ABSTRACT The aerial scene classification is one of the major tasks in the remote sensing community that automatically labels the corresponding semantic categories of aerial images. Recently, a lot of methods based on deep neural networks have been proposed, in which hierarchical internal features are extracted for representations. However, these presented methods often have complex structures and require large volume of memory, and a large number of labeled aerial scene images are difficult to obtain, hindering their implementation in practical applications. In this paper, we present the between-class similarity prior and adaptive knowledge mimic (BPKM) method for aerial scene classification. First, the method extracts the efficient prior relationship information of the scene images from large-scale network. Then, a compressed network is generated through learning the output and the intermediate representations of the large-scale network, and the compressed network achieves better feature description ability; in addition, an improved cross-entropy method with an adaptive threshold is applied to reduce the training time consumption. A large-scale data set (AID) and UC-Merced data set are considered for performance evaluation, and the experimental results indicate that the proposed method is about 24 × parameters saving compared to popular networks, e.g., AlexNet, and has outstanding classification performance in classes with similar features.

INDEX TERMS Aerial scene classification, deep learning, knowledge transfer, model compression, priori knowledge.

I. INTRODUCTION

Scene classification has been a significant topic in the remote sensing field. The category of each scene depends not only on the involved object but also determined by each semantic region and its hierarchical structure or spatial layout, which is different from the generic object classification task. The increasing of high-resolution aerial images, acquired by remote imaging sensors and aerial photography, provides an opportunity to recognize and analyze aerial scenes. A large number of works on this topic have appeared over the past few decades, which can be roughly cast into three categories. (1) The first category of methods is based on low-level features for describing the spectra, textures, and structural information of remote sensing images, for example, Scale Invariant Feature Transform (SIFT) [1], Local Binary Pattern (LBP) [2], and so on. (2) The second group of methods first extract the underlying visual features of the original image, and then encode the features to obtain a more reliable scene representation, e.g., Bag Of Visual Words (BoVW) [3], Spatial Pyramid Matching (SPM) [4], and Locality-constrained Linear Coding (LLC) [5], etc. (3) The third relies on deep learning (DL) to adaptively learn hierarchical features, so as to establish the mapping relationships from raw pixels to high-level semantics, such as AlexNet [6], ResNet [7], and so on. Among these methods, the DL-related studies are of great concern, which outperformed most of the other conventional methods on visual recognition.
However, the potential of CNN cannot be transferred straightforwardly to aerial image interpretation, mainly because of the following three aspects: (1) Networks with excellent classification performance are often with tens of millions of parameters, in order to evaluate the effectiveness of classification algorithms, these methods typically require large-scale annotated data. However, most people are inexperienced in labeling aerial images. A troublesome problem is that there has been a certain shortage of labeled data. (2) In contrast with object-level classification where the objects contained in different categories are completely different, scene classification aims to classify an aerial image into different semantic categories where some identical landcover types or object classes are frequently shared among different scene categories. The aerial image is impacted by changeable imaging conditions and highly overlapped classes that sometimes result in high diversities in the geometrical and spatial patterns of aerial scenes [8]. Such complexity of spatial and structural patterns in aerial scene images makes scene classification a fairly challenging problem to directly apply CNN which is suitable for object-level classification. (3) Most of the current CNN models are memory intensive and the process of training is time-consuming. Such complicated models are unwieldy to deploy in real applications. For this reason, efficient representation and recognition of the aerial images remain an open and challenging task.

To overcome these limitations, additional priori knowledge of aerial scenes should be incorporated to improve classification, which serves as a constraint for optimization objectives. In the meantime, the model Compression of scene classification is indispensable.

Accordingly, this paper aims to address the network compression problem in aerial scene classification by taking advantage of the priori knowledge of relationship between different scene classes. The major contributions of our work are twofold. Firstly, we compress a deep network for aerial scene clarification by setting priori knowledge collected from a large network into a simple network and using a training strategy which named hint-based training [9]. Secondly, In order to solve the problem that convergence on aerial image training is time-consuming, we propose an improved cross-entropy method based on adaptive threshold control which can decide whether to stop training according to the value of cross-entropy threshold.

The paper is organized as follows. Section II introduces preliminaries about the compression method, and the development of the Knowledge Transfer. The subsequent Section III presents the proposed classification approach based on priori of Between-class Similarity and knowledge mimic. Section IV provides the experimental results on AID. Finally, the conclusions are given in section VI.

II. PRELIMINARIES
A. REVIEW ON MODEL COMPRESSION
For being applied to mobile devices, reducing the storage and computational costs of large neural networks become critical.

Advanced techniques for compressing and accelerating deep neural networks attracted a lot of attention, and tremendous progress have been made in this area. These techniques are roughly categorized into two types: parameter pruning and structure compression.

The results of some earlier methods show that deleting unimportant parameters or reducing the bits number of each weight is effective for compressing the original network. The work in [10] applies pruning method and Huffman coding to reduce unimportant connections and to quantize weights. Recently, binary representation of each weight has been introduced for CNNs, for instance, BinaryConnect [11], BinaryNet [12], and XNOR-Networks [13]. In XNOR-Networks, the proposed method performs binarization on both the filter and the input of convolutional layers, so that the convolution operation is accelerated and the memory is reduced.

However, these methods ignore the effect on the accuracy loss and significantly reduced the classification effect compared to large networks.

Structural compression, that is, trying to replace original filters in the CNN with small and compact structural blocks or to reconstruct the computational process between the original CNN layers. Szegedy et al. [14] had taken advantage of 1 × 1 convolution kernels to enhance the functionality of the convolution module. And the main idea of SqueezeNet [15] is to replace the commonly used 3 × 3 convolution layer by a combination of two convolutional layers to achieve parameter reduction while maintaining classification accuracy.

Both Chollet [16] and Howard et al. [17] use depth-wise separable convolutions instead of the general convolution calculations to enhance the computing efficiency. It is observed that these structural compression methods can achieve remarkable reduction in parameters. In spite of that, the attribute of the aerial scene images, by contrast, is quite different from the one in the optical image because of the variations in the spatial arrangements and structural patterns as well as the lack of labeled data.

B. KNOWLEDGE TRANSFER
Another idea of network compression is to use the principle of knowledge transfer (KT) that transfer experience from
previous tasks to new tasks. In other words, the purpose of KT is to extract knowledge from one or more source tasks and then apply it to a target domain. The idea has been firstly proposed in Bucila et al. [18] to apply KT to compress a model. In their work, they demonstrate that the knowledge acquired by a variety of models can be transferred to a small and compact model. The author believes that any algorithm model can be simulated by a simple neural network through model compression. A newer model compression method proposed in Ba and Caruana [19] introduces a shallow net, which used the output before deep net’s softmax layer, called “logit,” as the labels for training, and try to imitate the expression function that deep networks have learned to get performance comparable to the complex model. It learns the function by entering original data into the deep network and getting the logits of these data, and then uses the original data without labels in conjunction with logits to train the shallow network.

This idea later evolved into knowledge distillation (KD) in [20], which guides the small network by the capabilities learned from the softmax function of an ensemble of complex model, keeping nice generalization ability and building a compact and fast network with little performance loss. The Simplified framework of it is shown in Fig. 1. The method uses the output class probability generated by the softmax layer, instead of its input, to guide the loss optimization. The output probability distribution is defined as the soft target, in the network that produces soft target, the information about the learned function resides in the multiple components of probabilities in the outputs, and parameter $T$ is introduced to mitigate the excessive probability difference between different categories. In other words, parameter $T$ is to soften the predicted output from the network, and the larger its value, the softer the output distribution of the softmax layer.

In this scheme, the softened output of the pre-trained cumbersome model (teacher network) was used as supervisory information to train a simple model (student network). This kind of training, which we call “distillation,” allows the student network to learn both the effective structure of the teacher network as well as the category information from the true labels.

The whole process of KD can be summarized as follows.

Concretely, given a teacher network $g$, writing $p^T_s$ for the soft target generated by softmax layer with parameter $T$. Let $s$ be a student network with output probability $q^T_s$ that are as similar meaning as in teacher network, and $CE$ refers to cross-entropy function.

In the student network, (1) cross-entropy is first calculated between the soft targets $p^T_s$ (with a high value for $T$) and $q^T_s$ (with the same high $T$). (2) Then cross-entropy is obtained for the standard output $q_s$ from softmax layer (T return to 1) and true label of data $y_{true}$.

$\mathcal{L}_1 = CE(p^T_s, q^T_s)$,
$\mathcal{L}_2 = CE(q_s, y_{true})$  

(3) the overall KD loss function that student network is trained to minimize is a weighted average of two different cross-entropies.

$\mathcal{L}_{KD} = \mathcal{L}_1 + \alpha \mathcal{L}_2$  

(2)

The role of parameter $\alpha$ is to weigh the effect of the two components on the overall loss function.

III. METHODOLOGY

As mentioned earlier, KT has been proven to be a powerful compression method with comparable performance, and the idea of distillation has given rise to a flurry of research in the NLP and the CV fields. However, in most of the recently proposed KT approaches, the two networks are often similar from the architectural point of view. And these approaches focused much on reduction in network complexity to obtain compact and fast models rather than taking advantage of depth.

In general, deep networks tend to have more robust feature learning capabilities. Considering that aerial scene images, which have a complex geometric structure, the depth of neural networks is essential in scene classification.

In our work, we add the pre-extracted between-class Similarity relationship in multistage training of compressed network which allows for training a student network that is deeper but simpler than the teacher network. In what follows, we first analysis the Priori Knowledge applying in the two-stage paradigm, the between-class similarity priori and adaptive knowledge mimic (BPKM) scheme we used for classifying aerial scene is then explicitly described. Finally, we detail the model structure of the network and an adaptive threshold cross-entropy method used to accelerate training convergence.
A. PRIORI OF BETWEEN-CLASS SIMILARITY

Prior knowledge refers to all information about the problem available in addition to the training data, it has been regrouped under two main categories: class-invariance and knowledge on the data. The latter form of priori knowledge concerns the data more specific and are thus of particular interest for real-world applications.

In this paper, considering that in the standard formulation of supervised learning the output is represented as a vector of features. Moreover, compared to normal network output, the output of the softmax layer is generally called “soft target.” For example, Fig. 2 shows the difference between the normal output and the soft output of an aerial scene image. We collect discriminatory information hidden in the network’s soft output as the priori knowledge of dataset, showing inter-class similarity of different classes, and transfer it to promoting parameter updates of networks by participating in the loss function.

As shown in Fig. 2(a), the traditional network output only sets the index of the playground as 1 and the rest as 0, which means all samples with various representation form in each category only map to the same point in the predicted output. In this way, intra-class variance and Inter-class distance, which are very helpful for training, will be lost. In contrast, supposing an image is similar to both the “Commercial” and the “Industrial,” it can theoretically provide supervision for both classes in the training stage. Based on this, the soft target in Fig. 2(b) gets the probability distribution of this image being predicted as each category: although the image is finally predicted as “Commercial”, it also has a probability of 0.21 of being an “Industrial.” This indicated the inter-class similarity between categories of “Commercial” and “Industrial.” Therefore, we generalized the probability between different classes hidden in the soft target as the Priori of Between-class Similarity, which has a large entropy and provides information about interrelation among various classes.

In general, the aerial scene images have characteristics of a high degree of inter-class similarity, which makes classification difficult. As to end-to-end learning that only concerned about the data and the output, the model gets insufficient information to learn. The importance of priori knowledge is suggested by its role in search and optimization.

The Priori of Between-class Similarity obtained by large-scale neural networks contains a rich similarity structure over the data and a small-scale neural network with a priori can converge with only a small amount of new scene data. Therefore, we exploit and integrate the priori information into the network as a strengthened constraint to guide the learning of networks. In our work, the Priori of Between-class Similarity is deduced purely from a “plump” model, and then adopted to participate in loss calculation of a “slim” network. As the “slim” network learns from both the ground-truth and the “plump” model, it will learn to balance when it should learn from data or prior knowledge.

Taking into account the characteristics of the aerial scene image, the way we use to obtain a more robust and efficient aerial scene classification is to combine a priori knowledge with the image data itself. And it is easier to handle details such as parameter adjustment and network structure optimization.

B. MIMIKING KNOWLEDGE IN LARGE NETWORK AND USING MULTISTAGE TRAINING

In order to apply the excellent feature expression ability of deep networks to achieve better results, we proposed a knowledge mimic method of the aerial scene by adding intermediate-layer supervision to deep architectures. The framework of our scheme based deep model is illustrated in Fig. 3. The original aerial images are first fed into a standard wide CNN, we call Plump Network, to generate a model with an effective expression of the aerial scene and to produce a softer probability distribution, i.e., soft targets, over classes. In the rest of the scheme, as for training on the small-scale network, defined as Slim Network, the process is divided into two phases. The first step is to extract the intermediate representations of aerial scene learned by the plump model hidden layer which named enlighten layer, and choose a middle layer in Slim Network, the absorbed layer, to match the feature maps from the enlighten layer. The training data are described as \( \{(x^{(1)}, y^{(1)}), ..., (x^{(N)}, y^{(N)})\} \), parameters from the first layer to the absorbed layer are trained to optimize following objective function with L2 loss:

\[
\min L_{MK}(W_s, W_r) = \frac{1}{N} \sum_{k=1}^{N} ||\sigma_g(x^{(i)}; W_p) - \gamma(\sigma_h(x^{(i)}; W_s); W_r)||^2
\]

(3)

where \( \sigma_h \) and \( \sigma_g \) are the forward pass functions. \( L(\cdot) \) represents the L2 loss function. Let \( W_p \) be the Plump Network’s parameters up to the enlighten layer \( g \), whereas \( W_s \) be the slim network’s parameters up to the absorbed layer \( h \). \( \gamma(\cdot) \) is the convolutional function on top of the absorbed layer with parameters \( W_r \), matching the size of output prediction between absorbed/enlighten layers. As a consequence of this stage, the student network is induced to reach a proper initialization state, updating only \( W_s \) and \( W_r \).

Then in the second stage, the Slim Network manages to fit not only the true labels but also the soft output of the Plump Network used to guide the training of the entire Slim Network. Beyond that, the proportion of the soft target in the total loss is gradually reduced, so that the ground truth helps identify difficult samples. We train the parameters of the whole student network to minimize Eq. (2). And to summarize, The goal was that even though the Slim Network is a smaller and less powerful network, it can “emulate” the Plump’s output i.e. learn from Plump Network.
C. IMPROVED CROSS-ENTROPY METHOD WITH ADAPTIVE THRESHOLD

In the knowledge mimic mentioned above, cross-entropy plays a greater role, since it trains the parameters of the entire network and modifies the optimization direction to find potentially better local minima. However, the time for convergence is mainly spent on the calculation of the cross-entropy, and when compared to traditional optical image, the aerial scene image shows smaller Interclass dissimilarity resulting in trouble with interpreting it. Consequently, the cross-entropy function calculated for aerial scenes images make the training process converge slowly.

In order to make our BPKM faster convergence, we set an adaptive threshold of cross-entropy in the process of training. The existing deep learning frameworks, e.g., Caffe [21], TensorFlow [22], etc., usually set a fixed number of training iterations at the beginning, this number is usually determined by multiple adjustments. In some cases, the network has already converged before meeting the given number of training times, with a waste of time and little help to network accuracy. In our implement, the convergent time of the network was improved by the dynamic cross-entropy method: Through observing the training process of aerial images, it is proved that there exists a threshold, which indicates the convergence of network, and the network will be aborted if the cross-entropy is less than the threshold. Specifically, the adaptive threshold is an experimental trick to get the most appropriate number of iterations, we added the code of monitor function to the experiment script, and the number of epochs was obtained by monitoring the error on the validation set when running this script: The iteration epochs are firstly set to be a large number, and an optional parameter, the monitoring channel, can be driven by any function, e.g., misclassification or cross-entropy loss. Taking cross-entropy loss for example, the method checks for the value of cross-entropy loss and stores the best parameters to disk every time the value decreases. After this stage, the relatively optimal loss and the corresponding iterative number that is most appropriate for training are obtained.

IV. EXPERIMENT AND ANALYSIS

In this section, we first introduce the optical remote sensing image set used for the experiments and then describe the parameter settings in the actual experiments. The results obtained for the scene recognition of an aerial image is then discussed.

A. EXPERIMENTAL DATA

This section first described two datasets used for aerial scene classification in the following experiments: UC Merced Land Use Dataset and the AID, and as illustrated in [23] that data augmentation can effectively improve performance, all data have been processed with global contrast normalization and whitening.

The UC – Merced dataset [3] is a 21-category remote sensing satellite image classification database with 100 images per category, which including: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. The images were manually extracted from large images from the USGS National Map Urban Area Imagery collection and were cropped into 256 × 256 pixels, the pixel resolution of the imagery is 1 foot. In the experiment,
75 images in each category are used for training, and the remaining 25 are used for testing.

To better assess the effectiveness of a scene classification method, Xia et al. [24] construct a new large-scale dataset, i.e., AID, for aerial scene classification. This dataset is selected from Google Earth imagery. There are a total of 10000 images labeled into 30 classes. It is worth noticing that the dataset contains a variety of spatial land-use patterns and the sample images of the same class are collected from different regions all around the world. Furthermore, aerial scenes might appear at varying scales, orientations, or imaging conditions.

The properties of AID could be summarized as: (1) It is the largest annotated aerial image datasets that cover a much wider range of scene types than the existed datasets. (2) Due to the high spatial resolutions and high complexity of the earth surface, there will be some disparities between the structural patterns of each image, even if they belong to the same class of scenes. (3) The dissimilarities between different scene classes, which are often small in real cases of aerial image classification, are essential indicators for datasets. In this regard, AID contains different scene classes with similar characteristics detail. For example, as shown in Fig. 5, the “stadium” and “playground” are almost indistinguishable, the main difference among them lies in whether there are stands around. In AID, many of these scene classes with small interclass dissimilarities have taken into account. AID is a finer data resource to evaluate the algorithms in aerial image analysis.

In Fig. 4 some sample images are shown, which are more representative types of scenes selected from the dataset, shows some of the example images, which are derived from the more representative types of scenes in the dataset. The number of each class varies from 220 to 420. They are cropped into regions of $600 \times 600$ pixels with various pixel resolutions. In our experiment, 10 kinds of scenes with more prominent features in AID were selected to consist a dataset, AID – 10: airport, bare land, beach, bridge, commercial, dense residential, desert, farmland, industrial, and mountain. We divided the dataset into 2.5k training examples and 0.5K validation examples.

**B. EXPERIMENT SETTINGS**

As the proposed aerial scene classification model involves two networks, we trained the maxout convolutional network reported in Goodfellow et al. [25] as the Plump Network. Moreover, the Slim Network is composed of 17 maxout convolutional layers of kernel size $3 \times 3$. And three main approaches of pooling are compared to apply in Slim Network that include max-pooling, average-pooling and stochastic-pooling in [26] [27], since the aerial scene images contain multiple elements full of semantic information that the task to distinguish various scene categories may complete better with max-pooling which preserves the most salient features while discarding irrelevant detail, and therefore a non-overlapping $2 \times 2$ max-pooling is applied after some of the convolutional hidden layers. At the end of the network, the 17th convolutional layer is followed by a maxout fully-connected layer and a top softmax layer. The construction of networks and the size of the convolutional layer & pooling layer are shown in Table. 1. Here, taking the parameter of “conv1” for example, “$3 \times 3, 1$” denotes $3 \times 3$ kernel size with 1 stride.
C. He et al.: Relationship Prior and Adaptive Knowledge Mimic Based Compressed Deep Network for Aerial Scene Classification

On UC — Merced and AID — 10, the student network was trained by stage as described in Section III. We trained stage 1 for Slim Network by minimizing Eq. 3, choosing the 11th layer in compact net as absorbed layer and 2nd in wide net as the enlightened layer, and stopped the training after reach the thresholds that stated in III-C. After that, we trained stage 2 by minimizing Eq. 2, all parameters in Slim Network were initialized randomly.

To fully evaluate the performance of the BPKM method, the famous compression method mentioned in Section II-A, SqueezeNet, is also tested on AID — 10 for comparison, its network structure is reported in [15].

C. RESULTS

We firstly compare the compression capabilities of our method to that of two representative networks, AlexNet and VGG-VD-16 [28], with performing experiments on two datasets. The efficiency is measured by computing the number of parameters. The results are reported in Table 2. When compared to state-of-the-art methods, AlexNet, our approach can achieve an amazing 24× parameters compression. In supervised learning, a confusion matrix is a visualization tool used to compare network outputs with actual categories of the datasets to assess the accuracy of classification results, and that of the two dataset are displayed in Fig. 6 and Fig. 7, each element in the confusion matrix is divided by the number of images for each class to get probability value. The matrixes show that, compared with traditional CNNs, the network using our method can do better on highly similar categories, and although the results on UC—Merced are limited by the amount of data, it also follow this conclusion. For testing with the adaptive cross-entropy threshold, our slim model outperforms the plump model, while twice as deep as the Plump Network and with roughly 4 times fewer parameters, confirming that depth is crucial to express finer aerial scene representation in the certain degree. Observing the loss of precision compared to AlexNet, we have trained SqueezeNet and original Slim Network without using the BPKM method, comparing their performance to verify the practical significance of the BPKM method for aerial scene images, Table 2 also reports the result, and although our method has caused a little loss of accuracy compared to AlexNet, our effect is slightly superior to other compression methods. The available model is so small that can be easily fitted into portable devices for aerial scene field.

V. DISCUSSION

One of the problems with deep neural networks is that they are not interpretable, the other problem is that they often need a
FIGURE 7. Confusion matrices for UC Merced dataset. It suggested that this method can be employed to help the network better distinguish similar categories, such as runway and freeway.

lot of training data. Therefore, our method hopes to introduce some knowledge in the deep neural network, which can help training on the one hand and improve the interpretability of the model on the other hand. The experimental results indicate that improvement is achieved by the BPKM method, the focus of which is Between-Class Similarity presented by the soft target and the adaptive threshold Cross-Entropy. The Between-Class Similarity is valuable information that defines a rich similarity structure over the data and includes much image feature information than a one-hot label (e.g., it says which stadium looks like a playground and which looks like a baseball field). The core idea of the soft target is to “break up” the supervised information that was originally compressed to a point, so that the network output contains more supervised information distributed across different components of it, and another main claims about using soft target instead of normal output is that a lot of helpful information can be carried in soft target that could prevent networks from overfitting.

The method also inspired by the discussion in [20] by Hinton where the distilled model uses the mnist dataset without the number 3 for training, but since the soft target of other numbers contains part of the information of the number 3 (such as the image of the number 2 which looks like the 3), some number 3 samples can be correctly classified during the test. Therefore, the method exploits and integrate the priori information into the network as a strengthened constraint to guide the learning of networks, reduce the search space of the neural network, and gain the generalization capability on aerial scene images. It also has the high scalability of the network structure that can easily adjust the network structure according to the application scenario.

The deep learning method will still be the first choice for image scene classification for some time. This is determined by the characteristics of the scene classification task itself and the advantages of the deep learning network. In the deep learning method, we will study the influence of multi-feature fusion on the accuracy of scene classification, such as combining the local detail features of the scene with the overall structural features. On the other hand, we will try a variety of ways to combine more network middle layer output to guide the learning of small networks. These will serve as our important research directions for the future. And image scene classification tasks will also be combined with engineering applications such as autonomous driving and robotics, and are widely used and developed.

VI. CONCLUSION

In order to make better use of aerial scene datasets and apply to the mobile terminal or embedded device, in this paper, we explore a between-class similarity priori and adaptive knowledge mimic (BPKM) method based on knowledge transfer to implement network compression on the scene classification. We use probability distributions to represent the likelihood that an image will be predicted for each category, and adopt knowledge mimic method that unites the intermediate representations and the output of the large network to guide the learning of the entire compressed network. We train a neural network, i.e., Slim Network, with a stage-wise strategy, and proposed an improved cross-entropy method which accelerates the convergence of the network. In contrast to previous works on aerial scene classification where the focus was on achieving higher classification accuracy, our method is capable of making great classification of similar categories in aerial scene images, as well as reducing the size of the network by about 24 ×. The model compression for aerial scene classification provides more possibilities and opportunities of loading very deep neural networks into practical mobile applications for the aerial scene. And compared to AlexNet, which cannot be used directly in many mobile scenarios or specific applications, our approach is more targeted to solve practical problems.
REFERENCES

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

[2] T. Ojala, M. Pietikäinen, and T. Maenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, Jul. 2002.

[3] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in Proc. 18th SGISPATIAL Int. Conf. Adv. Geograph. Inf. Syst., 2010, pp. 270–279.

[4] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2006, pp. 2169–2178.

[5] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained linear coding for image classification,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2010, pp. 3360–3367.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Adv. Neural Info. Process. Syst., 2012, pp. 1097–1105.

[7] 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.

[8] A. M. Cheriyadat, “Unsupervised feature learning for aerial scene classification,” IEEE Trans. Geosci. Remote Sens., vol. 52, no. 1, pp. 439–451, Jan. 2014.

[9] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “Fitnets: Hints for thin deep nets,” 2014, arXiv:1412.6550. [Online]. Available: https://arxiv.org/abs/1412.6550

[10] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding,” Fiber, vol. 56, no. 4, pp. 3–7, 2016.

[11] M. Courbariaux, Y. Bengio, and J.-P. David, “Binaryconnect: Training deep neural networks with binary weights during propagations,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 3123–3131.

[12] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binary neural networks: Training deep neural networks with weights and activations constrained to +1 or −1,” 2016, arXiv:1602.02830. [Online]. Available: https://arxiv.org/abs/1602.02830

[13] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, “XNOR-Net: Imagenet classification using binary convolutional neural networks,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 525–542.

[14] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in Proc. AAAI, vol. 4, 2017, p. 12.

[15] B. Wu, F. Iandola, P. H. Jin, and K. Keutzer, “SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 446–454.

[16] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” 2016, arXiv:1610.02357. [Online]. Available: https://arxiv.org/abs/1610.02357

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” 2017, arXiv:1704.04861. [Online]. Available: https://arxiv.org/abs/1704.04861

[18] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, “Model compression,” in Proc. 12th SIGKDD Int. Conf. Knowl. Discovery Data Mining, Philadelphia, PA, USA, Aug. 2006, pp. 535–541.

[19] J. Ba and R. Caruana, “Do deep nets really need to be deep?” in Proc. Adv. Neural Inf. Process. Syst., 2013, pp. 2654–2662.

[20] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” Comput. Sci., vol. 14, no. 7, pp. 38–39, 2015.

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in Proc. 22nd Int. Conf. Multimedia, 2014, pp. 675–678.

[22] M. Abadi et al., “TensorFlow: A system for large-scale machine learning,” in Proc. OSDI, vol. 16, 2016, pp. 265–283.

[23] S.-H. Wang, Y.-D. Lv, Y. Sui, S. Liu, S.-J. Wang, and Y.-D. Zhang, “Alcoholism detection by data augmentation and convolutional neural network with stochastic pooling,” J. Med. Syst., vol. 42, no. 1, p. 2, 2018.

[24] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, “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.

[25] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, “Maxout networks,” in Proc. 30th Int. Conf. Int. Conf. Mach. Learn., 2013, pp. 1319–1327.

[26] M. D. Zeiler and R. Fergus, “Stochastic pooling for regularization of deep convolutional neural networks,” 2013, arXiv:1301.3557. [Online]. Available: https://arxiv.org/abs/1301.3557

[27] Y.-D. Zhang, Z. Dong, X. Chen, W. Jia, S. Du, K. Muhammad, and S.-H. Wang, “Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation,” Multimedia Tools Appl., vol. 78, no. 3, pp. 3613–3632, 2019.

[28] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: https://arxiv.org/abs/1409.1556

CHU HE (M’09–SM’19) received the B.S., M.S., and Ph.D. degrees from Wuhan University, Wuhan, China, in 1996, 1999, and 2007, respectively. Since 1999, he has been with the School of Electronic Information, Wuhan University, where he has been a Full Professor, since 2014, and is currently with the State Key Laboratory for Information Engineering in Surveying, Mapping, and Remote Sensing. From 2007 to 2008, he was a Visiting Scholar with the Machine Vision Group, University of Oulu, Oulu, Finland. He has authored more than 70 papers in related journals and proceedings. He was selected in the talent program of Wuhan Yellow Crane Talent Program (Science and Technology), in 2015.

BOKUN HE received the M.Sc. degree from the School of Electronic Information, Wuhan University, China, in 2019. Her research interests include image processing and computer vision.

XIAOHUO YIN received the M.Sc. degree from the School of Electronic Information, Wuhan University, China, in 2018. His research interests include image processing and machine vision.

Wuhan University, China
WENWEI WANG received the Ph.D. degree from Duisburg University, Germany, in 2002. He is currently an Associate Professor of signal and information processing, communication, and information system with Wuhan University. He is also a member of the IAPR-TC11, the China Institute of Communications, and the China Society of Image and Graphics. He has published one book and more than 20 papers. His main research interests include target detection and recognition, and computer vision and image.

MINGSHENG LIAO received the M.Eng. degree in electronic engineering and the Ph.D. degree in photogrammetry and remote sensing from the Wuhan Technical University of Surveying and Mapping (WTUSM), Wuhan, China, in 1985 and 2000, respectively. From 1985 to 1992, he was with the Department of Electronic Engineering, WTUSM, which is one of the universities that formed Wuhan University, Wuhan, in 2000. From 1992 to 2000, he was with the State Key Laboratory for Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), WTUSM, where he obtained a Professorship, in 1997. From 1995 to 1996, he was a Visiting Ph.D. Candidate with the Department of Surveying and Photogrammetry, Technical University of Denmark, Lyngby, Denmark. From 2000 to 2002, he was a Research Associate with the Joint Laboratory for Geoinformation Science, The Chinese University of Hong Kong, Hong Kong. Since 2003, he has been with LIESMARS, Wuhan University, where he is leading the SAR Data Analysis and Processing Group. He has published over 60 research articles in scientific journals and books. His research interests include interferometric and polarimetric data processing, differential SAR interferometry, and multitemporal SAR image analysis.

***