

# *CAAI Transactions on Intelligence Technology*

## Special issue Call for Papers

---

**Be Seen. Be Cited.  
Submit your work to a new  
IET special issue**

Connect with researchers and experts in your field and share knowledge.

Be part of the latest research trends, faster.



[Read more](#)



The Institution of  
Engineering and Technology

## ORIGINAL RESEARCH

# Scale-wise interaction fusion and knowledge distillation network for aerial scene recognition

Hailong Ning<sup>1,2</sup>  | Tao Lei<sup>3</sup> | Mengyuan An<sup>1,2</sup> | Hao Sun<sup>4</sup> | Zhanxuan Hu<sup>1,2</sup> | Asoke K. Nandi<sup>5,6</sup> 

<sup>1</sup>School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an, China

<sup>2</sup>Xi'an Key Laboratory of Big Data and Intelligent Computing, Xi'an, China

<sup>3</sup>School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, China

<sup>4</sup>School of Computer, Central China Normal University, Wuhan, China

<sup>5</sup>Department of Electronic and Electrical Engineering, Brunel University London, London, UK

<sup>6</sup>Xi'an Jiaotong University, Xi'an, China

## Correspondence

Tao Lei.  
Email: leitao@sust.edu.cn

## Funding information

National Natural Science Foundation of China, Grant/Award Numbers: 62201452, 2271296, 62201453; Natural Science Basic Research Program of Shaanxi, Grant/Award Number: 2022JQ-592; Key Research and Development Program of Shaanxi Province, Grant/Award Number: 2021JC-47; Shaanxi Provincial Education Department, Grant/Award Number: 22JK0568

## Abstract

Aerial scene recognition (ASR) has attracted great attention due to its increasingly essential applications. Most of the ASR methods adopt the multi-scale architecture because both global and local features play great roles in ASR. However, the existing multi-scale methods neglect the effective interactions among different scales and various spatial locations when fusing global and local features, leading to a limited ability to deal with challenges of large-scale variation and complex background in aerial scene images. In addition, existing methods may suffer from poor generalisations due to millions of to-be-learned parameters and inconsistent predictions between global and local features. To tackle these problems, this study proposes a scale-wise interaction fusion and knowledge distillation (SIF-KD) network for learning robust and discriminative features with scale-invariance and background-independent information. The main highlights of this study include two aspects. On the one hand, a global-local features collaborative learning scheme is devised for extracting scale-invariance features so as to tackle the large-scale variation problem in aerial scene images. Specifically, a plug-and-play multi-scale context attention fusion module is proposed for collaboratively fusing the context information between global and local features. On the other hand, a scale-wise knowledge distillation scheme is proposed to produce more consistent predictions by distilling the predictive distribution between different scales during training. Comprehensive experimental results show the proposed SIF-KD network achieves the best overall accuracy with 99.68%, 98.74% and 95.47% on the UCM, AID and NWPU-RESISC45 datasets, respectively, compared with state of the arts.

## KEYWORDS

deep learning, image analysis, image classification, information fusion

## 1 | INTRODUCTION

Aerial scene is an organic combination of multiple ground objects and their contextual information. Aerial scene recognition (ASR) is to assign a specific semantic label for an aerial image, which can be applied to various practical applications, such as natural hazards detection [1], environmental monitoring [2], urban planning [3] and so on [4, 5]. Due to the exhibited huge value in practical applications, remarkable efforts

have been made to analyse and classify the scene of aerial images [6, 7].

In the past decades, a large collection of ASR methods have been proposed, which can be mainly divided into handcrafted feature-based and deep feature-based ASR methods. The early ASR methods are mainly based on handcrafted features, which are extracted by designing different feature descriptors, including scale-invariant descriptors [8], colour descriptors [9], texture descriptors [10], combinations of multiple feature

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

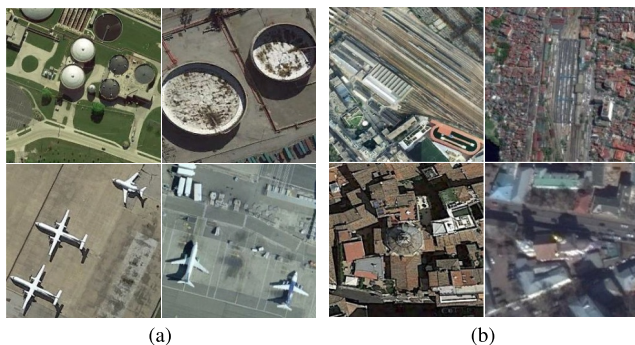
© 2023 The Authors. *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

descriptors etc. Although these methods perform well for ASR on simple images, they may offer limited use for complicated aerial scenes since hand-crafted features are difficult to represent the intrinsic properties of complex aerial scenes.

Recently, due to the tremendous success of deep learning [11, 12], deep learning feature-based ASR methods, especially the CNN-based methods, have exhibited powerful abilities to learn semantic information and have permeated many fields, involving ASR [3, 6, 7]. Among CNN-based ASR methods, the early ones mainly learn global semantic features by stacking multiple convolutional layers. For example, Nogueira et al. [13] applied several existing convolutional neural network (CNN) models to learn global features for ASR. Gong et al. [14] proposed a deep structural metric learning method to capture the structural information for ASR. Although these methods have yielded fruitful results for ASR, the local object-level features are ignored, which hinders the further improvements of ASR performance. To address this problem, extensive works focus on learning both global and local features for ASR. Specifically, Zheng et al. [15] proposed a deep scene representation to deal with the lack of geometric invariance of global CNNs. Sun et al. [16] proposed a gated bidirectional network to integrate hierarchical features and eliminate the interference information for ASR. Generally, these methods have reported promising results. However, the effective interactions among different scales and various spatial locations are not fully taken into account, which may limit the further improvement of the performance for ASR.

## 1.1 | Motivation

Practically, there exist two main challenges degenerating the performance of ASR, involving the large-scale variation and complex background, as shown in Figure 1. On the one hand, important ground objects are often with large-scale variation due to different spatial resolutions [17]. For instance, Figure 1a shows storage tanks (upper row) and aeroplanes (lower row) with various scales, which are important cues for ASR. On the other hand, the background of aerial scenes is quite complex because of the complicated spatial arrangement, for example,



**FIGURE 1** Large-scale variation and complex background are two major challenges that often degenerate the Aerial scene recognition (ASR) performance. (a) Large-scale variation. (b) Complex background.

multifarious ground objects (upper row) and dense buildings (lower row) in Figure 1b.

In the face of the above challenges, the intuitive method is to leverage multi-scale architecture learning discriminative features by fusing both global and local information. However, there remain two main issues to be addressed for existing multi-scale methods for ASR. First, most existing methods fuse global and local information by performing oversimplified concatenation or bilinear fusion operation. In fact, the human cerebral cortex enables multi-scale information interaction with a quite complex manner [18]. Hence, it is necessary to explore effective interactions among different scales and various spatial locations. Second, existing multi-scale methods for ASR do not take the consistent predictions between global and local features into account and therefore result in limited generalisations. Practically, the predictions between global and local features should be consistent [19]. For instance, the storage tanks in Figure 1a (upper row) can be distinguished by both global images and local objects simultaneously. The abovementioned two issues limit the further improvement of ASR accuracy and efficiency.

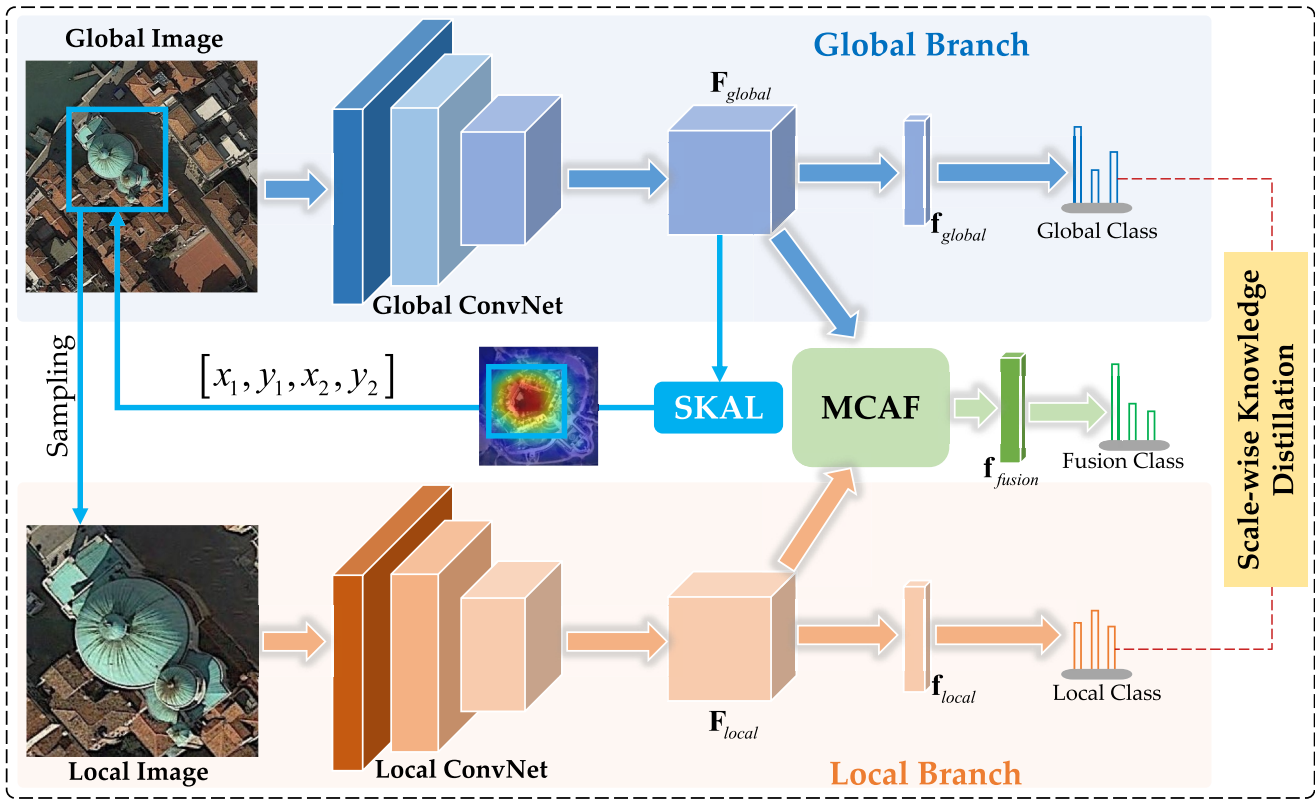
## 1.2 | Overview

To address the aforementioned issues, this study proposes a scale-wise interaction fusion and knowledge distillation (SIF-KD) network to learn robust and discriminative features with scale-invariance and background-independent information. As shown in Figure 2, the proposed SIF-KD network is composed of two main branches, for example, global branch (blue colour) and local branch (orange colour), which adopt the same network architecture but inputs of different scales. First, the global images are inputted into the global branch for learning global context features. Secondly, the local images are sampled by locating the most important cues in the global images with a structured key area localisation (SKAL) module [20]. Thirdly, the sampled local images are inputted into the local branch for learning local context features. Fourthly, the global and local context features are fed into the proposed multi-scale context attention fusion (MCAF) module to learn robust and discriminative representations with scale-invariance and background-independent information. Note that the scale-wise knowledge distillation (SKD) scheme is raised in the stage to boost the consistency on the predictive distributions between global and local context features. Finally, the fused robust and discriminative representations are employed for ASR.

## 1.3 | Contributions

To sum up, the main contributions of this study are threefold:

- To collaboratively fuse the context information between global and local features, a plug-and-play MCAF module is proposed by boosting the interactions among different scales and various spatial locations. Note that the proposed MCAF module can also be extended to other fusion tasks.



**FIGURE 2** The overall architecture of the proposed scale-wise interaction fusion and knowledge distillation (SIF-KD) network. The global and local branches are responsible for learning global and local context features, respectively. The structured key area localisation (SKAL) is leveraged to generate the local image as the input of the local branch. The multi-scale context attention fusion (MCAF) module is proposed to adaptively integrate the global and local context features. Finally, the proposed scale-wise knowledge distillation (SKD) scheme is applied to produce more consistent predictions by distilling the predictive distribution between global and local context features.

- To improve the generalisation of ASR, a SKD scheme is raised by distilling the predictive distribution between global and local context features, so as to produce more consistent predictions in a scale-wise manner.
- Extensive experimental results on three challenging ASR datasets demonstrate that the proposed SIF-KD network can achieve the state-of-the-art performance, and the raised MCAF module and SKD scheme are powerful to learn robust and discriminative multi-scale context features for ASR.

## 1.4 | Organisation

The remainder of this study is organised as follows. The related work about ASR is reviewed in Section 2. Section 3 describes the proposed SIF-KD network in details. The experimental results are shown in Section 4. Finally, Section 5 concludes the study.

## 2 | RELATED WORK

This study focusses on learning robust and discriminative feature representation for ASR. Consequently, the related work about ASR is reviewed according to the manner of

feature extraction, mainly including handcrafted feature-based methods and deep learning feature-based methods.

### 2.1 | Handcrafted feature-based ASR methods

The handcrafted feature-based methods extract image features by various human-engineering feature descriptors, involving local or global ones. The local structure descriptors are applied to model the local variations of structures in aerial scene images. For instance, Yang and Newsam [21] leveraged SIFT descriptors and Gabor texture features for ASR and compared their performance, finding SIFT features better. To depict the spatial arrangements in aerial scene images, scholars investigate different global distributions of certain spatial cues, such as colour and texture. In particular, Santos et al. [22] leveraged colour and texture image descriptors for ASR to comparatively evaluate their potential. To boost the recognition performance, various improved methods are developed by combining complementary features. For example, Avramović and Risojević [23] combined both Gist and SIFT descriptors for ASR. Chen et al. [24] integrated structure, texture and colour features with the bag-of-visual-words (BoVW) model [25] for ASR. Generally, the handcrafted feature-based methods perform well on

some aerial scene images with uniform structures, but it is limited or even impoverished to capture non-homogeneous and high-diversity spatial distributions due to poor ability to learn semantic information.

## 2.2 | Deep learning feature-based ASR methods

In recent years, deep learning, especially CNN, has broken the limits of traditional handcrafted feature-based methods and achieved great success in most fields, such as computer vision [26], speech recognition [27], medical image analysis [28], remote sensing [29] and so on. In the task of ASR, deep learning feature-based methods have become the mainstream due to the powerful ability to discover intricate structures and discriminative information hidden in aerial scene images [30].

In the beginning, most scholars concentrate on learning global features from aerial images for ASR. Specifically, Zhang et al. [31] introduced a saliency-guided sparse autoencoder to automatically learn the global features for ASR in an unsupervised manner. Xia et al. [32] reported different CNN-based model for ASR by using various global features. Zhang et al. [33] proposed an ASR architecture named CNN-CapsNet to capture the hierarchical structure and spatial information in images. To minimise the within-class diversity and enlarge the inter-class separation, Cheng et al. [34] developed a metric learning regularisation constraint on the global CNN features so as to improve the performance of ASR. Zhang et al. [35] presented a convolutional neural architecture search method to find the optimal network architecture and learn discriminative feature representations. To tackle the issue of difficulty in acquiring labelled data, Gu et al. [36] proposed a hierarchical prototype-based ensemble framework for ASR, which adopts a semi-supervised training manner. Considering that vanilla CNNs are with poor ability of geometric transformation, Liu et al. [37] proposed the contourlet CNNs to boost the ability of geometric transformations for better ASR. Zhao et al. [38] designed an enhanced attention module to learn more discriminative global features. To improve the defensive ability of ASR models against unknown attacks, Cheng et al. [39] developed a perturbation-seeking generative adversarial network for ASR. Although progress has been made by aforementioned global features-based ASR methods, it is not appropriate to utilise global features of a single scale since aerial scene images often have various scales of detailed textures. Specifically, some local object-level features should be integrated with global features to enhance the discrimination of features for ASR.

To solve the above problem, most of the subsequent methods adopt multi-scale architecture to synthetically leverage both global and local context information for ARS. In these works, some researchers adopted the decision-level fusion for integrating the global and local results. In particular, Xu et al. [40] designed a global–local dual-branch structure in which the decision-level fusion is adopted for fusing the global and local results. Wang et al. [20] presented joint global and local feature

representation to deal with the large-scale variation in aerial scene images. Note that a SKAL module is proposed in Ref. [20], which is applied in our study due to its efficiency in locating the most important area in aerial scene images. Actually, the decision-level fusion methods fail to consider the scale correlation, which may hinder the further improvement of ASR. To this end, the feature-level fusion manner is applied for integrating both global and local information. Specifically, Li et al. [41] raised an asymmetric filter bank, which can effectively capture key regions and retain global information simultaneously. To preserve more discriminative and local semantic features, Bi et al. [42] developed a densely connected CNN based on residual attention. Subsequently, Bi et al. [43] presented a multi-scale stacking attention pooling method for ASR, which can enhance the representation ability of local semantic information. Mei et al. [44] constructed a sparse representation framework to fuse multi-scale features. To improve feature discrimination capacity, Li et al. [45] developed a feature fusion framework for integrating multi-scale features in which a multi-scale Fisher kernel coding method is designed to extract middle-level feature representations. Lv et al. [46] proposed a multi-scale attentive region adaptive aggregation learning method, which can boost the semantic representation by combining cross-scale spatial semantic features.

Although existing feature-level fusion methods have achieved great classification performance, the effective interactions among different scales and various spatial locations are not fully explored, which may limit the further improvement of ASR accuracy. To this end, this study focuses on collaboratively fusing the context information between global and local features. In addition, existing multi-scale methods cannot fully model the problem of inconsistent predictions between global and local features, resulting in limited generalisations of ASR models. Therefore, this study aims to distill the predictive distribution between global and local context features.

## 3 | THE PROPOSED METHOD

As depicted in Figure 2, a SIF-KD network is proposed for the ASR task. In particular, a dual-branch architecture is adopted to learn both global and local context features with inputs of different scales. The MCAF module is designed to integrate both global and local context features, so as to enhance the interactions among different scales and various spatial locations. The SKD scheme is developed to produce more consistent predictions in a scale-wise manner. The details about each component and the optimisation strategy are elaborated in the following subsections.

### 3.1 | Global and local context features extraction

As shown in Figure 2, the dual-branch architecture consists of a global convolutional network for learning global scene

information, a SKAL module [20] for local object localization and a local convolutional network for capturing local discriminative information. The details about global branch, SKAL module and local branch are described as follows.

### 3.1.1 | Global branch and local branch

According to the previous works [47, 48], the global context feature plays an important role in the ASR task. The current mainstream deep learning methods extract the global context feature by stacking multiple neural network layers, which greatly improves the performance of ASR methods. However, the parameters to be learnt increase dramatically as the number of network layers increases, and huge amounts of data with manual annotations are required for learning parameters from scratch. Unfortunately, large-scale image dataset with manual annotation in the field of remote sensing is unavailable. To address the problem, the transfer learning strategy is adopted for extracting high-quality global context features. Specifically, this study exploits the convolutional layers in the original ResNet18 [49] with initial weights trained from ImageNet dataset as the main architecture of the global branch. Let  $\mathbf{I}_g$  represent the input global image, the global context feature map  $\mathbf{F}_{global} \in \mathbb{R}^{H \times W \times C}$  can be obtained by

$$\mathbf{F}_{global} = f_g(\mathbf{I}_g; \theta_g), \quad (1)$$

where  $f_g$  stands for the global convNet, and  $\theta_g$  means the to-be-learned parameter in  $f_g$ .

The local branch shares the same architecture with the global branch, and the local context feature map  $\mathbf{F}_{local} \in \mathbb{R}^{H \times W \times C}$  can be obtained by

$$\mathbf{F}_{local} = f_l(\mathbf{I}_l; \theta_l), \quad (2)$$

where  $\mathbf{I}_l$  represents the input local image,  $f_l$  stands for the local convnet and  $\theta_l$  means the to-be-learned parameter in  $f_l$ .

### 3.1.2 | Structured key area localization module

In order to sample the local input image  $\mathbf{I}_l$ , this study follows the previous work [20] and exploits the SKAL module to localise the local object. The SKAL module is composed of three successive steps, including energy aggregation, energy map structuration and greedy-like boundary search, for generating a bounding box and sampling the local input image. Firstly, the operation of energy aggregation is conducted to quantitatively describe the importance of the degree of each spatial element in the global context feature map  $\mathbf{F}_{global}$  as

$$\mathbf{M}_e = \sum_i^C \mathbf{F}_{global}(i, H, W), \quad (3)$$

where  $\mathbf{M}_e \in \mathbb{R}^{H \times W}$  represents the energy map,  $i$  indexes the channel and  $C$  is the channel number of  $\mathbf{F}_{global}$ . To remove the

interference of the negative element and achieve more accurate localisation, the elements of  $\mathbf{M}_e$  is scaled into the range of  $[0, 1]$  with min-max scaling and upsampled, obtaining a scaled energy map  $\hat{\mathbf{M}}_e$ .

Secondly, for the sake of optimisation, the operation of energy map structuration is performed to aggregate the 2-D  $\hat{\mathbf{M}}_e$  into 1-D structured energy vectors along the spatial height and width as

$$\begin{cases} V_b = \sum_{j=0}^W \hat{\mathbf{M}}_e(j, H) \\ V_w = \sum_{k=0}^H \hat{\mathbf{M}}_e(W, k), \end{cases} \quad (4)$$

where  $V_b$  and  $V_w$  denote the 1-D structured energy vectors along the spatial height and width, respectively.  $j$  and  $k$  index the spatial width and height of the scaled energy map  $\hat{\mathbf{M}}_e$ , respectively.

Thirdly, the greedy-like boundary search is completed to determine the bounding box of the local object. Let  $E_{[0:W]}$  stand for the energy sum of  $V_w$  and  $E_{[x_1:x_2]}$  means the energy from  $x_1$  to  $x_2$  along the spatial width, and they can be calculated by

$$\begin{cases} E_{[0:W]} = \sum_{s=0}^W V_w(s) \\ E_{[x_1:x_2]} = \sum_{s=x_1}^{x_2} V_w(s). \end{cases} \quad (5)$$

Here, the width boundary  $x_1$  and  $x_2$  can be solved by determining the smallest  $[x_1: x_2]$  area under the constraint  $E_{[x_1:x_2]}/E_{[0:W]} > \xi$ , where  $\xi$  represents a pre-defined energy threshold.

The height boundary  $y_1$  and  $y_2$  can be solved similarly. Once the entire boundary  $[x_1, x_2, y_1, y_2]$  is obtained, the input local image  $\mathbf{I}_l$  can be sampled from the input global image  $\mathbf{I}_g$ . Figure 3 shows the bounding box for sampled local images in the original images, which validates the efficiency of the SKAL module for local object localisation.

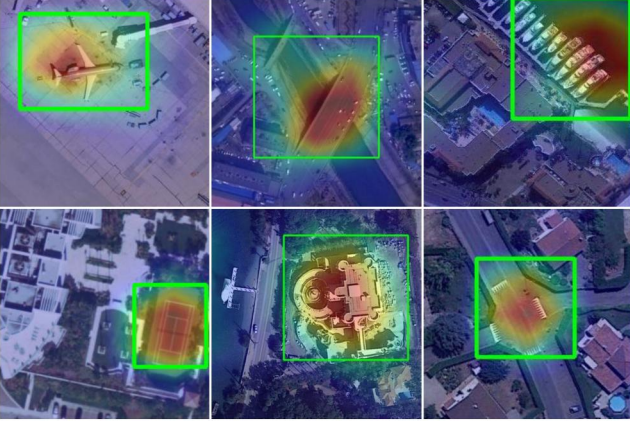
## 3.2 | Multi-scale context attention fusion

In order to boost the interactions among different scales and various spatial locations, a MCAF module is proposed for collaboratively aggregating the context information between global and local features. The workflow of the MCAF module is shown in Figure 4.

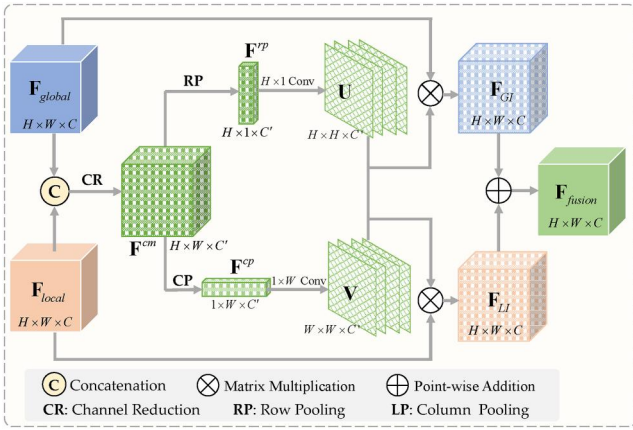
Firstly, the global context feature map  $\mathbf{F}_{global}$  and the local context feature map  $\mathbf{F}_{local}$  are concatenated and compressed via channel reduction by

$$\mathbf{F}^{cm} = \hat{h}([\mathbf{F}_{global}; \mathbf{F}_{local}]), \quad (6)$$

where  $\mathbf{F}^{cm} \in \mathbb{R}^{H \times W \times C'}$  is the compressed feature,  $\hat{h}(\cdot)$  means the  $1 \times 1$  convolutional layer followed by batch normalisation



**FIGURE 3** Visualisations of local object localisation by the structured key area localisation (SKAL) module.



**FIGURE 4** The workflow of the proposed multi-scale context attention fusion (MCAF) module.

and ReLU for channel reduction, and  $[\cdot; \cdot]$  denotes the concatenation operation of two features.

Secondly, row pooling and column pooling are conducted along the spatial width and height of  $\mathbf{F}^{cm}$ , respectively, obtaining row pooled feature  $\mathbf{F}^{rp} \in \mathbb{R}^{H \times 1 \times C'}$  and column pooled feature  $\mathbf{F}^{cp} \in \mathbb{R}^{1 \times W \times C'}$ . The operations are achieved by

$$\mathbf{F}_m^{rp} = \max \left\{ \mathbf{F}_{m,n}^{cm} \mid 1 \leq n \leq W \right\}, \quad (7)$$

$$\mathbf{F}_n^{cp} = \max \left\{ \mathbf{F}_{m,n}^{cm} \mid 1 \leq m \leq H \right\}, \quad (8)$$

Thirdly, based on the pooled features  $\mathbf{F}^{rp}$  and  $\mathbf{F}^{cp}$ , the transformation matrices  $\mathbf{U} \in \mathbb{R}^{H \times H \times C'}$  and  $\mathbf{V} \in \mathbb{R}^{W \times W \times C'}$  are estimated by

$$\mathbf{U} = \text{rearrange}(\mathbf{F}^{rp} * \mathbf{W}^{rp}), \quad (9)$$

$$\mathbf{V} = \text{rearrange}(\mathbf{F}^{cp} * \mathbf{W}^{cp}), \quad (10)$$

where  $*$  denotes the convolution operation,  $\mathbf{W}^{rp}$  represents the convolutional kernel with a size of  $H \times 1$ ,  $\mathbf{W}^{cp}$  represents the

convolutional kernel with a size of  $1 \times W$ , and rearrange means the reshape operation for adjusting the shape of generated matrix.

Fourthly, the global-aware interaction feature  $\mathbf{F}_{GI} \in \mathbb{R}^{H \times W \times C}$  and local-aware interaction feature  $\mathbf{F}_{LI} \in \mathbb{R}^{H \times W \times C}$  are acquired from the global context feature map  $\mathbf{F}_{global}$  and local context feature map  $\mathbf{F}_{local}$  respectively, using bilinear attentional transform by

$$\mathbf{F}_{GI} = \mathbf{U} \cdot \mathbf{F}_{global} \cdot \mathbf{V}, \quad (11)$$

$$\mathbf{F}_{LI} = \mathbf{U} \cdot \mathbf{F}_{local} \cdot \mathbf{V}, \quad (12)$$

where  $\cdot$  stands for the matrix product operation. For convenience, the matrix transposed in Equation (11) and Equation (12) are omitted.

Finally, the fused feature  $\mathbf{F}_{fuse} \in \mathbb{R}^{H \times W \times C}$  is learnt by combining the global-aware interaction feature  $\mathbf{F}_{GI}$  and local-aware interaction feature  $\mathbf{F}_{LI}$  as

$$\mathbf{F}_{fuse} = \mathbf{F}_{GI} \oplus \mathbf{F}_{LI}, \quad (13)$$

where  $\oplus$  denotes the point-wise addition operation of two matrices.

Based on the MCAF module, the interactions between the global context feature and local context feature can be effectively boosted. Note that the proposed MCAF module is plug and play and can be extended to various fusion tasks of multi-scale features. To validate the efficiency of the proposed MCAF module, the ablation study is conducted in Section 4.2.

### 3.3 | Scale-wise knowledge distillation and loss function

#### 3.3.1 | Scale-wise knowledge distillation

In order to mitigate the inconsistent predictions between global and local features, a SKD scheme is presented to produce more consistent predictions in a scale-wise manner. Specifically, let  $\mathbf{X}$  and  $\mathbf{y}$  denote the input image and the corresponding class label, respectively. The predictive distribution  $P(\mathbf{y}|\mathbf{X}; \mathbf{W}, T)$  can be expressed by

$$P(\mathbf{y}|\mathbf{I}; \mathbf{W}, T) = \frac{\exp(f_{\mathbf{y}}(\mathbf{I}; \mathbf{W})/T)}{\sum_{k=1}^K \exp(f_k(\mathbf{I}; \mathbf{W})/T)}, \quad (14)$$

where  $f_k$  represents the logit of the network for class  $k$  and  $\mathbf{W}$  is the parameter.  $T > 0$  indicates the temperature parameter of distillation. To match the predictive distributions between the local image and global of the same class, a scale-wise regularisation loss is proposed and written as

$$\mathcal{L}_{kd}(\mathbf{I}_g, \mathbf{I}_l; \mathbf{W}, T) = \text{KL}(P(\mathbf{y}|\mathbf{I}_l; \mathbf{W}_{lb}, T) \| P(\mathbf{y}|\mathbf{I}_g; \mathbf{W}_{gb}, T)), \quad (15)$$

where  $\text{KL}(\cdot\|\cdot)$  means the Kullback–Leibler (KL) divergence.  $\mathbf{W}_{lb}$  and  $\mathbf{W}_{gb}$  are the parameters in the local branch and global branch, respectively. In the implementation, the gradient is not propagated through  $\mathbf{W}_{lb}$  to avoid the problem of model collapse [50].

Under the constraint of the SKD loss, the inconsistent predictions between different scales can be effectively mitigated, and the ablation study is conducted in Section 4.2.

### 3.3.2 | Loss function

The SKD loss makes the model produce more consistent predictions between different scales during training. Except for the consistency of the predictive distribution, the accuracy of predictions also needs to be considered. As a result, the cross-entropy loss is applied on the outputs of the global branch, the local branch and the MCAF module, respectively. Formulaically,

$$\begin{aligned} \mathcal{L}_{ce}(\mathbf{I}_g, \mathbf{I}_l; \mathbf{W}_{gb}, \mathbf{W}_{lb}, \mathbf{W}_{g+l}) &= w_1 \text{CE}(y, \sigma(\mathbf{f}_{global})) \\ &+ w_2 \text{CE}(y, \sigma(\mathbf{f}_{local})) + w_3 \text{CE}(y, \sigma(\mathbf{f}_{fusion})) \\ &= w_1 \text{CE}(y, \sigma(f_{gb}(\mathbf{I}_g; \mathbf{W}_{gb}))) + w_2 \text{CE}(y, \sigma(f_{lb}(\mathbf{I}_l; \mathbf{W}_{lb}))) \\ &+ w_3 \text{CE}(y, \sigma(f_{g+l}(\mathbf{F}_{global}, \mathbf{I}_{local}; \mathbf{W}_{g+l}))), \end{aligned} \quad (16)$$

where  $\text{CE}(\cdot, \cdot)$  represents the cross-entropy loss for classification problems.  $\mathbf{f}_{global}$ ,  $\mathbf{f}_{local}$  and  $\mathbf{f}_{fusion}$  are the outputs of the global branch, the local branch and the MCAF module, respectively.  $\sigma$  means the softmax activation function.  $f_{gb}$ ,  $f_{lb}$ , and  $f_{g+l}$  indicates the global branch, local branch and MCAF module, respectively.  $\mathbf{W}_{gb}$ ,  $\mathbf{W}_{lb}$  and  $\mathbf{W}_{g+l}$  are the corresponding parameters.  $w_1$ ,  $w_2$  and  $w_3$  denote the trade-off coefficients.

With the above definitions, the final loss function of the proposed SIF-KD network is obtained as

$$\begin{aligned} \mathcal{L}_{final}(\mathbf{I}_g, \mathbf{I}_l; \mathbf{W}_{gb}, \mathbf{W}_{lb}, \mathbf{W}_{g+l}, T) &= w_1 \text{CE}(y, \sigma(f_{gb}(\mathbf{I}_g; \mathbf{W}_{gb}))) + w_2 \text{CE}(y, \sigma(f_{lb}(\mathbf{I}_l; \mathbf{W}_{lb}))) \\ &+ w_3 \text{CE}(y, \sigma(f_{g+l}(\mathbf{F}_{global}, \mathbf{I}_{local}; \mathbf{W}_{g+l}))) \\ &+ w_4 \text{KL}(P(y|\mathbf{I}_l; \mathbf{W}_{lb}, T) \| P(y|\mathbf{I}_g; \mathbf{W}_{gb}, T)), \end{aligned} \quad (17)$$

where  $w_4$  means the trade-off coefficient. Considering that the expensiveness to tune these trade-off coefficients, the strategy of multi-task learning with homoscedastic uncertainty [51] is adopted to learn optimal trade-off coefficients. Details of the proposed SIF-KD network are shown in Algorithm 1.

---

#### Algorithm 1 The proposed SIF-KD network

---

##### Input:

Training images  $\mathbf{I}$  and their corresponding labels  $y$ ;  
 Testing images  $\mathbf{I}_{test}$ ;  
 The pre-defined energy threshold  $\xi$

##### Output:

Testing aerial scene classes  $Y_{test}$ ;  
 All parameters  $\mathbf{W}_{gb}$  in the global branch;  
 All parameters  $\mathbf{W}_{lb}$  in the local branch;  
 All parameters  $\mathbf{W}_{g+l}$  in the MCAF module.

##### Initialisation:

The weights of the global convnet  $\theta_g$  and the local convnet  $\theta_l$  are initialised from the original ResNet18, and the remaining weights are randomly initialised.

##### Repeat:

- 1: Learn the global context feature  $\mathbf{F}_{global}$  based on Equation (1);
- 2: Generate the input local image  $\mathbf{I}_l$  based on Section 2;
- 3: Obtain the local context feature  $\mathbf{F}_{local}$  based on Equation (2);
- 4: Combine the global context feature  $\mathbf{F}_{global}$  and the local context feature  $\mathbf{F}_{local}$  to learn the fused feature  $\mathbf{F}_{fuse}$  based on Section 3.2;
- 5: Calculate the final loss  $\mathcal{L}_{final}$  according to Equation (17);
- 6: Update the parameters  $\mathbf{W}_{gb}$ ,  $\mathbf{W}_{lb}$  and  $\mathbf{W}_{g+l}$  by utilising the Adam optimiser.

**Until:** A fixed number of iterations.

- 7: Predict the testing aerial scene classes  $Y_{test}$  from the output of the MCAF module.

**Return:**  $Y_{test}$ ,  $\mathbf{W}_{gb}$ ,  $\mathbf{W}_{lb}$ ,  $\mathbf{W}_{g+l}$ .

---

## 4 | EXPERIMENT AND RESULTS

The proposed SIF-KD network is evaluated on three challenging ASR datasets, including UCM [52], AID [32] and NWPU-RESISC45 [30]. The ablation study is conducted to analyse the effectiveness of each component of the proposed method. In addition, several state of the arts are selected for comparison to demonstrate the superiority of the proposed SIF-KD network.

### 4.1 | Experimental setup

#### 4.1.1 | Datasets

This study adopts three widely used ASR datasets for verifying the performance of the proposed SIF-KD network. (1) The UCM dataset [52] is a small-scale ASR dataset, which includes 21 types of scene classes. Each scene class contains 100 images with the pixel size of  $256 \times 256$ . (2) The AID dataset [32] is a medium-scale ASR dataset, which consists of 30 categories of scene images with  $600 \times 600$  pixels. Each category includes 200–300 samples. (3) The NWPU-RESISC45 dataset [30] is a large-scale ASR dataset, which contains 45 types of scene classes. Each scene class contains 700 images with the pixel



size of  $256 \times 256$ . Note that the NWPU-RESISC45 dataset is more challenging due to the higher inter-class similarity.

#### 4.1.2 | Implementation details

Following the previous work [53], the training ratio is set to 80% for the UCM dataset, 20% and 50% for the AID dataset, 10% and 20% for the NWPU-RESISC45 dataset. To avoid the problem of overfitting, the data enhancement strategy is utilised, including random rotation, random flipping and so on. The input images are all resized to  $224 \times 224$ .

The proposed SIF-KD network is optimised using the Adam optimiser, and the learning rate is initialised to  $1e-4$ . The cosine decay with warmup strategy is adopted. The batch size is set to 64, and the training epoch is set to 100. The pre-defined energy threshold  $\xi$  in the SKAL module is set as 0.8. The learnt values of  $[w_1, w_2, w_3, w_4]$  on the UCM, AID and NWPU-RESISC45 datasets are  $[0.327, 0.397, 0.298, 0.959]$ ,  $[0.353, 0.470, 0.294, 1.119]$  and  $[0.373, 0.559, 0.298, 1.123]$ , respectively. The experiments are conducted with a 24 GB NVIDIA GeForce RTX 3090 GPU.

#### 4.1.3 | Evaluation metrics

Two typical evaluation metrics are utilised for quantitatively evaluating the experimental results. (1) Overall Accuracy (OA): OA is calculated via dividing the number of correctly classified samples by the number of total testing samples. The value of OA indicates the overall performance of classification models. (2) Confusion Matrix (CM): CM is a 2-D informative table in which each row represents the predicted class and each column indicates ground-truth class. Based on CM, it is easy for researchers to analyse the inter-class classification errors. In this study, the value of OA is reported by averaging three trials with standard deviation to eliminate the effects of random sampling.

## 4.2 | Ablation studies

In this subsection, three different variations except for the proposed SIF-KD network are performed for examining: (1) the effectiveness of the proposed MCAF module and (2) the importance of the raised SKD scheme. The detailed implementations of ablation studies are as follows.

First, the single-branch baseline without SIF-KD is implemented (Variation A). Second, the global and local context feature maps are combined with the proposed MCAF module, while the SKD scheme is not applied for distilling the predictive distribution (Variation B). Third, the SKD scheme is conducted between the predictive distributions, while the global and local context feature maps are aggregated in the concatenation manner instead of the proposed MCAF module (Variation C). Finally, the full version of the proposed SIF-KD network is implemented. The experiments are conducted on UCM, AID and NWPU-RESISC45 datasets, respectively, using 80%, 50% and 20% of them for training and the rest for

testing. Table 1 reports the experimental results, and two main observations can be made.

(1) The proposed MCAF module plays a significant role in boosting the feature representation ability for ASR. From the comparison in Table 1, it can be intuitively found that the recognition performance drops dramatically when the MCAF module is removed. The OA score drops from 99.68% to 98.34%, from 98.74% to 96.81% and from 95.47% to 93.71% on the UCM, AID and NWPU-RESISC45 datasets, respectively. This is because the proposed MCAF module can learn robust and discriminative representations with scale-invariance and background-independent information by boosting the interactions among different scales and various spatial locations. In addition, the comparison results between the single-branch baseline (Variation A) and the proposed SIF-KD network further demonstrate the effectiveness of the proposed MCAF module in aggregating global and local information for ASR.

(2) The raised SKD scheme can effectively integrate the predictive distributions between global and local context features. According to Table 1, the results of Variation B and the proposed SIF-KD networks differ substantially. The OA score improves from 99.68% to 98.62%, from 98.74% to 97.17% and from 95.47% to 94.26% on the UCM, AID and NWPU-RESISC45 datasets, respectively. This is mainly because the predictive distributions between the global and local context features are inconsistent sometimes when the SKD scheme is omitted, which results in the limited generalisations of ASR models. In contrast, the proposed SIF-KD network can achieve this effectually, which demonstrates that the proposed SKD scheme is able to produce more consistent predictions.

## 4.3 | Comparison with state of the arts

To demonstrate the advancement of the proposed SIF-KD network, 13 state-of-the-art networks are applied for

**TABLE 1** Evaluation results for ablation experiments.

Datasets	Networks	MCAF	SKD	OA (%)
UCM	Variation A			96.39 ± 0.20
	Variation B	✓		98.62 ± 0.24
	Variation C		✓	98.34 ± 0.26
	<b>SIF-KD</b>	✓	✓	<b>99.68 ± 0.17</b>
AID	Variation A			95.45 ± 0.16
	Variation B	✓		97.17 ± 0.14
	Variation C		✓	96.81 ± 0.16
	<b>SIF-KD</b>	✓	✓	<b>98.74 ± 0.13</b>
NWPU-RESISC45	Variation A			91.12 ± 0.15
	Variation B	✓		94.26 ± 0.11
	Variation C		✓	93.71 ± 0.19
	<b>SIF-KD</b>	✓	✓	<b>95.47 ± 0.10</b>

Note: The bold values indicate and highlight the results of our method.

comparison, including six single-branch and seven multi-branch networks. (1) D-CNN [34], SCCov [54], T-CNN [55], LGRIN [56], SCViT [53] and ET-GSNet [57] are single-branch networks. Specifically, the D-CNN [34] network leverages the metric learning scheme to learn discriminative features. SCCov [54] network adopts the skip-connection and covariance pooling strategy to tackle the problem of large-scale variance in aerial scene images. The T-CNN [55] network is based on meta networks for transferring knowledge from heterogeneous models. LGRIN [56] exploits the spatial priors in aerial scene images to construct a lightweight and robust model for ASR. The SCViT [53] network proposes a spatial-channel feature preserving vision transformer model for ASR. ET-GSNet [57] employs a vision transformer as a teacher to guide small networks for ASR. (2) MG-CAP [1], KFB [41], CNN-MS2AP [43], C-CNN [37], ACR-MLFF [58], MF<sup>2</sup>CNet [4] and SKAL-CNN [20] adopt multi-branch networks. MG-CAP [1] network introduces a multi-granularity canonical appearance pooling strategy for capturing the latent ontological structure of aerial scene images. The KFB [41] network presents a key filter bank without the attention mechanism for capturing the features from key regions in aerial scene images. The CNN-MS2AP [43] network proposes a multi-scale stacking attention pooling scheme for ASR. The C-CNN [37] network combines the contourlet transform with CNN to learn abundant information for ASR. The ACR-MLFF [58] network adopts the multilevel feature fusion network and adaptive channel dimensionality reduction mechanism for ASR. MF<sup>2</sup>CNet [4] proposes a multi-scale feature fusion covariance network to learn multi-scale and multi-frequency features to classify aerial scene images. The SKAL-CNN [20] network designs a SKAL module to locate the most important local area in aerial scene images.

First, the comparison experiments are conducted on the UCM dataset, and Table 2 reports the experimental results. It can be seen that the performance of the proposed SIF-KD network is better than all comparison single-branch networks. This is because the local object-level features and global features are integrated to enhance the discrimination of features for ASR. In addition, the OA score of the proposed SIF-KD network is superior to the ones of almost multi-branch networks. The OA score of KFB is higher than the one of SIF-KD network by about 0.20% (OA = 99.68%). In fact, the recognition accuracy of the proposed SIF-KD network is only slightly lower than KFB for the UCM dataset, but has better recognition performance on more challenging AID and NWPU-RESISC45 datasets, as shown in Tables 3 and 4. It demonstrates that the proposed SIF-KD network is more suitable for hard samples and has better generalisation ability.

Second, the comparison experiments are carried out on the AID dataset, and Table 3 reports the experimental results. It can be seen in Table 3 that the proposed SIF-KD network is better than all these comparison state-of-the-art networks. Especially, when the training ratio is set as 20%, the proposed SIF-KD network achieves 98.74% OA score, which surpasses KFB by 1.34%. It is notable that the proposed SIF-KD network is better than other methods with complex models

on the recognition performance, such as Transformer-based SCViT and ET-GSNet networks, even though merely small parameters network is employed. As a result, the proposed SIF-KD network has excellent ability of feature extraction since (1) it enhances the feature interaction among different scales and various spatial locations; (2) it can learn both global and local context features and boost their consistency on the predictive distributions.

**TABLE 2** Overall accuracy and standard deviation (%) on UCM Dataset.

Networks	Year	Published source	Training ratios 80%	
			20%	50%
D-CNN [34]	2018	IEEE TGRS	98.93 ± 0.10	
T-CNN [55]	2022	IEEE TGRS	99.33 ± 0.11	
LGRIN [56]	2022	IEEE TGRS	98.97 ± 0.31	
SCViT [53]	2022	IEEE TGRS	99.57 ± 0.31	
ET-GSNet [57]	2022	IEEE TGRS	99.29 ± 0.34	
MG-CAP [1]	2020	IEEE TIP	99.00 ± 0.10	
KFB [41]	2020	IEEE TGRS	99.88 ± 0.12	
CNN-MS2AP [43]	2021	Neurocomputing	99.01 ± 0.42	
C-CNN [37]	2021	IEEE TNNLS	98.97 ± 0.21	
MF <sup>2</sup> CNet [4]	2022	IEEE TGRS	99.52 ± 0.25	
SKAL-CNN [20]	2022	IEEE TNNLS	99.52 ± 0.24	
SIF-KD (proposed)	<b>2022</b>	–	<b>99.68 ± 0.17</b>	

Note: The bold values indicate and highlight the results of our method.

**TABLE 3** Overall accuracy and standard deviation (%) on AID Dataset.

Networks	Year	Published source	Training ratios	
			20%	50%
D-CNN [34]	2018	IEEE TGRS	90.82 ± 0.16	96.89 ± 0.10
SCCov [54]	2020	IEEE TNNLS	93.12 ± 0.25	96.10 ± 0.16
T-CNN [55]	2022	IEEE TGRS	94.55 ± 0.27	96.72 ± 0.23
LGRIN [56]	2022	IEEE TGRS	94.74 ± 0.23	97.65 ± 0.25
SCViT [53]	2022	IEEE TGRS	95.56 ± 0.17	96.98 ± 0.16
ET-GSNet [57]	2022	IEEE TGRS	95.58 ± 0.18	96.88 ± 0.19
MG-CAP [1]	2020	IEEE TIP	93.34 ± 0.18	96.12 ± 0.12
KFB [41]	2020	IEEE TGRS	95.50 ± 0.27	97.40 ± 0.10
CNN-MS2AP [43]	2021	Neurocomputing	92.19 ± 0.22	94.82 ± 0.20
C-CNN [37]	2021	IEEE TNNLS	–	97.36 ± 0.45
ACR-MLFF [58]	2022	IEEE GRSL	92.73 ± 0.12	95.06 ± 0.33
MF <sup>2</sup> CNet [4]	2022	IEEE TGRS	95.54 ± 0.17	97.02 ± 0.28
SKAL-CNN [20]	2022	IEEE TNNLS	93.89 ± 0.52	96.04 ± 0.68
SIF-KD (proposed)	<b>2022</b>	–	<b>96.53 ± 0.22</b>	<b>98.74 ± 0.13</b>

Note: The bold values indicate and highlight the results of our method.

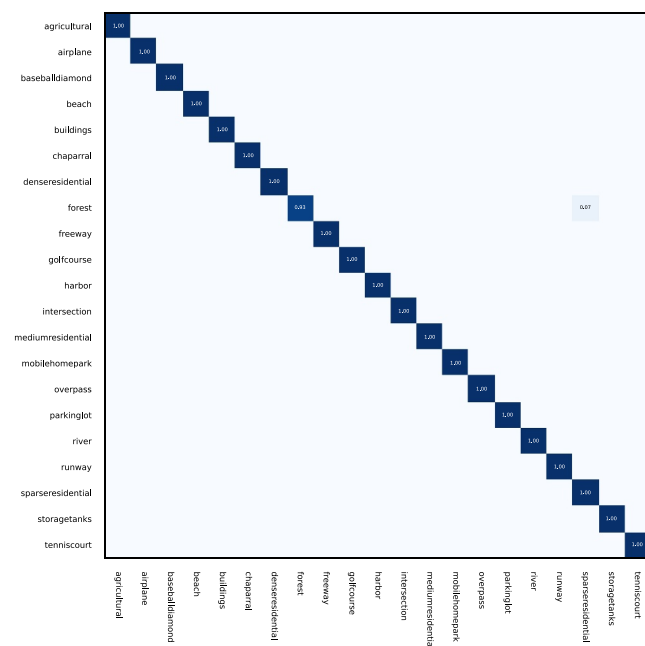
Finally, to comprehensively evaluate the superiority of the proposed SIF-KD network, the comparison experiments are also carried out on the most challenging NWPU-RESISC45 dataset, which has high intra-class diversity and inter-class similarity. Table 4 reports the experimental results. It can be seen that the proposed SIF-KD network makes the state-of-the-art performance with training ratios 10% and 20%.

Compared with KFBNet, the proposed SIF-KD network achieves 0.76% and 0.36% improvements under training ratios of 10% and 20%, respectively. This is because the proposed SIF-KD network not only captures global and local features but also enables multi-scale information interaction. In addition, it is worth mentioning that the proposed SIF-KD network is a follow-up method to SKAL-CNN, but our

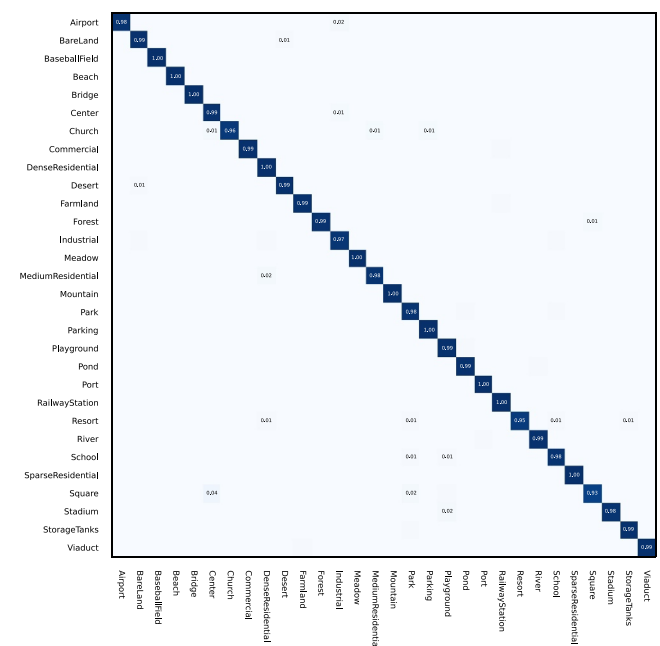
Networks	Year	Published source	Training ratios	
			10%	20%
D-CNN [34]	2018	IEEE TGRS	89.22 ± 0.50	91.89 ± 0.22
SCCov [54]	2020	IEEE TNNLS	89.30 ± 0.35	92.10 ± 0.25
T-CNN [55]	2022	IEEE TGRS	90.25 ± 0.14	93.05 ± 0.12
LGRIN [56]	2022	IEEE TGRS	91.91 ± 0.15	94.43 ± 0.16
SCViT [53]	2022	IEEE TGRS	92.72 ± 0.04	94.66 ± 0.10
ET-GSNet [57]	2022	IEEE TGRS	92.72 ± 0.28	94.50 ± 0.18
MG-CAP [1]	2020	IEEE TIP	90.83 ± 0.12	92.95 ± 0.13
KFB [41]	2020	IEEE TGRS	93.08 ± 0.14	95.11 ± 0.10
CNN-MS2AP [43]	2021	Neurocomputing	87.91 ± 0.19	90.98 ± 0.21
C-CNN [37]	2021	IEEE TNNLS	85.93 ± 0.51	89.57 ± 0.45
ACR-MLFF [58]	2022	IEEE GRSL	90.01 ± 0.33	92.45 ± 0.20
MF <sup>2</sup> CNet [4]	2022	IEEE TGRS	92.07 ± 0.22	93.85 ± 0.27
SKAL-CNN [20]	2022	IEEE TNNLS	90.04 ± 0.41	92.79 ± 0.11
SIF-KD (proposed)	<b>2022</b>	–	<b>93.84 ± 0.09</b>	<b>95.47 ± 0.10</b>

**TABLE 4** Overall accuracy and standard deviation (%) on NWPU-RESISC45 Dataset.

Note: The bold values indicate and highlight the results of our method.



**FIGURE 5** Confusion Matrix (CM) of the proposed scale-wise interaction fusion and knowledge distillation (SIF-KD) network on the UCM dataset under the training ratio of 80%.

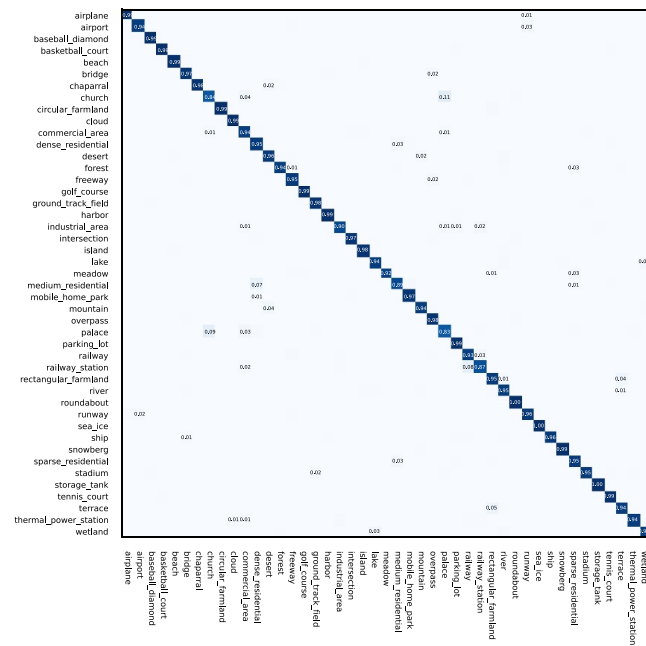


**FIGURE 6** Confusion Matrix (CM) of the proposed scale-wise interaction fusion and knowledge distillation (SIF-KD) network on the AID dataset under the training ratio of 50%.

network yields much better results than SKAL-CNN. This further demonstrates the superiority of the proposed SIF-KD network in terms of features fusion and improving inconsistent predictions.

### 4.4 | Confusion Matrix analysis and visualisation experiment

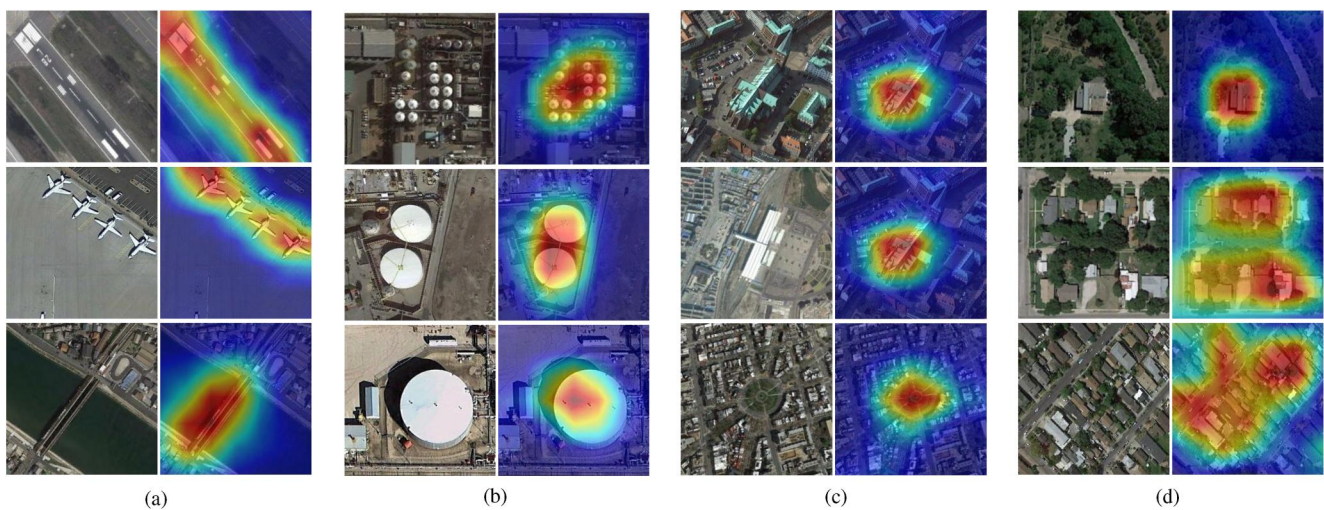
To further intuitively evaluate the confusion degree, the confusion matrices for the proposed SIF-KD network on three ASR



**FIGURE 7** Confusion Matrix (CM) of the proposed scale-wise interaction fusion and knowledge distillation (SIF-KD) network on the NWPU-RESISC45 dataset under the training ratio of 20%.

datasets are provided in Figures 5–7, respectively. As shown in Figure 5, accuracies of almost all categories on the UCM dataset reach 100%, except for the class of *forest*. The results demonstrate that the proposed SIF-KD network is significantly effective on small-scale ASR datasets. From Figure 6, the impressive performance is displayed on the AID dataset. Specifically, there are 26 categories with an accuracy rate above 98%, and the rest of the categories are all over 93%. The *Square* category is mistakenly identified as *Park* and *Centre* due to the similarity of these three classes in spatial and spectral characteristics. From the diagonal elements of the CM in Figure 7, the proposed SIF-KD network performs well in most categories, and the average OA reaches 91.2%. The most likely to be misclassified categories are the *church*, *medium\_residential*, *palace* and *railway\_station* because they contain overly confusing objects. On the whole, the performance of the proposed SIF-KD network on the three types of datasets are considerable.

To better explain the classification mechanism of the proposed SIF-KD network for ASR, the Grad-CAM++ [59] is applied to visualise the class-specific position of objects in aerial scene images. The fused feature  $F_{fuse}$  output by the proposed MCAF module is utilised for visualisation. Figure 8 exhibits four sets of samples, including simple samples, samples with large-scale variation, samples with complex background and ambiguous samples. From the visualisation results, it can be found that class-specific regions can be accurately located for various types of samples. Especially, class-specific regions in samples with large-scale variation and complex background are identified, which demonstrates that the proposed SIF-KD network can learn discriminative representations with scale-invariance and background-independent information. Furthermore, Figure 8d shows the visualisation results for *sparse\_residential*, *medium\_residential* and *dense\_residential*. It shows the proposed SIF-KD network can automatically locate the class-specific regions, such as the individual building for *sparse\_residential* samples, the



**FIGURE 8** Visualisation results with Grad-CAM++. (a) Simple samples. (b) Samples with large-scale variation. (c) Samples with complex background. (d) Ambiguous samples.

interface of buildings and trees for *medium\_residential* samples and arranged houses next to each other for *dense\_residential* samples. Overall, the proposed SIF-KD network performs well in locating class-specific regions for various scenes.

## 5 | CONCLUSIONS

In this study, a SIF-KD network is proposed for ASR. The proposed SIF-KD network consists of two important components, including the MCAF module and SKD scheme. The MCAF module can collaboratively aggregate global and local features by boosting the interactions among different scales and various spatial locations. The SKD scheme can mitigate the inconsistent predictions between global and local features by imposing a scale-wise regularisation constraint. Experimental results on the UCM, AID and NWPU-RESISC45 datasets demonstrate that the proposed MCAF module and SKD scheme are effective for learning robust and discriminative features with scale-invariance and background-independent information. In addition, the proposed SIF-KD network achieves superior performance compared with other state-of-the-art networks, manifesting its superiority.

Although the proposed SIF-KD network has achieved superior performance, it is not lightweight enough, which hinders its practical application. In the subsequent work, more lightweight and efficient ASR models will be explored to facilitate the deployment of ASR models on low-resource devices.

## ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 62201452, 2271296 and 62201453, in part by the Natural Science Basic Research Programme of Shaanxi under Grant 2022JQ-592, in part by the Special Construction Fund for Key Disciplines of Shaanxi Provincial Higher Education, in part by the Natural Science Basic Research Program of Shaanxi under Grant 2021JC-47, and in part by Scientific Research Program Funded by Shaanxi Provincial Education Department under Grant 22JK0568.

## CONFLICT OF INTEREST STATEMENT


None.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study.

## ORCID

Hailong Ning  <https://orcid.org/0000-0001-8375-1181>

Asoke K. Nandi  <https://orcid.org/0000-0001-6248-2875>

## REFERENCES

1. Wang, S., Guan, Y., Shao, L.: Multi-granularity canonical appearance pooling for remote sensing scene classification. *IEEE Trans. Image Process.* 29, 5396–5407 (2020). <https://doi.org/10.1109/tp.2020.2983560>
2. Li, B., et al.: Gated recurrent multiattention network for vhr remote sensing image classification. *IEEE Trans. Geosci. Rem. Sens.*, vol. 60, 1–13 (2021). <https://doi.org/10.1109/tgrs.2021.3093914>
3. Alem, A., Kumar, S.: Transfer learning models for land cover and land use classification in remote sensing image. *Appl. Artif. Intell.* 36(1), 2014192 (2022). <https://doi.org/10.1080/08839514.2021.2014192>
4. Bai, L., et al.: Remote sensing image scene classification using multiscale feature fusion covariance network with octave convolution. *IEEE Trans. Geosci. Rem. Sens.* 60, 1–14 (2022). <https://doi.org/10.1109/tgrs.2022.3160492>
5. Lu, X., Sun, H., Zheng, X.: A feature aggregation convolutional neural network for remote sensing scene classification. *IEEE Trans. Geosci. Rem. Sens.* 57(10), 7894–7906 (2019). <https://doi.org/10.1109/tgrs.2019.2917161>
6. Liu, Y., Zhong, Y., Qin, Q.: Scene classification based on multiscale convolutional neural network. *IEEE Trans. Geosci. Rem. Sens.* 56(12), 7109–7121 (2018). <https://doi.org/10.1109/tgrs.2018.2848473>
7. Zhu, Q., et al.: Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification. *IEEE Trans. Geosci. Rem. Sens.* 56(10), 6180–6195 (2018). <https://doi.org/10.1109/tgrs.2018.2833293>
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60(2), 91–110 (2004). <https://doi.org/10.1023/b:visi.0000029664.99615.94>
9. Li, H., et al.: Object-oriented classification of high-resolution remote sensing imagery based on an improved colour structure code and a support vector machine. *Int. J. Rem. Sens.* 31(6), 1453–1470 (2010). <https://doi.org/10.1080/01431160903475266>
10. Aptoula, E.: Remote sensing image retrieval with global morphological texture descriptors. *IEEE Trans. Geosci. Rem. Sens.* 52(5), 3023–3034 (2013). <https://doi.org/10.1109/tgrs.2013.2268736>
11. Zhang, Q., et al.: A robust deformed convolutional neural network (cnn) for image denoising. *CAAI Transactions on Intelligence Technology* (2022). <https://doi.org/10.1049/cit.2.12110>
12. Ahmad, F.: Deep image retrieval using artificial neural network interpolation and indexing based on similarity measurement. *CAAI Transactions on Intelligence Technology* 7(2), 200–218 (2022). <https://doi.org/10.1049/cit.2.12083>
13. Nogueira, K., Penatti, O.A., and Dos Santos, J.A.: Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recogn.* 61, 539–556 (2017). <https://doi.org/10.1016/j.patcog.2016.07.001>
14. Gong, Z., et al.: Diversity-promoting deep structural metric learning for remote sensing scene classification. *IEEE Trans. Geosci. Rem. Sens.* 56(1), 371–390 (2017). <https://doi.org/10.1109/tgrs.2017.2748120>
15. Zheng, X., Yuan, Y., Lu, X.: A deep scene representation for aerial scene classification. *IEEE Trans. Geosci. Rem. Sens.* 57(7), 4799–4809 (2019). <https://doi.org/10.1109/tgrs.2019.2893115>
16. Sun, H., et al.: Remote sensing scene classification by gated bidirectional network. *IEEE Trans. Geosci. Rem. Sens.* 58(1), 82–96 (2019). <https://doi.org/10.1109/tgrs.2019.2931801>
17. Bi, Q., et al.: Multiple instance dense connected convolution neural network for aerial image scene classification. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 2501–2505. IEEE (2019)
18. Keidel, J.L., et al.: Multiscale integration of contextual information during a naturalistic task. *Cerebr. Cortex* 28(10), 3531–3539 (2018). <https://doi.org/10.1093/cercor/bhx218>
19. Luo, X., et al.: Deep unsupervised hashing by global and local consistency. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2021)
20. Wang, Q., et al.: Looking closer at the scene: multiscale representation learning for remote sensing image scene classification. *IEEE Transact. Neural Networks Learn. Syst.* 33(4), 1414–1428 (2022). <https://doi.org/10.1109/tnnls.2020.3042276>
21. Yang, Y., Newsam, S.: Comparing sift descriptors and gabor texture features for classification of remote sensed imagery. In: 2008 15th IEEE International Conference on Image Processing, pp. 1852–1855 (2008)

22. Santos, J.A.D., Penatti, O.A.B., Torres, R.D.S.: Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification. In: VISAPP 2010 - Proceedings of the Fifth International Conference on Computer Vision Theory and Applications (2010)
23. Avramović, A., Risojević, V.: Block-based semantic classification of high-resolution multispectral aerial images. *Signal, Image and Video Processing* 10(1), 75–84 (2016). <https://doi.org/10.1007/s11760-014-0704-x>
24. Chen, L., et al.: Evaluation of local features for scene classification using vhr satellite images. In: 2011 Joint Urban Remote Sensing Event, pp. 385–388. IEEE (2011)
25. Kato, H., Harada, T.: Image reconstruction from bag-of-visual-words. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 955–962 (2014)
26. Zhao, B., et al.: Reconstructive sequence-graph network for video summarization. *IEEE Trans. Pattern Anal. Mach. Intell.* (2021)
27. Afouras, T., et al.: Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* (2018)
28. Chao, H., et al.: Deep learning predicts cardiovascular disease risks from lung cancer screening low dose computed tomography. *Nat. Commun.* 12(1), 1–10 (2021). <https://doi.org/10.1038/s41467-021-23235-4>
29. Zheng, X., et al.: Mutual attention inception network for remote sensing visual question answering. *IEEE Trans. Geosci. Rem. Sens.* 60, 1–14 (2021). <https://doi.org/10.1109/tgrs.2021.3079918>
30. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: benchmark and state of the art. *Proc. IEEE* 105(10), 1865–1883 (2017). <https://doi.org/10.1109/jproc.2017.2675998>
31. Zhang, F., Du, B., Zhang, L.: Saliency-guided unsupervised feature learning for scene classification. *IEEE Trans. Geosci. Rem. Sens.* 53(4), 2175–2184 (2014). <https://doi.org/10.1109/tgrs.2014.2357078>
32. Xia, G.-S., et al.: Aid: a benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Rem. Sens.* 55(7), 3965–3981 (2017). <https://doi.org/10.1109/tgrs.2017.2685945>
33. Zhang, W., Tang, P., Zhao, L.: Remote sensing image scene classification using cnn-capsnet. *Rem. Sens.* 11(5), 494 (2019). <https://doi.org/10.3390/rs11050494>
34. Cheng, G., et al.: When deep learning meets metric learning: remote sensing image scene classification via learning discriminative cnns. *IEEE Trans. Geosci. Rem. Sens.* 56(5), 2811–2821 (2018). <https://doi.org/10.1109/tgrs.2017.2783902>
35. Zhang, Z., Liu, S., Zhang, Y., and Chen, W.: Rs-darts: a convolutional neural architecture search for remote sensing image scene classification. *Rem. Sens.* 14(1), 141 (2022). <https://doi.org/10.3390/rs14010141>
36. Gu, X., et al.: A self-training hierarchical prototype-based ensemble framework for remote sensing scene classification. *Inf. Fusion* 80, 179–204 (2022). <https://doi.org/10.1016/j.inffus.2021.11.014>
37. Liu, M., et al.: C-cnn: contourlet convolutional neural networks. *IEEE Transact. Neural Networks Learn. Syst.* 32(6), 2636–2649 (2021). <https://doi.org/10.1109/tnnls.2020.3007412>
38. Zhao, Z., et al.: Remote sensing image scene classification based on an enhanced attention module. *Geosci. Rem. Sens. Lett. IEEE* 18(11), 1926–1930 (2021). <https://doi.org/10.1109/lgrs.2020.3011405>
39. Cheng, G., et al.: Perturbation-seeking generative adversarial networks: a defense framework for remote sensing image scene classification. *IEEE Trans. Geosci. Rem. Sens.* 60, 1–11 (2022). <https://doi.org/10.1109/tgrs.2021.3081421>
40. Xu, K., Huang, H., Deng, P.: Remote sensing image scene classification based on global-local dual-branch structure model. *Geosci. Rem. Sens. Lett. IEEE* 19, 1–5 (2022). <https://doi.org/10.1109/lgrs.2021.3075712>
41. Li, F., et al.: High-resolution remote sensing image scene classification via key filter bank based on convolutional neural network. *IEEE Trans. Geosci. Rem. Sens.* 58(11), 8077–8092 (2020). <https://doi.org/10.1109/tgrs.2020.2987060>
42. Bi, Q., et al.: Radc-net: a residual attention based convolution network for aerial scene classification. *Neurocomputing* 377, 345–359 (2020). <https://doi.org/10.1016/j.neucom.2019.11.068>
43. Bi, Q., Zhang, H., Qin, K.: Multi-scale stacking attention pooling for remote sensing scene classification. *Neurocomputing* 436, 147–161 (2021). <https://doi.org/10.1016/j.neucom.2021.01.038>
44. Mei, S., et al.: Remote sensing scene classification using sparse representation-based framework with deep feature fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens.* 14, 35867–5878 (2021). <https://doi.org/10.1109/jstars.2021.3084441>
45. Li, E., et al.: Integrating multilayer features of convolutional neural networks for remote sensing scene classification. *IEEE Trans. Geosci. Rem. Sens.* 55(10), 5653–5665 (2017). <https://doi.org/10.1109/tgrs.2017.2711275>
46. Lv, G., et al.: Multi-scale attentive region adaptive aggregation learning for remote sensing scene classification. *Int. J. Rem. Sens.* 42(20), 7742–7776 (2021). <https://doi.org/10.1080/01431161.2021.1963878>
47. Xu, K., et al.: Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing. *IEEE Transact. Neural Networks Learn. Syst.* 33(10), 5751–5765 (2021). <https://doi.org/10.1109/tnnls.2021.3071369>
48. Zhao, Q., et al.: Mgm1: multigranularity multilevel feature ensemble network for remote sensing scene classification. *IEEE Transact. Neural Networks Learn. Syst.* (2021). <https://doi.org/10.1109/tnnls.2021.3106391>
49. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
50. Miyato, T., et al.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 41(8), 1979–1993 (2018). <https://doi.org/10.1109/tpami.2018.2858821>
51. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7482–7491 (2018)
52. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 270–279 (2010)
53. Lv, P., et al.: Scvit: a spatial-channel feature preserving vision transformer for remote sensing image scene classification. *IEEE Trans. Geosci. Rem. Sens.* 60, 1–12 (2022). <https://doi.org/10.1109/tgrs.2022.3157671>
54. He, N., et al.: Skip-connected covariance network for remote sensing scene classification. *IEEE Transact. Neural Networks Learn. Syst.* 31(5), 1461–1474 (2020). <https://doi.org/10.1109/tnnls.2019.2920374>
55. Wang, W., Chen, Y., Ghamisi, P.: Transferring cnn with adaptive learning for remote sensing scene classification. *IEEE Trans. Geosci. Rem. Sens.* 60, 1–18 (2022). <https://doi.org/10.1109/tgrs.2022.3190934>
56. Xu, C., Zhu, G., Shu, J.: A lightweight and robust lie group-convolutional neural networks joint representation for remote sensing scene classification. *IEEE Trans. Geosci. Rem. Sens.* 60, 1–15 (2022). <https://doi.org/10.1109/tgrs.2020.3048024>
57. Xu, K., Deng, P., Huang, H.: Vision transformer: an excellent teacher for guiding small networks in remote sensing image scene classification. *IEEE Trans. Geosci. Rem. Sens.* 60, 1–15 (2022). <https://doi.org/10.1109/tgrs.2022.3152566>
58. Wang, G., et al.: Mfst: a multi-level fusion network for remote sensing scene classification. *Geosci. Rem. Sens. Lett. IEEE* 19, 1–5 (2022). <https://doi.org/10.1109/lgrs.2022.3205417>
59. Chattopadhyay, A., et al.: Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847 (2018)

**How to cite this article:** Ning, H., et al.: Scale-wise interaction fusion and knowledge distillation network for aerial scene recognition. *CAAII Trans. Intell. Technol.* 1–13 (2023). <https://doi.org/10.1049/cit2.12208>