. Thus, the angles of $\mathbf{M}_c$ and $\tilde{\mathrm{T}}(\mathbf{M}'_c)$ are unified. Here, $\tilde{\mathrm{T}}(\cdot)$ denotes the inverse operation of $\mathrm{T}(\cdot)$. In addition, we still suppose the sizes of feature maps are not changed by $\tilde{\mathrm{T}}(\cdot)$. Third, the mean square error (MSE) loss function is chosen to constrain the difference between $\mathbf{M}_c$ and $\tilde{\mathrm{T}}(\mathbf{M}'_c)$. For SAblock, we first use (4) to get the SA regions $\mathbf{M}_s$ and $\mathbf{M}'_s$. Then, similar to the operation for CAblock, we also reversely rotate $\mathbf{M}'_s$ to get $\tilde{\mathrm{T}}(\mathbf{M}'_s)$ for unifying the angles. Third, we use the MSE loss function to reduce the gap between two SA regions that correspond to two SA maps $\mathbf{M}_s$ and $\tilde{\mathrm{T}}(\mathbf{M}'_s)$. In the whole, the objective function of our attention consistent model is

$$J_{\text{different}} = \frac{1}{2HW}\sum_{i=1}^{H}\sum_{i'=1}^{W}\left(g_{ii'}\left(m_c, m'_c\right) + g_{ii'}\left(m_s, m'_s\right)\right)$$
$$g_{ii'}\left(m_c, m'_c\right) = \left\|\left(m_c\right)_{ii'} - \left(\tilde{\mathrm{T}}\left(m'_c\right)\right)_{ii'}\right\|_2^2$$
$$g_{ii'}\left(m_s, m'_s\right) = \left\|\left(m_s\right)_{ii'} - \left(\tilde{\mathrm{T}}\left(m'_s\right)\right)_{ii'}\right\|_2^2 \quad (6)$$

where $(m_c)_{ii'}$ and $(\tilde{\mathrm{T}}(m'_c))_{ii'}$ indicate the values in the position (i, i') of channel attention regions $\mathbf{M}_c$ and $\tilde{\mathrm{T}}(\mathbf{M}'_c)$, and $(m_s)_{ii'}$ and $(\tilde{\mathrm{T}}(m'_s))_{ii'}$ indicate the values in the position (i, i') of SA region $\mathbf{M}_s$ and $\tilde{\mathrm{T}}(\mathbf{M}'_s)$.

E. Classification Model

The main target of the classification model is to get the semantic labels for the input RS images $\mathbf{I}$ and $\mathrm{T}(\mathbf{I})$ according to their deep representation $\mathbf{f}$ and $\mathbf{f}'$. To this end, we add two FC layers and a softmax layer on the top of ACNet to transform $\mathbf{f}$ and $\mathbf{f}'$ into the predict labels $\mathbf{p}$ and $\mathbf{p}'$. Also, the cross-entropy loss function is selected to measure the predict labels. However, due to the specific architecture of our ACNet, the classification schemes of training and testing phases are different.

In the training phase, when we get the predicted labels $\mathbf{p}$ and $\mathbf{p}'$, the following objective function is developed to optimize our ACNet:

$$L = J_i + J_{i'} + \lambda J_{\text{different}} \quad (7)$$

where J_i and $J_{i'}$ represent the cross-entropy loss functions for two branches, $J_{\text{different}}$ means the objective function of the attention consistent model, and λ is a hyperparameter for controlling the contribution of $J_{\text{different}}$. In the testing phase, we directly combine the classification results $\mathbf{p}$ and $\mathbf{p}'$ together for the final classification results $\mathbf{P}$, and the formulation is

$$\mathbf{P} = \frac{\mathbf{p} + \mathbf{p}'}{2}. \quad (8)$$

IV. Experiments and Discussion

A. Testing Data Introduction

To testify the effectiveness of our ACNet, we select three RS image benchmarks. The first one is a small-scale aerial image dataset, which was published by the University of California Merced [50], and we name it UCM[1] in this article for short. There are 2100 aerial images in UCM that cover 20 U.S. regions, including Birmingham, New York, etc. These aerial images are divided into 21 scene classes, and each class contains 100 RS images. Their spatial resolution and sizes are one foot and 256×256. Some image examples and the semantic categories of the UCM dataset are displayed in Fig. 5. The second one is a medium-scale RS image dataset, which was proposed in the literature [51]. We record it AID[2] here for convenience. There are 30 scene classes (such as "Dense Residential" and "Viaduct") in AID, and the volume of images within each class varies from 220 to 420. The total number of images within AID is 10 000, and these aerial images cover different countries around the world. The spatial resolution of images changes from 0.5 to approximate 8 m, and the sizes of images are 600×600. The examples and their scene classes are displayed in Fig. 6. The last one is a large-scale RS image dataset, which was constructed in 2017 [31]. We name it NWPU[3] here for short. There are 31 500 images in NWPU with the spatial resolution from 0.2 to 30 m, which are collected by more than 100 countries and

[1][Online]. Available: http://vision.ucmerced.edu/datasets/landuse.html
[2][Online]. Available: http://captain.whu.edu.cn/project/AID/
[3][Online]. Available: http://www.escience.cn/people/gongcheng/NWPU-RESISC45.html

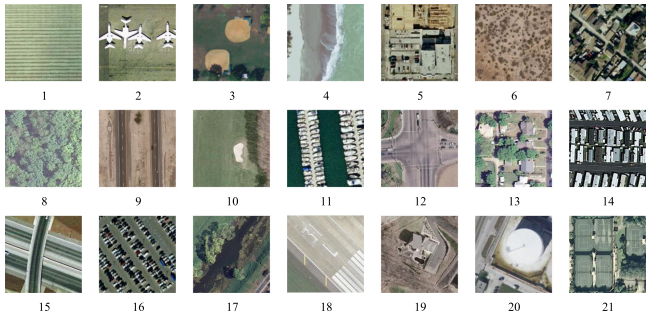

Fig. 5. Examples of different scenes of the UCM dataset. The scene numbers and names are summarized as follows. 1-Agricultural, 2-Airplane, 3-Baseball Diamond, 4-Beach, 5-Buildings, 6-Chaparral, 7-Dense Residential, 8-Forest, 9-Freeway, 10-Golf Course, 11-Harbor, 12-Intersection, 13-Medium Density Residential, 14-Mobile Home Park, 15-Overpass, 16-Parking Lot, 17-River, 18-Runway, 19-Sparse Residential, 20-Storage Tanks, and 21-Tennis Courts.

Fig. 6. Examples of different scenes of the AID dataset. The scene numbers and names are summarized as follows. 1-Airport, 2-Bare Land, 3-Baseball Field, 4-Beach, 5-Bridge, 6-Center, 7-Church, 8-Commercial, 9-Dense Residential, 10-Desert, 11-Farmland, 12-Forest, 13-Industrial, 14-Meadow, 15-Medium Residential, 16-Mountain, 17-Park, 18-Parking, 19-Playground, 20-Pond, 21-Port, 22-Railway Station, 23-Resort, 24-River, 25-School, 26-Sparse Residential, 27-Square, 28-Stadium, 29-Storage Tanks, and 30-Viaduct.

regions around the world. All of the images are equally grouped into 45 scene categories, including "River," "Water," etc. The image examples and the scene classes of the NWPU dataset are displayed in Fig. 7.

B. Experimental Settings

In this article, we use an HP-Z840-Workstation with Xeon(R) CPU E5-2630, NVIDIA GTX TITAN Xp, and 128 G RAM to complete the experiments. As mentioned in Section III-B, the intermediate feature extraction model of our network is initialized with the ImageNet's pretrained weights. The rest parts of our ACNet are initialized by a set of random parameters that follows a normal distribution with a standard deviation of 0.1. To train ACNet, we choose the Adam algorithm with the learning rate of 0.001 and the weight decay of 0.0001. Furthermore, the batch size and epochs are equal to 64 and 120. The training process is accomplished by the PyTorch platform [52]. Here, due to the structure of the intermediate feature extraction model, we resize the input RS images into 224×224. Two free parameters impact the performance of our ACNet, i.e., the rotation angle θ for building the image pairs and the hyperparameter λ for

Fig. 7. Examples of different scenes of the NWPU dataset. The scene numbers and names are summarized as follows. 1-Airplane, 2-Airport, 3-Baseball Diamond, 4-Basketball Court, 5-Beach, 6-Bridge, 7-Chaparral, 8-Church, 9-Circular Farmland, 10-Cloud, 11-Commercial Area, 12-Dense Residential, 13-Desert, 14-Forest, 15-Freeway, 16-Golf Course, 17-Ground Track Field, 18-Harbor, 19-Industrial Area, 20-Intersection, 21-Island, 22-Lake, 23-Meadow, 24-Medium Residential, 25-Mobile Home Park, 26-Mountain, 27-Overpass, 28-Palace, 29-Parking Lot, 30-Railway, 31-Railway Station, 32-Rectangular Farmland, 33-River, 34-Roundabout, 35-Runway, 36-Sea Ice, 37-Ship, 38-Snowberg, 39-Sparse Residential, 40-Stadium, 41-Storage Tank, 42-Tennis Court, 43-Terrace, 44-Thermal Power Station, and 45-Wetland.

controlling the contributions of different terms in (7). We use the fivefold cross-validation method to obtain their optimal values for different datasets. Their influence would be discussed in Section IV-E.

To validate our model's performance, we choose two widely used assessment criteria, i.e., overall accuracy (OA) [53] and the confusion matrix (CM) [54]. OA is defined as the number of correctly classified images divided by the number of the total testing images. CM is an informative table in which the column indicates the ground-truth and the row denotes the prediction. From the observation of CM, it is easy for researchers to find if the predicted labels of the test data are correct or not.

C. Performance of ACNet

To validate our ACNet extensively, we compare it with the following five RS scene classification networks.

1) The discriminative CNN (D-CNN): The D-CNN model was proposed in the paper [54], where a new objective function is developed to replace the common cross-entropy loss for considering the issues of intraclass diversity and interclass similarity. The positive results counted on three RS image datasets demonstrate the usefulness of D-CNN.
2) The FACNN: FACNN was introduced in the literature [26], in which a CNN feature-oriented encoding module and a feature fusion scheme are developed to fully explore the semantic information from the RS images. Then, the classification results can be obtained in an end-to-end manner.

TABLE I
OVERALL ACCURACIES AND STANDARD DEVIATIONS (%) OF THE PROPOSED ACNET AND THE COMPARED NETWORKS ON THE UCM DATASET

Networks	OA (8:2)
D-CNN [54]	98.81±0.30
FACNN [26]	99.05±0.24
S-CNN [29]	98.81±0.16
GLANet [40]	99.29±0.24
RAN [39]	98.81±0.30
ACNet (Ours)	**99.76±0.10**

The entry with the highest values is bold-faced.

3) The Siamese CNN (S-CNN): Based on the dual-channel framework, S-CNN was developed for the RS scene classification task [29]. There are two submodels within S-CNN, including the identification and verification blocks. Along with the specific metric learning loss function, the classification results can be obtained.
4) The GLANet: GLANet was presented in the literature [40], in which the FC layers of VGGNet are replaced by the attention blocks to explore the global and local information from RS images. Also, two auxiliary loss functions are adopted in this model to complete the scene classification.
5) The residual attention network (RAN): To highlight the useful information and eliminate the redundant information during the feature learning, the RAN model was introduced in the paper [39]. With the residual units and attention techniques, the promising scene classification results can be obtained.

Note that, all of the methods are accomplished by ourselves. In addition, for the sake of the fairness, the experimental settings of the compared methods are the same as the contents discussed in Section IV-B.

1) Results of the UCM Dataset: For the UCM dataset, we select 80% RS images randomly to construct the training data, and the rest of the images are regarded as the testing data. The optimal values of λ and θ for UCM are 0.7 and 180°, respectively. The OA and Kappa values of different methods are summarized in Table I, where we can find that the performance of all methods is good and our network has the strongest behavior. Compared with other methods, the improvements in OA values obtained by our ACNet are 0.95% (D-CNN, S-CNN, and RAN), 0.71% (FACNN), and 0.47% (GLANet). The reasons for the superior performance of our method are threefold. First, due to the dual-network architecture and the specific loss function, not only the global features but also the similarities between RS images can be learned by our ACNet. Second, with the help of the attention mechanisms, the diverse land cover information within RS images can be fully explored. Third, the developed attention consistent model could help ACNet to unify the important regions in the RS images, which is beneficial to highlight the target-level information further. Apart from the OA values, the superiority of ACNet is also confirmed by CM

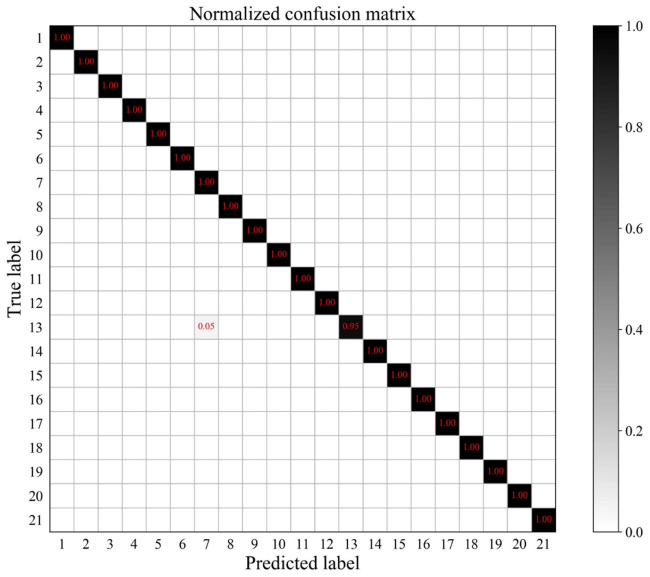

Fig. 8. CM of the UCM dataset under the training ratio of 80% using our ACNet. The semantic of each number can be found in Fig. 5.

TABLE II
OVERALL ACCURACIES AND STANDARD DEVIATIONS (%) OF THE PROPOSED ACNET AND THE COMPARED NETWORKS ON THE AID DATASET

Networks	OA (2:8)	OA (5:5)
D-CNN [54]	92.05±0.16	94.62±0.10
FACNN [26]	92.48±0.21	95.10±0.11
S-CNN [29]	92.38±0.13	95.24±0.18
GLANet [40]	91.80±0.28	94.16±0.19
RAN [39]	92.18±0.42	93.66±0.28
ACNet (Ours)	**93.33±0.29**	**95.38±0.29**

The entries with the highest values are bold-faced.

that is exhibited in Fig. 8. Here, due to the space limitation, we only show CM generated by ACNet. From the observation, it is apparent that the confusion is only appeared between "Medium Density Residential" and "Dense Residential." The encouraging results discussed above demonstrate that our model is useful to classify the scenes within the UCM archive.

2) Results of the AID Dataset: To study the performance of our network to the AID dataset deeply, we establish two training sets, respectively. In the first set, the proportion of the numbers of training and testing data is 2:8 and we set $\lambda = 0.8$ and $\theta = 180°$. In the second set, this proportion is changed into 5:5 and the λ and θ are equal to 0.8 and 90°. The OA values and their standard deviations are exhibited in Table II. Similar to the results counted on the UCM dataset, the performance of our model is the best among all methods in any case. When there are 20% RS images that can be used to train different networks, the enhancements achieved by ACNet are 1.28% (D-CNN), 0.85% (FACNN), 0.95% (S-CNN), 1.53% (GLANet), and 1.15% (RAN). When the percentage of the number of RS images in the training set equals to 50%, the improvements obtained by our model are 0.76% (D-CNN), 0.28% (FACNN), 0.14% (S-CNN), 1.22% (GLANet), and 1.72% (RAN). Different from the results of the UCM dataset, the behavior of the attention-based methods

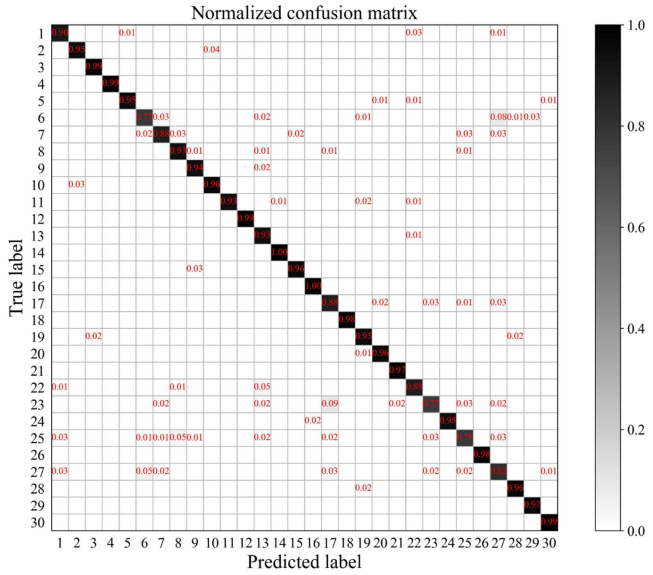

Fig. 9. CM of the AID dataset under the training ratio of 20% using our ACNet. The semantic of each number can be found in Fig. 6.

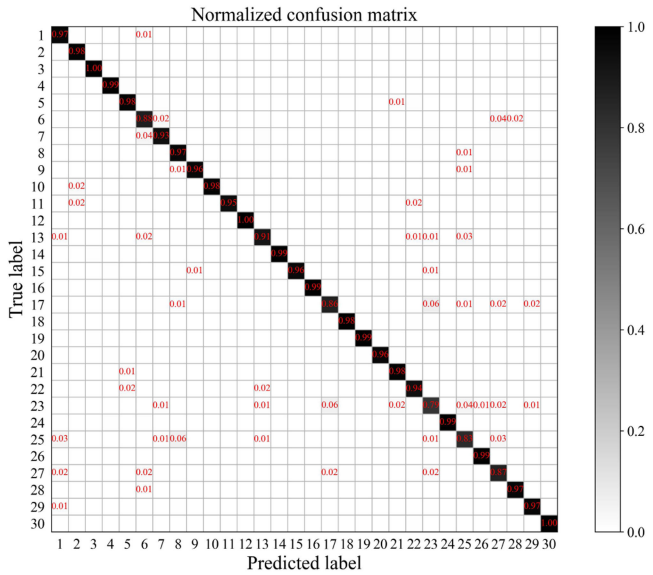

Fig. 10. CM of the AID dataset under the training ratio of 50% using our ACNet. The semantic of each number can be found in Fig. 6.

(GLANet and RAN) is weaker than that of others. On the one hand, the size of RS images within the AID dataset are 600×600, which are larger than that of the UCM dataset. The areas highlighted by the attention technique used in GLANet and RAN maybe not important to the classification. On the other hand, the semantics of the AID dataset are more diverse than that of the UCM archive. The relationships between the RS images would be impacted by the improper salient regions generated by the attention method, which harms the classification results. The CMs of ACNet counted by the AID dataset under different training sets are exhibited in Figs. 9 and 10. Through observing these matrices, we can find that our ACNet is good at distinguishing the RS images belonging to "Baseball Field," "Beach," and "Viaduct." However, for the images from "Resort" and "School," our model's performance is not as good as expected.

TABLE III
OVERALL ACCURACIES AND STANDARD DEVIATIONS (%) OF THE PROPOSED ACNET AND THE COMPARED NETWORKS ON THE NWPU DATASET

Networks	OA (1:9)	OA (2:8)
D-CNN [54]	89.09±0.50	91.68±0.22
FACNN [26]	90.87±0.66	91.38±0.21
S-CNN [29]	88.05±0.78	90.99±0.16
GLANet [40]	89.50±0.26	91.50±0.17
RAN [39]	88.79±0.53	91.40±0.30
ACNet (Ours)	**91.09±0.13**	**92.42±0.16**

The entries with the highest values are bold-faced.

3) Results of the NWPU Dataset: The NWPU dataset is the largest one among three archives. Thus, we only select 10% and 20% RS images from NWPU to train different models, respectively. Then, the rest of 90% and 80% images are used as the testing data. Here, the value of λ is set to be 0.7 for two scenarios, and θ equals to $90°/180°$ when the proportion of the training set is 10%/20%. The OA values and their standard deviations are exhibited in Table III, in which we can find that the strongest network is the proposed ACNet. Compared with other methods, the enhancements achieved by ACNet under the training ratio of 10% are 2.00% (D-CNN), 0.22% (FACNN), 3.04% (S-CNN), 1.59% (GLANet), and 2.3% (RAN). The improvements obtained by our model under the training ratio of 20% are 0.74% (D-CNN), 1.04% (FACNN), 1.43% (S-CNN), 0.92% (GLANet), and 1.02% (RAN). These encouraging results illustrate that our method is useful to the scene classification task even though the dataset is diverse and complex. Besides, ACNets' CMs under different training sets are displayed in Figs. 11 and 12. From the observation of CMs, it is easy to find that ACNet is effective for most categories. Taking Fig. 12 as an example, the accuracies of ACNet are higher than 90% for 35 out of 45 categories and are higher than 85% for 42 out of 45 categories. Especially for the "Chaparral" class, there is no incorrect prediction. These promising results prove the effectiveness of our model again.

D. Ablation Study

As mentioned in Section III, our ACNet mainly contains an intermediate feature extraction model, a parallel-attention model, and an attention consistent model. To study their influence on ACNet, we conduct the following ablation experiments. First, three networks are constructed, as follows.

1) *Net-0*: Intermediate feature extraction model.
2) *Net-1*: Intermediate feature extraction model + Parallel-attention model.
3) *Net-2*: Intermediate feature extraction model + Parallel-attention model + attention consistent model.

Here, Net-0 is the VGG16 net, Net-1 is the single version of ACNet, and Net-2 is our ACNet. Net-0 and Net-1 are single-branch networks, whereas Net-2 is a multibranch network. Then, we count their classification performance on three datasets for studying different models' behavior. The experimental settings

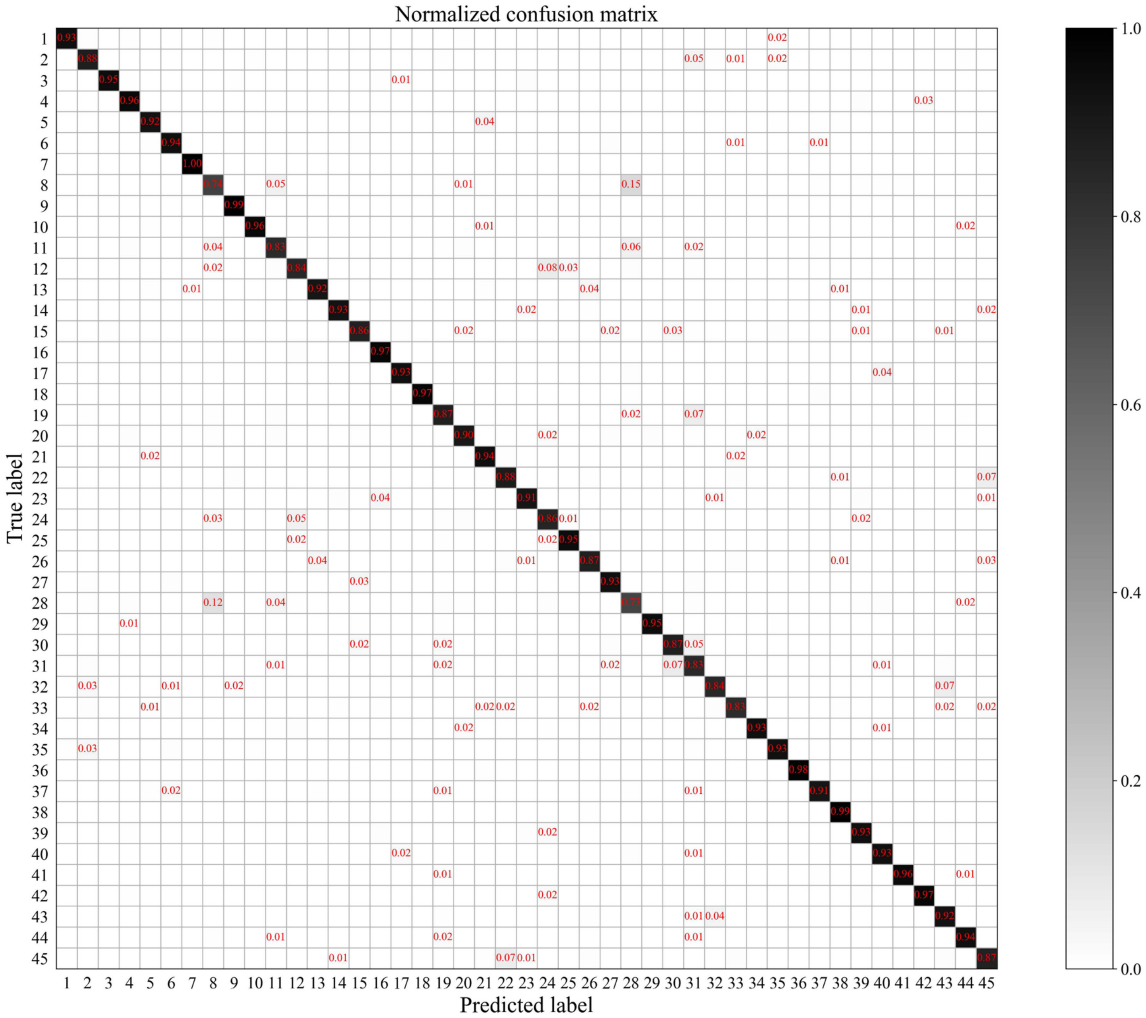

Fig. 11. CM of the NWPU dataset under the training ratio of 10% using our ACNet. The semantic of each number can be found in Fig. 7.

are the same as the contents mentioned in Section IV-B. The results of the three networks are shown in Fig. 13(a). From the observation, we can find some conclusions. First, the performance of Net-0 is the weakest among the three networks since only the global features learned by VGG16 are utilized to complete the classification. After adding the attention models, the behavior of Net-1 is stronger than that of Net-0. Moreover, there is a distinct performance gap between Net-1 and Net-0. These indicate the usefulness of the local features for scene understanding. When we expand the structure of the network into the dual-branch, Net-2's performance is enhanced for different datasets. This denotes that the data augmentation achieved by the spatial rotation and the intraimage compaction achieved by the attention consistent model play a positive role in the scene classification task.

Besides the above ablation experiments, we also want to study the contributions of two attention blocks. To this end, two single-branch networks are constructed based on Net-1 and two dual-branch networks are constructed based on Net-2. In detail, given as follows.

1) *Net-1-CA*: Intermediate feature extraction model + CA-block.
2) *Net-1-SA*: Intermediate feature extraction model + SAblock.
3) *Net-2-CA*: Intermediate feature extraction model + CA-block + attention consistent model.
4) *Net-2-SA*: Intermediate feature extraction model + SAblock + attention consistent model.

The performance of different networks is displayed in Fig. 13(b), in which the results of Net-1 and Net-2 are also exhibited for reference. It is easy to find that 1) the networks with channel-wise attention model (Net-1-CA and Net-2-CA) outperform the networks embedding spatial-wise attention model (Net-1-SA and Net-2-SA), and 2) the dual-branch networks that contain two kinds of attention models (Net-1 and Net-2) outperform the single-branch networks. The contents discussed above confirm the contributions of different attention models and the superiority of our ACNet.

Here, we want to further explain why Net-1-CA and Net-2-CA outperform Net-1-SA and Net-2-SA. The reasons behind this phenomenon can be summarized as follows. In general, channel-wise attention methods aim to emphasize the significant channels and suppress the insignificant channels for capturing the salient targets from the RS images, whereas spatial-wise

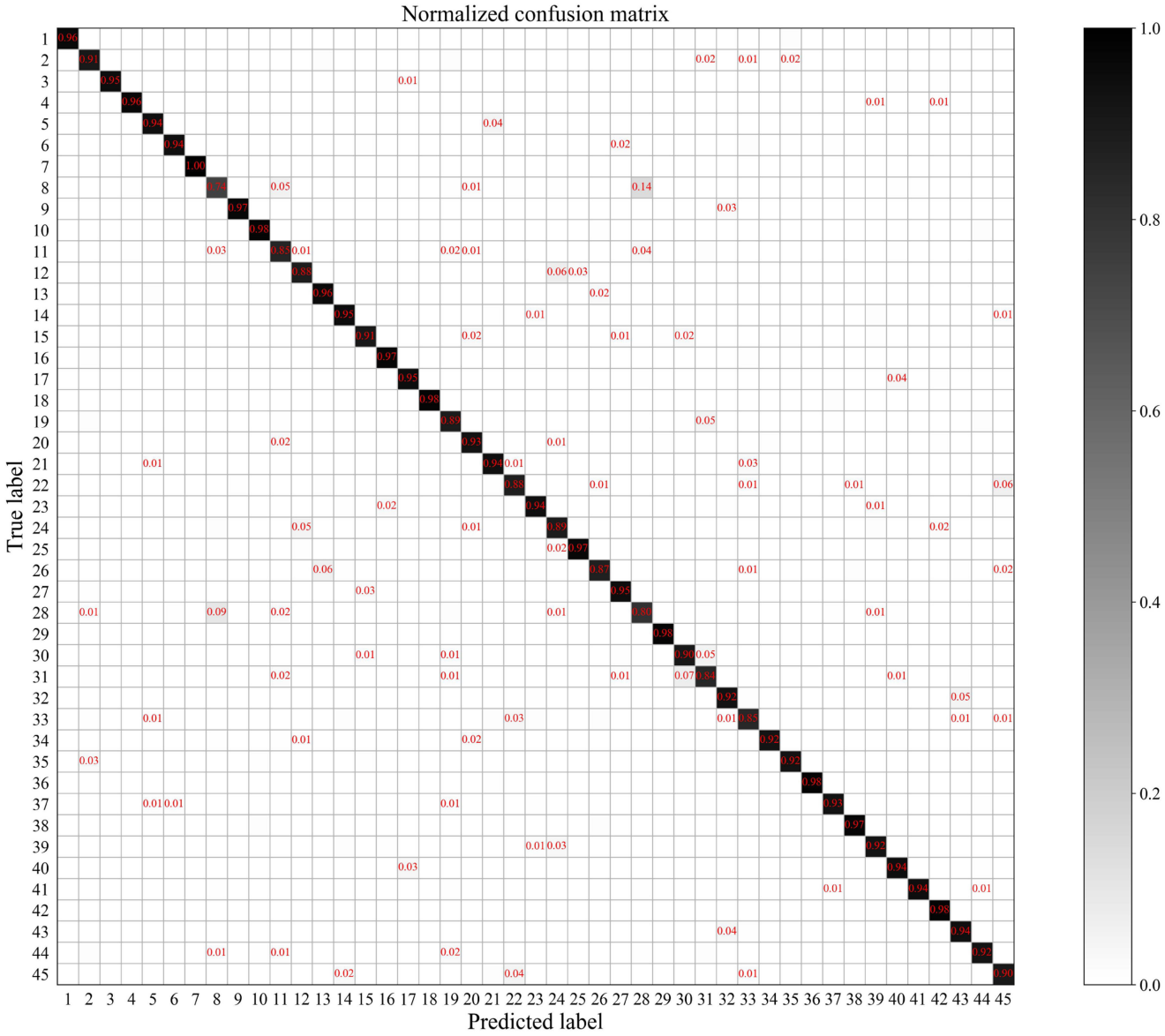

Fig. 12. CM of the NWPU dataset under the training ratio of 20% using our ACNet. The semantic of each number can be found in Fig. 7.

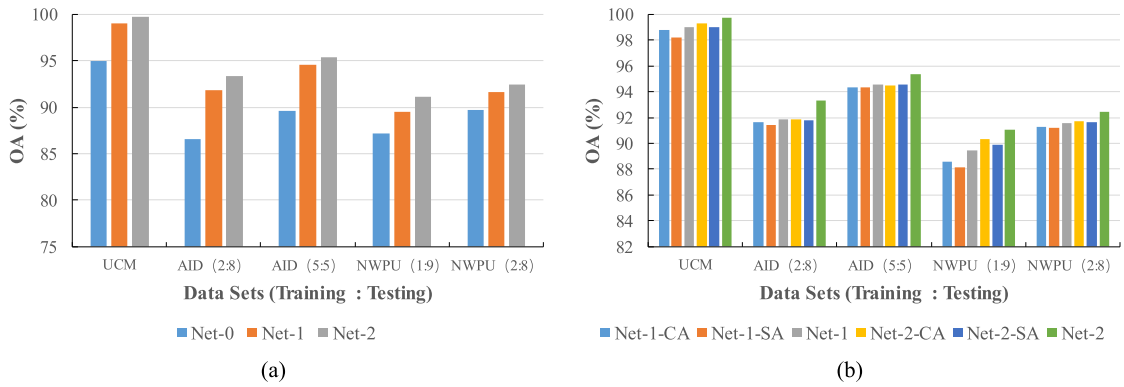

Fig. 13. Ablation experimental results counted on different RS image datasets with the different ratios of training data. (a) Performance of networks with different models. (b) Performance of networks with different attention mechanisms.

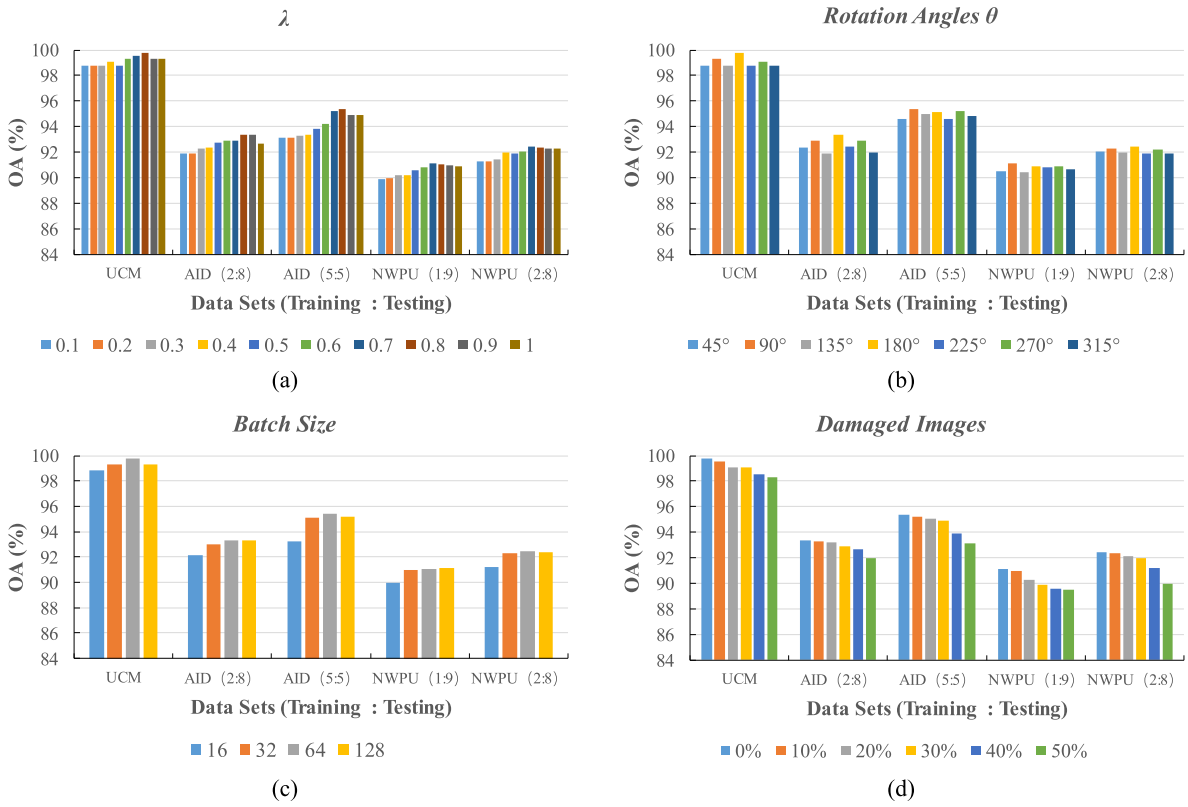

Fig. 14. Sensitivity study of our ACNet, including overall accuracies of ACNets based on different (a) λ, (b) rotation angle θ, (c) batch size, and (d) percentage of damaged RS images within training set.

attention methods focus on transforming the feature maps into a specific space with the consideration of spatial information so that the important targets within RS images can be highlighted.

To decide the importance of each channel within the feature map, the selected SE model (CAblock) [48] learns the weights of different channels by GAP, FC layers, and the self-gating mechanism (sigmoid activation). Then, the attention maps are generated through the average, BN, and ReLU activation operations, which are conducted on the updated feature maps (which are obtained by multiplying the original feature maps by the learned weights). This leads that the obtained attention maps contain much useful information for our task.

Different from SE, the adopted SA model (SAblock) [49] first transforms the feature map into a spatial space. Then, the relationships between feature map pixels are mined for emphasizing the useful local information so that the SA maps can be generated. Although the selected SA model completes the attention areas extraction under the paradigm of spatial-wise attention methods, its shortcomings should not be ignored. First, the spatial-wise dependencies between feature map pixels are not fully mined. Second, the SA maps are learned directly, which would be influenced by the training data. Due to the limitation discussed above, the target information may not be captured from the complex RS images accurately. Consequently, the performance of networks based on the SA model is weaker than that of networks based on the SE model. We have to admit that many other advanced spatial-wise attention methods can be used [55]. If we choose one of them, the behavior of networks based on it could be stronger. How to select or develop a proper SA model could be our future work.

E. Sensitivity Analysis

In this section, we study the sensitivity of our ACNet from the four aspects, including the influence of two free parameters [the rotation angle θ and the hyperparameter λ within (7)], the impact of different batch size in the training process, and the performance variation if there are some damaged RS images in the training set.

First, the value of λ is varied from 0.1 to 1, and then the OAs of ACNets obtained by three datasets are shown in Fig. 14(b). From the observation of figures, we can find the following points. First, the trend of our networks' performance is upward with λ is increased. This demonstrates the importance of the attention consistent model. Second, when $\lambda \in [0.6, 0.9]$ the behavior of ACNets is strong and stable. The peak values of model's performance appear at $\lambda = 0.7$ (UMC and NWPU) and $\lambda = 0.8$ (AID). Therefore, we suggest that the value of λ can be tuned at a range of $[0.6, 0.9]$.

Second, we change θ from 45° to 315° with the interval of 45°, and then the OAs of ACNets counted on different datasets are exhibited in Fig. 14(a). We can find that the difference of ACNets is not big, which denotes that our network is not sensitive to the rotation angles. Taking the results counted on "NWPU (2:8)" as examples, the OAs are 92.05% (45°), 92.27% (90°), 91.96% (135°), 92.42% (180°), 91.87% (225°), 92.19% (270°),

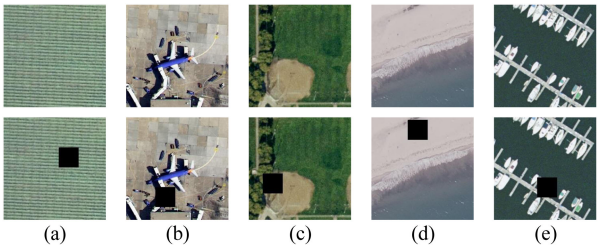

Fig. 15. Examples of RS images and their damaged version. All of the RS images are randomly selected from the UCM dataset. The images exhibited in the first row are the original RS images, and the images displayed in the second row are the damaged RS images. (a) Agricultural. (b) Airplane. (c) Baseball Diamond. (d) Beach. (e) Harbor.

and 91.92% (315°), respectively. In addition, an interesting observation is that the behavior of ACNets under 90°, 180°, and 270° is stronger than that of ACNets under 45°, 135°, and 225°. The reason behind this is that there is information loss when the rotation angles are arbitrary. Fortunately, the information loss impacts our model slightly. Note that, the reason why the arbitrary rotation angles (45°, 135°, 225°) are adopted in our experiments is that we want to study if our ACNet works or not when one of the input RS images loses some contents.

Third, the values of batch size are varied from 16 to 128 for studying its influence, and the results of ACNets are exhibited in Fig. 14(c). It is easy to find that the performance of ACNets is enhanced with the batch size increases. Taking the UCM dataset as an example, the OAs rise from 98.81% to 99.76% when batch size is varied from 16 to 64. When the value of batch size equals 128, the OAs of different ACNets are almost remained (compared with the results of batch size equals 64). Therefore, we suggest that the batch size can be tuned around 64 for some other RS image datasets.

Last, we further discuss the influence of damaged images on our ACNet. To construct the damaged images, we cut out a rectangular region with the size of 50×50 from the original RS images (which have not been resized to 224×224). In detail, for an RS image, we first select a 50×50 rectangular region randomly. Then, the contents of the rectangular region are wiped off from the RS image and the pixels within this region are set to be 0. The examples of RS images and their damaged version are displayed in Fig. 15. When we input the damaged RS images into our ACNet, their sizes would be resized to 224×224. Here, the percentage of damaged RS images within the training set is varied from 10% to 50% to observe the performance of the proposed ACNet. The classification results of three datasets are shown in Fig. 14(d). From the observation of bars, it is easy to find that the performance of ACNet is decreased when the percentage of numbers of damaged RS images increase. Fortunately, the degrees of decline for three datasets are acceptable. For example, the OA values of ACNets are decreased from 99.76% to 98.29% when the volume of damaged RS image within the training set is increased from 0% to 50%. The positive results prove that ACNet is not sensitive to the damaged images.

Apart from the cases discussed above, the convergence of ACNet is also studied in this section. As mentioned in Section IV-B, the epochs for training our model are set to be 120 for different

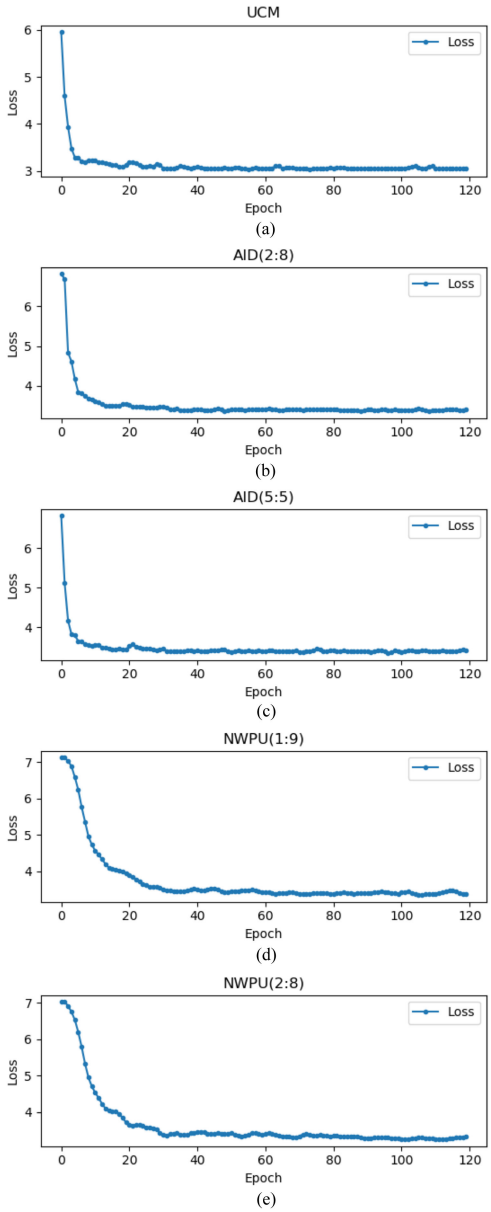

Fig. 16. Loss curves of ACNets counted on different RS image datasets with the different ratios of training data. (a) UCM dataset. (b) AID dataset with 20% training data. (c) AID dataset with 50% training data. (d) NWPU dataset with 10% training data. (e) NWPU dataset with 20% training data.

RS image datasets. To observe if this setting is suitable or not, we count the loss values for three datasets with different ratios of training data. The results are exhibited in Fig. 16. It is easy to find that all ACNets are convergent when epochs equal around $[80, 100]$. The reason why we set epoch to be 120 is that a little more training epochs could make our model more stable.

F. Time Costs

In this section, we study the time costs of our model. The time consumption of training ACNet using different datasets with the different ratios of training data is recorded in Table IV. The compared methods' training times are counted as well for the reference. Through observing the results, we can find that 1)

TABLE IV
TRAINING TIMES (MIN) OF THE PROPOSED ACNET AND THE COMPARED NETWORKS ON THREE DATASETS UNDER DIFFERENT TRAINING RATIOS

Networks	UCM (8:2)	AID (2:8)	AID (5:5)	NWPU (1:9)	NWPU (2:8)
D-CNN [54]	64.6	72.7	177	112	226
FACNN [26]	65.0	72.3	179	121	225
S-CNN [29]	56.8	68.2	168	104	215
GLANet [40]	33.7	42	106	70.7	133
RAN [39]	70.1	82.4	189	120	241
ACNet (Ours)	69.1	84.6	225	232	459

all of the models' training times are acceptable, and 2) the most time-saving method is GLANet and the most time-consuming network is our model. The time costs of GLANet are low since it is constructed by a feature extraction network with a simple attention block. The main reason is that the structure of ACNet is the most complex, in which an intermediate feature extraction model, a parallel-attention model, an attention consistent model, and a classification model are combined. The function of them is learning the global and local information from RS images, unifying the local areas, and compacting the RS images with same semantics. Compared with D-CNN, FACNN, and S-CNN, our ACNet has extra attention models. Compared with GLANet and RAN, our ACNet is a dual-branch model. Therefore, it is not surprising that training ACNet needs more times. Fortunately, training ACNet is an offline process that can be completed only once. When ACNet is trained, the time costs of predicting an RS image are low, which only needs several milliseconds. Moreover, the encouraging experimental results illustrate that the comparatively high time cost of ACNet is acceptable.

V. CONCLUSION

In this article, we propose a dual-branch network (ACNet) to accomplish the RS scene classification task. It consists of four parts, including the intermediate feature extraction model, the parallel-attention model, the attention consistent model, and the classification model. The input image pairs' (constructed by the spatial rotation) global features are learned by the intermediate feature extraction model. Then, two attention techniques are run concurrently to explore the local information from RS images deeply. To eliminate the influence of the spatial rotation in the generation of salient regions, the attention consistent model is developed based on the reversed rotation and the specific loss function. This step can also impact the samples within the same categories and separate the samples from different categories. Finally, the results are obtained by the classification model. The positive results counted on three popular benchmarks demonstrate that our model is useful to the RS scene classification task.

REFERENCES

[1] N. Zhu *et al.*, "Deep learning for smart agriculture: Concepts, tools, applications, and opportunities," *Int. J. Agricultural Biol. Eng.*, vol. 11, no. 4, pp. 32–44, 2018.

[2] J. Marçais and J.-R. de Dreuzy, "Prospective interest of deep learning for hydrological inference," *Ground Water*, vol. 55, pp. 688–692, 2017.

[3] X. Zou, M. Cheng, C. Wang, Y. Xia, and J. Li, "Tree classification in complex forest point clouds based on deep learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2360–2364, Dec. 2017.

[4] L. Chen, W. Yang, K. Xu, and T. Xu, "Evaluation of local features for scene classification using VHR satellite images," in *Proc. IEEE Joint Urban Remote Sens. Event*, 2011, pp. 385–388.

[5] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2395–2412, 2012.

[6] L. Jiao, X. Tang, B. Hou, and S. Wang, "SAR images retrieval based on semantic classification and region-based similarity measure for earth observation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 3876–3891, Aug. 2015.

[7] X. Tang and L. Jiao, "Fusion similarity-based reranking for SAR image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 242–246, Feb. 2017.

[8] X. Tang, L. Jiao, W. J. Emery, F. Liu, and D. Zhang, "Two-stage reranking for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5798–5817, Oct. 2017.

[9] X. Tang, L. Jiao, and W. J. Emery, "SAR image content retrieval based on fuzzy similarity and relevance feedback," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1824–1842, May 2017.

[10] Q. Zhu, Y. Zhong, L. Zhang, and D. Li, "Scene classification based on the fully sparse semantic topic model," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5525–5538, Oct. 2017.

[11] R. Xu, Y. Tao, Z. Lu, and Y. Zhong, "Attention-mechanism-containing neural networks for high-resolution remote sensing image classification," *Remote Sens.*, vol. 10, no. 10, p. 1602, 2018.

[12] Q. Zhu, Y. Zhong, L. Zhang, and D. Li, "Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 6180–6195, Oct. 2018.

[13] Q. Zhu, Y. Zhong, S. Wu, L. Zhang, and D. Li, "Scene classification based on the sparse homogeneous-heterogeneous topic feature model," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2689–2703, May 2018.

[14] X. Tang, X. Zhang, F. Liu, and L. Jiao, "Unsupervised deep feature learning for remote sensing image retrieval," *Remote Sens.*, vol. 10, no. 8, p. 1243, 2018.

[15] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.

[16] X. Tang, C. Liu, J. Ma, X. Zhang, F. Liu, and L. Jiao, "Large-scale remote sensing image retrieval based on semi-supervised adversarial hashing," *Remote Sens.*, vol. 11, no. 17, p. 2055, 2019.

[17] C. Liu, J. Ma, X. Tang, F. Liu, X. Zhang, and L. Jiao, "Deep hash learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3007533.

[18] S. R. Gunn *et al.*, "Support vector machines for classification and regression," *ISIS Tech. Rep.*, vol. 14, no. 1, pp. 5–16, 1998.

[19] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[20] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Statist. Learn. Comput. Vis.,* Prague, Czech Republic, 2004, vol. 1, pp. 1–2.

[21] A. Liaw *et al.*, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[22] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.

[23] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2018–2025.
[24] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2528–2535.
[25] Y. Liu, Y. Zhong, F. Fei, Q. Zhu, and Q. Qin, "Scene classification based on a deep random-scale stretched convolutional neural network," *Remote Sens.*, vol. 10, no. 3, p. 444, 2018.
[26] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7894–7906, Oct. 2019.
[27] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York, NY, USA: Elsevier, 2013.
[28] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 737–744.
[29] X. Liu, Y. Zhou, J. Zhao, R. Yao, B. Liu, and Y. Zheng, "Siamese convolutional neural networks for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1200–1204, Aug. 2019.
[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[31] 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.
[32] 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.
[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[34] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," 2013, *arXiv:1312.6229*.
[35] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.
[36] 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.*, 2009, pp. 248–255.
[37] X. Han, Y. Zhong, L. Cao, and L. Zhang, "Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification," *Remote Sens.*, vol. 9, no. 8, p. 848, 2017.
[38] X. Lu, T. Gong, and X. Zheng, "Multisource compensation network for remote sensing cross-domain scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2504–2515, Apr. 2020.
[39] R. Fan, L. Wang, R. Feng, and Y. Zhu, "Attention based residual network for high-resolution remote sensing imagery scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 1346–1349.
[40] Y. Guo, J. Ji, X. Lu, H. Huo, T. Fang, and D. Li, "Global-local attention network for aerial scene classification," *IEEE Access*, vol. 7, pp. 67200–67212, 2019.
[41] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 14, no. 1, pp. 1–20, 2017.
[42] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep Siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017.
[43] Z. Gong, P. Zhong, Y. Yu, and W. Hu, "Diversity-promoting deep structural metric learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 371–390, Jan. 2018.
[44] K. Ma, L. Wu, L. Tao, W. Li, and Z. Xie, "Matching descriptions to spatial entities using a Siamese hierarchical attention network," *IEEE Access*, vol. 6, pp. 28064–28072, 2018.
[45] Y. LeCun *et al.*, "LeNet-5, convolutional neural networks," vol. 20, p. 5, 2015. [Online]. Available: http://yann.lecun.com/exdb/lenet
[46] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[49] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5659–5667.
[50] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.
[51] 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.
[52] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.
[53] X.-Y. Tong *et al.*, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, 2020, Art. no. 111322.
[54] 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.
[55] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6688–6697.

Xu Tang (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electronic circuit and system from Xidian University, Xi'an, China, in 2007, 2010, and 2017, respectively. From 2015 to 2016, he was a Joint Ph.D. Candidate along with Prof. W. J. Emery with the University of Colorado at Boulder, Boulder, CO, USA.

He is currently an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University. His research interests include remote sensing image content-based retrieval and reranking, hyperspectral image processing, remote sensing scene classification, object detection, etc.

Qiushuo Ma received the B.Eng. degree in electronic and information engineering from Yanshan University, Qinhuangdao, China, in 2018. He is currently working toward the master's degree in computer science with the Institute of Artificial Intelligence, Xidian University, Xi'an, China.

His research interests include machine learning and remote scene classification.

Xiangrong Zhang (Senior Member, IEEE) received the B.S. and M.S. degrees in computer science from the School of Computer Science, Xidian University, Xi'an, China, in 1999 and 2003, respectively, and the Ph.D. degree in pattern recognition from the School of Electronic Engineering, Xidian University, in 2006.

She is currently a Professor with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University. From January 2015 to March 2016, she was a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. Her research interests include pattern recognition, machine learning, and remote sensing image analysis and understanding.

Fang Liu (Member, IEEE) was born in China, in 1990. She received the B.S. degree in information and computing science from Henan University, Kaifeng, China, in 2012, and the Ph.D. degree in intelligent information processing from Xidian University, Xi'an, China, in 2018.

She is currently a Lecturer with Nanjing University of Science and Technology, Nanjing, China. Her research interests include deep learning, object detection, and polarimetric SAR image classification and change detection.

Jingjing Ma (Member, IEEE) received the B.S. and Ph.D. degrees in electronics science and technology from Xidian University, Xi'an, China, in 2004 and 2012, respectively.

She is currently an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education, Xidian University. Her research interests include computational intelligence and image understanding.

Licheng Jiao (Fellow, IEEE) received the B.S. degree in high voltage from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

From 1984 to 1986, he was an Assistant Professor with the Civil Aviation Institute of China, Tianjin, China. From 1990 to 1991, he was a Postdoctoral Fellow with the Key Laboratory for Radar Signal Processing, Xidian University, Xi'an, China, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China. He has authored or coauthored more than 200 scientific articles. His research interests include signal and image processing, nonlinear circuits and systems theory, wavelet theory, natural computation, and intelligent information processing.

Dr. Jiao is a member of the IEEE Xian Section Executive Committee and an Executive Committee Member of the Chinese Association of Artificial Intelligence. He is the Chairman of the Awards and Recognition Committee.