An Inverted Residual Cross Head Knowledge Distillation Network for Remote Sensing Scene Image Classification

Cuiping Shi, *Member, IEEE*, Mengxiang Ding, and Liguo Wang, *Member, IEEE*

***Abstract*—In recent years, remote sensing scene classification (RSSC) has achieved notable advancements. Remote sensing scene images exhibit greater complexity in terms of land features, with large intra class differences and high inter class similarity, posing challenges in effectively extracting discriminative features. Convolutional neural networks are extensively used in RSSC tasks, where convolution focuses more on the high-frequency components of the image. Unlike convolution, transformer can model long-distance feature dependencies and mine contextual information in remote sensing scene images. Moreover, in traditional knowledge distillation methods, conflicts sometimes arise between teacher predictions and true labels, which hinder the training of the model. To enable the model to obtain sufficient supervision information while avoiding information conflicts, in this paper, an inverted residual cross head knowledge distillation network (IRCHKD) is proposed. First, an inverted residual attention module is designed to extract and leverage both local and global information effectively, enhancing the model's ability to capture complex details while retaining contextual information. Then, a multiscale spatial attention module is constructed to further extract global and local features of the image through multiple dilated convolutions, using spatial attention to weight important features in each dilated convolution branch. Finally, a cross head knowledge distillation structure is carefully designed to avoid conflicts between real labels and teacher predictions. The experimental results indicate that the proposed IRCHKD outperforms than some state-of-the-art RSSC approaches with a large margin in lower computational complexity.**

***Index Terms*—Remote sensing scene classification (RSSC), convolutional, transformer, knowledge distillation.**

I. Introduction

RSSC as a key application field of RS technology plays an irreplaceable role in interpreting earth surface information, urban planning, agricultural monitoring [1], [2], [3], [4], and other aspects. With the proposal of convolutional neural networks (CNNs) and transformers, the field of image classification is facing new opportunities and challenges.

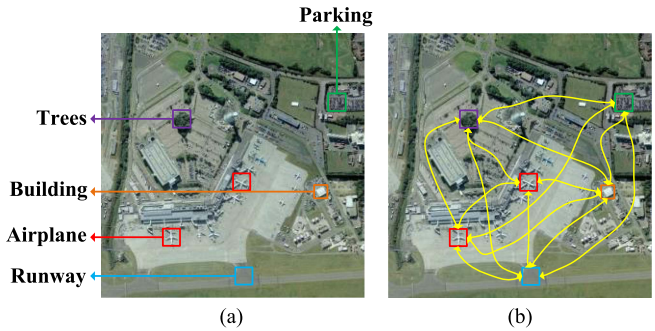

Fig. 1. A RS scene image of "Airport." (a) Local information and (b) extensive contextual details within the image.

Traditional RSSC method predominantly relied on manually crafted features, including but not limited to texture characteristics [5], [6], spectral characteristics [7], [8], color characteristics [9], [10], and shape characteristics [11], [12], coupled with machine learning classifiers such as support vector machines [13] and decision trees [14]. These methods often involve a large amount of domain expert knowledge and are often challenging to obtain rich feature information in complex RS scenes. Although this type of method solves the problem of RSSC partially, its performance is gradually limited as the volume of data grows and task complexity.

CNNs have attained notable success in the realm of image processing. Convolutional layers can effectively capture local features in images, while simultaneously reducing the parameter count and enhancing the model's computational efficiency through weight sharing. CNNs have exhibited outstanding performance in RS image processing [15], [16], [17].

However, traditional CNNs have certain limitations in processing global information and sequential data, ignoring long-distance contextual information. As depicted in Fig. 1(a), it is manifest that the scene comprises various land covers including "Airplane," "Runway," "Parking," and "Trees." If the model solely relies on local structural features, there is a likelihood that the "Airport" scene may be incorrectly classified as a different scene. Consequently, the anticipated classification model should consider not only the local land cover but also its contextual

Received 28 July 2024; revised 28 December 2024; accepted 15 January 2025. Date of publication 3 February 2025; date of current version 12 February 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42271409, in part by the Science and Technology Plan Project of Huzhou under Grant 2024GZ36, and in part by the Fundamental Research Funds in Heilongjiang Provincial Universities under Grant 145409340. *(Corresponding author: Cuiping Shi.)*

Cuiping Shi is with the College of Information Engineering, Huzhou University, Huzhou 313000, China (e-mail: shicuiping@qqhru.edu.cn).

Mengxiang Ding is with the College of Communication and Electronic Engineering, Qiqihar University, Qiqihar 161000, China (e-mail: 2021910321@qqhru.edu.cn).

Liguo Wang is with the College of Information and Communication Engineering, Dalian Nationalities University, Dalian 116000, China (e-mail: wangliguo@hrbeu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2025.3535437

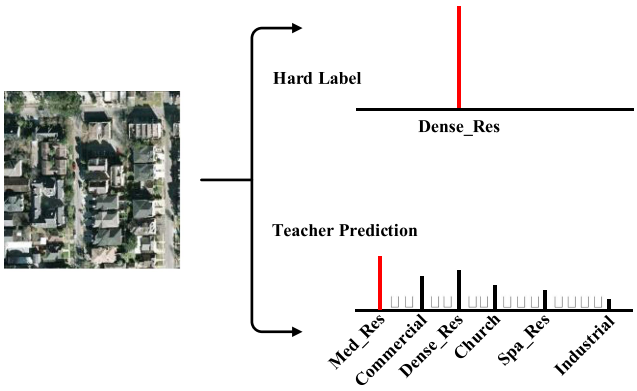

Fig. 2. Differences between hard labels and teacher predictions.

relationships, particularly those spanning longer distances. As shown in Fig. 1(b), when the relationship between "Airplane," "Runway," and "Building" is well explained, the scene can be correctly classified as "Airport." At this point, vision transformer [18] has attracted widespread attention as a powerful sequence modeling tool. The transformer has attained remarkable success in domains like natural language processing by employing the self-attention mechanism to achieve global modeling of sequences. The introduction of transformer in remote sensing image classification provides a new approach for dealing with global and long-distance dependency relationships.

By reason of the multiscale characteristics of RS scene images, dilated convolution [19] has emerged as an important variant of convolution. Dilated convolution introduces an adjustable dilation rate into the convolutional kernel, allowing the network to perceive contextual information more widely while maintaining local information. This provides a successful method for tackling information loss that traditional convolutions are prone to when processing multiscale features.

To fully leverage the strengths of transformer and convolution, researchers have begun to explore their fusion methods. The model that combines transformer and convolution not only effectively processes local and global information, but also has better generalization ability and adaptability. This article proposes an IRAM module that combines convolution and transformer to capture local and global information of remote sensing scene images, to better classify the scene. A multiscale parallel spatial attention module was also proposed for further extracting image features.

Distillation learning, a cutting-edge model compression technique, has been widely applied. In 2015, Hinton et al. [20] proposed the principle of knowledge distillation (KD). The KD method transfers the knowledge of a complex model to a relatively simple model, where the complex model is commonly referred to as the "teacher model" and the relatively simple model is as the "student model." However, sometimes there is a conflict between the real label and the distillation target. As shown in Fig. 2, the true label is "Dense_Res," while the prediction probability of "Med_Res" in the teacher's prediction output is the highest. The student's predictions mimic both the true labels and the teacher's predictions. This will affect the learning of student models to a certain extent. In order to avoid conflicting learning objectives in student networks, we propose a cross head knowledge distillation framework that enhances the model's supervision ability while avoiding the aforementioned problem of learning objective conflicts.

Overall, some existing methods for remote sensing scene image classification suffer from insufficient feature extraction and supervision. For the method proposed in this article, the supervision ability of the model is strengthened through cross head knowledge distillation structure. This article introduces an inverted residual cross head knowledge distillation neural network (IRCHKD) to efficiently extract image features, attain high classification accuracy, and address the limitations of conventional distillation methods.

The article makes the following contributions:

1) An inverted residual attention module (IRAM) is constructed, which first designs an inverted residual multi head self-attention structure carefully to establish long-range dependencies of images, and then devises a deep separable convolution with residual structure to further extract features and leverage the local inductive bias of the convolution.
2) A MSSA is designed, which further extracts global and local features through multiple serial and parallel dilated convolutions, and then weighs important features using spatial attention in each branch.
3) A CHKD structure is proposed for the first time, which input the features of the student classifier into the teacher classifier, and the resulting cross head prediction imitates the prediction of the teacher network. It can ensure the classifier of the student model no longer receives conflicting supervision signals from real labels and teacher predictions.
4) The introduced IRCHKD can effectively reduce computational complexity. A lot of experimental results and ablation experiments have demonstrated the prominent classification performance of the proposed IRCHKD method.

The rest of this article is organized as follows. Section II introduces the current development status of RSSC. Section III introduces the proposed method. Section IV presents the experimental results and discussion. Section V concludes this article.

II. Related Work

A. Methods Based on CNN for RSSC

CNNs have demonstrated excellent performance in feature extraction and are extensively utilized in RS scene image evaluation. For instance, Shi et al. [21] introduced a branch characteristic integration convolutional neural network, which merges characteristic data from two segments and employs depthwise separable convolution (DSC) to minimize model complexity. Zhao et al. [22] introduced a framework that integrates global texture characteristics, local structural characteristics, and spectral characteristics. Li et al. [23] integrated knowledge transfer and CNNs to generate robust visual features, which improved classification precision under data constraints. Singh et al. [24]

proposed a heterogeneous convolution method, which introduced two convolution kernels, $k \times k$ convolution filter for some channels, and 1×1 convolution filter for the remaining channels. Adjust the ratio between channels through the hyperparameter p. The objective of this technique is to boost performance of image processing by optimizing the convolutional structure. In the dynamic convolution proposed by Chen et al. [25], attention mechanism is used to dynamically aggregate multiple small-sized convolution kernels, improving characteristic depiction in a nonlinear manner. The self-calibration convolution method introduced by Liu et al. [26] dynamically establishes extensive spatial and channel dependencies tailored to individual spatial positions. Generate more discriminative features through self-calibration operations. From a frequency perspective, Chen et al. [27] introduced the concept of octave convolution. This method splits input characteristic into high-pass and low-pass components through hyperparameters α control the ratio of the two to reduce parameters and computational complexity, and effectively improve feature expression ability. When addressing redundant information, Han et al. [28] enhanced conventional convolution by introducing ghost convolution. This method extracts feature information via traditional convolution and generates redundant information through linear transformation, effectively decreasing the computational intricacy of the network. In conventional convolution, the convolutional parameters are common across all samples. Lu et al. [29] proposed an energy based remote sensing scene image classification method, which obtains the feature values of the weight tensor through singular value decomposition. The energy of eigenvalues can reflect redundant information. A filter pruning framework based on energy is proposed to reduce the complexity of the model by calculating the energy of each layer through the ratio of lower eigenvalues. Wang et al. [30] proposed a reverse multiscale feature fusion model for extracting features of objects of different scales in remote sensing scene images and enhancing the contextual understanding of the model. This model uses convolution to capture texture details and channel attention as a complement to features. Adopting a combination of local and global feature extraction in deeper stages of the network. A reverse cross scale interaction module was proposed to fuse features from different stages.

B. Methods Based on Attention for RSSC

In the realm of RSSC, CNN-based approaches have achieved significant outcomes. Nonetheless, when confronted with intricate RS scene images, the efficient extraction of crucial information and the capture of contextual dependencies remain pressing issues. To tackle these challenges, visual attention technology has been integrated into the domain of RS image processing.

Shi et al. [31] introduced a multiple branch features fusion technique leveraging attention mechanisms to comprehensively extract deep image features through collaborative multiconvolution. Tang et al. [32] suggested an ACNet tailored for RSSC. This network incorporates an attention consistency model intended to harmonize prominent areas for more precise extraction of discriminative characteristic. Chen et al. [33] devised a multiple branches local attention model and introduced a convolutional local attention module to efficiently derive weights of attention in channel and spatial dimensions. This method can automatically emphasize or suppress important or redundant information in remote sensing scenes.

Dosovitskiy et al. [18] suggested ViT, which converts natural images into a series of image blocks and constructs long-distance connections using self-attention. Subsequently, transformers were within the realm of natural image processing [34], [35], [36]. Based on ViT, Chen et al. [37] proposed multiple scales cross attention for image classification. In addition, a token merging unit based on cross-attention was designed for the final classification. Liu et al. [38] introduced a Swin transformer for diverse image processing tasks. Reduce the time cost caused by self-attention mechanism by dividing windows and calculating self-attention within the window. Meanwhile, increasing information exchange between windows with variable windows. Xia et al. [39] proposed a dual branch global local attention network to solve the problem of high inter class similarity and large intra class difference in remote sensing scene images. Designed a global and local attention module to focus on global and local features in images. At the end of the model, use a fusion module to combine the complete features. Chen et al. [40] proposed a transformer framework based on patch based hierarchical feature fusion to better utilize intermediate layer features. Zhao et al. [41] proposed local and long-distance joint learning for remote sensing scene image classification. First, a dual stream structure was designed to extract local and long-distance features. Second, a cross feature joint module is designed to improve the representation of fused feature maps. Finally, a joint loss is proposed to enhance the extraction of dual stream features and further fuse feature maps. Tang et al. [42] introduced a multiscale transformer to mine the context-related data of RS scene content. They used a cross-layer attention mechanism to fuse hierarchical characteristic representations together, and then used a classification score fusion module for final classification, modeling intrinsic relationships with lower time costs. Chen et al. [43] introduced a network that integrates convolution and Transformer to classify smoke scenes, combining local and global features to achieve high classification accuracy.

C. Knowledge Distillation

KD aims to guide the student models through the prediction of teacher network. During the training procedure, the student model will continuously optimize its own parameters.

To effectively facilitate knowledge transfer, Kullback-Leibler (KL) divergence [44] is used to assess the disparity between two distributions. Its calculation formula is expressed as follows:

$$\mathrm{KL}(P \parallel Q) = \sum\nolimits_i P(i) * \log \frac{P(i)}{Q(i)} \tag{1}$$

where i represents the input tensor, $P(i)$ denotes the predicted distribution by the teacher model, while $Q(i)$ signifies the predicted distribution by the student model.

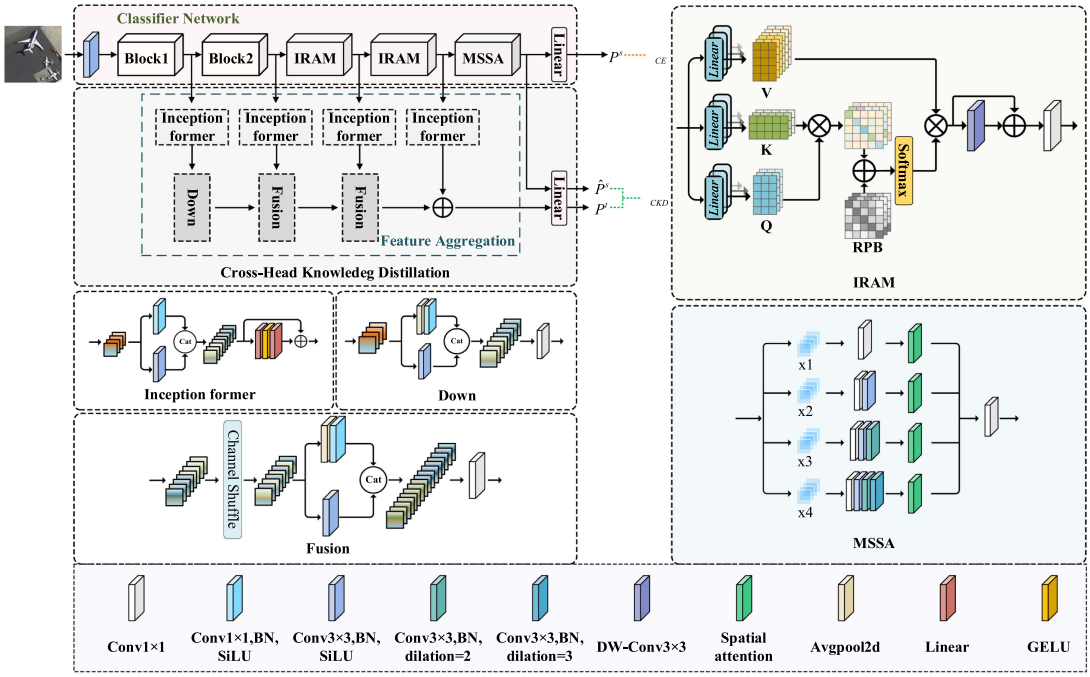

Fig. 3. Overall structure of the proposed IRCHKD.

Generally speaking, teacher models should be complex and accurate enough. Student models are usually lighter to deploy on resource constrained devices. As research deepens, researchers have designed many improved KD methods. For example, Romero et al. [45] aligned the outcome between the teacher and student model, introducing the concept of middle layer alignment. Park et al. [46] used relational modeling to optimize the effectiveness of KD, making knowledge transfer more accurate and effective. Heo et al. [47] designed a new distillation function that pays special attention to the characteristic before the ReLU function and maintains below zero results of these features during the distillation process. Ahn et al. [48] introduced a probabilistic information transfer method, which facilitates the transfer of information acquired by convolutional networks to fully connected layer, maximizing mutual information between two neural networks.

Considering the challenges of selecting appropriate teacher networks, some research has shifted towards self-distillation algorithms. Zhang et al. [49] introduced a self-distillation architecture that employs ResNet as the foundational model, which downsamples deep features through bottleneck structures and ultimately outputs soft labels through fully connected layers. Used to monitor the distribution of shallow networks, enabling them to learn deep features. Ji et al. [50] used horizontal convolution to improve feature maps and achieve self KD. Hu et al. [51] presented a self-learning feature refinement system, where the spread produced by low-level networks is guided by the spread from deep networks. They also introduced a gradient separation fusion module to ensure that ultimate class assignment gradient does not backpropagate to the main model. These studies provide new approaches and techniques for the realm of KD, aiding enhancement of learning efficiency and correctness of models.

Shi et al. [52] introduced a feature enhanced self-distillation CNNs, which first uses ResNet34 to extract multilevel features of the image, then uses a feature enhancement pyramid module to improve the characteristic, and finally uses feature distillation and logits distillation to supervise the model. Hu et al. [53] introduced a variational self-distillation approach, employing VKT to progressively distill both deep and shallow features, layer by layer, utilizing the predicted entanglement data vectors for classes as additional class data. Li et al. [54] devised a dual KD technique, incorporating attention mechanisms and spatial structures, along with two robust objective functions. Liu et al. [55] presented a cross-model KD strategy, leveraging pretrained RGB image models as teacher networks to assist in classifying multispectral images.

III. Methodology

The overall architecture of the IRCHKD approach is depicted in Fig. 3, which contains a classifier network in the pink region and a cross head knowledge distillation structure in the gray region. The classifier network mainly includes Block1 and Block2, IRAM and MSSA. Among them, Block1 and Block2 are used to extract shallow features, IRAM is used for the fusion of self-attention and convolution, and MSSA is used to capture features of multiscale targets. The CHKD structure mainly includes inception former, down, and fusion modules. These three modules further extract and fuse features, forming a feature aggregation branch. The classifier network is used for RS scene image classification, and the cross head knowledge distillation structure is utilized to provide supervised information for the classifier network.

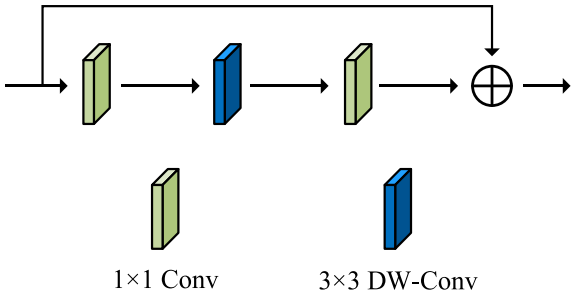

Fig. 4. Inverted residual structure used in Block1 and Block2 in the model.

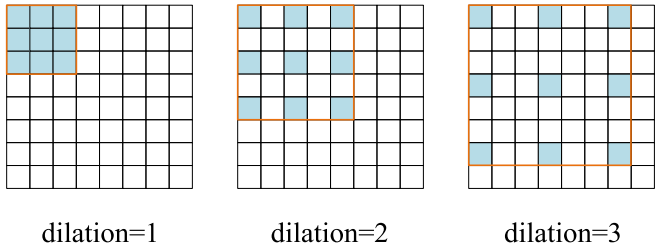

Fig. 5. Schematic diagram of dilated convolution, with dilation rates of 1, 2, and 3 from left to right. The orange box represents the size of the equivalent convolution kernel.

A. IRAM

Deep separable convolution consists of a dual-stage process; the initial stage is known as depthwise convolution, where an individual filter is applied to each input channel for the purpose of filtering. The second stage is 1×1 convolution, also known as point convolution, which is utilized to linearly combine features between channels. The DSC can directly replace the standard convolution, producing an effect equivalent to the standard convolution, while the computational cost is only $h_i \cdot w_i \cdot d_i(k^2 + d_j)$, effectively reducing the cost.

As illustrated in Fig. 4, the model's first two blocks employ an inverted residual structure. Initially, 1×1 pointwise convolution expands the count of feature channels, followed by DSC for further feature extraction. Finally, use residual connections. This procedure can be outlined as follows:

$$\mathrm{Conv} = \mathrm{Residual}(\mathrm{Conv}_{1\times1}(\mathrm{Conv}_{\mathrm{dsc}}(\mathrm{Conv}_{1\times1}(\mathrm{x})))) \quad (2)$$

where $\mathrm{Conv}_{1\times1}$ represents 1×1 convolution, $\mathrm{Conv}_{\mathrm{dsc}}$ represents depth-separable convolution, and $\mathrm{Residual}$ represents residual connection.

To leverage the benefits of self-attention for long-range modeling, we improve the first 1×1 convolution in Fig. 4. Specifically, as shown in the IRAM module in Fig. 3, the process can be expressed as follows:

$$F(\cdot) = \mathrm{Conv}(\mathrm{IRSA}(\cdot)). \quad (3)$$

Among them, IRSA represents the inverted residual self-attention, and Conv represents the combination of 3×3 DSC and 1×1 convolution. First, mapping the feature map to obtain three tensors Q, K, V. Double the channel count in V to provide richer feature representations. Using Q and K to obtain a self-attention map, and then combining relative position bias to obtain a global attention map, the attention map is multiplied by V to achieve attention interaction. Add deep convolution and residual structure after self-attention. The self-attention approach is capable of capturing global context, and the deep convolution layer can focus on local feature extraction to better capture details. Finally, a 1×1 convolution is utilized for interchannel information interaction and integrate features between different channels.

The self-attention method used in the IRSA structure is window based self-attention. Using this method can reduce the complexity of calculating self-attention. The computational complexity can be expressed using the following formula:

$$\Omega(\mathrm{SA}) = 4hwC^2 + 2(hw)^2C \quad (4)$$

$$\Omega(\mathrm{WSA}) = 4hwC^2 + 2M^2hwC. \quad (5)$$

Among them, $\Omega(\mathrm{SA})$ express the complexity of global self-attention, $\Omega(\mathrm{WSA})$ express the complexity of window based self-attention, M express the size of the divided window, C represents the count of channels, and hw express the size of the input feature map. The complexity of global self-attention is quadratic with the size of the feature map, and window self-attention is linear with a fixed value (set to 7). Window self-attention is the process of dividing a feature map into individual windows and then performing local self-attention within each window. From formulas (4) and (5), the first term of the formula is the same, with the difference in the second term. Assuming the size of the feature map is 64×64 and the window size is 7, formula (5) saves about 83 times the number of parameters compared to formula (4). It can be seen that the window self-attention can significantly reduce computational complexity.

B. MSSA

CNNs have seen widespread adoption in deep learning, largely owing to their powerful feature extraction abilities. However, the constraints of traditional convolution have led to the development of various alternative convolution techniques. One such technique is dilated convolution. Dilated convolution enlarges the receptive field (RF) of the convolutional kernel without adding extra parameters. This allows each convolution to encompass a broader range of information while maintaining the same output feature size. The process of dilation convolution is shown in Fig. 5. RF increases as the expansion rate increases.

When the expansion factor is set to 1, the RF of dilated convolution matches that of regular convolution. At an expansion factor of 3, the RF of a 3×3 dilated convolution kernel corresponds to that of a 7×7 standard convolution kernel. The calculation process of RF is

$$R_{i+1} = R_i + (r' - 1) S_i \quad (6)$$

$$\mathrm{S}_i = \prod_{i=1}^{i} \mathrm{stride}(i). \quad (7)$$

Among them, i express the count of dilated convolution layers used in total, r' represents the size of the equivalent convolution

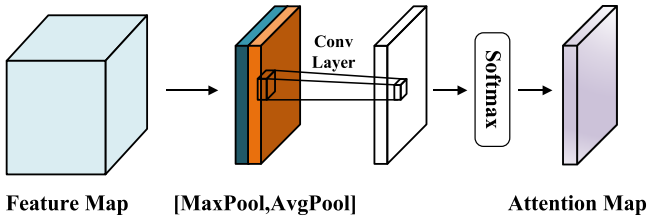

Fig. 6. Schematic diagram of spatial attention.

kernel of the $i+1$th layer, R_i represents the RF of the ith layer, R_{i+1} represents the RF of the $i+1$th layer, and S_i represents the cumulative product of all strides from the preceding i layers

The spatial attention approach can further improve the model's representation capacity. The spatial attention approach concentrates on key areas within an image, emphasize or suppress certain features by assigning different weights to different regions. Helps improve the model's perception of images, enabling it to better understand image content. Especially for objects or targets of different sizes, as well as recognition tasks in complex backgrounds. The spatial attention we use is shown in Fig. 6 This process can be expressed as follows:

$$M_s(F) = \mathrm{Softmax}\left(f^{3\times3}\left(\left[\mathrm{AvgPool}(F); \mathrm{MaxPool}(F)\right]\right)\right) \tag{8}$$

where F represents the feature map and $f^{3\times3}$ represents the 3×3 convolution.

The multiscale spatial attention module comprises four branches. Connect the features obtained from the four branches together. By using multiple parallel dilated convolution branches, features at different scales can be captured, enabling the model to have a more comprehensive understanding of image content. By using convolution kernels with different RF sizes, it is possible to effectively fuse contextual information from different ranges in an image.

C. CHKD

The cross head KD we propose is the gray block in Fig. 3. One can observe that the cross head KD structure primarily comprises two parts: the feature aggregation part and the cross head distillation part. The feature aggregation section mainly consists of three components: inception former, down, and fusion. In addition, the feature aggregation section is also known as the teacher network. In RS scenes, features at different levels contain different spatial structural information. Shallow neural networks typically acquire basic features of data, such as edges, textures, and other basic patterns. Deeper neural networks can learn more advanced and abstract features. For these different layers of features, we adopt a progressive aggregation strategy to gradually fuse the low-level feature maps with the deep feature maps.

Use four inception formers to further extract features from the feature maps of four distinct stages. The detailed framework of the Inception model is shown in Fig. 3. In the inception architecture, the feature map undergoes division into two branches. We use two parallel convolutions to imitate the self-attention structure in transformer, while preserving the MLP in transformer. This structure can extract spatial features and fuse feature maps in the channel dimension. Choosing convolution to replace self-attention mainly considers the computational efficiency.

The process of feature aggregation is as follows. First, the features obtained by the first Inception former are input into the downsampling structure. The downsampling structure is shown in the "Down" part in Fig. 3. The feature map is input into two parallel branches. Then, the feature map acquired by the down structure and the feature map acquired by the second Inception former are input into the first fusion structure at the same time. The fusion structure is shown in Fig. 3. In the fusion structure, the two input feature maps are first spliced along the channel dimension. Next, use the channel shuffle method to shuffle the channels, and then input the feature map into two branches. The concatenate feature map is input to the 1×1 convolution layer, BN and SiLU activation function in sequence, thus completing the second downsampling of the feature map. Then, input the feature map acquired from the first fusion and the feature map acquired from the third inception former into the second fusion structure simultaneously. Like the first fusion structure, after the second fusion structure, the feature map is downsampled again. it is worth mentioning that during the training process of the model, both the student network and the teacher network will be trained together. In the deployment phase after training, we will remove the cross head knowledge distillation part shaded in gray in Fig. 3 to reduce the model's parameter count and computational complexity.

As shown in Fig. 2, we observe that directly imitating the teacher's predictions will face the issue of goal conflict. To alleviate this challenge, we propose a CHKD structure. CHKD, like ordinary knowledge distillation, performs predictive simulations. Differently, CHKD passes the intermediate features of the student network to the classification head of the teacher network and generates cross head predictions for distillation. We denote the network outputs of teachers and students as P^t and P^s, respectively, and the cross head predictions as $\hat{P}^s$. We use the knowledge distillation loss between cross head prediction $\hat{P}^s$ and teacher prediction P^t as the goal of CHKD. The objective function for knowledge distillation can be expressed as follows:

$$\mathrm{L_{CKD}}(T) = D_{KL}\left(\mathrm{softmax}\left(\frac{\hat{P}^s}{T}\right)||\mathrm{softmax}\left(\frac{P^t}{T}\right)\right). \tag{9}$$

Among them, T express the temperature hyperparameter, and D_{KL} express the KL divergence between the two distributions of P^t and $\hat{P}^s$.

Apart from the distillation loss, the student network also employs cross-entropy loss to learn from real labels. The cross-entropy loss is expressed as follows:

$$L_{\mathrm{CE}}(x;\theta_s) = -\sum_{i=1}^{N} y_i \log(p_i^S(x;\theta_s)) \tag{10}$$

N denotes the training sample, x denotes the input tensor, θ_s denotes the parameters in the classifier network, y_i expresses

TABLE I
DATA INFORMATION FOR THREE DATASETS

Datasets	Count of images in each class	Count of scene categories	Total Count of images	Spatial resolution	Image size
UC-Merced	100	21	2100	0.3	256×256
AID	200–400	30	10000	0.5–0.8	600×600
NWPU-45	700	45	31500	0.2–30	256×256

Fig. 7. Examples of the UC merced dataset.

Fig. 8. Examples of the AID dataset.

the true labels, and p_i^S represents the outcome of the student network.

In total, the model's supervised loss comprises two parts. The overall supervised loss can be defined as follows:

$$\text{Loss} = L_{\text{CKD}}(T) + L_{\text{CE}}(x;\theta_s). \tag{11}$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

To validate the efficiency of our IRCHKD, we conducted multiple experiments on three common and challenging datasets. The three datasets are the UC Merced dataset [56], AID [57], and NWPU-RESISC45 dataset [58].

Fig. 9. Examples of the NWPU dataset.

A. Datasets

We introduced three datasets used in the experiment. The following samples from the UC Merced dataset are shown in Fig. 7, some samples from the AID dataset are shown in Fig. 8, and some samples from the NWPU dataset are shown in Fig. 9. Table I provides relevant information for three datasets.

B. Experimental Details

All experiments were conducted on a platform with GeForce RTX 3070 Ti, using the Pytorch framework. An Adam is utilized for model optimization, and the initial parameters of the network are randomly initialized data. Set the initial learning rate of the network to 0.0001 and train 150 epochs. Cosine decay is utilized for learning rate adjustment. The image size is adjusted to 224 × 224, and random horizontal and vertical flipping are used to augment the image. The main experiment in the article was repeated five times and the average results were reported. In the experiment, we utilized overall accuracy (OA) and confusion matrix (CM) to assess the efficacy of our proposed approach.

C. Experimental Results and Analysis

To assess the capability of the IRCHKD method, a set of experiments was carried out on three datasets. Table II lists the methods and publication years used for comparison in this article.

1) Experimental Results on the UC Merced Dataset: The experimental results can be seen in Table III. With a training ratio

TABLE II
DIFFERENT APPROACHES AND PUBLICATION YEARS IN THE EXPERIMENT

Methods	Year
RANet [59]	JSTARS2021
EFPN-DSE-TDFF [60]	TGRS2021
DFAGCN [61]	TNNLS2021
ViT [18]	ICLR2021
SCViT [62]	TGRS2022
EMTCAL [42]	TGRS2022
TECN [43]	TGRS2022
MLF2Net_SAGM [63]	RS2022
VSDNet-ResNet34 [53]	TGRS2022
CFDNN [64]	RS2022
MBFANet [65]	GRSL2023
SAGN [66]	TGRS2023
HFAM [67]	TGRS2024
CASD [68]	JSTARS2024
CGINet [69]	TGRS2024
GLR-CNN [70]	TGRS2024
LSMNet [71]	JSTARS2024
MMPC-Net [72]	GRSL2024
PFFGCN [73]	JSTARS2024
PSCLI-TF [74]	GRSL2024
SAF-Net [75]	GRSL2024
TAKD [76]	TCSVT2024

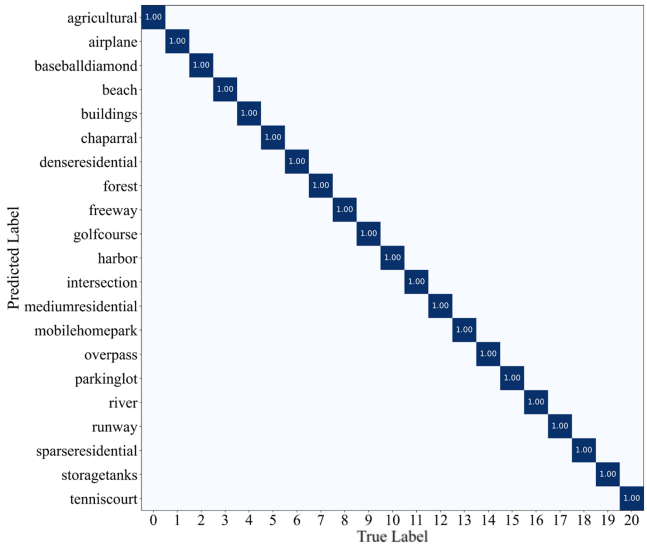

Fig. 10. CM at 80% training ratio on UC merced dataset.

TABLE III
COMPARISON BETWEEN OUR APPROACH AND THE APPROACHES PROPOSED IN RECENT YEARS ON UC MERCED DATASET

Methods	OA (50%)	OA (80%)
EFPN-DSE-TDFF [60]	96.19 ± 0.13	99.14 ± 0.22
RANet [59]	97.80 ± 0.19	99.27 ± 0.24
DFAGCN [61]	–	98.48 ± 0.42
ViT [18]	98.75 ± 0.39	99.29 ± 0.21
EMTCAL [42]	98.52 ± 0.18	99.28 ± 0.32
TECN [43]	–	99.52
SCViT [62]	98.90 ± 0.19	99.57 ± 0.31
MBFANet [65]	–	99.66 ± 0.19
SAGN [66]	–	99.82 ± 0.10
VSDNet-ResNet34 [53]	98.49 ± 0.18	99.67 ± 0.18
HFAM [67]	97.46 ± 0.31	98.67 ± 0.21
CASD-ViT [68]	99.07 ± 0.09	99.70 ± 0.11
CGINet [69]	–	99.84 ± 0.16
GLR-CNN [70]	97.52 ± 0.33	99.14 ± 0.26
LSMNet [71]	97.71 ± 0.14	99.29 ± 0.15
MMPC-Net [72]	98.98 ± 0.18	99.82 ± 0.12
PFFGCN [73]	99.04 ± 0.19	99.67 ± 0.21
PSCLI-TF [74]	98.70	99.62
SAF-Net [75]	–	99.32 ± 0.16
TAKD [76]	95.41	97.33
IRCHKD (ours)	99.33 ± 0.16	**99.90 ± 0.10**

TABLE IV
COMPARISON BETWEEN OUR PROPOSED METHOD AND RECENT METHODS ON THE AID DATASET

Methods	OA (20%)	OA (50%)
EFPN-DSE-TDFF [60]	94.02 ± 0.21	94.50 ± 0.30
DFAGCN [61]	–	94.88 ± 0.22
ViT [18]	94.90 ± 0.29	96.49 ± 0.18
EMTCAL [42]	94.33 ± 0.16	96.12 ± 0.28
MBFANet [65]	93.98 ± 0.15	96.93 ± 0.16
SAGN [66]	95.17 ± 0.12	96.77 ± 0.18
SCViT [62]	95.56 ± 0.17	96.98 ± 0.16
TECN [43]	95.45	97.40
VSDNet-ResNet34 [53]	96.00 ± 0.18	97.28 ± 0.14
HFAM [67]	93.53 ± 0.21	95.94 ± 0.10
CASD-ResNet50 [68]	95.72 ± 0.13	96.96 ± 0.16
CGINet [69]	95.35 ± 0.14	97.10 ± 0.24
GLR-CNN [70]	93.09 ± 0.16	95.64 ± 0.11
LSMNet [71]	94.31 ± 0.10	96.78 ± 0.16
MMPC-Net [72]	95.46 ± 0.21	99.52 ± 0.13
PFFGCN [73]	95.88 ± 0.23	97.40 ± 0.14
PSCLI-TF [74]	96.28	97.52
SAF-Net [75]	94.20 ± 0.12	96.72 ± 0.14
TAKD [76]	91.25	94.80
IRCHKD (ours)	96.30 ± 0.16	**97.84 ± 0.12**

of 80% , the OA of the proposed approach reached 99.90% , surpassing all comparison approaches. Specifically, our method's OA surpassed EMTCAL by 0.33% , TECN by 0.38% , and VSDNet-ResNet by 0.33% .

With a training ratio of 50% , our suggested approach attains a classification accuracy of 99.33% , outperforming all comparative approaches. In particular, our method's OA surpasses that of EMTCAL by 0.66% , ViT by 0.58% , SCViT by 0.43% , TAKD by 3.92% , and VSDNet-ResNet using the distillation method by 0.84% .

The CM diagram obtained with an 80% training ratio is shown in Fig. 10. From the figure, it is obvious that each category is well classified. Our proposed method achieved outstanding classification performance, effectively distinguishing different categories.

2) Classification Results on AID: On AID, the experimental results are shown in Table IV. Utilizing a 20% training ratio on the AID dataset, the proposed IRCHKD approach attains an OA of 96.30% . This performance surpasses ViT by 1.2% , EMTCAL by 1.41% , SCViT by 0.54% , TECN, which integrates convolution and transformer by 0.65% , TAKD by 5.05% , and SAGN by 0.93% .

When the training percentage is increased to 50% , the OA of our method reaches 97.84% , outperforming all comparison approaches significantly. It is 2.96% higher than DFAGCN, 1.35% higher than ViT, 1.43% higher than EMTCAL, 0.86% higher than SCViT, 3.04% higher than TAKD, and 0.44% higher than TECN.

The CM for the 50% training ratio is illustrated in Fig. 11. Out of the 30 categories, 27 have accuracies exceeding 90% , while only three classes fall below this threshold. These three

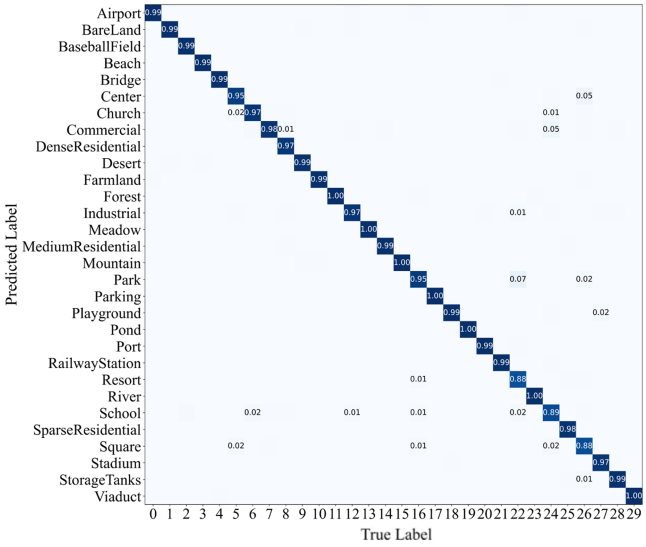

Fig. 11. CM diagram at 50% training ratio on AID dataset.

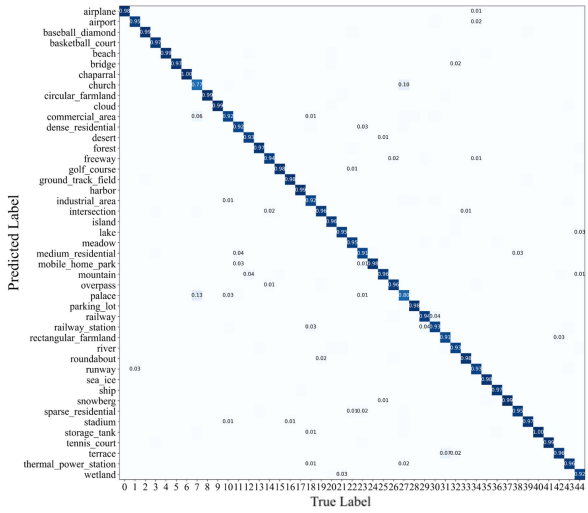

Fig. 12. Confusion matrix diagram at 20% training ratio on NWPU dataset.

TABLE V
COMPARISON BETWEEN PROPOSED METHOD AND RECENT METHODS ON THE NWPU

Methods	OA (10%)	OA (20%)
DFAGCN [61]	–	89.29 ± 0.28
ViT [18]	91.59 ± 0.19	93.90 ± 0.20
EMTCAL [42]	91.05 ± 0.17	93.36 ± 0.24
MBFANet [65]	91.61 ± 0.14	94.01 ± 0.08
SAGN [66]	91.73 ± 0.18	93.49 ± 0.10
SCViT [62]	92.72 ± 0.04	94.66 ± 0.10
VSDNet-ResNet34 [53]	92.13 ± 0.16	94.68 ± 0.13
TECN [43]	93.05	95.08
HFAM [67]	90.81 ± 0.11	–
CASD-ResNet50 [68]	92.28 ± 0.23	94.75 ± 0.14
CGINet [69]	92.28 ± 0.17	94.38 ± 0.13
GLR-CNN [70]	89.35 ± 0.25	92.11 ± 0.22
LSMNet [71]	90.80 ± 0.15	93.16 ± 0.13
MMPC-Net [72]	92.75 ± 0.19	94.88 ± 0.15
PFFGCN [73]	92.91 ± 0.15	94.89 ± 0.12
PSCLI-TF [74]	92.92	94.86
SAF-Net [75]	90.94 ± 0.08	93.62 ± 0.10
TAKD [76]	87.96	91.96
IRCHKD(ours)	93.13 ± 0.14	**95.29 ± 0.12**

categories are “Resort,” “School,” and “Square.” This misclassification is attributed to the high similarity between certain scene categories. In the AID dataset, categories like “Commercial” and “Dense residential” often share similar building layouts, textures, and spectral characteristics, making it challenging for the model to distinguish between them. For instance, both may contain a high density of man-made structures, and without additional contextual information, they can easily be confused.

3) Classification Results on the NWPU Dataset: Table V presents an evaluation of classification effectiveness across different methods.

The proposed IRCHKD method achieved an overall accuracy of 93.13% with a 10% training ratio and 95.29% with a 20% training ratio. Specifically, at a 10% training ratio, our IRCHKD outperforms TAKD by 5.17% , ViT by 1.54% , EMTCAL by 1.50% , SAGN by 1.40% , and VSDNet-ResNet34 by 1.0% . At a 20% training ratio, IRCHKD exceeds TAKD by 3.33% , ViT by 1.39% , EMTCAL by 1.64% , SAGN by 1.80% , SCViT by 0.63% , and VSDNet-ResNet34 by 0.61% .

TABLE VI
ANALYSIS OF MODEL STRUCTURE COMPLEXITY

Conditions	Parameter
SA	8.4M
WSA	6.2M
CHKD	2.4M

The CM on the NWPU dataset is illustrated in Fig. 12. At a 20% training ratio, only two out of 45 categories failed to attain an accuracy of 90% , while 31 categories achieved an accuracy of over 95% . The two categories with a classification accuracy of less than 90% are “church” and “palace,” both of which have highly similar architectural styles. The “palace” category is mainly misclassified as the “church” category. There are also many cases where the “church” category is mistakenly classified as the “palace” category. Churches and palaces usually have complex architectural designs, magnificent appearances, and symmetrical structures, which can be easily confused in remote sensing images. Many churches and palaces use large area pitched roofs or dome structures, which may appear as similar geometric shapes in remote sensing images. The facades of both buildings are often made of stone, marble, or brick, which show similar spectral reflectance characteristics in remote sensing images.

D. Evaluation of Size of Models

A complexity analysis was conducted on IRAM modules with and without window self-attention. At the same time, a complexity analysis was conducted on the distillation structure of cross head knowledge. The specific results are shown in Table VI.

From Table VI, it can be seen that using window attention reduces the number of model parameters by 2.2 million compared to not using window attention. The crosshead knowledge distillation structure contains 2.4 million parameters that can

TABLE VII
COMPARISON OF OA, PARAMETERS, AND FLOPS ON AID

The Network Model	OA(%)	Number of Parameter	FLOPs
GoogLeNet [57]	85.84	7M	1.5G
CaffeNet [57]	88.25	60.97M	715M
VGG-VD-16 [57]	87.18	138M	15.5G
DenseNet121 [77]	93.93	8.0M	11.5G
LCNN-BFF [21]	94.64	6.1M	24.6M
Contourlet CNN [78]	95.54	12.6M	2.1G
SE-MDPMNet [79]	97.14	5.17M	3.27G
EMTCAL [42]	96.41	27.8M	4.3G
TECN [43]	97.40	146.9M	24.9G
KFBNet [80]	97.40	216.3M	41.1G
IRCHKD (training)	97.84	8.6M	1.3G
IRCHKD (deployment)	97.84	6.2M	0.98G

TABLE VIII
ABLATION EXPERIMENTAL RESULTS OF THE IRCHKD ON THE NWPU DATASET

Conditions	MobileNetv2	IRAM	MSSA	CHKD	NWPU
1	✓				92.00 ± 0.13
2	✓	✓			94.33 ± 0.16
3	✓		✓		93.33 ± 0.28
4	✓			✓	94.01 ± 0.18
5	✓	✓	✓		94.53 ± 0.12
6	✓	✓		✓	94.82 ± 0.14
7	✓	✓	✓	✓	95.29 ± 0.12

be removed during the model inference phase. This knowledge distillation structure improves model performance while saving resource consumption.

We have listed the FLOPs and parameter quantities of some methods. The results are shown in Table VII.

Our model achieves OA of 0.44% higher than TECN and KFBNet, with significantly lower parameter count and complexity. Compared with methods such as VGG-VD-16, Contourlet CNN, and EMTCAL, the proposed approach not only excels in parameter amount and FLOPs, but also greatly outperforms these models in classification accuracy. Compared with GoogLeNet, SE-MDPMNet, DenseNet121, and LCNN-BFF, the proposed IRCHKD has slightly higher parameter count than these methods, but greatly outperforms these models in accuracy. As depicted in Table VII, the parameter count of the proposed approach during deployment is 2.4 million lower than during training, while FLOPs decrease by 0.32 billion compared to training.

E. Discussion

We conducted a variety of experiments, including ablation experiment, heat map analyses, and T-SNE visualizations. Ablation experiments were carried out on the NWPU dataset with a training ratio of 20% to validate the efficacy of the proposed three components. The results of the ablation experiment are shown in Table VIII. In the first scenario, the network is MobileNetv2, and the resulting model has the worst classification performance. In the second scenario, adding IRAM on the basis of MobileNetv2 improves classification accuracy. The third case is to add the MSSA module to the MobileNetv2 network, which improves the model classification performance by 1.33% . The fourth case is to add CHKD on the basis of MobileNetv2. After adding CHKD, the classification accuracy improved by 2.01% . These four cases indicate that the three modules we propose are effective when used alone. The fifth case is to add IRAM and MSSA, resulting in improved accuracy as opposed to adding only IRAM. The sixth case is to add IRAM and CHKD, and the accuracy is also higher than that of adding only IRAM. The seventh scenario involves incorporating IRAM, MSSA, and CHKD. When incorporating these three modules into the network, the highest OA of 95.25% is achieved. This indicates that the proposed module performs better when used in combination than when used alone. In comparison to the initial scenario, the OA of the seventh case increased by 3.29% . Hence, the ablation experiment effectively showcases the effectiveness of the proposed IRCHKD method.

TABLE IX
ABLATION EXPERIMENTAL RESULTS OF THE EXPANSION RATE ON THE AID DATASET

Conditions	Expansion rate	Accuracy
Baseline	1	97.26 ± 0.18
Single Rate	2	97.48 ± 0.20
Single Rate	3	97.44 ± 0.16
Multi Rate	[1,2,3]	97.84 ± 0.12
Multi Rate	[1,2,4]	97.65 ± 0.14
Multi Rate	[1,3,5]	97.58 ± 0.12

We also discussed how to choose the dilation rate in the MSSA module to achieve the best classification performance of the model. Due to the grid effect in dilated convolution, we did not use a very large dilation rate to avoid gaps between RFs, which would result in discontinuous feature extraction. In the ablation experiment, the performance of a fixed expansion rate was first tested, and the effects of different single expansion rates were analyzed. Then, the effect of using different combinations of dilation rates in the parallel dilation convolution module was tested. Using regular convolution (with a dilation rate of 1) as the baseline, compare the improvement of dilated convolution. The experiment was conducted using the AID dataset, and the results are shown in Table IX.

From Table IX, it can be seen that under the condition of a single dilation rate, using dilation convolution yields better results than not using dilation convolution. Under conditions of multiple expansion rates, the effect is better than that with using a single expansion rate, so we choose to use a combination of multiple expansion rates. Under the condition of multiple expansion rates, the combination of expansion rates [1,2,3] are chosen as the hyperparameters of the model.

Grad-CAM can be adopted to generate attention maps, also called heat maps, which can display areas of interest to the model. We randomly selected some scenes such as “Church,” “ commercial_area,” “ freeway,” and “island” in the NWPU dataset to generate heat maps. The heat maps generated by the MobileNetv2 model and the proposed IRCHKD model are shown in Fig. 13.

As depicted in Fig. 13, for “island,” the MobileNetv2 approach cannot accurately focus on the island area, while the suggested approach can efficiently focus on the island area. For the “church,” “commercial_area,” and “highway” scenarios, the

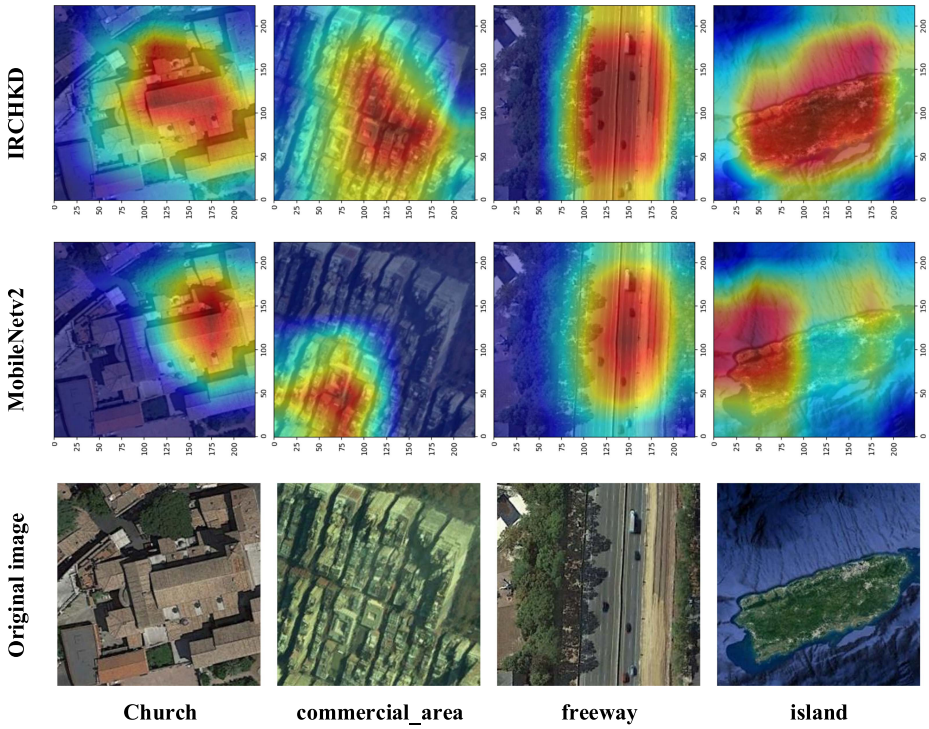

Fig. 13. Heat map of some scenarios in NWPU dataset.

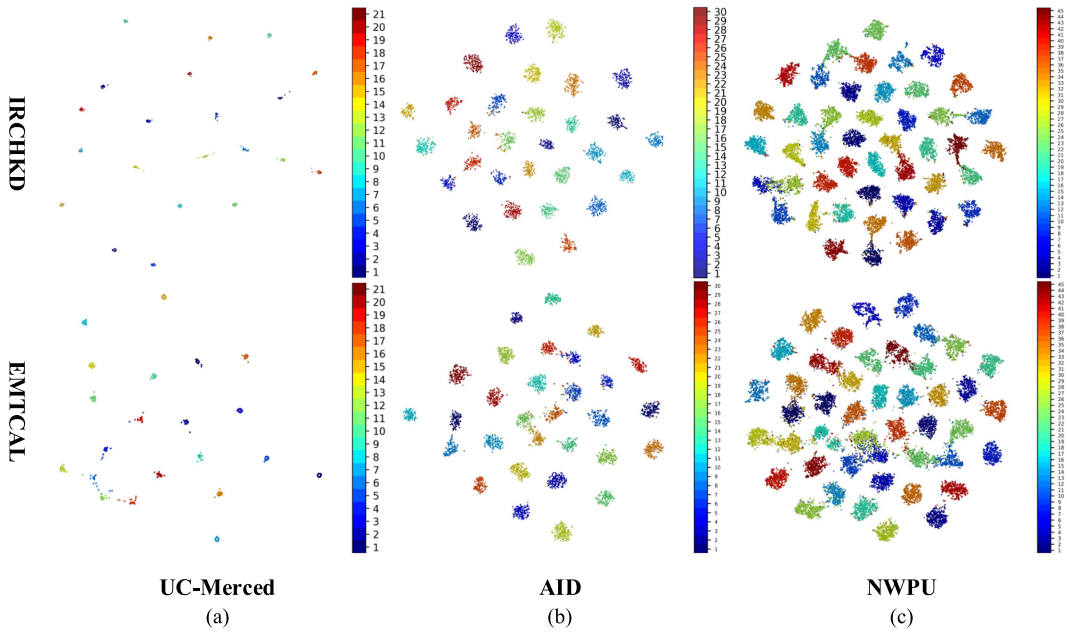

Fig. 14. T-SNE visualization outcomes on different datasets. (a) T-SNE visualization outcomes at 80% training ratio on UCM dataset. (b) T-SNE visualization outcomes at 50% training ratio on AID. (c) T-SNE visualization outcomes under 20% training ratio on the NWPU dataset.

MobileNetv2 network's focus area is not comprehensive and ignores surrounding targets. The method we proposed can obtain the complete region of interest, which fully proves our method's focus on global information, thereby increasing classification accuracy.

The T-SNE approach is implemented to visualize the classification maps on the UCM, AID, and NWPU datasets. T-SNE transforms high-dimensional features into a lower dimensional space for visualization. T-SNE attempts to keep dissimilar data points away in low dimensional space. From Fig. 14, in low

dimensional space, the proposed method can bring similar data points closer and dissimilar data points further away, making it easier to distinguish different categories.

V. Conclusion

This article introduces a new RSSC method, named IRCHKD. This method primarily comprises of four key components: inverted residual convolution block, IRAM, MSSA, and CHKD structure. First, two inverted residual convolutional blocks are introduced to acquire the shallow features of the image. Next, an IRAM module was designed, which first utilizes self-attention mechanism to learn long-range dependency information and increase the count of feature channels, and then further extracts features through convolution. Then, a MSSA module was constructed, which achieved the fusion of contextual information from different ranges in the image by using dilated convolutional with different RF sizes, thereby comprehensively capturing the multi-scale features of the target object. Finally, using the CHKD module, the intermediate features of the student classification head are input into the teacher classification head, and the resulting cross head prediction imitates the teacher's prediction. This effectively reduces the impact of conflicting supervision signals between true labels and teacher predictions on student networks. Verified on three widely used RSSC datasets, the experimental results show that the proposed IRCHKD method exhibits excellent performance. In the future, our focus will be on lightweight RS scene image classification, striving to further reduce network complexity to meet the needs of fast processing applications.

References

[1] R. K. Jaiswal, R. Saxena, and S. Mukherjee, "Application of remote sensing technology for land use/land cover change analysis," *J. Indian Soc. Remote Sens.*, vol. 27, pp. 123–128, 1999.

[2] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[3] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[4] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[5] S. E. Grigorescu, N. Petkov, and P. Kruizinga, "Comparison of texture features based on Gabor filters," *IEEE Trans. Image Process.*, vol. 11, no. 10, pp. 1160–1167, Oct. 2002.

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

[7] S. Mei, J. Ji, J. Hou, X. Li, and Q. Du, "Learning sensor-specific spatial-spectral features of hyperspectral images via convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4520–4533, Aug. 2017.

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

[9] S. Sergyan, "Color histogram features based image classification in content-based image retrieval systems," in *Proc. 6th Int. Symp. Appl. Mach. Intell. Inform.*, 2008, pp. 221–224.

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

[11] H. Soltanian-Zadeh and F. Rafiee-Rad, "Comparison of multiwavelet, wavelet, haralick, and shape features for microcalcification classification in mammograms," *Pattern Recognit.*, vol. 37, no. 10, pp. 1973–1986, 2004.

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

[13] L. Zhang, W. Zhou, and L. Jiao, "Wavelet support vector machine," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 1, pp. 34–39, Feb. 2004.

[14] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote Sens. Environ.*, vol. 61, no. 3, pp. 399–409, 1997.

[15] X. Tang et al., "Unsupervised deep feature learning for remote sensing image retrieval," *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1243.

[16] H. Wu et al., "A cross-channel dense connection and multi-scale dual aggregated attention network for hyperspectral image classification," *Remote Sens.*, vol. 15, no. 9, 2023, Art. no. 2367.

[17] C. Shi, H. Wu, and L. Wang, "A feature complementary attention network based on adaptive knowledge filtering for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5527219.

[18] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[19] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

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

[21] C. Shi, T. Wang, and L. Wang, "Branch feature fusion convolution network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5194–5210, 2020.

[22] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.

[23] W. Li et al., "Classification of high-spatial-resolution remote sensing scenes method using transfer learning and deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1986–1995, 2020.

[24] P. Singh, V. K. Verma, P. Rai, and V. P. Namvoodiri, "HetConv: Heterogeneous kernel-based convolutions for deep cnns," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4835–4844.

[25] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11030–11039.

[26] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10093–10102.

[27] Y. Chen et al., "Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3434–3443.

[28] K. Han, Y. Wang, Q. Tian, J. Gou, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1580–1589.

[29] Y. Lu, M. Gong, Z. Hu, W. Zhao, Z. Guan, and M. Zhang, "Energy-based CNN pruning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3000214.

[30] W. Wang, Y. T. Shi, and X. Wang, "RMFFNet: A reverse multi-scale feature fusion network for remote sensing scene classification," in *Proc. Int. Joint Conf. Neural Netw.*, 2024, pp. 1–8.

[31] C. Shi, X. Zhao, and L. Wang, "A multi-branch feature fusion strategy based on attention mechanism for remote sensing image scene classification," *Remote Sens.*, vol. 13, no. 10, 2021, Art. no. 1950.

[32] X. Tang, Q. Ma, X. Zhang, F. Liu, J. Ma, and L. Jiao, "Attention consistent network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2030–2045, 2021.

[33] 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, 2022.

[34] J. Lu et al., "Soft: Softmax-free transformer with linear complexity," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 21297–21309, 2021.

[35] F. Zhu, Y. Zhu, L. Zhang, C. Wu, Y. Fu, and M. Li, "A unified efficient pyramid transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2667–2677.

[36] D.-J. Chen, H.-Y. Hsieh, and T.-L. Liu, "Adaptive image transformer for one-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12242–12251.

[37] C. F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 357–366.
[38] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[39] J. Xia, Y. Zhou, and L. Tan, "DBGA-Net: Dual-branch global–local attention network for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 7502305.
[40] X. Chen et al., "Hierarchical feature fusion of transformer with patch dilating for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4410516.
[41] M. Zhao, Q. Meng, L. Zhang, X. Hu, and L. Bruzzone, "Local and long-range collaborative learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5606215.
[42] X. Tang, M. Li, J. Ma, X. Zhang, F. Liu, and L. Jiao, "EMTCAL: Efficient multiscale transformer and cross-level attention learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5626915.
[43] S. Chen, W. Li, Y. Cao, and X. Lu, "Combining the convolution and transformer for classification of smoke-like scenes in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4512519.
[44] H. J. I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[45] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*.
[46] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3962–3971.
[47] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1921–1930.
[48] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9163–9171.
[49] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3712–3721.
[50] M. Ji, S. Shin, S. Hwang, G. Park, and I. C. Moon, "Refine myself by teaching myself: Feature refinement via self-knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10664–10673.
[51] Y. Hu et al., "Hierarchical self-distilled feature learning for fine-grained visual categorization," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 15, 2021, doi: 10.1109/TNNLS.2021.3124135.
[52] C. Shi, M. Ding, L. Wang, and H. Pan, "Learn by yourself: A feature-augmented self-distillation convolutional neural network for remote sensing scene image classification," *Remote Sens.*, vol. 15, no. 23, 2023, Art. no. 5620.
[53] Y. Hu, X. Huang, X. Luo, J. Han, X. Cao, and J. Zhang, "Variational self-distillation for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5627313.
[54] D. Li, Y. Nan, and Y. Liu, "Remote sensing image scene classification model based on dual knowledge distillation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4514305.
[55] H. Liu, Y. Qu, and L. Zhang, "Multispectral scene classification via cross-modal knowledge distillation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5409912.
[56] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.
[57] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[58] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[59] X. Wang, L. Duan, C. Ning, and H. Zhou, "Relation-attention networks for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 422–439, 2022.
[60] X. Wang, S. Wang, C. Ning, and H. Zhou, "Enhanced feature pyramid network with deep semantic embedding for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7918–7932, Sep. 2021.
[61] 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.*, vol. 33, no. 10, pp. 5751–5765, Oct. 2022.
[62] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4409512.
[63] Q. Meng, M. Zhao, L. Zhang, W. Shi, C. Su, and L. Bruzzone, "Multilayer feature fusion network with spatial attention and gated mechanism for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6510105.
[64] P. Deng, H. Huang, and K. Xu, "A deep neural network combined with context features for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8000405.
[65] C. Shi, X. Zhang, and L. Wang, "A lightweight convolutional neural network based on channel multi-group fusion for remote sensing scene classification," *Remote Sens.*, vol. 14, no. 1, 2021, Art. no. 9.
[66] Y. Yang, X. Tang, Y.-M. Cheung, X. Zhang, and L. Jiao, "SAGN: Semantic-aware graph network for remote sensing scene classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1011–1025, 2023.
[67] Q. Wan, Z. Xiao, Y. Yu, Z. Liu, K. Wang, and D. Li, "A hyperparameter-free attention module based on feature map mathematical calculation for remote-sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5600318.
[68] B. Wu, S. Hao, and W. Wang, "Class-aware self-distillation for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2173–2188, 2024.
[69] Y. Zhao, Y. Chen, S. Xiong, X. Lu, X. X. Zhu, and L. Mou, "Co-enhanced global-part integration for remote-sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4702114.
[70] L. Liu, Y. Wang, J. Peng, and L. Zhang, "GLR-CNN: CNN-based framework with global latent relationship embedding for high-resolution remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5633913.
[71] K. Zhang, T. Cui, W. Wu, X. Zheng, and G. Cheng, "Large kernel separable mixed convnet for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4294–4303, 2024.
[72] S. Li, M. Dai, and B. Li, "MMPC-Net: Multi-granularity and multi-scale progressive contrastive learning neural network for remote-sensing image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 2502505.
[73] C. Zhang and B. Wang, "Progressive feature fusion framework based on graph convolutional network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3270–3284, 2024.
[74] D. Li, R. Liu, Y. Tang, and Y. Liu, "PSCLI-TF: Position-sensitive cross-layer interactive transformer model for remote sensing image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 5001305.
[75] W. Song, Y. Zhang, C. Wang, Y. Jiang, Y. Wu, and P. Zhang, "Remote sensing scene classification based on semantic-aware fusion network," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 2505805.
[76] J. Wu, L. Fang, and J. Yue, "TAKD: Target-aware knowledge distillation for remote sensing scene classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 9, pp. 8188–8200, Sep. 2024.
[77] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.
[78] 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, Jun. 2021.
[79] B. Zhang, Y. Zhang, and S. Wang, "A lightweight and discriminative model for remote sensing scene classification with multidilation pooling module," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2636–2653, Aug. 2019.
[80] F. Li, R. Feng, W. Han, and L. Wang, "High-resolution remote sensing image scene classification via key filter bank based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8077–8092, Nov. 2020.

Cuiping Shi (Member, IEEE) received the M.S. degree in signal and information processing from Yangzhou University, Yangzhou, China, in 2007, and the Ph.D. degree in information and communication engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2016.

From 2017 to 2020, she held postdoctoral research with the College of Information and Communications Engineering Harbin Engineering University, Harbin. She is currently a Professor with the Department of Communication Engineering, Qiqihar University, Qiqihar, China. From 2024 she works with the College of Information Engineering, Huzhou University, Huzhou, China. Her main research interests include remote sensing image processing pattern recognition and machine learning. She has authored or coauthored two academic books about remote sensing image processing and more than 90 papers in journals and conference proceedings.

Dr. Shi was the recipient of the nomination award of Excellent Doctoral Dissertation of HIT in 2016 for her doctoral dissertation.

Mengxiang Ding received the bachelor's degree in biomedical engineering from WeiFang Medical University, WeiFang, China, in 2021, and the master's degree in communication and information systems from Qiqihar University, Qiqihar, China, in 2024.

His research interests include remote sensing image processing, machine learning, and deep learning.

Liguo Wang (Member, IEEE) received the M.S. and Ph.D. degrees in signal and information processing from the Harbin Institute of Technology, Harbin, China, in 2002 and 2005, respectively.

From 2006 to 2008, he held a postdoctoral research position with the College of Information and Communications Engineering, Harbin Engineering University, Harbin, where he is currently a Professor. Since 2020, he has been working with the College of Information and Communication Engineering, Dalian Nationalities University, Dalian, China. His main research interests include remote sensing image processing and machine learning. He has authored or coauthored two books about hyperspectral image processing and more than 130 papers in journals and conference proceedings.