Multi-level Feature Fusion Networks with Adaptive Channel Dimensionality Reduction for Remote Sensing Scene Classification

Xin Wang, Lin Duan, Aiye Shi, and Huiyu Zhou

***Abstract*—Scene classification in very high resolution (VHR) remote sensing (RS) images is a challenging task due to complex and diverse content of the images. Recently, convolution neural networks (CNNs) have been utilized to tackle this task. However, CNNs cannot fully meet the needs of scene classification due to clutters and small objects in VHR images. To handle these challenges, this letter presents a novel multi-level feature fusion network with adaptive channel dimensionality reduction for RS scene classification. Specifically, an adaptive method is designed for channel dimensionality reduction of high dimensional features. Then, a multi-level feature fusion module is introduced to fuse the features in an efficient way. Experiments on three widely used data sets show that our model outperforms several state-of-the-art methods in terms of both accuracy and stability.**

***Index Terms*—Remote sensing scene classification, convolutional neural networks (CNNs), adaptive channel dimensionality reduction, multi-level feature fusion.**

I. Introduction

REMOTE sensing scene classification, which focuses on assigning specific semantic labels to RS scene images, has attracted wide attention in the field of remote sensing. However, the lack of rich well-labeled training data as well as the high intraclass variations and low interclass dissimilarity make the scene classification a challenging task [1].

In general, RS scene classification methods are divided into two categories according to the features they used: handcrafted/deep learning (DL) features. Compared with the first category, DL methods, especially CNNs have exhibited a powerful ability for RS image feature extraction and achieved remarkable performance for scene classification [2]-[4].

Recently, with the rapid development of RS instruments, VHR RS images have become available, which contain very detailed and complex object information. A VHR scene image may include diverse objects, each of which plays an important role in determining the semantic label for the scene. As shown in Fig. 1, the 'ship' objects in the 'Bridge' scene hardly appear in the 'Commercial' scene. If a deep neural network identifies 'ship' objects from a scene image, it has a small probability to assign the scene as 'Commercial', while it has a large probability to assign it as 'Bridge'. Unfortunately, in CNNs, due to the downsampling operations, the resolutions of feature maps gradually decrease, leading to the ignorance of some small objects which are very important for the classification of scenes. In general, small-scale information is often preserved in shallow high-resolution feature maps, while large-scale information is more concentrated in the higher-level features because of the larger receptive field in the deeper convolutional layers. Shallow low-level features usually contain spatial contextual and location information, while deeper high-level features always include more abstract semantic information.

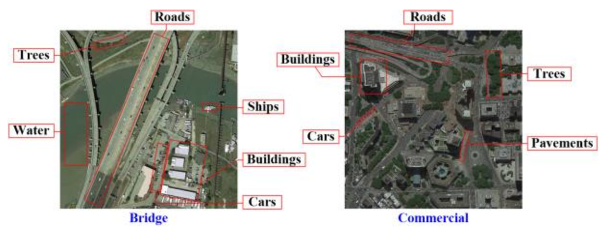

Fig. 1. VHR RS scenes contain a diversity of objects, each of which plays an important role in determining the semantic labels for the scenes.

Therefore, to improve the classification performance for VHR scenes, fusing the CNN features at different levels has attracted much attention. For example, Li *et al.* [5] proposed an aggregated deep Fisher feature to fully use deep convolutional features for VHR RS scene classification. Ma *et al.* [6] proposed a CNN-based multilayer feature fusion method to explore potential information from additional layers. Although these methods increase the scene classification performance through multiple feature combination, they struggle to handle different scenes based on the shallow CNN models. Ji *et al.* [7] proposed to integrate multiple VGG-Nets to deepen the original network and gain an advantage over a shallower VGG-Net, which also indirectly proves that the shallow network is not as good as the deeper network for RS scene classification.

In fact, for deeper CNN networks, taking ResNet [8] as an example, they can capture more abundant spatial and higher-level semantic information. Combining these features together will improve the performance greatly. Nevertheless, directly fusing multi-level features extracted from such deeper

X. Wang is supported in part by the Fundamental Research Funds for the Central Universities under Grant B210202077, Six Talents Peak Project of Jiangsu Province under Grant XYDXX-007, Jiangsu Province Government Scholarship for Studying Abroad. H. Zhou is supported in part by Royal Society-Newton Advanced Fellowship under Grant NA160342.
(Corresponding author: Xin Wang.)

X. Wang, L. Duan, and A. Shi are with the College of Computer and Information, Hohai University, Nanjing 211100, China (e-mail: wang_xin@hhu.edu.cn).

H. Zhou is a full Professor at School of Informatics, University of Leicester, Leicester LE1 7RH, United Kingdom (e-mail: hz143@leicester.ac.uk).

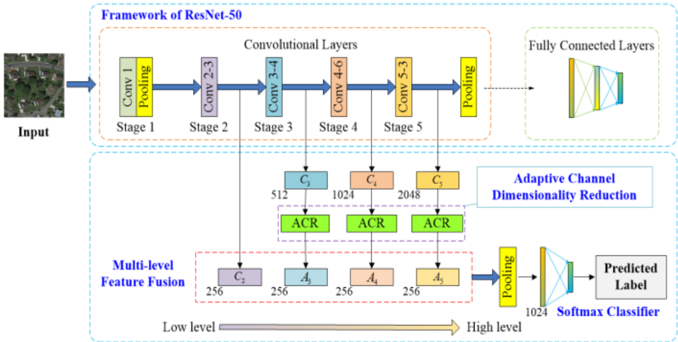

Fig. 2. Framework of the proposed method. It is composed of four modules: 1) multi-level features extraction by ResNet-50; 2) ACR; 3) MLFF; and 4) scene classification by a softmax classifier.

CNNs will increase the computational cost. Meanwhile, since the feature level span of deeper CNNs is often large, the direct integration may also lead to semantic gaps [9], decreasing the classification ability. Therefore, to effectively fuse mid-level and high-level features with the low-level ones, it is necessary to reduce the channel dimensionalities of higher-level features, so that the final fused features have smaller dimensions and more coherent semantic information, which is more conducive to classification. Currently, for channel dimensionality reduction in CNNs, the widely used approaches mainly rely on 1×1 convolution. However, a large amount of information may be lost by doing so. Furthermore, 1×1 convolution cannot adaptively mix channel information [10]. These drawbacks will affect the fused features, eventually affecting the performance of classification.

To overcome the above problems, in this letter, we propose novel multi-level feature fusion networks with adaptive channel dimensionality reduction for RS scene classification. The main contributions can be summarized as follows.

1) A multi-level feature fusion network with adaptive channel dimensionality reduction (ACR-MLFF), which makes full use of multi-level features of deep convolutional neural networks, is proposed for remote sensing scene classification.
2) An adaptive channel dimensionality reduction (ACR) module is proposed to solve the information loss problem caused by 1×1 convolution channel dimensionality reduction. It contains a trunk branch for dimensionality reduction and two side branches for enhancing the channel attention and supplementing the semantic information, respectively. The adaptivity is achieved by reweighting the features through the channel attention calculation of the top branch.
3) A multi-level feature fusion (MLFF) module is designed to integrate multi-level and adaptive channel reduction features together for classifying the complex scenes.

II. Proposed Method

Fig. 2 illustrates the overall architecture of the proposed method for RS scene classification. As can be seen, it mainly consists of four modules: 1) multi-level features extraction using a pretrained ResNet-50 model; 2) ACR; 3) MLFF; and 4) scene classification by a softmax classifier. The details of this framework are described below.

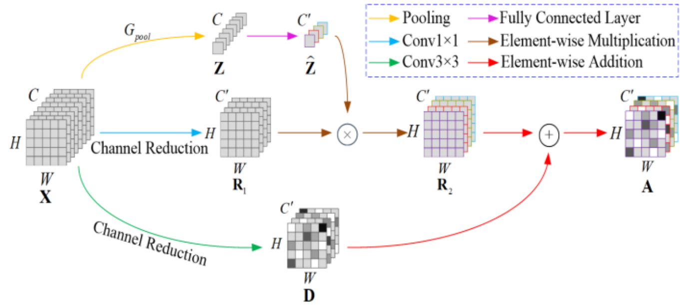

Fig. 3. Adaptive channel dimensionality reduction (ACR) module.

A. Multi-level Features Extraction

In this letter, ResNet-50 is chosen as a feature extractor, as it possesses great capabilities of extracting features at different levels, such as basic features at the shallow level and complex features at the high level, avoiding the influence of gradient disappearance in the backpropagation process.

In the feature extraction process of ResNet-50, the layers which generate feature maps of the same sizes are generally defined as a stage, and then there are total five stages, as shown in Fig. 2. Since 'Stage 1' only contains a convolutional layer of size 7×7, we do not use its outputs as multi-level features. Thus, we obtain four level features: $\mathbf{C}_2$, $\mathbf{C}_3$, $\mathbf{C}_4$, and $\mathbf{C}_5$ at Stages 2~5, the number of channels of which are 256, 512, 1024 and 2048, respectively.

B. ACR

With the ResNet-50 model, we have collected multi-level features. Next, ACR, as shown in Fig. 3, is constructed for adaptive channel dimensionality reduction, which is made up of three branches: one trunk branch (i.e., the middle branch) and two side branches (i.e., the top and the down branches).

1) The Middle Branch: Given the feature maps $\mathbf{X}\in\mathbb{R}^{H\times W\times C}$ before dimensionality reduction, where H, W and C represent the height, width and channel dimension of the feature maps, we perform a 1×1 convolution on them to get $\mathbf{R}_1\in\mathbb{R}^{H\times W\times C'}$ ($C'<C$), which are shown as follows:

$$\mathbf{R}_1=\delta(G_{conv1\times1}(\mathbf{X})) \quad (1)$$

where $\mathbf{R}_1$ represent the feature maps with low-dimensional channels, $G_{conv1\times1}$ denotes a two-dimensional (2-D) convolution of kernel size 1×1 and δ represents the rectified linear unit (ReLU) activation function.

2) The Top Branch: This branch is designed for enhancing the channel attention for the reduced features $\mathbf{R}_1$. The idea is inspired by the squeeze operation in the SENet [11], which exploits the global embedding information to model channel relationships and modulate feature maps on the channel-wise level. By decoupling the relationships between channels, the channel attention could be generated, and such attention is beneficial for emphasizing informative features and suppressing less useful ones in the channel dimension.

First, we adopt global average pooling (GAP) to generate the global features for each channel dimension. We rewrite the feature maps $\mathbf{X}$ as:

$$\mathbf{X}=[\mathcal{X}^1,\mathcal{X}^2,\cdots,\mathcal{X}^C] \quad (2)$$

where $\mathcal{X}^k \in \mathbb{R}^{H\times W} (k=1,2,\cdots,C)$ refers to the feature map of the k-th channel. The pooling operation can be expressed as:

$$G_{pool}(\mathcal{X}) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathcal{X}(i,j) \tag{3}$$

After the pooling operation, the obtained results are stacked to produce the global pooling result $\mathbf{Z}\in\mathbb{R}^C$ of $\mathbf{X}$:

$$\mathbf{Z} = [G_{pool}(\mathcal{X}^1), G_{pool}(\mathcal{X}^2), \cdots, G_{pool}(\mathcal{X}^C)] \tag{4}$$

Then, the attention of channels, $\mathbf{Z}\in\mathbb{R}^{C'}$, is calculated through a fully connected (FC) layer. The calculated attention is quite similar to the gate mechanism of a recurrent neural network. It uses parameter $\mathbf{W}$ in FC to generate the attention, i.e., the weight for each feature channel after dimensionality reduction. We express this process as:

$$\mathbf{Z} = \mathrm{FC}(\mathbf{Z},\mathbf{W}) = \sigma(\mathbf{W}\cdot\mathbf{Z}) \tag{5}$$

where $\mathbf{Z}\in\mathbb{R}^{C'}$ is the attention vector, σ refers to the Sigmoid activation function, and $\mathbf{W}\in\mathbb{R}^{C\times C'}$ is learned to explicitly model the correlation between channels.

Third, with the attention $\mathbf{Z}$, which reflects the importance of each feature channel, the feature maps $\mathbf{R}_1$ of the middle branch can be reweighted in the channel dimension:

$$\mathbf{R}_2 = \mathbf{Z}\otimes\mathbf{R}_1 \tag{6}$$

where $\mathbf{R}_2\in\mathbb{R}^{H\times W\times C'}$ denote the channel attention-enhanced feature maps and $\otimes$ denotes the element-wise multiplication.

3) The Down Branch: The down branch can be treated as the calculation of saliency maps, which further extracts the feature information and compensates in the spatial dimension for information lost after dimension reduction.

First, a 3×3 convolution is performed on the input feature maps $\mathbf{X}$ to produce feature maps $\mathbf{D}\in\mathbb{R}^{H\times W\times C'}$ with low-dimensional channels:

$$\mathbf{D} = \delta(G_{conv3\times 3}(\mathbf{X})) \tag{7}$$

where $G_{conv3\times 3}$ is a 2-D convolution of kernel size 3×3.

Next, the information in $\mathbf{D}$ is used as a complement to the semantic information of $\mathbf{R}_2$. The final output of ACR is obtained by:

$$\begin{aligned}\mathbf{A} &= \mathbf{R}_2\oplus\mathbf{D}\\ &= \mathbf{Z}\otimes\mathbf{R}_1\oplus\mathbf{D}\\ &= \sigma(\mathbf{W}\cdot\mathbf{Z})\otimes\delta(G_{conv1\times 1}(\mathbf{X}))\oplus\delta(G_{conv3\times 3}(\mathbf{X}))\end{aligned} \tag{8}$$

where $\oplus$ denotes the element-wise addition.

C. MLFF

By applying ACR to $\mathbf{C}_3$, $\mathbf{C}_4$ and $\mathbf{C}_5$, the multi-level and adaptive channel feature maps $\mathbf{A}_3$, $\mathbf{A}_4$ and $\mathbf{A}_5$ can be obtained. Considering that the spatial resolutions of these maps are different, we use global average pooling to spatially normalize the sizes of $\mathbf{A}_3$, $\mathbf{A}_4$, $\mathbf{A}_5$ and $\mathbf{C}_2$. This pooling operation can effectively retain salient features, preserve the directional invariance of feature maps, and reduce the spatial dimension of feature maps for better classification.

Then, the features processed by global average pooling are fused by the concatenation method:

$$\mathbf{M} = \mathrm{Concat}[G_{pool}(\mathbf{C}_2), G_{pool}(\mathbf{A}_3), G_{pool}(\mathbf{A}_4), G_{pool}(\mathbf{A}_5)] \tag{9}$$

where Concat refers to the concatenation operation and $\mathbf{M}$ is the multi-level fused features, which will be fed into the Softmax classifier for the final scene classification.

III. Experiments and Analysis

A. Data Sets

We evaluate our method on three well-known data sets: UC-Merced Data Set (UCM) [12], Aerial Image Data Set (AID) [13], and NWPU-RESISC45 (NWPU) [14].

1) UCM: This dataset contains 2100 images of 256×256 pixels and 1ft/pixel spatial resolution, covering 21 land-use scene classes with 100 images per class.

2) AID: This is a collection of 10000 images divided into 30 categories. The number of images in each category varies from 220 to 420. Each image is of 600×600 pixels in size, and the spatial resolution ranges from 0.5 to 8 m.

3) NWPU: This is a larger RS data set, which contains 31500 images distributed into 45 classes. Its size is 256×256 pixels, and the spatial resolution varies from 0.2 to 30 m.

B. Implement Details

To make a fair comparison between our method and the state-of-the-arts, we choose the training-test ratios as 8:2 & 5:5 for UCM, 5:5 & 2:8 for AID, and 2:8 & 1:9 for NWPU. The training samples are randomly selected and all the experiments are performed ten times, independently. During training, all images are randomly cropped to 224×224 pixels with random horizontal flipping. For testing, the images are also resized to 224×224 pixels.

All the experiments are performed on NVIDIA RTX 2080Ti GPU with the PyTorch deep learning framework. The backbone parameters are initialized by a pretrained ResNet-50 on ImageNet. The optimization is performed using Adam with the weight decay penalty of 10^{-5} and a batch size of 64. The learning rate is set to 3×10^{-4} and the total training epoch number is 200.

C. Ablation Study

To show the effect of ACR, a series of ablation experiments are conducted with different architecture designs. As shown in Fig. 4, for each architecture, we use ResNet-50 to extract the multi-level features $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$, $\mathbf{C}_4$, and $\mathbf{C}_5$. As described in Section II-A, similarly, we only consider the fusion of four multi-level features $\mathbf{C}_2$, $\mathbf{C}_3$, $\mathbf{C}_4$, and $\mathbf{C}_5$. For the three multi-level features $\mathbf{C}_3$, $\mathbf{C}_4$, and $\mathbf{C}_5$, we use ordinary 1×1 convolution to replace our ACR module to achieve channel dimensionality reduction, as shown in Fig. 4(a). Then, we gradually increase the number of the ACR modules in the dimensionality reduction paths, as shown in Fig. 4(b) and Fig. 4(c). Fig. 4(d) corresponds to our method.

We compare the above four cases with the experiments on 'backbone', which only uses the output of the last layer of ResNet-50 for scene classification, with the goal of verifying our proposed multi-level fusion idea.

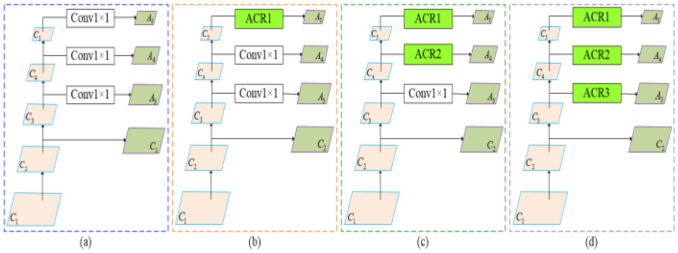

Fig. 4. Illustration of various architectures. (a) Without ACR. (b) With one ACR. (c) With two ACRs. (d) With three ACRs.

The overall accuracy of different architectures on the UCM, AID and NWPU data sets is listed in Tables I, II and III, respectively, and the following observation come from the results.

First, by comparing the results of Schemes (a), (b), (c) and (d), we witness with the increasing number of the ACR modules in the architectures, the overall accuracy becomes higher. Evidently, ACR provides better channel reduction performance than that of 1×1 convolution.

Second, by comparing our method (Scheme (d)) with Scheme 'backbone', we observe that our method achieves largely improved performance for all the data sets, which specifies the effectiveness of our multi-level fusion scheme.

Third, the comparison between Schemes (a) and 'backbone' shows that, the direct dimension reduction by 1×1 convolution leads to inferior performance. The reason is that, for Scheme (a), the feature $\mathbf{C}_5$ has the highest level semantic information and the largest number of channels. When the feature $\mathbf{C}_5$ with the channel of 2048 is reduced to 256 dimension by 1×1 convolution, most of the semantic information may be lost. Lower level features cannot help to boost the system performance. This further demonstrates the effectiveness of our ACR module.

TABLE I
OVERALL ACCURACY (%) OF DIFFERENT ARCHITECTURES ON UCM DATASET (THE BEST RESULT IS IN BOLD, WHICH IS SIMILAR TO THE FOLLOWING TABLES).

Scheme	ACR1	ACR2	ACR3	Training set : Test set	
				8 : 2	5 : 5
backbone				98.33 ±0.47	97.05 ±0.71
(a)				98.04 ±0.52	96.67 ±0.89
(b)	✓			98.75 ±0.30	97.48 ±0.55
(c)	✓	✓		99.18 ±0.23	97.71 ±0.42
(d)(Ours)	✓	✓	✓	**99.37 ±0.15**	**97.99 ±0.26**

TABLE II
OVERALL ACCURACY (%) OF DIFFERENT ARCHITECTURES ON AID DATASET.

Scheme	ACR1	ACR2	ACR3	Training set : Test set	
				5: 5	2 : 8
backbone				94.48 ±0.62	92.04 ±0.70
(a)				94.22 ±0.97	91.98 ±0.89
(b)	✓			94.76 ±0.30	92.36 ±0.51
(c)	✓	✓		94.94 ±0.25	92.61 ±0.42
(d)(Ours)	✓	✓	✓	**95.06 ±0.33**	**92.73 ±0.12**

TABLE III
OVERALL ACCURACY (%) OF DIFFERENT ARCHITECTURES ON NWPU DATASET.

Scheme	ACR1	ACR2	ACR3	Training set : Test set	
				1 : 9	2: 8
backbone				88.99 ±0.61	91.36 ±0.57
(a)				88.72 ±0.55	91.15 ±0.73
(b)	✓			89.34 ±0.52	91.66 ±0.48
(c)	✓	✓		89.78 ±0.36	92.01 ±0.42
(d)(Ours)	✓	✓	✓	**90.01 ±0.33**	**92.45 ±0.20**

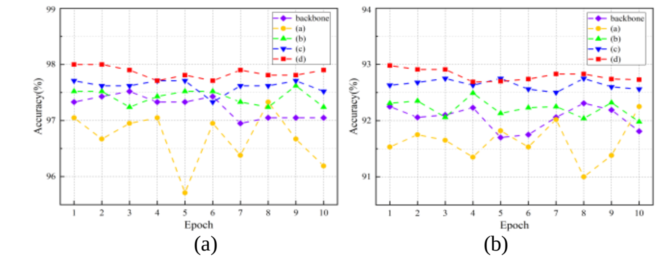

Fig. 5. Comparison of accuracy achieved after convergence of different architectures. (a) Under the 50% training ratio on the UCM dataset. (b) Under the 20% training ratio on the AID dataset.

Moreover, we report the accuracy comparison achieved after the convergence of different architectures on the UCM and AID data sets. The results are shown in Fig. 5, where the X-axis represents the domain of 10 epochs since the training processing has become stable. It can be seen that, the accuracy of our method is much higher than that of the other four architectures.

D. Comparative Study

To measure the classification performance of the proposed method in depth, the classification accuracy for each class of each data set is given by confusion matrix (CM), as shown in Figs. 6-8. It can be seen that our method yields more than 90% classification accuracy for most of the categories (e.g., 21 of the 21 classes for the UCM data set, 22 of the 30 classes for the AID data set, and 35 of the 45 classes for the NWPU data set) even with a small number of training samples.

We also compare our method (referred to as ACR-MLFF) with several state-of-the-art algorithms. The results are presented in Tables IV~ and VI. As can be seen, our method achieves superior performance with higher overall accuracies and smaller standard deviations compared to most of the competing algorithms. Compared to the SFCNN, our method yields a 0.1% lower accuracy with the 20% training ratio, whereas obtains 0.12% higher accuracy with the 10% training ratio, for the NWPU data set, as shown in Table VI.

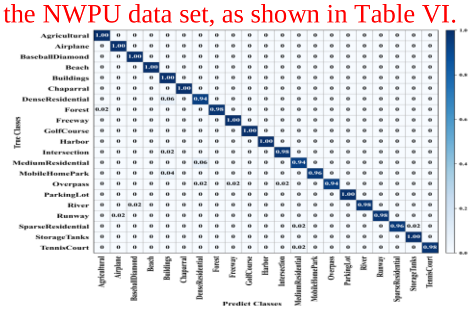

Fig. 6. CM of our method under the 50% training ratio on the UCM dataset.

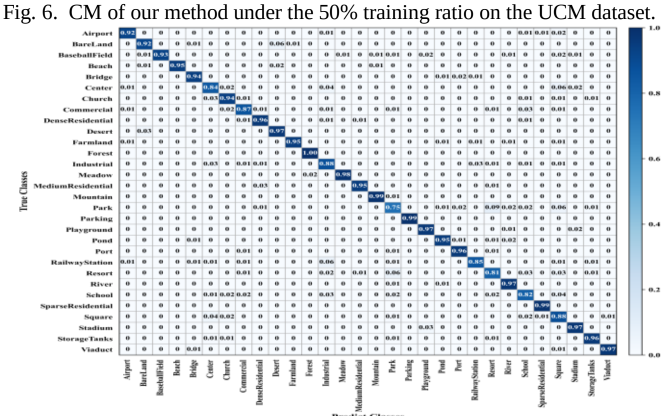

Fig. 7. CM of our method under the 20% training ratio on the AID dataset.

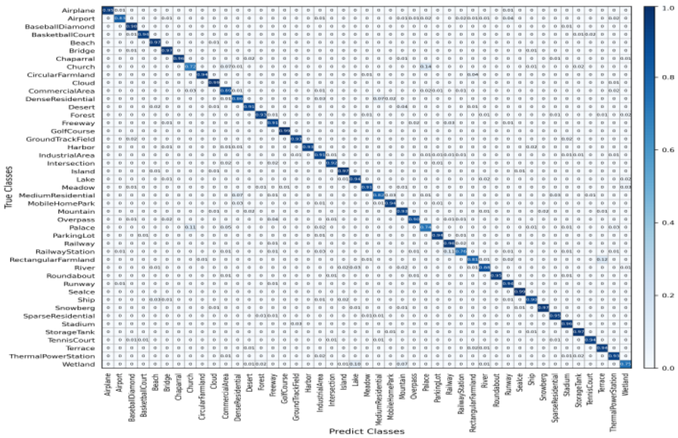

Fig. 8. CM of our method under the 10% training ratio on the NWPU dataset.

TABLE IV
OVERALL ACCURACY (%) ON THE UCM DATASET.

Method	80% for training	50% for training
TEX-Net-LF [2]	96.62 ±0.49	95.89 ±0.37
MCNN [3]	96.66 ±0.90	-
Two-Stream Fusion [15]	98.02 ±1.03	96.97 ±0.75
MSCP [4]	98.36 ±0.58	-
ADFF [5]	98.81 ±0.51	-
MLFF-WWA [6]	98.46	-
ARCNet-VGG16 [16]	99.12 ±0.40	96.81 ±0.14
ResNet_LGFFE [17]	98.62 ±0.88	-
LCPB [18]	96.66 ±1.36	-
LCPP [18]	97.54 ±1.02	-
ORRCNN [19]	96.42	96.58
ACR-MLFF(Ours)	**99.37 ±0.15**	**97.99 ±0.26**

TABLE V
OVERALL ACCURACY (%) ON THE AID DATASET.

Method	50% for training	20% for training
MCNN [3]	91.80 ±0.22	-
TEX-Net-LF [2]	92.96 ±0.18	90.87 ±0.11
ARCNet-VGG16 [16]	93.10 ±0.55	88.75 ±0.40
Two-Stream Fusion [15]	94.58 ±0.25	92.32 ±0.41
MSCP [4]	94.42 ±0.17	91.52 ±0.21
ResNet_LGFFE [17]	94.46 ±0.48	90.83 ±0.55
LCPB [18]	91.33 ±0.36	87.68 ±0.25
LCPP [18]	93.12 ±0.28	90.96 ±0.33
ORRCNN [19]	92.00	86.42
ACR-MLFF(Ours)	**95.06 ±0.33**	**92.73 ±0.12**

TABLE VI
OVERALL ACCURACY (%) ON THE NWPU DATASET.

Method	10% for training	20% for training
MSCP [4]	88.07 ±0.18	90.81 ±0.13
SFCNN [20]	89.89 ±0.16	**92.55 ±0.14**
Siamese ResNet50 [21]	-	92.28 ±3.78
SDAResNet [22]	89.40	91.15
VGG_VD16+SAFF [23]	84.38 ±0.19	87.86 ±0.14
SCCov [24]	89.30 ±0.35	92.10 ±0.25
ACR-MLFF(Ours)	**90.01 ±0.33**	92.45 ±0.20

IV. CONCLUSION

This letter has proposed a RS scene classification method by multi-level feature fusion network with adaptive channel dimensionality reduction. We designed ACR to reduce the high dimensionality of features. Then, the multi-level features are fused together for precise classification. The proposed method is evaluated on three RS data sets, and the results show that it outperforms several baseline algorithms. For the future work, we will work on designing more plug-and-play modules, like the proposed ACR, and embedding them into different CNN architectures to further boost the networks' ability for RS scene classification.

REFERENCES

[1] G. Cheng, X. Xie, J. Han, L. Guo and G. Xia, "Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735-3756, 2020.

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

[3] Y. Liu, Y. Zhong and Q. Qin, "Scene Classification Based on Multiscale Convolutional Neural Network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7109-7121, Dec. 2018.

[4] N. He, L. Fang, S. Li, A. Plaza and J. Plaza, "Remote Sensing Scene Classification Using Multilayer Stacked Covariance Pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6899-6910, Dec. 2018.

[5] B. Li et al., "Aggregated Deep Fisher Feature for VHR Remote Sensing Scene Classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3508-3523, Sept. 2019.

[6] C. Ma, X. Mu, R. Lin and S. Wang, "Multilayer Feature Fusion With Weight Adjustment Based on a Convolutional Neural Network for Remote Sensing Scene Classification," *IEEE Geosci. Remote Sens. Lett.*, doi: 10.1109/LGRS.2020.2970810.

[7] J. Ji, T. Zhang, L. Jiang, W. Zhong and H. Xiong, "Combining Multilevel Features for Remote Sensing Image Scene Classification With Attention Model," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1647-1651, Sept. 2020.

[8] 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.

[9] Y. Lu, L. Zhang, J. Liu and Q. Tian, "Constructing Concept Lexica With Small Semantic Gaps," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 288-299, June 2010.

[10] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011-2023, 1 Aug. 2020.

[11] Z. Lu, K. Deb and V. Naresh Boddeti, "MUXConv: Information Multiplexing in Convolutional Neural Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12041-12050.

[12] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM SIGSPATIAL GIS*, 2010, pp. 270-279.

[13] G. Xia et al., "AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965-3981, July 2017.

[14] G. Cheng, J. Han and X. Lu, "Remote Sensing Image Scene Classification: Benchmark and State of the Art," *Proceedings of the IEEE*, vol. 105, no.10, pp. 1865-1883, Oct. 2017.

[15] Y. Yu and F. Liu, "A Two-Stream Deep Fusion Framework for High-Resolution Aerial Scene Classification," *Comput. Intell. Neurosci.*, vol. 2018, p. 8639367, 2018/01/18 2018.

[16] Q. Wang, S. Liu, J. Chanussot and X. Li, "Scene Classification With Recurrent Attention of VHR Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155-1167, Feb. 2019.

[17] Y. Lv, X. Zhang, W. Xiong, Y. Cui, and M. Cai, "An end-toend local-global-fusion feature extraction network for remote sensing image scene classification," *Remote Sens.*, vol. 11, no. 24, p. 3006, Dec. 2019.

[18] X. Sun, Q. Zhu and Q. Qin, "A Multi-Level Convolution Pyramid Semantic Fusion Framework for High-Resolution Remote Sensing Image Scene Classification and Annotation," *IEEE Access*, vol. 9, pp. 18195-18208, 2021.

[19] Z. Li, Q. Wu, B. Cheng, L. Cao and H. Yang, "Remote Sensing Image Scene Classification Based on Object Relationship Reasoning CNN," *IEEE Geosci. Remote Sens. Lett.* doi: 10.1109/LGRS.2020.3017542, 2020.

[20] J. Xie, N. He, L. Fang and A. Plaza, "Scale-Free Convolutional Neural Network for Remote Sensing Scene Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916-6928, Sept. 2019.

[21] 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.

[22] D. Guo, Y. Xia and X. Luo, "Scene Classification of Remote Sensing Images Based on Saliency Dual Attention Residual Network," *IEEE Access*, vol. 8, pp. 6344-6357, 2020.

[23] R. Cao, L. Fang, T. Lu and N. He, "Self-Attention-Based Deep Feature Fusion for Remote Sensing Scene Classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 43-47, Jan. 2021.

[24] N. He, L. Fang, S. Li, J. Plaza and A. Plaza, "Skip-Connected Covariance Network for Remote Sensing Scene Classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1461-1474, May 2020.