This article has been accepted for publication in IEEE Geoscience and Remote Sensing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LGRS.2022.3173473

Multilayer Feature Fusion Network with Spatial Attention and Gated Mechanism for Remote Sensing Scene Classification

Qingyan Meng, Maofan Zhao, Linlin Zhang, Wenxu Shi, Chen Su, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—Remote sensing (RS) scene classification has attracted extensive attention due to its large number of applications. Recently, convolutional neural networks (CNNs) methods have shown impressive ability of feature learning in RS scene classification. However, the performance is still limited by large-scale variance and complex background. To address these problems, we present a multilayer feature fusion network with spatial attention and gated mechanism (MLF2Net_SAGM) for RS scene classification. At first, the backbone is employed to extract multilayer convolutional features. Then, a residual spatial attention module (RSAM) is proposed to enhance discriminative regions of the multilayer feature maps, and key areas can be harvested. Finally, the multilayer spatial calibration features are fused to form the final feature map, and a gated fusion module (GFM) is designed to eliminate feature redundancy and mutual exclusion (FRME). To verify the effectiveness of the proposed method, we conduct comparative experiments based on three widely used RS image scene classification benchmarks. The results show that the direct fusion of multilayer features via element-wise addition leads to FRME, whereas our method fuses multilayer features more effectively and improves the performance of scene classification.**

***Index Terms*—scene classification, multilayer feature fusion, spatial attention, gated mechanism, remote sensing.**

I. Introduction

With the rapid development of satellite imaging technology, the resolution of remote sensing (RS) has been continuously improved to sub-meter level. Accordingly, RS image analysis mode gradually transforms from pixel-level to scene-level, such as scene classification. Compared with pixel-level classification [1], scene classification can obtain semantic information [2]. So RS scene classification received more and more attention in the RS applications [3]–[6].

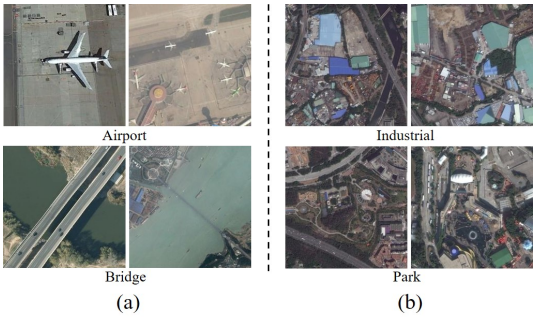

Fig. 1. Challenges of RS scene classification. (a) large-scale variance. (b) complex background.

The diversity of RS image acquisition platforms and sensors leads to large-scale variance of objects in RS scenes, as shown in Fig. 1(a). In addition, RS scenes contain complex ground objects especially urban scenes, as shown in Fig. 1(b). So the objects of interest may only occupy a small part of the image, which is easily disturbed by useless objects contained in complex backgrounds. Therefore, although the CNN-based method has greatly improved the performance of RS scene classification, the classification performance is still limited. In particular, the continuous pooling operation in CNNs causes feature map size reduction, which also results in ignoring key areas of the scenes.

To address these problems, many studies try improve spatial representation, or aggregate multilayer features in CNNs to exploit detailed information. On the one hand, inspired by CBAM [7], Zhao *et al.* [8] and Chen *et al.* [9] propose enhanced spatial attention module and local spatial attention module (LSAM) respectively. But they relies on local convolution kernels, making it difficult to obtain global correlations. On the other hand, Xu *et al.* [10] aggregate multilayer features based on dictionary learning. Xu *et al.* [11] fuse multilayer convolutional features based on the transferred VGGNet-16 model. But multilayer feature fusion of CNNs is prone to feature redundancy and mutual exclusion (FRME), which is not considered in them [10], [11].

In this letter, we propose the multilayer feature fusion network with spatial attention and gated mechanism (MLF2Net_SAGM) to solve the above problems. The main contributions of MLF2Net_SAGM can be summarized as follows:

This work was supported in part by Hainan Provincial Department of Science and Technology under Grant ZDKJ2019006; in part by the National Natural Science Foundation of China under Grant 42171357; in part by the Major Projects of High Resolution Earth Observation Systems of National Science and Technology under Grant 05-Y30B01-9001-19/20-1; and in part by the China Scholarship Council under Grant 202104910460. (*Corresponding author: Maofan Zhao.*)

Qingyan Meng, Linlin Zhang are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China, and with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Key Laboratory of Earth Observation of Hainan Province, Hainan Research Institute, Aerospace Information Research Institute, Chinese Academy of Sciences, Sanya, 572029, China (e-mail: mengqy@radi.ac.cn; zhangll@radi.ac.cn).

Maofan Zhao is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China, and with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: mfzhao1998@163.com).

Wenxu Shi, Chen Su are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: shiwenxu20@mails.ucas.edu.cn; suchen21@mails.ucas.edu.cn).

Lorenzo Bruzzone is with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: lorenzo.bruzzone@unitn.it).

This article has been accepted for publication in IEEE Geoscience and Remote Sensing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LGRS.2022.3173473

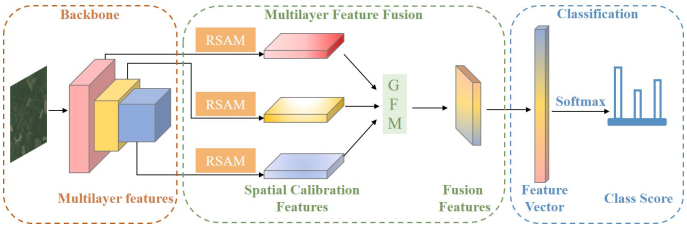

Fig. 2. Illustration of MLF2Net_SAGM.

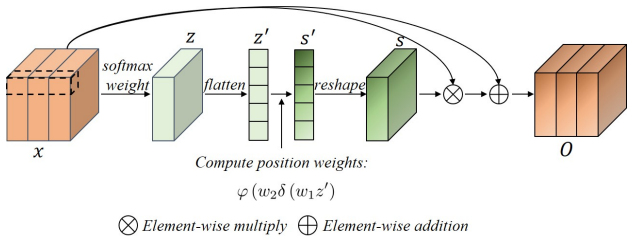

Fig. 3. Detailed structure diagram of RSAM.

(1) Proposing a multilayer feature fusion strategy with spatial attention and gated mechanism which can integrate into an end-to-end network.

(2) In order to focus on the key areas in RS images, a residual spatial attention module (RSAM) is specially designed to obtain global correlations and calibrate features in spatial dimension.

(3) To eliminate the FRME, a gated fusion module (GFM) is proposed to select and fuse multilayer features. GFM can make more effective use of multilayer features, enhancing the complementarity.

II. Methodology

As shown in Fig. 2, we propose MLF2Net_SAGM, that consists of three modules: CNN backbone, multilayer feature fusion and classification. The CNN backbone is used to obtain multilayer convolution features. In our study, ResNet50 is used to get multilayer features. Then, multilayer features are fed into RSAM for spatial feature calibration respectively, that makes the model focus on the key areas and ignore the background information. Next, the multilayer spatial calibration features are fused through the designed GFM which can effectively avoid FRME. Finally, the RS scene is classified by softmax classifier.

A. Multilayer features

The multilayer features of CNNs contain various information, so the effective use of multilayer features is essential to improve feature representation. This study extracts multilayer features based on ResNet50. More specifically, the output features of conv3_x, conv4_x and conv5_x are used for multilayer features fusion.

B. Residual Spatial Attention Module (RSAM)

Due to the large-scale variance and complex background of RS images, CNNs often fail to obtain good spatial representation. Therefore, we design RSAM to generate the corresponding spatial attention weights and improve the spatial representation, as shown in Fig. 3.

Firstly, the position-wise statistic $z \in \mathbb{R}^{H\times W}$ is generated by softmax weighting of $x \in \mathbb{R}^{C\times H\times W}$ through the channel dimensions C. H, W are the width and height of feature map respectively, and x is the input of RSAM. $z(i,j)$ is calculated by

$$w_c(i,j) = \frac{e^{x_c(i,j)}}{\sum_{c=1}^{C} e^{x_c(i,j)}}, \tag{1}$$

$$z(i,j) = \sum_{c=1}^{C} w_c(i,j) \times x_c(i,j). \tag{2}$$

Then, a nonlinear bottleneck is constructed. More specifically, it includes two fully connected layers, relu and softmax functions. So the spatial weights are generated as follows:

$$z' = flatten(z), \tag{3}$$

$$s' = \sigma\left(g\left(z', w\right)\right) = \varphi\left(w_2\delta\left(w_1 z'\right)\right), \tag{4}$$

$$s = reshape(s'), \tag{5}$$

where δ is the relu nonlinear unit, φ is the softmax function and $z' \in \mathbb{R}^{M}$, $w_1 \in \mathbb{R}^{\frac{M}{r_1}\times M}$, $w_2 \in \mathbb{R}^{M\times\frac{M}{r_1}}$, $s \in \mathbb{R}^{H\times W}$. The dimensionality reduction ratio is r_1.

The spatial calibration feature O is obtained by recalibrating x with residual:

$$O = F_{reweight}(x, s) = x \otimes s + x, \tag{6}$$

where $\otimes$ represents elements-wise multiply.This module essentially introduces a dynamic adaptive strategy based on the input, which can get global correlation and not limited to the convolution kernel with the local receptive field. Through RSAM, key regions can be effectively emphasized in multilayer features.

C. Gated Fusion Module (GFM)

To effectively utilize the hierarchical information of the convolution structure, widely used fusion methods include element-wise addition and concatenation in channel. However, these methods are prone adding some disturbing information, such as FRME among multilayer. In order to consider the importance of multilayer features, this letter designs GFM, as shown in Fig. 4. O_1, O_2, O_3 denote spatial calibration features based on conv3_x, conv4_x, conv5_x respectively. It should be specially noted that O_1 and O_2 adopt channel transformation similar to GhostNet [12] to unify their channel number with O_3, which can reduce the number of parameters compared to 1×1 convolution. gp denotes the concatenation of the global average pooling (GAP) results for O_1, O_2, O_3. p, q and l denote three groups of weights for O_1, O_2, O_3. fc_1, fc_2, fc_3 and fc_4 are fully connected layers. I represents the fusion feature. The details of GFM are introduced as follows.

We fuse the results from RSAM. The feature map $O \in \mathbb{R}^{C\times H\times W}$ is squeezed to generate channel-wise statistics $gp_o \in \mathbb{R}^{C}$ by GAP, and initial fusion feature gp' is obtained

This article has been accepted for publication in IEEE Geoscience and Remote Sensing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LGRS.2022.3173473

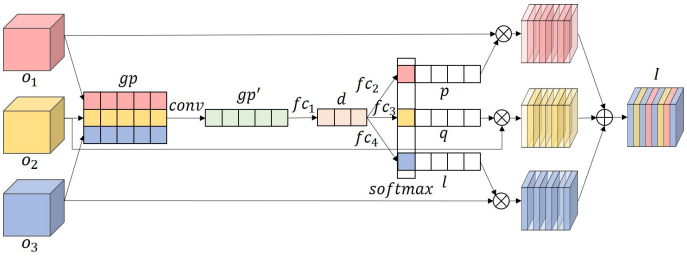

Fig. 4. Detailed structure diagram of GFM.

by

$$gp_{o_c} = F_{gp}(O_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} O_c(i,j), \tag{7}$$

$$gp = cat[gp_1, gp_2, gp_3], \tag{8}$$

$$gp' = conv(gp), \tag{9}$$

where O_c denotes the c-th channel of O, gp_{o_c} represents the c-th of gp_o. gp_1, gp_2, gp_3 represent the GAP results of O_1, O_2, and O_3, respectively. cat and $conv$ represent concatenation and convolution operations, respectively. gp represents combined results by the gp_1, gp_2, gp_3 via concatenation.

To improve efficiency and generalization, a compact feature $d \in \mathbb{R}^{\frac{c}{r_2}}$ is obtained by a fully connected layer:

$$d = fc_1(gp'). \tag{10}$$

Furthermore, the three groups of importance weights (p, q and l) are computed as follows:

$$\begin{vmatrix} p \\ q \\ l \end{vmatrix} = \varphi \begin{vmatrix} fc_2(d) \\ fc_3(d) \\ fc_4(d) \end{vmatrix}, \tag{11}$$

where fc_2, fc_3 and fc_4 have the same structure but don't share parameters. φ means softmax operation on the same channel of p, q and l, so that the elements of p, q and l are in the range of 0 to 1, and the sum of the corresponding positions in them is 1. The fusion feature I can be computed by

$$I = p \otimes O_1 + q \otimes O_2 + l \otimes O_3. \tag{12}$$

The gated mechanism adaptively selects the effective information in the multilayer features for fusion, so that the fused features have the complementarity of the multilayer features and avoid FRME.

III. Experimental Results

A. Data Sets Description

We analyze the performance of the MLF2Net_SAGM on three datasets (AID [13], NWPU-RESISC45 [14], RSSCN7 [15]) whose details are shown in Table I.

TABLE I
Description of three different remote sensing scene data sets

Data set	Total images	Scene classes	Images per class	Resolution / scale	Image size
AID	10000	30	220-400	0.5-0.8m	600×600
NWPU-RESISC45	31500	45	700	0.2-30m	256×256
RSSCN7	2800	7	400	1:700, 1:1300, 1:2600, 1:5200	256×256

B. Design of Experimentals

The SGD is chosen as the optimizer. All models are trained for 150 epochs with an initial learning rate of 0.01, and the learning rate is multiplied by 0.5 every 30 epochs, and the batch size is set to 64. The backbone is initialized with ImageNet-based pretrained parameters. The experiments are based on the Pytorch framework with the hardware environment of NVIDIA 3090 GPU, I9-10980XE CPU and a memory size of 64GB.

The image sizes of all three data sets are resized to 224 × 224, and common data augmentation strategies are adopted. In all experiments, r_1 is set to twice the square root of M and r_2 is set to 64. In addition, different training ratios (Trs) are adopted for each data set. For AID and RSSCN7 data sets, the Trs are set to 20% and 50%; for NWPU-RESISC45 data set, the Trs are set to 10% and 20%. All experiments are repeated five times to get more reliable experimental results by randomly selecting the training samples, and the remain samples are used for testing. Overall accuracy (OA) is used to quantitatively evaluate the experimental results, which reflects the overall performance of the model.

C. Ablation Experiment

To understand the effects of RSAM and GFM clearly, ablation experiments are carried out on the RSSCN7 data set, the params, flops, OAs of different settings are shown in Table II.

TABLE II
Comparison of OA (%) by employing different settings in the ablation study on RSSCN7 data set

Method	Params (MB)	Flops (G)	Trs	
			20%	50%
ResNet50	23.52	4.12	92.07±0.53	94.87±0.21
ResNet50+Addition	23.53	4.12	91.89±0.40	94.71±0.19
ResNet50+RSAM	23.56	4.12	92.20±0.68	95.20±0.38
ResNet50+GFM	23.79	4.12	92.93±0.68	95.70±0.28
MLF2Net_SAGM	23.82	4.12	**93.28±0.36**	**96.01±0.23**

The fine-tuned ResNet50 obtains 92.07% and 94.87% OAs on the RSSCN7 data set at 20% and 50% Trs, respectively. The OAs decreases by 0.18% and 0.16% respectively by ResNet50+Addition (fusion via element-wise addition), confirming that direct feature fusion via element-wise addition causes FRME. By using RSAM for multilayer feature before fusion, the OAs are improved by 0.31% and 0.49% respectively compared with ResNet50+Addition. The results show that RSAM can effectively calibrate spatial features. Using GFM in the element-wise fusion process, the OAs are improved by 1.04% and 0.99% compared to ResNet50+Addition, indicating that GFM can capture the complementary information in multilayer features and avoid FRME. Furthermore, MLF2Net_SAGM obtains 93.28% and 96.01% OAs with RSAM and GFM, which improves 1.21% and 1.14% compared to ResNet50, showing the effectiveness of MLF2Net_SAGM. In addition, our method brings a very limited number of parameters and computations.

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

1) AID Data Set: The comparison results between MLF2Net_SAGM and the state-of-the-art methods are shown

This article has been accepted for publication in IEEE Geoscience and Remote Sensing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LGRS.2022.3173473

in Table III. The OAs of MLF2Net_SAGM reaches 95.44% and 97.08% at 20% and 50% Trs, respectively. Comparison with other three attention mechanism methods, such as ResNet50+EAM, ResNet101+SENet, and ResNet101+CBAM, the OA of the proposed method improved by about 1.80%, 1.75% and 1.83% at 20% Tr, respectively. EAM and CBAM are mixed attention mechanisms, including spatial attention and channel attention. However, the spatial attention depends on the local convolution kernel, which is difficult to obtain the global relationship, which limits their performance. The classification results of all three attention mechanisms are significantly lower than the MLF2Net_SAGM, which also verifies the effectiveness of RSAM.

TABLE III
COMPARISON OF OA (%) WITH SOME STATE-OF-THE-ART METHODS ON AID DATA SET

Method	Trs	
	20%	50%
SIFT [14]	13.50±0.67	16.76±0.65
BoVW(SIFT) [14]	61.40±0.41	67.65±0.49
CaffeNet [14]	86.86±0.47	89.53±0.31
SAFF [16]	90.25±0.29	93.83±0.28
Two-Stream Deep Fusion [10]	92.32±0.41	94.58±0.25
TFADNN [17]	93.21±0.32	95.64±0.16
ResNet101+CBAM [8]	93.51±0.22	96.56±0.21
ResNet50+EAM [8]	93.64±0.25	96.62±0.13
ResNet101+SENet [8]	93.69±0.35	96.61±0.21
MF2Net [11]	93.82±0.26	95.93±0.23
MINet-ResNet50 [18]	-	95.93±0.22
MLF2Net_SAGM	**95.44±0.25**	**97.08±0.17**

2) NWPU-RESISC45 Data Set: Table IV reports the classification results of the considered methods. The MLF2Net_SAGM significantly improves the OA compared to five other feature fusion methods, namely Two-Stream Deep Fusion, SAFF, TFADNN, MF2Net, and MINet-ResNet50. At 20% Tr, the improvements are 12.13%, 4.57%, and 2.18% compared to Two-Stream Deep Fusion, TFADNN, and MF2Net, respectively. At 50% Tr, the improvements compared to Two-Stream Deep Fusion, SAFF, TFADNN, MF2Net, and MINet-ResNet50 are 11.68%, 6.98%, 3.98%, 2.11%, and 0.88%, respectively. SAFF, MF2Net, and MINet-ResNet50 all adopt the multilayer features of CNNs. However, they do not enhance the spatial representation of feature maps and consider the FMRE in feature fusion. So their OAs are dramatically lower than the proposed method. This confirms the effectiveness of MLF2Net_SAGM.

TABLE IV
COMPARISON OF OA(%) WITH SOME STATE-OF-THE-ART METHODS ON NWPU-RESISC45 DATA SET

Method	Trs	
	10%	20%
GIST [13]	15.90±0.23	17.88±0.22
Two-Stream Deep Fusion [10]	80.22±0.22	83.16±0.18
SAFF [16]	84.38±0.19	87.86±0.14
TFADNN [17]	87.78±0.11	90.86±0.24
MF2Net [11]	90.17±0.25	92.73±0.21
ResNet50+EAM [8]	90.87±0.15	93.51±0.12
ResNet101+SENet [8]	91.36±0.25	93.52±0.11
ResNet101+CBAM [8]	91.63±0.15	93.86±0.13
MINet-ResNet50 [18]	-	93.96±0.12
MLF2Net_SAGM	**92.35±0.17**	**94.84±0.09**

3) RSSCN7 Data Set: The performance comparison between MLF2Net_SAGM and some sate-of-the-art methods at 20% and 50% Trs is shown in the Table V. The OAs of MLF2Net_SAGM has reached 93.28% and 96.01% with a huge improvement compared with existing methods.

TABLE V
COMPARISON OF OA (%) WITH SOME STATE-OF-THE-ART METHODS ON RSSCN7 DATA SET

Method	Trs	
	20%	50%
BoVW(SIFT) [14]	76.33±0.88	81.34±0.55
Tex-Net-LF_VGG-M [19]	88.61±0.46	91.25±0.57
Resnet50 [19]	90.23±0.43	93.12±0.55
WSPM-CRC-ResNet152 [20]	-	93.90
Tex-Net-LF_Resnet50 [19]	92.45±0.45	94.00±0.57
DFAGCN [21]	-	94.14±0.44
SE-MDPMNet [22]	92.65±0.13	94.71±0.15
Contourlet CNN [23]	-	95.54±0.71
MLF2Net_SAGM	**93.28±0.36**	**96.01±0.23**

E. Visualization

The convergence of the proposed MLF2Net_SAGM is visualized by using AID data set with 20% Tr, as show in Fig. 5. In the early stage of training, the loss decreases rapidly and the accuracy increases rapidly. Around the 15th epoch, they start to fluctuate wildly. But as the learning rate is halved every 30 epochs, the loss further decreases and gradually converges.

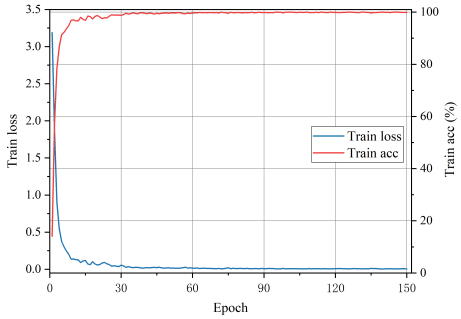

Fig. 5. The train loss and train accuracy of MLF2Net_SAGM for training the AID data set with 20% Tr.

Fig. 6 shows the confusion matrix of MLF2Net_SAGM on the AID data set with 20% Tr. The accuracy is higher than 90% in 24 categories and higher than 98% in 13 categories . The categories with less than 90% accuracy are school (86%), park (87%), resort (87%), commercial (88%), church (89%), and square (89%). The confusion of these six categories with other categories provides an important contribution to the OA.

For qualitative analysis, we applied ScoreCAM [24] to the baseline and MLF2Net_SAGM using images from the NWPU-RESISC45 test set. The visualization results of the final output feature map of ResNet-50 and MLF2Net_SAGM are compared, as shown in Fig. 7. We can find that the ScoreCAM mask for MLF2Net_SAGM covers the key regions better compared to the baseline. In other words, MLF2Net_SAGM can emphasize key regions in multilayer features, aggregate effective features and eliminate redundant features. In summary, MLF2Net_SAGM effectively utilizes the hierarchical convolution features.

This article has been accepted for publication in IEEE Geoscience and Remote Sensing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LGRS.2022.3173473

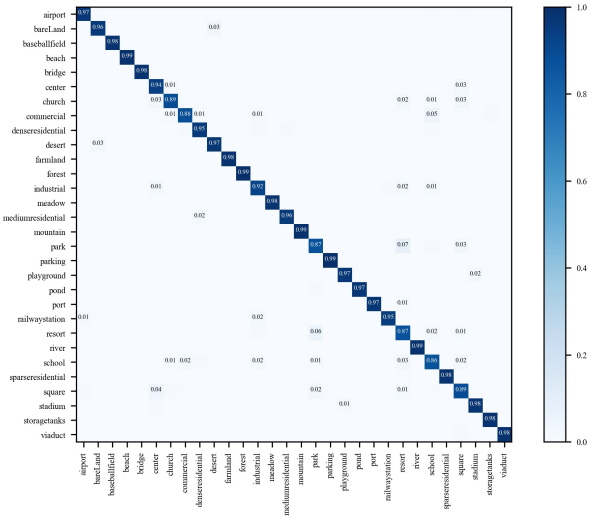

Fig. 6. Confusion matrix of the MLF2Net_SAGM on the AID data set with 20% Tr.

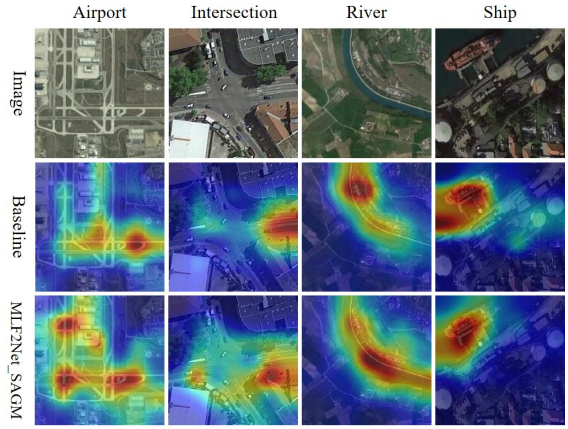

Fig. 7. ScoreCAM visualization results

IV. Conclusion

In this letter, MLF2Net_SAGM is proposed for RS scene classification. MLF2Net_SAGM addresses the issues of large-scale variance and complex background of RS scenes. Firstly, a multilayer feature fusion strategy is used to effectively utilize detailed and semantic information in hierarchical features of CNNs. Then, RSAM is designed to obtain global correlations and emphasize the key areas of the images. In addition, GFM is designed avoid FRME in the feature fusion process. In order to verify the effectiveness and robustness of MLF2Net_SAGM, we have carried out lots of comparative experiments on three benchmarks. The experimental results show that RSAM and GFM are effective in the multilayer feature fusion process. Although RSAM can calibrate spatial features, it is still difficult to extract long-range features compared to visual transformer. Therefore, it is necessary to extract the local and long-range features of images combined with visual transformer and CNNs in the future.

References

[1] F. Luo, Z. Zou, J. Liu, and Z. Lin, "Dimensionality reduction and classification of hyperspectral image via multistructure unified discriminative embedding," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[2] K. Xu, P. Deng, and H. Huang, "Vision transformer: An excellent teacher for guiding small networks in remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[3] J. Li, X. Huang, and X. Chang, "A label-noise robust active learning sample collection method for multi-temporal urban land-cover classification and change analysis," *ISPRS-J. Photogramm. Remote Sens.*, vol. 163, pp. 1–17, 2020.

[4] L. Ding, H. Tang, and L. Bruzzone, "Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, 2021.

[5] X. Huang, J. Yang, J. Li, and D. Wen, "Urban functional zone mapping by integrating high spatial resolution nighttime light and daytime multi-view imagery," *ISPRS-J. Photogramm. Remote Sens.*, vol. 175, pp. 403–415, 2021.

[6] X. Huang, A. Liu, and J. Li, "Mapping and analyzing the local climate zones in china's 32 major cities using landsat imagery based on a novel convolutional neural network," *Geo-Spat. Inf. Sci.*, pp. 1–30, 2021.

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

[8] Z. Zhao, J. Li, Z. Luo, J. Li, and C. Chen, "Remote sensing image scene classification based on an enhanced attention module," *IEEE Geosci. Remote Sens. Lett.*, pp. 1–5, 2020.

[9] S.-B. Chen, Q.-S. Wei, W.-Z. Wang, J. Tang, B. Luo, and Z.-Y. Wang, "Remote sensing scene classification via multi-branch local attention network," *IEEE Trans. Image Process.*, vol. 31, pp. 99–109, 2021.

[10] Y. Yu and F. Liu, "A two-stream deep fusion framework for high-resolution aerial scene classification," *Comput. Intell. Neurosci.*, vol. 2018, no. 4-5, pp. 1–13, 2018.

[11] K. Xu, H. Huang, Y. Li, and G. Shi, "Multilayer feature fusion network for scene classification in remote sensing," *IEEE Geosci. Remote Sens. Lett.*, vol. PP, no. 99, pp. 1–5, 2020.

[12] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet more features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1580–1589, 2020.

[13] C. Gong, 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, 2017.

[14] 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, 2017.

[15] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, 2015.

[16] 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, 2021.

[17] K. J. Xu, H. Huang, P. F. Deng, and G. Y. Shi, "Two-stream feature aggregation deep neural network for scene classification of remote sensing images," *Inf. Sci.*, vol. 539, pp. 250–268, 2020.

[18] J. Hu, Q. Shu, J. Pan, J. Tu, Y. Zhu, and M. Wang, "Minet: Multilevel inheritance network-based aerial scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[19] R. M. Anwer, F. S. Khan, V. Joost, 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, no. APR., pp. 74–85, 2017.

[20] B.-D. Liu, J. Meng, W.-Y. Xie, S. Shao, Y. Li, and Y. Wang, "Weighted spatial pyramid matching collaborative representation for remote-sensing-image scene classification," *Remote Sens.*, vol. 11, no. 5, p. 518, 2019.

[21] K. Xu, H. Huang, P. Deng, and Y. Li, "Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2021.

[22] B. Zhang, Y. Zhang, and S. Wang, "A lightweight and discriminative model for remote sensing scene classification with multidilation pooling module," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2636–2653, 2019.

[23] M. Liu, L. Jiao, X. Liu, L. Li, F. Liu, and S. Yang, "C-cnn contourlet convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2636–2649, 2020.

[24] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 24–25, 2020.