# Local and Long-Range Collaborative Learning for Remote Sensing Scene Classification

Maofan Zhao, Qingyan Meng, Linlin Zhang, Xinli Hu, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—With the development of high-resolution satellites, more and more attention has been paid to remote sensing (RS) scene classification. Convolutional neural networks (CNNs), which replace the traditional handcrafted features with a learning-based feature extraction mechanism, are widely used in scene classification. But CNNs are less effective in deriving long-range contextual relations, which limits the further improvement. Visual transformer (VT), an emerging image processing method, provides a new perspective for RS scene classification by directly acquiring long-range features. Although there have been limited works combining CNN and VT through simple concatenation, the collaborations between them are insufficient. To address these issues, we propose a local and long-range collaborative framework (L2RCF). First, we design a dual-stream structure to extract the local and long-range features. Second, a cross-feature calibration (CFC) module is designed for them to improve representation of the fusion features. Then, combining deep supervision (DS) and deep mutual learning (DML), a novel joint loss is proposed to enhance the dual-stream feature extractor and further improve the fused features. Finally, a two-stage semi-supervised training strategy is designed to improve performance with unlabeled samples. To demonstrate the effectiveness of L2RCF, we conducted experiments on three widely used RS scene classification data sets: RSSCN7, AID, and NWPU. The results show that L2RCF performs significantly better compared with some state-of-the-art scene classification methods.**

***Index Terms*—Scene classification, convolutional neural network, visual transformer, cross-feature calibration, deep supervision, deep mutual learning, semi-supervised, remote sensing.**

## I. Introduction

LAND-use/land-cover information interpretation is a crucial research area in remote sensing (RS). However, most of the previous studies focused on land-cover and only few of them on land-use representation. In recent years, with the development of high-resolution satellite sensors, the RS interpretation mode has gradually developed from pixel level and object level to scene level, with the goal to obtain higher-level semantic information. Therefore, RS scene classification has gained more attention, as it can be used in land-use classification [1], urban functional zone identification [2], and other related fields [3]. However, due to the complexity and large-scale variance of geographic objects in high-resolution RS scenes, how to extract more discriminative features in the scene remains an important and challenging task.

Commonly used features can be summarized into two main types: handcrafted features and deep features. Handcrafted features include low-level features, such as spectra and textures, and mid-level features that are encoded based on low-level features, such as bags of visual words (BoVW) [4]. Compared with handcrafted features, deep features are more abstract but contain richer semantic information. In particular, convolutional neural networks (CNNs), which are commonly used in image processing, have achieved a dominant role and state-of-the-art performance in the field of scene classification.

Since CNNs rely on local convolution kernels, they have very good local feature representation, but suffer in representing long-range information in the images. Some methods have been gradually proposed to overcome this problem, such as feature pyramids [5], and multiscale strategy [6]. However, they still have limitations in solving this problem. Recently, transformer [7], a structure widely used in natural language processing, has been gradually used in image processing [8], [9]. It can directly obtain the long-range information in the images and provide a new perspective for RS image scene classification.

Visual Transformers (VTs) can capture the long-range features, that are difficult to model with CNNs. However, due to their origin in natural language processing, there are still some shortcomings to use them in image processing. First of all, VTs directly expand the image into a one-dimensional vector, which is not effective to model the local structure information of the images. Second, the redundant attention module brings a computational burden.

The class activation maps of VTs and CNNs for local and long-range ground-objects are shown in Fig.1, where river, freeway are long-range ground-objects, and airplane, overpass are local objects. VTs successfully capture the non-local contextual information of long-range ground-objects. CNNs make the model focus on local ground-objects through the convolution kernel. To sum up, CNNs have good representation of local structural features, but it is difficult to capture long-range

Manuscript received xxx; revised xxx; accepted. Date of publication xxx; date of current version xxx. (*Corresponding author: Qingyan Meng.*)

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

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

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

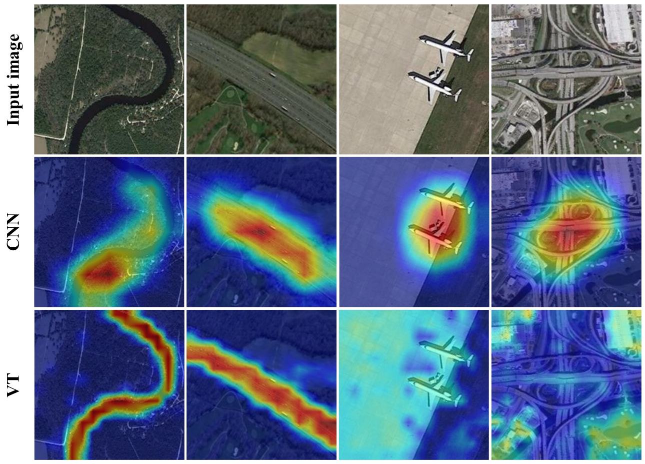


Fig. 1. Class attention maps produced by CNN and VT. CNNs and VTs are good at extracting local features and long-range features, respectively.

information, whereas VTs are good at extracting long-range information and easily ignore local features. Therefore, it is promising to study how combining CNNs and VTs effectively.

Currently, there are limited works combining CNNs and VTs for RS scene classification. Deng *et al.* [10] combine CNN and VT via concatenation for RS scene classification. But there are the following problems: i) The features are simply concated, and there is no interaction to form more discriminative fused features; ii) CNNs and VTs do not cooperate with each other to enhance their respective feature extraction capabilities.

On the basis of the above mentioned limitations, we enhance collaborative learning between them in three ways: i) Local and long-range features interact with each other to calibrate and fuse features; ii) Local and long-range feature extractors learn from each other to compensate for their shortcomings; iii) Local, long-range and fused features collaborate to exploit unlabeled samples to further improve performance. Accordingly, the main contributions of this paper can be summarized as follows:

1) We introduce a local and long-range collaborative framework (L2RCF) for RS scene classification that can fully collaborates CNNs and VTs.

2) To effectively improve representation of the fused features, we design a cross-feature calibration (CFC) module for feature fusion.

3) We propose a novel joint loss based on deep supervision (DS) and deep mutual learning (DML), which can not only further improve the fused features, but also effectively enhance the dual-stream feature extractor.

4) To further exploit potential of unlabeled data, we design a two-stage semi-supervised training strategy.

The rest of this paper is organized as follows. Related works such as CNNs, VTs and deep learning-based RS scene classification are reviewed in Section II. Section III describes the steps in detail and theory of proposed L2RCF. Section IV presents the data sets, the evaluation metrics, the experimental settings used to evaluate the performance of L2RCF. Moreover, it analyzes the experimental results in detail through ablation experiments and visualization methods. Section V not only summarizes paper, but also prospects the possible future research directions.

## II. Related Work

### A. CNN

CNNs are generally composed of convolution layers, pooling layers, activation functions, and fully connected layers [11]. And features are extracted repeatedly through continuous convolution - pooling operation to form a series of feature maps. Shallow feature maps contain a large number of low-level and mid-level features, such as structure and texture information, while deep features contain high-level semantic features. CNNs are widely used in many visual tasks, such as image classification [12], image fusion [13] and change detection [14], due to their good robustness to translation and scaling.

Since AlexNet [15] was proposed in 2012, CNNs have become more and more important in computer vision. Then, models such as VGGNet [16] and ResNet [17] have been proposed and widely used.

### B. Visual Transformer

Transformers [7] gradually replace recurrent neural networks (RNNs) and are widely used in natural language processing. The multi-head self-attention (MSA) structure in it greatly improves the computational efficiency, allowing the transformer to achieve fast parallelism. VTs have gradually migrated to many tasks, such as image classification [18], semantic segmentation [19], object detection [20], etc.

DERT [20] is the early architecture that applies transformer to computer vision tasks. It effectively combines CNNs and VTs. More specifically, it first uses CNNs to extract features and then uses them as input to the transformer. Although DERT successfully combines CNNs and VTs into an end-to-end structure, it has relevant shortcomings such as high computational cost and tendency to ignore small objects. In order to avoid the attention computation for all pixels in the image, vision transformer (ViT) [8] splits the image into fixed-size patches and feeds them into the transformer encoder using patch embedding to compute the self-attention (SA) between the patches. ViT has demonstrated great performance, but it is still far behind the most advanced CNNs. In addition, it has large model parameters and relies on pre-training on very large data sets. Due to the excessive parameters and long training time of ViT, DeiT [21], which is based on ViT, introduces the teacher-student strategy to greatly improve the performance of VTs. Unlike previous transformers that compute global attention, swin-transformer [9] first computes the local attention within patches, and then extends the computation of attention to the global scale by gradually merging patches. This improves the ability of the transformer to extract local features.

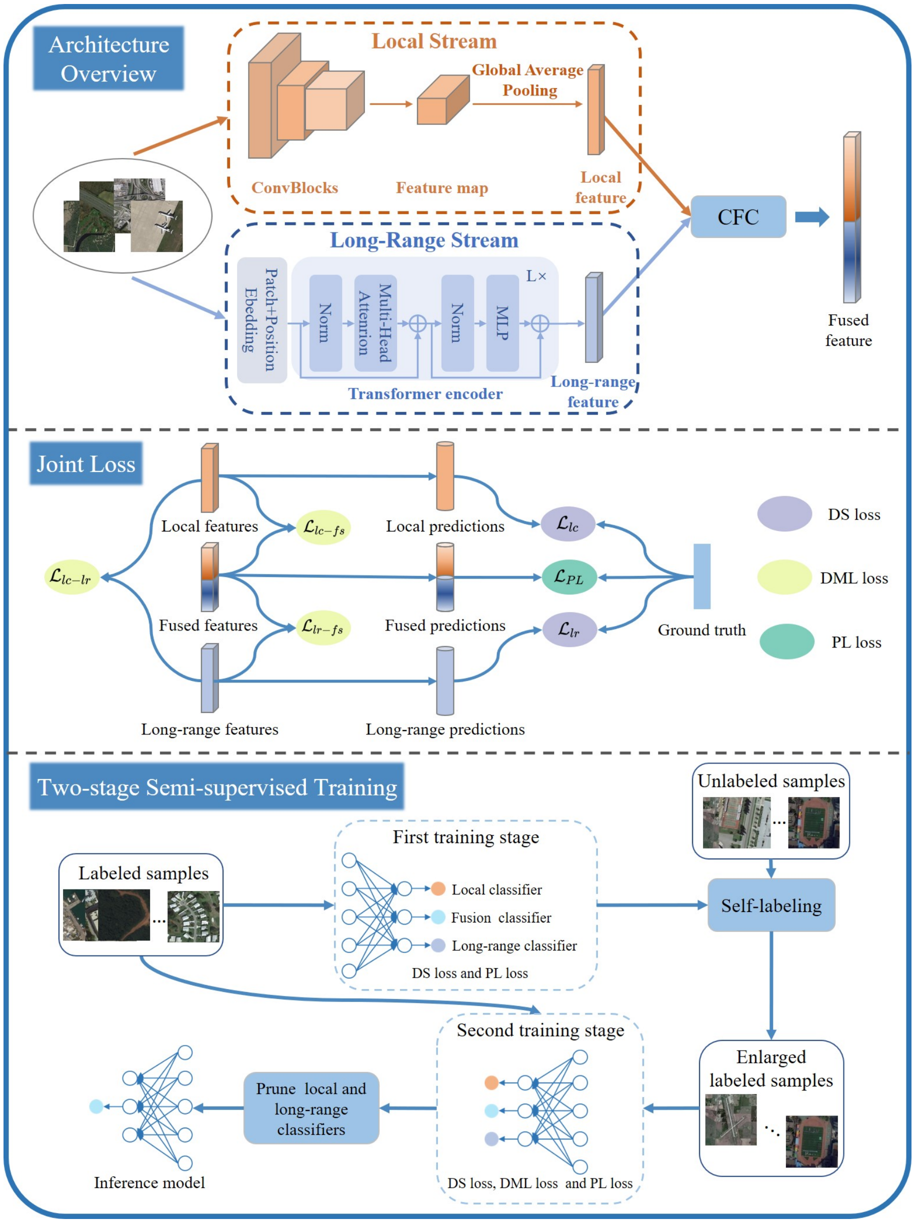


Fig. 2. Illustration of L2RCF. (a) Architecture overview: a dual-stream structure is designed to extract local and long-range features, and then the CFC enhances the representation of the fused feature. (b) Joint loss: DS and DML are combined to form joint loss for further improving the fused features and enhancing the dual-stream feature extractor. (c) Semi-supervised two-stage training: the first training stage is based on labeled samples, and enlarged labeled samples are obtained by self-labeling; the second training stage is based on labeled samples and enlarged labeled samples; finally, the inference model is obtained by pruning local classifier and long-range classifier. Note that DML is not used in the first training stage.

### C. Deep Learning-based Remote Sensing Scene Classification

Before the rise of deep learning, handcrafted features were widely used for RS scene classification by considering global and local feature descriptors. Global features such as color histograms and texture descriptors directly generate a feature representation of images. Sande *et al.* [22] builds color histograms based on the HSV color space for the representation of the scene. Local feature descriptors such as histogram of oriented gradients (HOG) [23] and scale-invariant feature transform (SIFT) [24] are often used for feature encoding to generate mid-level features of the scene. Common feature coding methods include latent dirichlet allocation (LDA) [25] and BoVW. Zhu *et al.* [4] explore RS scene tasks based on BoVW. In contrast to handcrafted features, deep learning methods have been recently used for RS scene classification.

Autoencoders play an important role in RS scene classification in the early stage of deep learning. Othman *et al.* [26] combine convolutional features and sparse autoencoders for RS scene classification tasks. CNNs are the most widely used deep learning method in RS image processing (including scene classification). Hu *et al.* [27] use transfer learning to directly obtain the feature representation of the scene. Liu *et al.* [28] propose a multi-scale CNN combined with fixed-scale net and varied-scale net for modeling the scale variation of the objects in RS images. Lu *et al.* [29] construct a feature aggregation network to generate an accurate representation of RS scenes. In addition, various attention mechanisms are continuously combined with CNNs for RS scene classification tasks [30]–[33].

Emerging deep learning methods such as graph convolutional networks (GCNs) [34], neural architecture search (NAS) [35], generative adversarial networks (GANs) [36], local-global learning [37] [38] and others [39]–[41] have also been used in scene classification. Xu *et al.* [42] design a deep feature aggregation framework based on GCN. Wang *et al.* [43] propose a RS neural network framework using an automatic search strategy, which can be effectively used for scene classification and semantic segmentation tasks. Ma *et al.* [44] develop a RS scene classification NAS framework based on multi-objective neural evolution. Yu *et al.* [45] propose an attention GAN in combination with the attention mechanism, which improves the performance in RS scene classification by improving the representation of the discriminator. Cheng *et al.* [46] propose an effective defense framework named PSGANs for RS scene classification. In order to improve global representation of CNNs, Lv *et al.* [37] propose the local-global-fusion feature extraction network, which leverages RNNs to capture contextual information. And Chen *et al.* [38] propose the local–global mutual learning (LML) method to obtain different features and learn from each other through KL. However, they are still difficult to improve the extraction of CNN for long-range features. To support the inference recognition of unseen RS image scenes, the remote sensing knowledge graph (RSKG) [40] and asymmetric collaborative network (SCN) for lifelong RS image classification [39] are designed. And Li *et al.* [41] propose the error tolerance deep learning method for the negative impact of error labels in the data set.

In Addition, there have been limited works utilizing VTs for RS scene classification. Bazi *et al.* [47] and Kaselimi *et al.* [48] use the VT to classify RS scenes using transfer learning. The spatial-channel feature preserving ViT (SCViT) model is proposed to consider the contribution of different channels [49]. Tang *et al.* [50] the propose efficient multiscale transformer and cross-level attention learning (EMTCAL) model. Ma *et al.* [51] propose homo–heterogenous transformer learning (HHTL) to more effectively distinguish intra-class/inter-class samples. Yu *et al.* [52] design a cross-context and cross-scale capsule vision transformer which combines combining capsule networks and VTs.

## III. Methodology

The objects in RS scenes are complex, including overpasses, stadiums and other objects with specific local structural features, as well as long-range objects such as rivers and highways. Therefore, a classification model needs to be sensitive to the properties of different types of objects. But CNNs or VTs can not be sensitive to both local objects and long-range objects. Although limited work has combined CNNs and VTs, the collaboration between them is inadequate. To address these issues, the L2RCF for RS scene classification is proposed, as shown in Fig.2. L2RCF includes four main parts: i) a dual-stream structure based on CNNs and VTs designed to extract local features and long-range features; ii) the CFC for calibration of local features and long-range features to obtain more discriminative fusion features; iii) a joint loss based on DS and DML to enhance the dual-stream feature extractor and further improve the fused features; iv) a two-stage semi-supervised training strategy for improving the performance by exploiting potential of unlabeled samples.

### A. CNN

The Convolutional blocks are the basic structure of CNNs, which contain convolutional layers, pooling layers, activation functions. The input of the convolutional block is the original image or the feature map $\mathbf{I} \in \mathbb{R}^{C_i \times H_i \times W_i}$ from the previous convolutional block, and the output is the feature map $\mathbf{O} \in \mathbb{R}^{C \times H \times W}$, which is computed as follows:

$$\mathbf{O} = \mathrm{ConvBlocks}(\mathbf{I}). \tag{1}$$

Then, the global average pooling (GAP) computes an average value for each channel of $\mathbf{O}$. After GAP, the feature vector $\mathbf{GP} \in \mathbb{R}^{C}$ is obtained which can be written as follows:

$$\mathbf{GP}_c = \mathrm{F}_{gp}\left(\mathbf{O}_c\right) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{O}_c(i,j), \tag{2}$$

where $\mathbf{GP}_c$ and $\mathbf{O}_c$ denote the $c$-th channel of $\mathbf{GP}$ and $\mathbf{O}$, respectively. And $\mathrm{F}_{gp}$ represents the GAP function.

### B. Visual Transformer

As shown in the long-range stream in Fig. 2, VTs mainly include embedding and encoder, which are briefly described below.

The image $\mathbf{I}_m \in \mathbb{R}^{C' \times H' \times W'}$is divided into patches with dimension of $C' \times P \times P$. Then each patch is mapped into the dimension $D$ through a linear transformation. In order to preserve the position information between patches, learnable 1-$D$ positional embedding is used. Patch embeddings and positional embedding are directly added as the input of transformer encoder. Unlike CNNs, which usually use GAP to obtain features for image classification, VTs add a special class token as input. And the related output $\mathbf{T}_L^0$ represents classification head, which are generally connected to a linear classifier. In summary, the vector $\mathbf{T}_0$ after linear embedding is specifically expressed as follows:

$$\mathbf{T}_0 = \left[\mathbf{x}_{class}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \ldots; \mathbf{x}_p^N E\right] + \mathbf{E}_{pos}, \tag{3}$$

where $\mathbf{E} \in \mathbb{R}^{\left(P^2 \times C'\right) \times D}$, $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$, $\mathbf{x}_{class}$ is the initial value of the class token ($\mathbf{T}_0^0 = x_{class}$), $\mathbf{x}_p^1$, $\mathbf{x}_p^2$, $\mathbf{x}_p^N$ represent different patches, $N$ represents the number of patches ($N = H' \times W'/P^2$) , $\mathbf{E}$ is the linear transformation matrix of the patch and $\mathbf{E}_{pos}$ is the learnable position matrix.

The transformer encoder is the core of VTs, which is stacked by a unified structure composed of MSA and multi-layer perception (MLP). The MSA uses multiple attention heads in parallel, which allows the model to learn correlation weights in different representation subspaces. And the MLP consists of two linear layers with a GELU [53] activation. In addition,

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3265346

both MSA and MLP use residual connections, and have a normalization layer (LN) in front of them, as follows:

$$\mathbf{T}'_l = \mathrm{MSA}\left(\mathrm{LN}\left(\mathbf{T}_{l-1}\right)\right) + \mathbf{T}_{l-1}, \quad l = 1 \ldots L, \tag{4}$$

$$\mathbf{T}_l = \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{T}'_l\right)\right) + \mathbf{T}'_l, \quad l = 1 \ldots L, \tag{5}$$

where $\mathbf{T}_{l-1}$ and $\mathbf{T}_l$ are the output of the $l-1$ layer and the $l$ layer transformer encoder, respectively.

## C. Feature Fusion via CFC

Previous feature fusion methods in related work usually adopt direct concatenation. But it is not effective due to lack of interaction with each other and limits the representation of fusion features. The ”Squeeze-Excitation-Reweight” paradigm in SENet [54] has been shown to be effective for feature recalibration. For local feature and long-range feature fusion, it is crucial to fully consider the correlation between them to achieve more effective information interaction. Therefore, we design the CFC module for modeling the calibration and fusion between local and long-range features, as shown in Fig. 3.

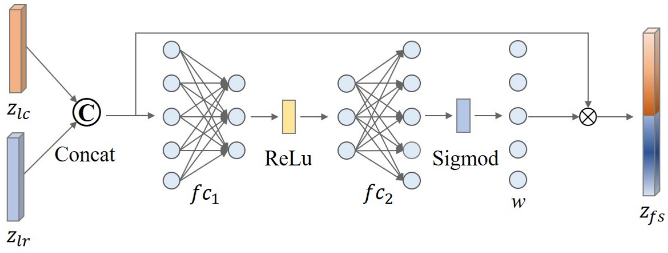


Fig. 3. Structure of CFC. It can effectively enhance the representation of fusion features.

First we concat local feature and long-range feature as follows:

$$\mathbf{z}_{concat} = \mathrm{Concat}(\mathbf{z}_{lc}, \mathbf{z}_{lr}), \tag{6}$$

where $\mathbf{z}_{lc}$ and $\mathbf{z}_{lr}$ represent local features and long-range features, respectively.

Then the calibration weight $\mathbf{w}$ is obtained through a nonlinear unit, which contains two fully connected layers, namely dimensionality-reduction layer $fc_1$ and dimensionality-increasing layer $fc_2$. $\mathbf{w}$ can be expressed as:

$$\mathbf{w} = \sigma\left(\mathbf{w}_2 \delta\left(\mathbf{w}_1 \mathbf{z}_{concat}\right)\right), \tag{7}$$

where $\mathbf{w}_1 \in \mathbb{R}^{d' \times d}$ and $\mathbf{w}_2 \in \mathbb{R}^{d \times d'}$ represent the weight parameters of $fc1$ and $fc2$, respectively, $d$ is the dimension of $\mathbf{z}_{concat}$, $\delta$ and $\sigma$ represent the ReLu and Sigmod activation functions, respectively. A dimensionality reduction ratio $r$ is used to control the value of $d'$:

$$d' = \max(d/r, L_{dn}), \tag{8}$$

where $L_{dn}$ represents the minimum value of $d'$. $\mathbf{z}_{fs}$ is obtained by multiplying $\mathbf{w}$ and $\mathbf{z}_{concat}$. By controlling calibration weights, the features with great representation are further enhanced and some redundant features are diminished.

## D. Joint Loss

We designed a novel joint loss for boosting the dual-stream feature extractor and further improving the fused feature in L2RCF. It consists of three losses: the DS loss $\mathcal{L}_{DS}$, the proposed DML loss $\mathcal{L}_{DML}$ within the network, the common prediction loss $\mathcal{L}_{PL}$ based on the final feature (i.e. fusion feature). The diagram of DS loss and prediction loss are shown in Fig. 4. The diagram of DML loss is shown in Fig. 5 (c). Joint loss enhances the representation of the L2RCF in three ways: i) DS helps the dual-stream feature extractor to drive more discriminative features; ii) DML helps local stream and long-range stream to compensate for their shortcomings and refine the fused features; iii) DS and DML are complementary. On the one hand, DML can provide some auxiliary DS information in the form of ‘soft label’. On the other hand, DS can help to correct mis-knowledge that may be introduced by DML. The DS loss, DML loss, and joint loss functions are described as follows.

### 1) DS loss

The deep models are usually used to extract features from images to obtain a discriminative representation of RS scene. However, when the gradients are passed from the deep layers to the shallow layers during training, they become very small or disappear, which makes it difficult to converge. In addition, we use CFC when fusing the features, which further increases the length of the gradient back-propagation.

So we use the DS strategy to help learn more discriminative representation, as shown in Fig. 4. More specifically, two auxiliary classifiers are added based on the local features and long-range features. DS loss is computed from the prediction results of the auxiliary classifier with the labels.

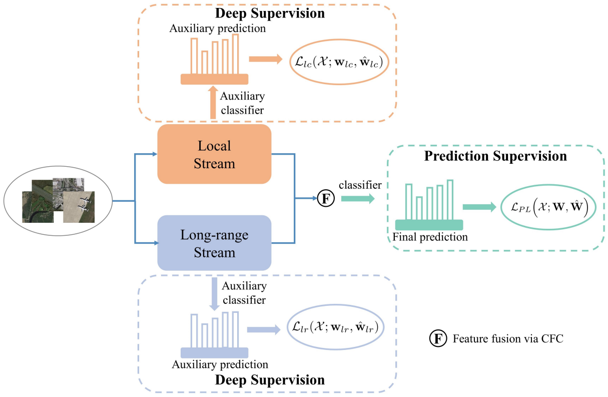


Fig. 4. Diagram of DS loss and prediction loss in L2RCF.

Let us consider the sample set $\mathcal{X} = \{\boldsymbol{x}_k\}_{k=1}^{K}$ with $K$ samples and $M$ classes, and the corresponding label set $\mathcal{Y} = \{y_k\}_{k=1}^{K}$, $y_k \in \{1, 2, \ldots, M\}$. The probability of the sample $\boldsymbol{x}_k$ belongs to class $m$ is computed as:

$$p(m \mid \boldsymbol{x}_k; \mathbf{w}_{md}) = \frac{exp(\mathbf{v}_m)}{\sum_{m=1}^{M} exp(\mathbf{v}_m)}, \tag{9}$$

where $\mathbf{w}_{md}$ is the model parameter and $\mathbf{v}$ is the output of the model.

We use the cross-entropy loss to represent the error between the prediction and the label. It is defined as follows:

$$\mathcal{L}_{CE} = -\sum_{k=1}^{K}\sum_{m=1}^{M} I\{\bullet\}\log\left(p(m \mid \boldsymbol{x}_k; \mathbf{w}_{md})\right), \tag{10}$$

where $I\{\bullet\}$ is an indicator function, defined as:

$$I\{\bullet\} = \begin{cases} 1 & y_k = m \\ 0 & y_k \neq m, \end{cases} \tag{11}$$

and $y_k$ is the label of the sample $\boldsymbol{x}_k$.

Therefore, the DS loss based on the two auxiliary classifiers is computed as follows:

$$\mathcal{L}_{lc}(\mathcal{X};\mathbf{w}_{lc},\hat{\mathbf{w}}_{lc}) = -\sum_{k=1}^{K}\sum_{m=1}^{M} I\{\bullet\}\log(p(m|\boldsymbol{x}_k;\mathbf{w}_{lc},\hat{\mathbf{w}}_{lc})), \tag{12}$$

$$\mathcal{L}_{lr}(\mathcal{X};\mathbf{w}_{lr},\hat{\mathbf{w}}_{lr}) = -\sum_{k=1}^{K}\sum_{m=1}^{M} I\{\bullet\}\log(p(m|\boldsymbol{x}_k;\mathbf{w}_{lr},\hat{\mathbf{w}}_{lr})), \tag{13}$$

where $\mathcal{L}_{lc}$ and $\mathcal{L}_{lr}$ represent the DS losses for local and long-range streams respectively, $\mathbf{w}_{lc}$ and $\mathbf{w}_{lr}$ represent the local stream and long-range stream parameters respectively, $\hat{\mathbf{w}}_{lr}$ and $\hat{\mathbf{w}}_{lr}$ represent the parameters which bridge the features to predictions. And the DS loss $\mathcal{L}_{DS}$ can be represented as:

$$\mathcal{L}_{DS} = (\mathcal{L}_{lc} + \mathcal{L}_{lr})/2. \tag{14}$$

### 2) DML loss

DML comes from knowledge distillation [55], which distills the knowledge contained in the teacher network into the student network. In knowledge distillation, the teacher network is pre-trained and tends to be larger than the student network. The gradient generated by the mimicry loss between the teacher network and the student network is only back-propagated to the student network, as shown in Fig. 5 (a). In other words, the teacher network has fixed weights.

In DML, two or more networks learn from each other without teacher network. The gradient from mimicry loss between them is back-propagated to all networks, as shown Fig. 5 (b). We extended DML into a single network, as shown in Fig. 5 (c). Local, long-range, fusion streams are viewed as sub-networks learning from each other.

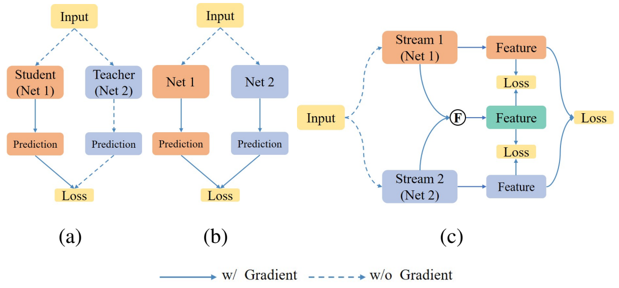


Fig. 5. Diagram of knowledge distillation and DML. (a) Knowledge distillation. (b) DML proposed in [56]. (c) DML proposed in L2RCF. 'w/ Gradient' means 'with gradient', and 'w/o Gradient' means 'without gradient'.

To compute the mimicry loss of the features generated by the three 'networks', we use the $L2$ loss:

$$\mathcal{L}_2 = \|\mathbf{z}_1 - \mathbf{z}_2\|^2, \tag{15}$$

where $\mathbf{z}_1$ and $\mathbf{z}_2$ are the features generated by the different 'networks'.

Thus the DML loss of local stream and long-range stream with fusion stream can be represented as:

$$\mathcal{L}_{lc-fs}(\mathcal{X};\mathbf{w}_{lc},\mathbf{W}) = \|\mathbf{z}_{lc} - \mathbf{z}_{fs}\|^2, \tag{16}$$

$$\mathcal{L}_{lr-fs}(\mathcal{X};\mathbf{w}_{lr},\mathbf{W}) = \|\mathbf{z}_{lr} - \mathbf{z}_{fs}\|^2, \tag{17}$$

where $\mathbf{z}_{fs}$ represents the fusion features, $\mathbf{W}$ represent the parameters of the dual-stream and CFC modules. Similarly, DML loss between the local stream and long-range stream can be represented as:

$$\mathcal{L}_{lc-lr}(\mathcal{X};\mathbf{w}_{lc},\mathbf{w}_{lr}) = \|\mathbf{z}_{lc} - \mathbf{z}_{lr}\|^2. \tag{18}$$

In this way, the three features can learn from each other to mitigate their shortcomings and improve the representation of the framework. And the DML loss $\mathcal{L}_{DML}$ can be represented as:

$$\mathcal{L}_{DML} = (\mathcal{L}_{lc-fs} + \mathcal{L}_{lr-fs} + \mathcal{L}_{lc-lr})/3. \tag{19}$$

### 3) Loss Function

In summary, there are three types of loss functions in our network (DS loss, DML loss, and prediction loss). Among them, the DS loss and prediction loss employ the cross entropy loss. DML loss employs the $L2$ loss. So the prediction loss $\mathcal{L}_{PL}$ can be represented as:

$$\mathcal{L}_{PL}\left(\mathcal{X};\mathbf{W},\hat{\mathbf{W}}\right) = -\sum_{k=1}^{K}\sum_{m=1}^{M} I\{\bullet\}\log\left(p(m|\boldsymbol{x}_k;\mathbf{W},\hat{\mathbf{W}})\right), \tag{20}$$

where $\hat{\mathbf{W}}$ represent the parameters of the fusion feature-based classifier. The joint loss $\mathcal{L}$ is the average value of them and is represented as follows:

$$\mathcal{L} = (\mathcal{L}_{PL} + \mathcal{L}_{DS} + \mathcal{L}_{DML})/3, \tag{21}$$

that is, $\mathcal{L}_{PL}$, $\mathcal{L}_{DS}$ and $\mathcal{L}_{DML}$ are set to equal weight.

## E. Two-stage Semi-supervised Training

As described in previous subsections, local features, long-range features, and fusion features are obtained. And three predictions are derived based on them. This enables the use of semi-supervised strategy that can exploit the large quantity of unlabeled. The semi-supervised paradigm can effectively extract the knowledge hidden in a large number of unlabeled samples.

In semi-supervised learning, it is crucial to obtain pseudo-labels of unlabeled samples. We design a simple self-labeling strategy based on three classifiers for assigning pseudo-labels to unlabeled samples that may be added to the enlarged sample set $\mathcal{S}_{el}$ with high confidence.

Mathematically, given an unlabeled sample $\boldsymbol{x}_u \in \mathcal{U}$, its predicted labels and class scores are first obtained using different classifiers:

$$y_n, p_n = f_n(\boldsymbol{x}_u), \tag{22}$$

where $y_n$ and $p_n$ are the predicted labels and class scores, and $n \in \{local, long-range, fusion\}$.

Then need to determine whether predcitions meet the criteria for a high confidence. We check it the different classifiers are consistent, i.e. they provide the same predicted labels based on local features, long-range features, and fusion features.

In addition, we set that the three class scores need to satisfy the following criterion:

$$\min(p_{local}, p_{long-range}, p_{fusion}) \geq \lambda, \tag{23}$$

where $\lambda$ is a constant, $p_{local}$, $p_{long-range}$, and $p_{fusion}$ represent the class scores based on local features, long-range features, and fusion features, respectively. Samples in $\mathcal{U}$ satisfying are added to the enlarged sample set $\mathcal{S}_{el}$.

We design a two-stage semi-supervised training strategy, as shown in Fig. 2. In the first training stage, the joint loss $\mathcal{L}$ is set to contain DS loss and prediction loss. The labeled sample set $S_{lb}$ is fed into the model for training, and three initial classifiers are obtained. Further, the enlarged sample set $S_{el}$ is obtained based on the designed self-labeling strategy. In the second training stage, the joint loss $\mathcal{L}$ is set to contain DS loss, DML loss and prediction loss. The $S_{lb}$ and $S_{el}$ are given as input to the model training. In addition, local and long-range classifiers are pruned to reduce the size of the inference model.

## IV. Experimental Results

### A. Data Sets Description

we analyzed the experiment of the proposed L2RCF on the three data sets described below.

*1) RSSCN7 data set:* It contains 2800 images from 7 categories, with the size of 400 × 400 pixels. Each category contains four different scales (1:700, 1:1300, 1:2600, and 1:5200) and 100 images. And images were collected at a variety of times, weather and scales [57].

*2) AID data set:* It contains 10,000 images from 30 categories, with the size of 600 × 600 pixels. The number of samples for each category is different, ranging between 220 and 400. Images are all obtained from Google Images with defferent sensors and in many countries under different imaging conditions. And the resolutions are between 0.5 and 8 m [58].

*3) NWPU data set:* It contains 31,500 images from 45 categories with the image size of 256 × 256 pixels. The number of images for each category is 700. The acquisition area is widely distributed and also has variable imaging conditions with a resolution range between 0.2 and 30 m. Compared to the first two data sets, it has more categories and samples, and is more challenging [59].

### B. Evaluation Metrics

The quantitative measures of accuracy used in the experiment are: i) overall accuracy (OA), which is an indicator that reflects the overall performance of the model and ii) the confusion matrix (CM), which is used to assess the accuracy of different categories and the degree of confusion between them. Floating point operations (FLOPs) and model size are also used to measure computational complexity and model complexity. In addition, GradCAM [60] and T-SNE are considered to visualize and analyze the experimental results.

TABLE I
COMPARSION OF OAs (%) PROVIDED BY THE PROPOSED AND SOME STATE-OF-THE-ART METHODS (RSSCN7 DATA SET)

| Type | Method | RSSCN7 | |
|---|---|---|---|
| | | 20% | 50% |
| △ | BoVW(SIFT) [58] | 76.33±0.88 | 81.34±0.55 |
| | BoVW(LBP) [58] | 76.98±0.90 | 81.69±1.11 |
| ○ | Tex-Net-LF [61] | 92.45±0.45 | 94.00±0.57 |
| | SE-MDPMNet [62] | 92.65±0.13 | 94.71±0.15 |
| | Contourlet CNN [63] | - | 95.54±0.71 |
| Proposed | L2RCF-18-T | 94.22±0.72 | 95.56±0.30 |
| | L2RCF-18-S | 94.50±0.71 | 95.70±0.38 |
| | L2RCF-34-T | 94.34±0.45 | 95.64±0.38 |
| | L2RCF-34-S | 94.47±0.62 | 95.93±0.44 |
| | L2RCF-50-T | 94.42±0.45 | 95.94±0.50 |
| | L2RCF-50-S | **94.70±0.55** | **96.00±0.12** |

### C. Implementation Details

We chose ResNet (ResNet18, ResNet34, ResNet50) and DeiT (DeiT-T, DeiT-S) as the backbones for the local stream and the long-range stream, respectively. So L2RCF-18-T, L2RCF-18-S, L2RCF-34-T, L2RCF-34-S, L2RCF-50-T, L2RCF-50-S are formed. The image of all three data sets were resized to 224 × 224 pixels. Two data enhancement strategies, random vertical flipping and horizontal flipping, were used to prevent overfitting. Both $r$ and $L_{dn}$ in Equation (12) were set to 32. The value of $\lambda$ in the semi-supervised strategy is set to 0.6. The training ratios (Trs) for RSSCN7 and AID were set to 20% and 50%, and for NWPU to 10% and 20%. In addition, to enhance the confidence of the results, all experiments were repeated five times and the training samples were randomly selected.

SDG was used as the model optimizer. The total number of training epochs was 60 with a learning rate of 0.01 for the first 30 epochs and 0.001 for the last 30 epochs. The batch size was set to 32. Backbones were initialized using ImageNet-based pre-trained parameters. In addition, the experiments were implemented by Pytorch in the computing environment of Intel i9-10980XE CPU, NVIDIA RTX 3090 Graphics Card, and 64-GB memory.

### D. Comparison With State-of-the-Art Approaches

To effectively evaluate the performance of L2RCF, some state-of-the-art methods are used for comparison on three considered data sets. The handcrafted feature-based methods (△), CNN-based methods (○), VT-based methods (∗) and CNN-VT-based methods (⊛) are shown separately in Table I and Table II. Due to the limited application of VTs in RS, CNN-based methods present more in the comparison.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3265346

TABLE II
COMPARSION OF OAS (%) PROVIDED BY THE PROPOSED AND SOME STATE-OF-THE-ART METHODS (AID AND NWPU DATA SETS)

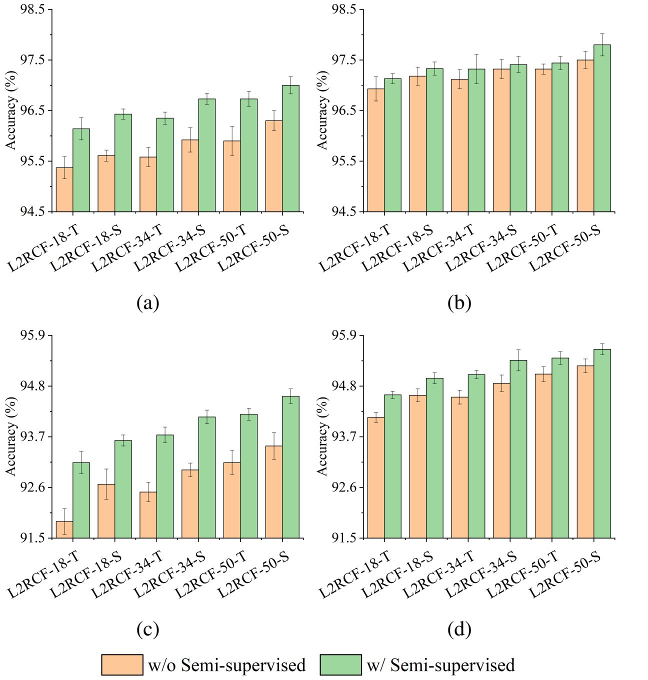
| Type | Method | AID | | NWPU | |
|---|---|---|---|---|---|
| | | 20% | 50% | 10% | 20% |
| △ | GIST [58], [59] | 30.61±0.63 | 35.07±0.41 | 15.90±0.23 | 17.88±0.22 |
| | BoVW(SIFT) [58], [59] | 62.49±0.53 | 68.37±0.40 | 41.72±0.21 | 44.97±0.28 |
| ○ | SCCoV [64] | 93.12±0.25 | 96.10±0.16 | 89.30±0.35 | 92.10±0.25 |
| | MG-CAP [65] | 93.34±0.18 | 96.12±0.12 | 90.83±0.12 | 92.95±0.11 |
| | ResNet101+CBAM [66] | 93.51±0.22 | 96.56±0.21 | 91.63±0.15 | 93.86±0.13 |
| | ResNet50+EAM [66] | 93.64±0.25 | 96.62±0.13 | 90.87±0.15 | 93.51±0.12 |
| | ResNet18$_{\text{local+global}}$ [67] | 94.38±0.10 | 96.76±0.20 | 90.04±0.15 | 92.79±0.11 |
| | ResNet50-FSoI-Net2 [68] | 95.49±0.31 | 97.16±0.07 | 92.49±0.31 | 94.40±0.21 |
| | MBLANet [69] | 95.60±0.17 | 97.14±0.03 | 92.32±0.15 | 94.66±0.11 |
| | GCSANet [70] | 95.96±0.38 | 97.53±0.32 | 93.39±0.39 | 94.95±0.36 |
| ∗ | V16_21k [224 × 224] [47] | 94.97±0.01 | - | 92.60±0.10 | - |
| | SCViT [49] | 95.56±0.17 | 96.98±0.16 | 92.72±0.04 | 94.66±0.10 |
| | HHTL [51] | 95.62±0.13 | 96.88±0.21 | 92.07±0.44 | 94.21±0.09 |
| | V16_21k [384 × 384] [47] | 95.86±0.28 | - | 93.83±0.46 | - |
| | C$^2$-CapsViT [52] | 96.05±0.11 | 97.57±0.15 | 93.32±0.05 | 95.28±0.08 |
| ⊛ | EMTCAL [50] | 94.69±0.14 | 96.41±0.23 | 91.63±0.19 | 93.65±0.12 |
| | CTNet(MobileNet_v2-ViT_B) [10] | 96.25±0.10 | 97.70±0.11 | 93.90±0.14 | 95.40±0.15 |
| | CTNet(ResNet34-ViT_B) [10] | 96.35±0.13 | 97.56±0.20 | 93.86±0.22 | 95.49±0.12 |
| Proposed | L2RCF-18-T | 96.14±0.22 | 97.13±0.10 | 93.14±0.24 | 94.61±0.08 |
| | L2RCF-18-S | 96.43±0.10 | 97.33±0.13 | 93.62±0.12 | 94.97±0.12 |
| | L2RCF-34-T | 96.35±0.12 | 97.32±0.29 | 93.74±0.17 | 95.05±0.09 |
| | L2RCF-34-S | 96.73±0.11 | 97.41±0.16 | 94.13±0.15 | 95.36±0.23 |
| | L2RCF-50-T | 96.73±0.15 | 97.44±0.13 | 94.19±0.13 | 95.41±0.14 |
| | L2RCF-50-S | **97.00±0.17** | **97.80±0.22** | **94.58±0.16** | **95.60±0.12** |

*1) RSSCN7 data set:* For the RSSCN7 data set, Table I shows the comparisons between L2RCF and some state-of-the-art methods under 20% and 50% Trs. To our best knowledge, VT-based methods and CNN-VT-based methods have not been used for this data set, so the results of the comparisons of handcrafted feature-based methods and CNN-based methods are presented in Table I. The accuracies of CNN-based methods are much better than those of handcrafted feature-based methods. CNN-based methods achieve 92.65% and 95.54% OAs under 20% and 50% Trs, respectively. While the best OAs of handcrafted feature-based methods are only 76.98% and 81.69%. Note that Tex-Net-LF [61] encodes CNN features, and SE-MDPMNet [62] uses atrous convolution and channel attention mechanism based on MobileNet-V2. Contourlet CNN [63] incorporates spectral features and statistical information on the basis of spatial features. However, their accuracies are far lower than those of L2RCF. Under different backbone settings, L2RCF achieves 94.22%-94.70% and 95.56%-96.00% OAs under 20% and 50% Trs, respectively.

*2) AID data set:* Table II shows the results of the comparisons with some state-of-the-art methods on the AID data set. The handcrafted feature-based methods include the direct use of low-level features (GIST) and the encoding low-level features to form mid-level features (BoVW). We can find that the mid-level feature methods perform much better than the low-level feature methods. CNN-based methods achieve OAs of 95.96% and 97.53% under 20% and 50% Trs, respectively. SCCoV [64] exploits second-order information in multi-resolution feature maps based on covariance pooling, and MG-CAP [65] mines latent ontologies of RS images based on multi-granularity canonical appearance pooling. ResNet50-FSoI-Net2 [68] exploits self-attention-based second-order pooling to obtain second-order information. ResNet18$_{\text{local+global}}$ [67] constructs a dual-stream network to extract local and global features of images respectively. ResNet101+CBAM [66], ResNet50+EAM [66], MBLANet [69], GCSANet [70] all use the attention mechanism, which can effectively improve the performance of CNNs. Therefore, these selected CNN-based methods are very representative. The V16_21k [47] achieves high accuracy based on ViT using transfer learning for RS scene classification. SCViT [49], HHTL [51] and C$^2$-CapsViT [52] are improved on the basis of ViT [8] for RS scene classification tasks. CTNet [10] builds a dual-stream network of CNN and VT through concatenation fusion. EMTCAL [50] uses transformer to obtain context information based on multi-level convolution features. It is more lightweight than the

VT-based method mentioned above, but the computational complexity is still higher than that of L2RCF-18-T (4.23G vs 2.89G). L2RCF achieves 96.14%-97.00% and 97.56%-97.80% OAs under 20% and 50% Trs, respectively. Under 20% Tr, L2RCF significantly outperforms these state-of-the-art methods. Under 50% Tr, L2RCF performs significantly better than handcrafted feature-based methods, CNN-based methods, VT-based methods, and slightly better than the CNN-VT-based method (CTNet). And, the model and computational complexity of the VT backbone (ViT_B [8]) used in CTNet, SCViT, HHTL and C$^2$-CapsViT is much higher than that of L2RCF-50-S.

*3) NWPU data set:* The results of the comparisons between L2RCF and some state-of-the-art methods on the NWPU data set are shown in Table II. The NWPU data set is one of the most complex RS scene classification task data set that is widely used, with large-scale and variety of categories. Under 10% and 20% Trs, the OAs of the proposed L2RCF reached 93.14%-94.58% and 94.61%-95.60%, respectively. It is worth nothing that the proposed L2RCF improves more the OAs of existing methods at lower Trs, so L2RCF performs better in the case of fewer samples. In order to fully explore the performance of L2RCF in the case of fewer labeled samples, we conduct more experiments (see Section IV-F for details).



(a) (b)

(c) (d)

Fig. 6. Comparison results with the two-stage semi-supervised training on different data sets. (a) AID under Tr=20%. (b) AID under Tr=50%. (c) NWPU under Tr=10%. (d) NWPU under Tr=20%

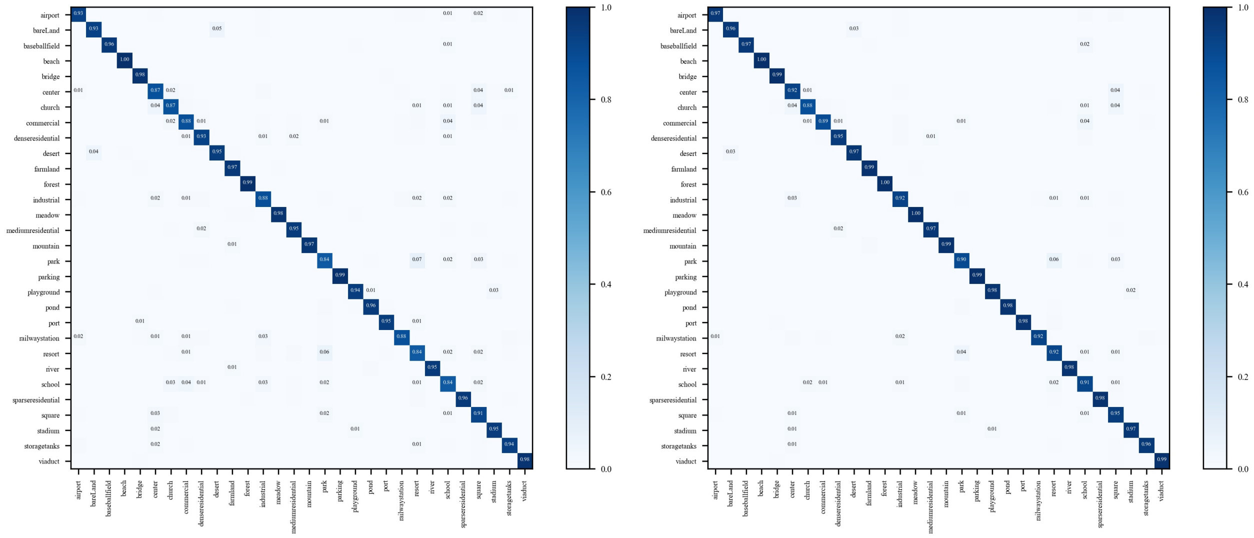
## *E. Ablation Experiment*

The proposed L2RCF in this paper mainly includes three strategies, namely CFC, joint loss and two-stage semi-supervised training strategy. In order to illustrate the role of each strategy, we conducted very complete ablation experiments.

*1) Effect of CFC:* In this ablation experiment, we tested the performance on RSSCN7 and NWPU data sets at 20% and 10% Trs, respectively, based on all backbones (ResNet18, ResNet34, ResNet50, DeiT-T and DeiT-S). We take direct concatenation of local features and long-range features as the baseline for this study, since concatenation is the most common method for dual-stream network fusion. Therefore the model performance based on concatenation and CFC is tested, as shown in Table III. Note that the joint loss and the semi-supervised strategy are not used for all experimental settings. First of all, DeiT-T and ResNet50 in the backbone show the worst and best performance, respectively, which is in line with their own model and computational complexity. Concatenation can improve accuracy compared to the backbone with better performance in the dual stream. Furthermore, CFC outperforms concatenation in all experimental settings. These results show that concatenation can improve the classification accuracy compared to the backbones, and CFC can further significantly improve the accuracy on the basis of concatenation, proving the effectiveness of CFC.

*2) Effect of Joint Loss:* Besides the effectiveness of the joint loss, the effectiveness of DS and DML is also fully explored in this ablation experiment. Therefore, the performance of different combinations of DS and DML is fully tested based on two data sets (AID, NWPU) and four backbones (ResNet18, ResNet50, DieT-T, DieT-S), as shown in Table IV. Note that CFC is used for all experimental settings. In all experiments, models using only DS or DML also show significant improvement. This illustrates the effectiveness of DS and DML, respectively. In addition, although using both DS and DML performs better than using alone, the improvement is not significant. On the one hand, DML can also provide DS information. Different from DS, DML uses the supervision information of fusion features, which can play a role similar to ‘soft labels’. However, using only DML introduce some misinformation from fusion features. On the other hand, although the supervision information of DS is completely correct, it uses hard labels and does not perform as well as soft labels. Therefore, they can provide some complementary information to each other. In other words, this ablation experiment not only fully demonstrates the effectiveness of the joint loss, but also illustrates the complementarity between DS and DML.

*3) Effect of Two-Stage Semi-supervised Training:* We further exploit the two-stage semi-supervised training strategy. The performance comparisons of L2RCF w/o SF (without semi-supervised strategy) and L2RCF w/ SF (with semi-supervised strategy) on the AID and NWPU data sets are shown in Fig. 6. L2RCF w/ SF performs better in all experiments. Furthermore, the improvement tends to be more significant at lower Trs. For example, in the experiment on the AID data set, the improvement can reach a maximum of 0.83% under 20% Tr, while it can only reach a maximum of 0.2% under 50% Tr. When using the NWPU data set, the improvement can reach up to 1.28% under 10% Tr, which is much higher than other experimental settings. When the Tr is higher, the number of enlarged samples brought by self-



(a)

(b)

Fig. 7. CM obtained by different models on the AID data set under Tr=20%. (a) Concat-18-T. (b) L2RCF-18-T.

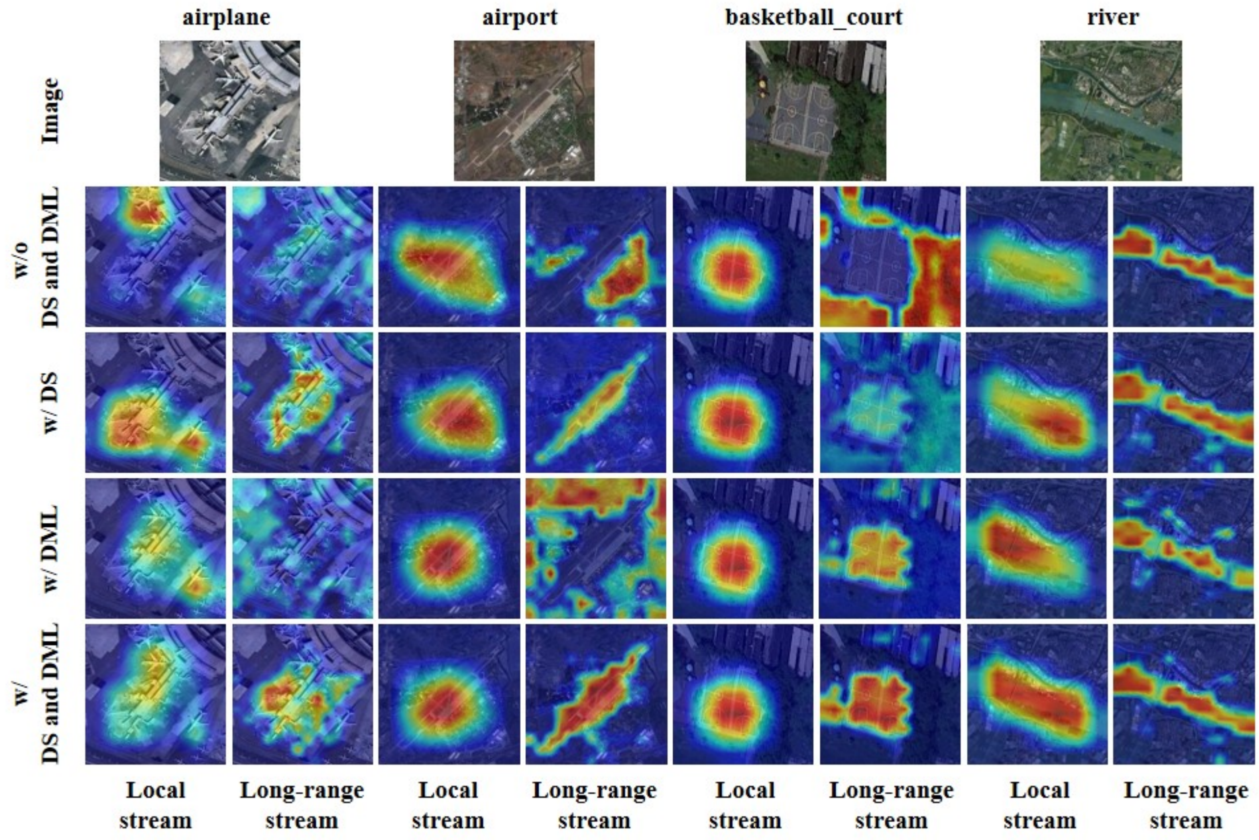


Fig. 8. Visualization of the class attention maps. Four images were randomly selected from the NWPU data set. From top to bottom, the original image, class attention map obtained without DS and DML, class attention map obtained with DS, class attention map obtained with DML, and class attention map obtained with DS and DML are represented, respectively.

TABLE III
OAs(%) OF ABLATION COMPARSION EXPERIMENTS WITH CFC (RSSCN7 AND NWPU DATA SETS)

| Backbone | | Data sets | Local backbone | Long-range backbone | w/ Concat | w/ CFC |
|---|---|---|---|---|---|---|
| CNN | VT | | | | | |
| ResNet18 | DieT-T | RSSCN7 | 90.21±0.98 | 89.79±0.61 | 91.30±0.86 | 91.88±1.10 (0.58↑) |
| | | NWPU | 89.11±0.13 | 86.29±0.22 | 89.41±0.23 | 89.93±0.31 (0.50↑) |
| ResNet18 | DieT-S | RSSCN7 | 90.21±0.98 | 91.45±0.94 | 91.53±0.67 | 92.26±0.54 (0.73↑) |
| | | NWPU | 89.11±0.13 | 90.02±0.34 | 90.36±0.27 | 90.90±0.23 (0.54↑) |
| ResNet34 | DieT-T | RSSCN7 | 90.33±0.49 | 89.79±0.61 | 91.51±0.53 | 92.02±0.56 (0.51↑) |
| | | NWPU | 88.97±0.24 | 86.29±0.22 | 89.44±0.29 | 90.01±0.17 (0.57↑) |
| ResNet34 | DieT-S | RSSCN7 | 90.33±0.49 | 91.45±0.94 | 91.88±0.97 | 92.44±0.47 (0.56↑) |
| | | NWPU | 88.97±0.24 | 90.02±0.34 | 90.43±0.21 | 90.95±0.25 (0.52↑) |
| ResNet50 | DieT-T | RSSCN7 | 91.79±0.82 | 89.79±0.61 | 91.84±0.73 | 92.36±0.55 (0.52↑) |
| | | NWPU | 90.82±0.29 | 86.29±0.22 | 90.62±0.25 | 91.11±0.13 (0.49↑) |
| ResNet50 | DieT-S | RSSCN7 | 91.79±0.82 | 91.45±0.94 | 92.19±0.65 | 92.99±0.68 (0.80↑) |
| | | NWPU | 90.82±0.29 | 90.02±0.34 | 91.31±0.15 | 91.82±0.14 (0.51↑) |

TABLE IV
OAs(%) OF ABLATION COMPARSION EXPERIMENTS WITH JOINT LOSS (AID AND NWPU DATA SETS)

| Backbone | | DS | DML | AID | | NWPU | |
|---|---|---|---|---|---|---|---|
| CNN | VT | | | 20% | 50% | 10% | 20% |
| ResNet18 | Deit-T | | | 94.13±0.39 | 96.15±0.14 | 89.93±0.31 | 92.69±0.18 |
| | | ✓ | | 95.18±0.21 (1.05↑) | 96.81±0.24 (0.66↑) | 91.44±0.37 (1.51↑) | 93.80±0.19 (1.11↑) |
| | | | ✓ | 95.08±0.25 (0.95↑) | 96.75±0.21 (0.60↑) | 91.77±0.32 (1.84↑) | 93.86±0.09 (1.17↑) |
| | | ✓ | ✓ | 95.37±0.22 (1.24↑) | 96.93±0.24 (0.78↑) | 91.86±0.28 (1.93↑) | 94.12±0.11 (1.43↑) |
| ResNet18 | Deit-S | | | 94.97±0.23 | 96.48±0.30 | 90.90±0.23 | 93.29±0.17 |
| | | ✓ | | 95.55±0.11 (0.58↑) | 97.03±0.26 (0.55↑) | 92.25±0.17 (1.35↑) | 94.37±0.15 (1.08↑) |
| | | | ✓ | 95.48±0.19 (0.51↑) | 96.99±0.16 (0.51↑) | 92.26±0.21 (1.36↑) | 94.38±0.13 (1.09↑) |
| | | ✓ | ✓ | 95.61±0.11 (0.64↑) | 97.18±0.18 (0.70↑) | 92.67±0.33 (1.77↑) | 94.60±0.14 (1.31↑) |
| ResNet50 | Deit-T | | | 94.96±0.31 | 96.68±0.20 | 91.11±0.13 | 93.57±0.08 |
| | | ✓ | | 95.78±0.15 (0.82↑) | 97.18±0.22 (0.50↑) | 92.53±0.25 (1.42↑) | 94.73±0.10 (1.16↑) |
| | | | ✓ | 95.76±0.20 (0.80↑) | 97.29±0.24 (0.61↑) | 93.03±0.19 (1.92↑) | 94.94±0.21 (1.37↑) |
| | | ✓ | ✓ | 95.90±0.29 (0.94↑) | 97.32±0.10 (0.64↑) | 93.14±0.26 (2.03↑) | 95.06±0.16 (1.49↑) |
| ResNet50 | Deit-S | | | 95.46±0.19 | 96.96±0.24 | 91.82±0.14 | 93.90±0.10 |
| | | ✓ | | 96.23±0.12 (0.77↑) | 97.44±0.24 (0.48↑) | 93.08±0.25 (1.26↑) | 95.03±0.15 (1.13↑) |
| | | | ✓ | 96.01±0.13 (0.55↑) | 97.41±0.17 (0.45↑) | 93.41±0.25 (1.59↑) | 95.17±0.13 (1.27↑) |
| | | ✓ | ✓ | 96.30±0.20 (0.84↑) | 97.50±0.17 (0.54↑) | 93.50±0.29 (1.68↑) | 95.24±0.15 (1.34↑) |

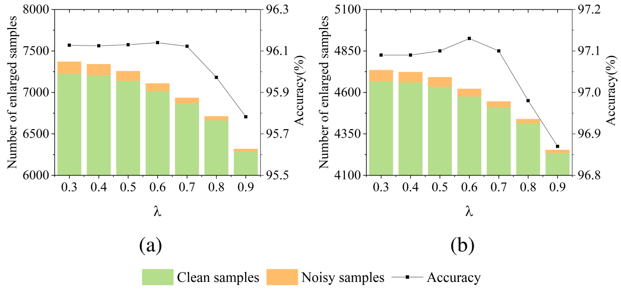
labeling tends to be lower. And RS data can often be obtained in large quantities, but their labeling is often manual and time-consuming. Therefore, labeled samples are often lacking in various RS tasks. This ablation experiment demonstrates the effectiveness of the proposed two-stage semi-supervised training strategy, especially in the few labeled sample case.

### *F. Visualization and Analysis*

*1) CM Analysis:* In order to show the accuracy of each category and the confusion between categories, we compare the CMs of the baseline (Concat-18-T) and the proposed L2RCF-18-T, as shown in Fig. 7. Experiments are performed on the AID data set under 20% Tr. The CM of Concat-18-T shows 8 categories with accuracy below 90%, while the CM of L2RCF-18-T shows only two categories with accuracy below 90%. Compared with Concat-18-T, L2RCF-

TABLE V
ANALYSIS OF THE OAS (%) AND PARAMETER SIZE (MB) ON THE AID DATA SET WITH DIFFERENT REDUCTION RATIO $r$

| $r$ | AID |  | Params |
|---|---|---|---|
|  | 20% | 50% |  |
| 8 | 95.23 | 96.74 | 46.65 |
| 16 | 95.28 | 96.85 | 45.91 |
| 32 | 95.46 | 96.96 | 45.54 |
| 64 | 95.34 | 96.81 | 45.36 |

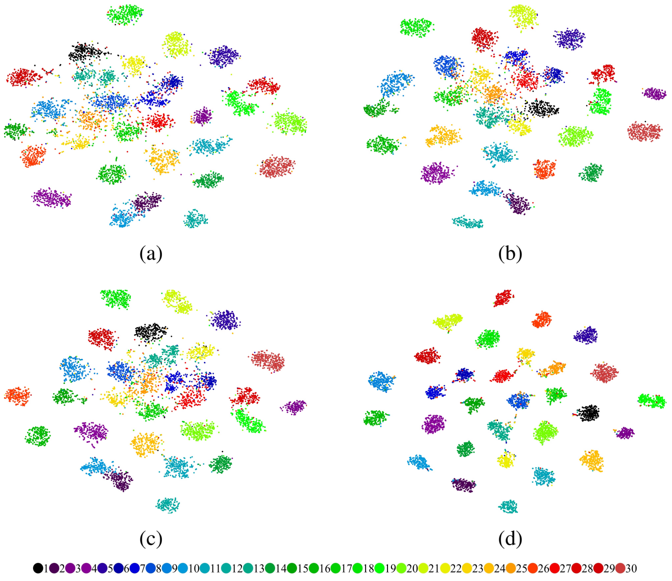


(a) (b)

Fig. 9. Analysis of the OAs and the samples in $\mathcal{S}_{el}$ on the AID data set with $\lambda$ ranging from 0.3 to 0.9. (a) Tr=20%. (b) Tr=50%.

18-T improves the accuracy of 'resort' from 84% to 92%, the accuracy of 'school' from 84% to 91%, and the accuracy of 'park' from 84% to 90%. Furthermore, we can find that the proportion of L2RCF-18-T misclassifying 'school' into 'church', 'commercial', 'industrial', 'park' and 'square' is reduced by 1%-2%. To sum up, L2RCF can effectively reduce the confusion between categories, which proves the effectiveness of the proposed method.

*2) Hyperparameter Analysis:* In this study, we conduct experiments to investigate the impact of hyperparameters $r$ and $\lambda$. $r$ represents the reduction ratio in the CFC and is used to regulate capacity and model complexity. $\lambda$ represents the confidence constant in the two-step semi-supervised training. We evaluate the performance of a range of $r$ based on L2RCF-50-S and AID data set. The results of our experiments, including accuracy and parameter size, are presented in Table V. And we can find that the performance of the model is robust across a range of $r$. The optimal results are achieved when $r$ is set to 32, which only adds limited parameters compared to $r$=64. And we conduct comparative experiments on the L2RCF-18-T and AID data set with varying values of $\lambda$ between 0.3-0.9, with stride of 0.1. The Fig. 9 shows the change of the number of samples in $\mathcal{S}_{el}$ and accuracy with $\lambda$. The composition of clean and noisy samples in is also presented. As $\lambda$ increases, both the number of samples in $\mathcal{S}_{el}$ and the proportion of noise samples decrease. The performance of L2RCF-18-T remains robust when $\lambda$ is between 0.3-0.7, but drops significantly when $\lambda$ greater than 0.7. And the best performance is achieved by setting $\lambda$ to 0.6. These demonstrate that the proportion of noise samples and the number of samples in $\mathcal{S}_{el}$ both affect the performance.



(a) (b) (c) (d)

Fig. 10. Feature visualization by T-NSE of different methods on the AID data set under Tr=20%. (a) DeiT-T. (b) ResNet18. (c) Concat-18-T. (d) L2RCF-18-T.

*3) Joint Loss Analysis using Grad-CAM Visualization:* In order to better demonstrate the role of DS and DML for enhancing dual-stream feature extractors, we use Grad-CAM to visualize class attention maps based on local stream and long-range stream respectively, as shown in Fig. 8. Grad-CAM provides the spatial response related to a specific category according to the gradient, which helps to understand the area of concern of the model. We can find from Fig. 8 that DS and DML can effectively help local stream and long-range stream to identity key regions for airplanes and airports, combining the special feature extraction strengths of local stream and long-range stream. For ground-objects with obvious local features or long-range features, such as 'basketball_court' and 'river', DS and DML can help local stream and long-range stream to obtain long-range and local features, respectively. In summary, Fig. 8 combined with the ablation study fully demonstrates the effectiveness of DS and DML.

*4) Feature Visualization Analysis With T-SNE:* We use t-SNE, a linear dimensionality reduction method, to visualize feature separability. Experiments are based on the AID data set under 20% Tr. Fig. 10 shows the feature distribution based on DeiT-T, ResNet18, Concat-18-T, L2RCF-18-T. DeiT-T has the worst feature separability, and many categories are seriously confused together. This is caused by the fact that its model and computational complexity are much lower than those of the other three networks. The feature separability of ResNet18 and Concat-18-T are similar, both are better than that of Deit-T. However, some categories are still confused together. Compared with the other three types of networks, the feature separability of L2RCF-18-T is generally better than the backbone network (Deit-T, ResNet18) and the baseline (Concat-18-T). However, some samples are sporadically clustered into other categories and are close to the center of the wrong categories, which may be caused by the inevitable generation of wrongly self-labeled samples when using semi-supervised strategies.

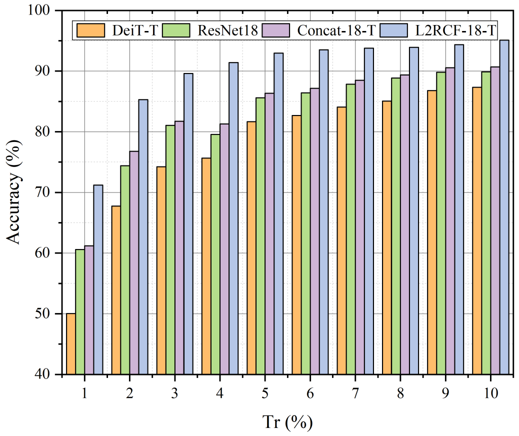


Fig. 11. Analysis of the OAs on the AID data set with Trs ranging from 1% to 10%.

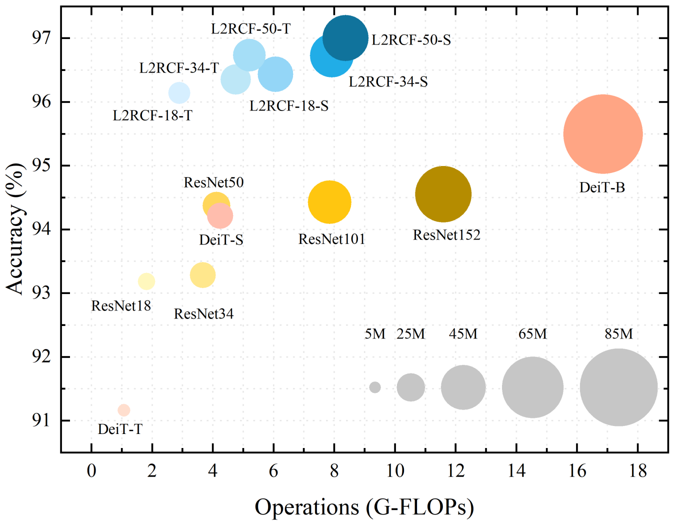


Fig. 12. Accuracy vs computational complexity and model complexity.

*5) Few Labeled Samples Analysis:* As described in Section IV-D, L2RCF improves more significantly under fewer labeled samples. Because of the massive amount of RS data and the difficulty in obtaining labels, few labeled samples is suitable for practical situations. To further explore the performance of L2RCF under fewer labeled samples, we set the Trs to 1%-10% and the stride to 1% on the AID data set. The results of L2RCF-18-T, the corresponding backbones and baseline are shown in Fig. 11. When Trs are less than 4%, the accuracy of all models improved significantly by increasing Trs. However, when the Tr is 4%, the accuracy growth rate of DeiT-T is significantly weakened, and the accuracy of ResNet18 and Concat-18-T is even reduced. This may be due to that the Tr is too low, resulting in greater randomness and unrepresentative training samples. The performance of L2RCF-18-T is the best because the semi-supervised strategy introduces more training samples and enhances the robustness. When the Trs are lower, the improvement of L2RCF is more significant. For example, under Trs of 1% and 10%, compared with the baseline, the improvement is 10.01% and 4.41%, respectively. Furthermore, L2RCF with 4% training samples outperforms the baseline with 10% training samples. Therefore, L2RCF can be effectively applied to the situation of insufficient labeled samples for RS tasks.

*6) Computational Complexity and Model Complexity Analysis:* The ball chart shown in Fig. 12 shows the relationship between the accuracy and the complexity of L2RCF, local network (ResNet), and long-range network (DeiT) on the AID data set under 20% Tr. The horizontal axis represents computational complexity, the vertical axis represents accuracy, and the size of the sphere represents model complexity. And L2RCF-50-S achieves the highest accuracy, but its computational complexity and model complexity are in the middle level. The highest accuracy among local networks and long-range networks is obtained by ResNet152 and DeiT-B, which is consistent with their computational complexity and model complexity performance. L2RCF-18-T achieved the lowest accuracy in L2RCF, but it is still far better than ResNet151 and DeiT-B. However, the computational complexity and model complexity of L2RCF-18-T are much lower than those of ResNet152 and DeiT-B. Overall, L2RCF is far superior to local networks and long-range networks in terms of accuracy, computational complexity and model complexity. In addition, within each series of L2RCF, local network, and long-range network, the accuracy has a strong correlation with computational complexity and model complexity.

## V. Conclusion

In this paper, we designed the L2RCF for the RS scene classification task to address the issue that CNNs and VTs are difficult to be sensitive to both local and long-range geo-objects. First, local and long-range streams are used to extract local and long-range features, respectively; then, they are efficiently fused by the designed CFC. Second, a joint loss combining DS and DML is proposed to enhance the dual-stream feature extractor and further improve the fused features. Finally, to fully expand the potential of unlabeled samples, we proposed a two-stage semi-supervised training strategy. Experiments on three widely used data sets show that our method is comparable to state-of-the-art methods. Extensive ablation experiments and visualization analysis demonstrate the effectiveness of each proposed strategy. L2RCF can perform better than CNNs and VTs with less computation and model complexity. Furthermore, compared with other methods, L2RCF is more suitable for the lack of labeled samples, which is very common in RS tasks. In the future, it is necessary to explore the interaction between CNNs and VTs in the middle layers, which may further effectively combine the advantages of the two approaches.

## References

[1] H. Zhao, S. Liu, Q. Du, L. Bruzzone, Y. Zheng, K. Du, X. Tong, H. Xie, and X. Ma, "Gcfnet: Global collaborative fusion network for multispectral and panchromatic image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

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

[3] D. Hu, Q. Meng, L. Zhang, and Y. Zhang, “Spatial quantitative analysis of the potential driving factors of land surface temperature in different “centers” of polycentric cities: A case study in tianjin, china,” *Science of The Total Environment*, vol. 706, p. 135244, 2020.

[4] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, “Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery,” *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 6, pp. 747–751, 2016.

[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017.

[6] C. Zhang, G. Li, and S. Du, “Multi-scale dense networks for hyperspectral remote sensing image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9201–9222, 2019.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” *arXiv preprint arXiv:2103.14030*, 2021.

[10] P. Deng, K. Xu, and H. Huang, “When cnns meet vision transformer: A joint framework for remote sensing scene classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[11] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” *Insights into imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[12] L. Bai, S. Du, X. Zhang, H. Wang, B. Liu, and S. Ouyang, “Domain adaptation for remote sensing image semantic segmentation: An integrated approach of contrastive learning and adversarial learning,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[13] K. Zhang, A. Wang, F. Zhang, W. Diao, J. Sun, and L. Bruzzone, “Spatial and spectral extraction network with adaptive feature fusion for pansharpening,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[14] L. Wang, L. Wang, Q. Wang, and L. Bruzzone, “Rscnet: A residual self-calibrated network for hyperspectral image change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[16] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[18] X. Huang, M. Dong, J. Li, and X. Guo, “A 3-d-swin transformer-based hierarchical contrastive learning method for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[19] L. Ding, D. Lin, S. Lin, J. Zhang, X. Cui, Y. Wang, H. Tang, and L. Bruzzone, “Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images,” *arXiv preprint arXiv:2106.15754*, 2021.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European Conference on Computer Vision*, pp. 213–229, Springer, 2020.

[21] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *International Conference on Machine Learning*, pp. 10347–10357, PMLR, 2021.

[22] K. van de Sande, T. Gevers, and C. Snoek, “Evaluating color descriptors for object and scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[23] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, pp. 886–893 vol. 1, 2005.

[24] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[26] E. Othman, Y. Bazi, N. Alajlan, H. Alhichri, and F. Melgani, “Using convolutional features and a sparse autoencoder for land-use scene classification,” *International Journal of Remote Sensing*, vol. 37, no. 10, pp. 2149–2167, 2016.

[27] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, “Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery,” *Remote Sensing*, vol. 7, no. 11, pp. 14680–14707, 2015.

[28] Y. Liu, Y. Zhong, and Q. Qin, “Scene classification based on multiscale convolutional neural network,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 7109–7121, 2018.

[29] X. Lu, H. Sun, and X. Zheng, “A feature aggregation convolutional neural network for remote sensing scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7894–7906, 2019.

[30] Q. Wang, S. Liu, J. Chanussot, and X. Li, “Scene classification with recurrent attention of vhr remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1155–1167, 2018.

[31] Q. Bi, K. Qin, H. Zhang, J. Xie, Z. Li, and K. Xu, “Apdc-net: Attention pooling-based convolutional network for aerial scene classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 9, pp. 1603–1607, 2020.

[32] B. Li, Y. Guo, J. Yang, L. Wang, Y. Wang, and W. An, “Gated recurrent multiattention network for vhr remote sensing image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[33] 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 Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[34] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.

[35] B. Zoph and Q. V. Le, “Neural architecture search with reinforcement learning,” *arXiv preprint arXiv:1611.01578*, 2016.

[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.

[37] Y. Lv, X. Zhang, W. Xiong, Y. Cui, and M. Cai, “An end-to-end local-global-fusion feature extraction network for remote sensing image scene classification,” *Remote Sensing*, vol. 11, no. 24, p. 3006, 2019.

[38] X. Chen, X. Zheng, Y. Zhang, and X. Lu, “Remote sensing scene classification by local–global mutual learning,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[39] D. Ye, J. Peng, H. Li, and L. Bruzzone, “Better memorization, better recall: A lifelong learning framework for remote sensing image scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[40] Y. Li, D. Kong, Y. Zhang, Y. Tan, and L. Chen, “Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 179, pp. 145–158, 2021.

[41] Y. Li, Y. Zhang, and Z. Zhu, “Error-tolerant deep learning for remote sensing image scene classification,” *IEEE transactions on cybernetics*, vol. 51, no. 4, pp. 1756–1768, 2020.

[42] 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 Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[43] J. Wang, Y. Zhong, Z. Zheng, A. Ma, and L. Zhang, “Rsnet: The search for remote sensing deep neural networks in recognition tasks,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2520–2534, 2020.

[44] A. Ma, Y. Wan, Y. Zhong, J. Wang, and L. Zhang, “Scenenet: Remote sensing scene classification deep learning network using multi-objective neural evolution architecture search,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 172, pp. 171–188, 2021.

[45] Y. Yu, X. Li, and F. Liu, “Attention gans: Unsupervised deep feature learning for aerial scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 519–531, 2019.

[46] G. Cheng, X. Sun, K. Li, L. Guo, and J. Han, “Perturbation-seeking generative adversarial networks: A defense framework for remote

sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[47] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing*, vol. 13, no. 3, p. 516, 2021.

[48] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, "A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–9, 2022.

[49] 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 Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[50] 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 Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[51] J. Ma, M. Li, X. Tang, X. Zhang, F. Liu, and L. Jiao, "Homo–heterogenous transformer learning framework for rs scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2223–2239, 2022.

[52] Y. Yu, Y. Li, J. Wang, H. Guan, F. Li, S. Xiao, E. Tang, and X. Ding, "$C^2$-capsvit: Cross-context and cross-scale capsule vision transformers for remote sensing image scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[53] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[54] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[55] G. Hinton, O. Vinyals, J. Dean, *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[56] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4320–4328, 2018.

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

[58] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[59] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[60] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

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

[62] B. Zhang, Y. Zhang, and S. Wang, "A lightweight and discriminative model for remote sensing scene classification with multidilation pooling module," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 2636–2653, 2019.

[63] M. Liu, L. Jiao, X. Liu, L. Li, F. Liu, and S. Yang, "C-cnn: Contourlet convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2636–2649, 2020.

[64] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-connected covariance network for remote sensing scene classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1461–1474, 2020.

[65] S. Wang, Y. Guan, and L. Shao, "Multi-granularity canonical appearance pooling for remote sensing scene classification," *IEEE Transactions on Image Processing*, vol. 29, pp. 5396–5407, 2020.

[66] Z. Zhao, J. Li, Z. Luo, J. Li, and C. Chen, "Remote sensing image scene classification based on an enhanced attention module," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 11, pp. 1926–1930, 2020.

[67] Q. Wang, W. Huang, Z. Xiong, and X. Li, "Looking closer at the scene: Multiscale representation learning for remote sensing image scene classification," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2020.

[68] E. Li, A. Samat, C. Zhang, P. Du, and W. Liu, "First and second-order information fusion networks for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[69] 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 Transactions on Image Processing*, vol. 31, pp. 99–109, 2021.

[70] W. Chen, S. Ouyang, W. Tong, X. Li, X. Zheng, and L. Wang, "Gcsanet: A global context spatial attention deep learning network for remote sensing scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1150–1162, 2022.