

Vision Transformer With Contrastive Learning for Remote Sensing Image Scene Classification

Meiqiao Bi, Minghua Wang , *Member, IEEE*, Zhi Li, and Danfeng Hong , *Senior Member, IEEE*

Abstract—Remote sensing images (RSIs) are characterized by complex spatial layouts and ground object structures. ViT can be a good choice for scene classification owing to the ability to capture long-range interactive information between patches of input images. However, due to the lack of some inductive biases inherent to CNNs, such as locality and translation equivariance, ViT cannot generalize well when trained on insufficient amounts of data. Compared with training ViT from scratch, transferring a large-scale pretrained one is more cost-efficient with better performance even when the target data are small scale. In addition, the cross-entropy (CE) loss is frequently utilized in scene classification yet has low robustness to noise labels and poor generalization performances for different scenes. In this article, a ViT-based model in combination with supervised contrastive learning (CL) is proposed, named ViT-CL. For CL, supervised contrastive (SupCon) loss, which is developed by extending the self-supervised contrastive approach to the fully supervised setting, can explore the label information of RSIs in embedding space and improve the robustness to common image corruption. In ViT-CL, a joint loss function that combines CE loss and SupCon loss is developed to prompt the model to learn more discriminative features. Also, a two-stage optimization framework is introduced to enhance the controllability of the optimization process of the ViT-CL model. Extensive experiments on the AID, NWPU-RESISC45, and UCM datasets verified the superior performance of ViT-CL, with the highest accuracies of 97.42%, 94.54%, and 99.76% among all competing methods, respectively.

Index Terms—Joint loss function, remote sensing, scene classification, supervised contrastive (SupCon) loss, vision transformer.

I. INTRODUCTION

THANKS to the rapid development of Earth observation (EO) technology, a massive amount of remote sensing (RS) images with a high spatial resolution (HSR) are being generated every day. Interpreting these RS images, which contain sufficient land-cover/land-use information, has practical significance in many fields, such as object detection [1], land planning [2], and traffic management [3]. Among the many image interpretation tasks, RS images have received increasing attention. RS images aim to allocate a semantic label to the input RS image, where

the label is from a predefined label set that refines the content of the RS images [4], [5], [6].

Scene classification is done in feature space so that the description ability of features extracted by the model directly affects the classification performance. In the beginning, scene classification methods are mainly based on hand-crafted features, which can be divided into low-level and high-level features. Low-level features [7], [8], [9] are usually constructed by visual attributes such as color [10], texture [11], and shape. And mid-level features are generated by encoding the low-level features through some encoding methods, such as bag-of-visual-words (BoVW) [12], vectors of locally aggregated descriptors (VLAD) [13], and improved Fisher kernel (IFK) [7]. These hand-crafted features heavily depend on the expertise of designers, and the capacity of information expression is limited.

With the rise of deep learning, data-driven feature extraction methods that do not rely on prior knowledge are born. Especially in supervised deep learning, models learn deep features by training themselves on a large number of the labeled dataset so that they can fully exploit category information to extract high-level semantic features. Among them, convolutional neural networks (CNNs) have shown powerful capability of feature learning in visual applications. Several classical CNNs have been proposed, such as AlexNet [14], VGGNet [15], GoogLeNet [16], ResNet [17], and U-Net [18]. Concerning scene classification, CNN-based methods can be divided into three branches depending on how they are used: employing a pretrained model as a feature extractor, fine-tuning a pretrained model, and training a new model from scratch.

In the first branch, pretrained CNNs are considered feature extractors, and then, the resulting features are fused or combined to capture more visual information. Studies [19] use different pretrained CNNs to extract vision features and fuse the result features. The results show that fused features are more discriminated against. In [20], the CNN model is used to extract multilayer feature maps, and these feature maps are combined by calculating their covariance matrix of them after being stacked. Finally, the result covariance matrices are used for classification. The aforementioned models demonstrate that CNNs have good generalization capability for scene classification.

In contrast with using pretrained CNNs as extractors, fine-tuning pretrained CNNs on target datasets, which can lead to an end-to-end model, is more straightforward and more effective. In addition, when training data are insufficient, fine tuning takes precedence over training from scratch. The emphasis of

Manuscript received 24 November 2022; revised 10 December 2022; accepted 16 December 2022. Date of publication 20 December 2022; date of current version 29 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42271350 and Grant 62201552 and in part by the China Postdoctoral Science Foundation under Grant 2022M713223. (Corresponding author: Minghua Wang.)

The authors are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: meiqiaobi2022@163.com; minghuawang1993@163.com; lizhi21@mails.ucas.ac.cn; hongdf@aircas.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3230835

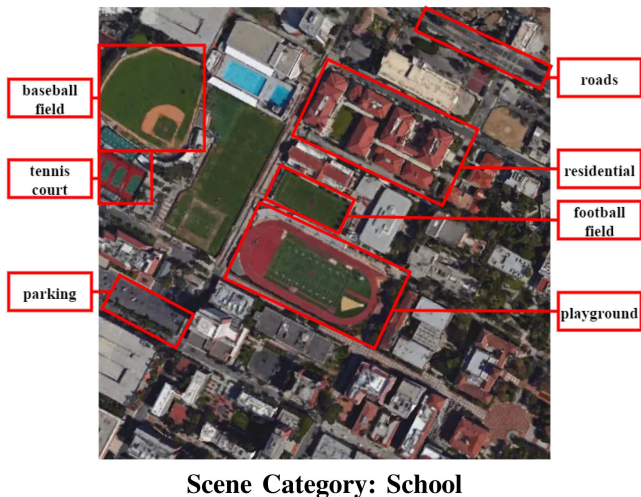


Fig. 1. Coexistence of multiple ground objects. Multiple ground objects related to the “School” scene are distributed over different locations of the RSI. The image is from the AID dataset [32].

optimizations in the second branch is placed on fine adjusting the networks [21], [22], [23] and loss functions [24], [25].

Although fine tuning pretrained CNNs can achieve effective classification performance, pretrained CNNs still have some limitations. Due to the gap between the nature dataset and the RSI dataset, features learned by the model trained on the nature dataset are not perfectly suitable for the RSI dataset. Moreover, modifying the pretrained model is not that convenient. Many studies of the third branch focused on either improving the CNN structure [26], [27], [28], [29] or constructing a hybrid model framework [30], [31] according to the characteristics of remote sensing datasets.

Overall, the CNN is a multilayer structure, where the convolutional layer plays a prominent role in extracting features from images. Thanks to the convolutional operations, the model can learn the local spatial information of the input images. And though progressively expanding the receptive field of the convolutional kernels in each layer, it can acquire the features of a global view. Sacking multiple convolutional layers can boost the classification performance significantly. However, the CNN cannot capture long-range relationships limited by the local receptive field.

Due to the complex spatial distribution of ground objects and the bird’s eye view of RS imaging equipment, it is very generic for multiple ground objects to coexist in one single RS image [33], [34]. In the RS image shown in Fig. 1, which is of the “school” scene, there are multiple ground objects, a baseball field, a tennis court, a playground, roads, and so on. Furthermore, these objects are distributed in all directions of the image. The coexistence and dispersed distribution of multiple ground objects bring challenges to scene classification. So capturing global long-range interactions for these ground objects has vital practical significance in scene classification. Besides the CNN, transformer [35] is another deep learning structure that has taken off in the natural language processing (NLP) domain. The transformer benefiting from the self-attention mechanism,

can capture long-range interactions on input sequence data and learn a global representation. Encouraged by the success of the transformer in NLP, Dosovitskiy et al. [36] have extended the standard transformer structure to visual applications and proposed the ViT model that demonstrates the enormous potential for image classification.

Although ViT has shown excellent feature learning ability, its performance on RS images has not yet reached saturation. Further improving its performance without increasing the parameter scale or integrating with additional depth structure is possible. In addition to the characteristics of the coexistence of multiple ground objects mentioned above, intraclass diversity (shown in Fig. 2) and interclass similarity (shown in Fig. 3) are also two nonnegligible challenges. In recent years, contrastive learning has received considerable attention due to its great potential for visual representation learning ability. Since 2019, research on comparative learning (CL) has developed rapidly, resulting in many excellent methods, such as SimCLR [38], SimCLR V2 [39], MoCo [40], and MoCo V2 [41]. Among them, the supervised contrastive (SupCon) loss, a batch contrastive approach for a supervised setting, has the intrinsic ability to perform hard positive/negative mining [42]. So, it is possible to employ SupCon loss to help the ViT model to learn more discriminative features.

Given the appealing properties of ViT and CL, in this article, a novel two-stage end-to-end framework for the scene classification is proposed, named ViT-CL. ViT-CL aims to combine the advantages of the transformer structure and the principle of contrastive learning to improve the performance of scene classification. First of all, considering that the scale of RS image datasets is hardly sufficient to train ViT models from scratch, transferring a large-scale pretrained ViT model to the target dataset, which can help ViT surpass inductive bias, is preferred. Second, as a combination of SupCon loss and CE loss, a joint loss is proposed to fine tune the pretrained ViT model. In this way, the two loss functions complement each other, forcing the model to learn more discriminating high-level semantic features and further making the model more robust. Finally, considering ViT is hard to optimize and sensitive to hyperparameters, we develop a two-stage optimization. In the first stage, only CE loss is adopted to fine tune the pretrained ViT model on the target dataset. In the second stage, the proposed joint loss is utilized to fine tune the model produced in the first stage. After the two-stage fine tuning, the optimized model is obtained, but only the cross-entropy loss part of the model is retained for the following inference.

II. RELATED WORK

In the last two years, some studies have begun to explore how ViT performs in RS images. Bazi et al. [43] introduced the ViT model into the RS images and improved the classification accuracy through data augmentation such as CutMix and Cutout. Also, they proved that the model performance could be maintained even if half of the layers were pruned to compress the network. Then, Bashmal et al. [44] proposed the data-efficient image transformers (DeiT), a ViT-based model trained



Fig. 2. Intra-class diversity.



Fig. 3. Interclass similarity. (a) Similarity between scene “freeway” and “runway.” (b) Similarity between scene “railway station” and “industrial area.” These images are from the NWPU-RESISC45 dataset [37].

by knowledge distillation with fewer data, and proved that the performance of ViT was superior to the CNN-based method on the remote sensing datasets AID and NWPU-RESISC. In [45], SCViT is proposed to overcome the disadvantage that the original model can only capture global spatial features. By improving the structure of ViT, the model not only considers the detailed geometric information of high spatial resolution images but also considers the contribution of different channels of the class token.

In addition, as the advantages of convolution structure and transformer structure complement each other, some studies have explored ways to combine these two network structures. Deng et al. [46] designed a joint loss function to build the joint framework CTNet. In this framework, the ViT model is used to capture semantic features, while the CNN model is used to extract local structure information. In [47], the advantages of the two models are integrated without improving the computational complexity by knowledge distillation, in which the ViT is worked as a teacher to guide the student model ResNet18. Besides classifying tasks, this article also proves that this method has good generalization ability for different tasks.

The remainder of this article is organized as follows: Section III introduces ViT and the supervised contrastive loss, then describes the proposed method ViT-CL in detail. Section IV contains both contrast and ablation experiments. The former compares our models with several classical CNN models and ViT-based methods on three well-known datasets, and the latter analyzes how the optimized model works. Finally, Section V concludes this article.

III. PROPOSED METHOD

Let $D = \{X_i, y_i\}_{i=1}^r$ denote an SRI dataset of size r , where X_i represents the i th image and y_i is its corresponding category label. $X_i \in \mathbb{R}^{h \times w \times c}$, where h , w , and c represent the height, the width, and the number of channels, respectively. $y_i \in \{1, 2, \dots, m\}$, where m is the predetermined number of categories.

A. Vision Transformer

A vision transformer is proposed to apply the vanilla transformer to the image task. The main goal is to generalize it to visual applications without integrating any data-specific architecture. ViT only retains the encoder module of the standard transformer, and the complete end-to-end architecture is shown in Fig. 4.

First, the input image is subdivided into nonoverlapping 2-D patches with dimensions $p \times p \times c$ before being passed to the transformer encoder to adapt the standard transformer structure. The patch size p is usually set to 16 or 32, and a smaller patch size will lead to a longer sequence and vice versa. Then, the n 2-D patches are flattened and passed to a linear layer to generate a patch sequence $P \in \mathbb{R}^{n \times (p^2 \cdot c)}$, where $n = \frac{h \times w}{p^2}$ is the length of P . In the linear layer, a learnable matrix $E \in \mathbb{R}^{(p^2 \cdot c) \times d}$ is utilized to embed these patches into a d -dimensional space. After that, like most classification tasks with transformer structure, the embedded patch sequence is concatenated with a learnable classification token P_0 . Finally, the patch’s spatial arrangement E_{pos} ,

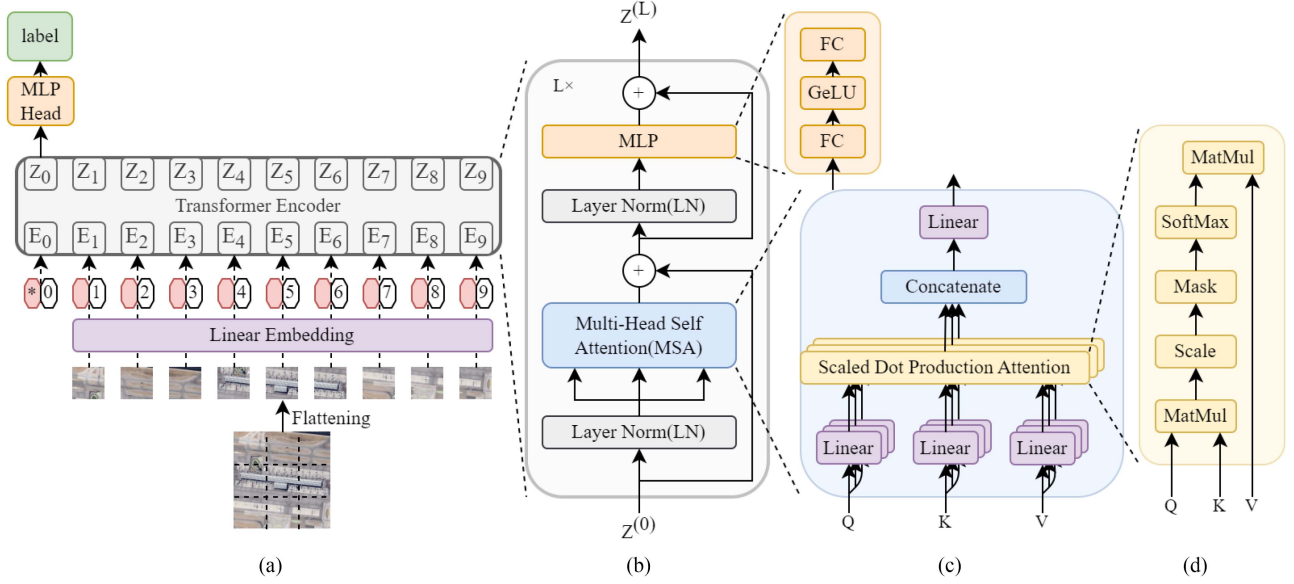


Fig. 4. Illustration of the proposed ViT architecture. (a) Main architecture of the model. (b) Transformers encoder module. (c) Multihead self-attention (MSA). (d) Self-attention head (SA).

which helps the transformer to distinguish them, is encoded and added to the embedded patch sequence to obtain the embedding sequence $Z^{(0)}$. The aforementioned process is formulated as follows:

$$Z^{(0)} = \{P_0; P_1E; P_2E; \dots; P_mE; \} + E_{\text{pos}}$$

$$E \in \mathbb{R}^{(p^2-c) \times d}, E_{\text{pos}} \in \mathbb{R}^{(n+1) \times d}. \quad (1)$$

Next, the embedding sequence $Z^{(0)}$ is entered into the transformer encoder that contains L blocks. As shown in Fig. 4(b), there are two main subcomponents in each block: multihead self-attention (MSA) [see (2)] and multilayer perceptron (MLP) [see (3)]. Before entering these two components separately, the input needs to be preceded by a normalization layer (LN), which can stabilize the gradient of the loss to the input during backpropagation. And both the output of the two subcomponents employ residual skip connections to obtain a result as the input of the next subcomponent. The calculation process is as follows:

$$Z^{(l)'} = \text{MSA} \left(\text{LN} \left(Z^{(l-1)} \right) \right) + Z^{(l-1)}, l = 1 \dots L \quad (2)$$

$$Z^{(l)} = \text{MLP} \left(\text{LN} \left(Z^{(l)'} \right) \right) + Z^{(l)'}, l = 1 \dots L. \quad (3)$$

Here, notice that the output of the L th layer Z^L is the final result of the encoder. For classification, the first token of Z^L can be regarded as the final feature representation f of an input image after an LN processing. The calculation is as follows:

$$f = \text{LN} \left(Z_0^{(L)} \right). \quad (4)$$

Then, f is passed into an MLP head, which is composed of a full connection layer (FC) and the softmax loss function to predict the class label

$$y = \text{softmax} \left(\text{FC}(f) \right). \quad (5)$$

The construction of MSA, the core of the transformer encoder, is shown in Fig. 4(c). Attention can be understood as the weight of interaction between tokens, and self-attention means these tokens belong to one single sequence. For each token in the sequence Z , first, calculate the attention scores between itself and all the tokens of Z . And second, calculate the sum over all token embeddings weighted by these attention scores to obtain a new embedding for the current token. Before calculating self-attention, the sequence Z is mapped to three different sequences $Q \in \mathbb{R}^{(n+1) \times d_Q}$, $K \in \mathbb{R}^{(n+1) \times d_K}$, and $V \in \mathbb{R}^{(n+1) \times d_V}$ by multiplying a learned matrix M_{QKV} , where Q , K , and V represent query, key, and value, and d_K , d_Q , and d_V are their dimensions. In theory, it just requires $d_K = d_Q$, and for convenience, there are $d_K = d_Q = d_V$. The formula is as follows:

$$[Q, K, V] = Z M_{QKV}, M_{QKV} \in \mathbb{R}^{d \times 3d_K}. \quad (6)$$

Then, it comes to the SA block, shown in Fig. 4(c). The dot production $Q \cdot K^T$ is calculated to measure the pairwise similarity between tokens in sequence Z . And to alleviate the problem of vanishing gradient, the result needs to be divided by $\sqrt{d_K}$. After a softmax operation, the final scaled dot attention is obtained. The entire procedure is as follows:

$$SA(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_K}} \right) V. \quad (7)$$

Suppose that the number of heads is h , and the MSA block computes the scaled dot attention h times separately, using (7) with h different values for Q , K , and V , respectively. These result h SA values will be concatenated, and then, passed to a linear layer with parameter W^0 to ensure that the dimensions of the input and output of each MSA block stay the same. The formula is as follows:

$$\text{MSA} = \text{Concat} \left(SA_1, SA_2, \dots, SA_h \right) W^0$$

TABLE I
PARAMETER STATISTICS FOR THE SMALL BASE AND LARGE VARIANTS OF
VISION TRANSFORMER

Model	Number of Layers	MLP size	Heads
small ViT	8	2358	12
base ViT	12	3072	16
large ViT	24	4096	16

$$W^0 \in \mathbb{R}^{(h \cdot d_k) \times d}. \quad (8)$$

When it comes to the MLP block, there are two dense layers and a GeLU activation in between. It is too simple to expand on.

From the calculation process of MSA, it can be seen that this mechanism can capture long-distance dependencies between tokens. The generated feature representation, not only contains the information of all patches but also their higher order spatial interaction information. However, as mentioned previously, the transformer lacks some inductive biases, so it cannot generalize well when there is no adequate data. The authors in [48] has analyzed the effects of pretraining data scale, data augmentation, model size, and compute budget on the performance of the ViT model. And they proved that for most practical purposes, compared with training a ViT model from scratch, fine tuning a large-scale pretrained ViT model on the target dataset is both more cost effective and can produce better results. The author also gives some suggestions on how to choose a pretraining model.

- 1) The larger the pretrain dataset, the more generic the obtained model, and the larger the model, the longer the inference time;
- 2) The validation score obtained in a pretrain stage can be the direct reference index. And there is no need to transfer all available pretrained models to the target data and choose the model by comparing verification scores in a fine-tuning stage.

The pretrained ViT model mainly contains three versions with different scales of parameters: small ViT, base ViT, and large ViT. And each version usually owns two different patch sizes, 16 and 32. See Table I for some vital parameters of the three version models.

Following the aforementioned suggestions and the recommendations of the official ViT documents,¹ The model B/16_21 k, which means base ViT with patch size 16, is chosen as the backbone network. More specifically, the model is pretrained on a large-scale dataset ImageNet-21 k (including 13 M images) [49], applying varying amounts of AugReg strategies [48].

B. Supervised Contrastive Learning

Most classification tasks usually employ CE loss as the objective function. But some studies have shown that this loss has drawbacks such as not being robust to noisy labels [49] and may produce poor margins [50], which can reduce the model's

generalization ability and further affect the classification accuracy. As mentioned in the introduction section, RS images are characterized by big intraclass diversity and high interclass similarity. That is to say, RS images of the same class may be very different (shown in Fig. 2), while RS images of different classes may be very similar (shown in Fig. 3). So, poor margins, which means CE loss does not explicitly encourage discriminative learning of features, will be a loss for scene classification. However, SupCon loss can promote model learning discriminative feature representations by pulling together the clusters of similar samples in feature space while pushing apart the clusters of dissimilar samples. It can be employed to compensate for the drawbacks of CE loss.

SupCon loss is produced by extending the self-supervised contrast learning [38] to the fully supervised setting [42]. In a supervised setting, the loss' selection criteria of positive samples changed to "whether it belongs to the same class," from that "whether it is from the same picture" in a self-supervised setting. Thus, the number of positive sample pairs in the comparison loss is expanded. It has been proved that this change can encourage the model better depict the intraclass similarity. For each anchor, the SupCon loss first calculates the similarity scores between it and all the other positive samples, and then, weighted sum these scores. The calculation formula is as follows:

$$l_i^{\text{sup}} = \frac{-1}{|p(i)|} \sum_{p \in P(i)} \log \frac{\exp(f_i \cdot f_p / \tau)}{\sum_{a \in A(i)} \exp(f_i \cdot f_a / \tau)}. \quad (9)$$

Here $i \in I \equiv \{1 \cdots N\}$ is the index of the anchor. $A(i) \equiv I \setminus \{i\}$ represents the overall sample set besides the sample i , and $P(i) \equiv \{p \in A(i) : \tilde{y}_p = \tilde{y}_i\}$ is the positive sample set, in which the samples have the same label as the anchor. The symbol "." represents the inner (dot) product and $\tau \in \mathbb{R}^+$ is the temperature parameter. It can be seen from (9) that the contrastive denominator contains the summation over negative samples, and this form improves the model's ability to discriminate between signal and noise (negative samples). Overall, the significance of SupCon loss lies in narrowing the distance between the samples from the same class in the feature space, while widening the distance between samples from different classes. However, for each anchor, only positive samples in the batch contribute to the numerator of (9), so the batch size should be larger than the number of classes to ensure that there are enough positive samples in the batch. Meanwhile, a larger batch size call also guarantees enough negatives to form a sharp contrast with positive pairs.

C. ViT-CL

In this article, a method named ViT-CL is proposed to combine the advantages of SupCon loss and ViT. ViT-CL utilizes ViT as a backbone network, and then, optimizes the backbone network by a two-stage optimizing framework with a joint loss. The framework of ViT-CL is shown in Fig. 5.

After encoding by the ViT, each image X_i can obtain its embedding feature f_i . In the proposed framework, image features of one input batch will be passed to a joint loss, which is constituted by the CE loss and SupCon loss. In Fig. 5, each loss function is visualized as a task.

¹https://github.com/google-research/vision_transformer

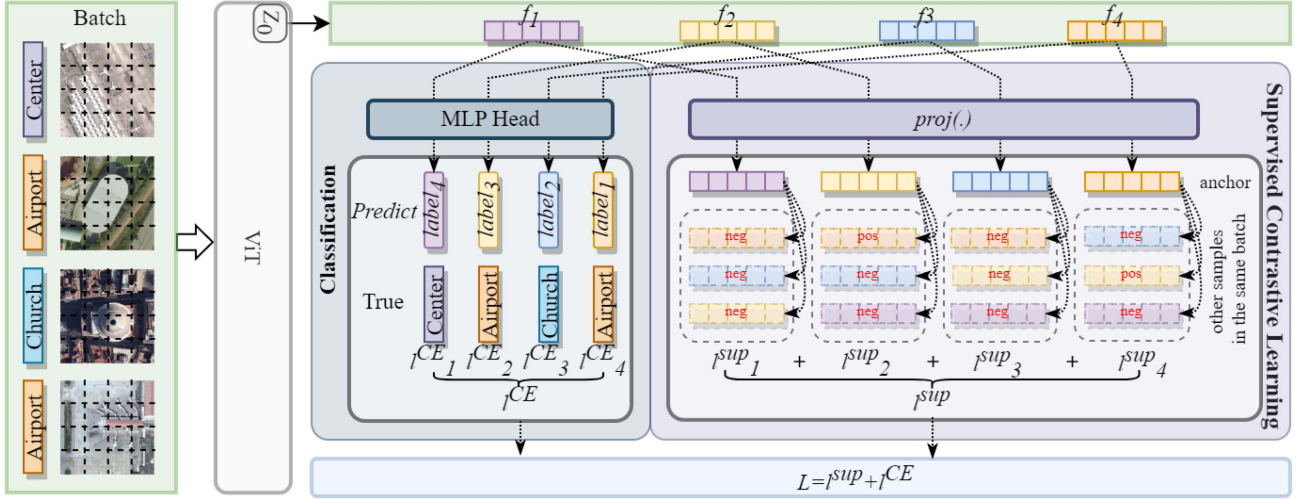


Fig. 5. Framework of the proposed ViT-CL.

The first task is the classification task, corresponding to the left one of the two tasks in Fig. 5. Here, the classifier structure of ViT is directly applied: First, feature f_i is mapped to a new feature space for classification by the MLP head, and then, the CE loss is calculated. The calculation formulas are as follows:

$$f_i^{\text{CE}} = \text{FC}(f_i) \quad (10)$$

$$l_i^{\text{CE}} = L_{\text{CE}}(\text{softmax}(f_i^{\text{CE}}), y_i). \quad (11)$$

The second task is supervised contrast learning, corresponding to the right one of the two tasks in Fig. 5. Referring to the contrast learning framework [38], [42], a project network $\text{proj}(\cdot)$, which plays the same role as the MLP head in the classification task, is introduced. Some studies on contrast learning have shown that the project network is necessary and can help to improve the model's performance [38]. Here, $\text{proj}(\cdot)$ is instantiated as a two-layer MLP, whose hidden layer size is 2048 and output layer size is 128. Formulaic the calculations as follows:

$$f_i^{\text{SupCon}} = \text{FC}(\text{ReLU}(\text{FC})) \quad (12)$$

$$l_i^{\text{SupCon}} = \frac{-1}{|p(i)|} \sum_{p \in P(i)} \log \frac{\exp(f_i \cdot f_p / \tau)}{\sum_{a \in A(i)} \exp(f_i \cdot f_a / \tau)}. \quad (13)$$

It should be pointed out that in the actual minibatch optimization process, the positive sample set $P(i)$ of each anchor is limited in the batch where the anchor is, so the optimization parameter batch size would impact the performance of the model. Finally, the joint loss of one input batch is calculated as follows:

$$L = \sum_{i=1}^n (l_i^{\text{CE}} + \lambda l_i^{\text{SupCon}}) \quad (14)$$

where λ acts as a tradeoff between these two losses, which needs to be judiciously tuned to control the distinctiveness of learned features. Along with ViT's sensitivity to optimizer hyperparameters [51], tuning these hyperparameters including λ is time consuming. Instead, a simpler but effective strategy

TABLE II
PARAMETER STATISTICS FOR THE SMALL BASE AND LARGE VARIANTS OF THE VISION TRANSFORMER

Datasets	AID	NWPU-RESISC45	UCM
Classes	30	45	21
Images per class	220 ~ 420	700	100
Total images	10000	31500	2100
Images size	600 × 600	256 × 256	256 × 256
Data source	Google Earth	Google Earth	USGS
Published year	2017	2017	2010

is proposed to tackle this problem: a two-stage optimization method. In the first stage, the pretrained ViT model selected in Section III-A is initially fine tuned only by the CE loss on the target RSI dataset. In the second stage, the result fine-tuned model of the first stage is fine tuned by the joint loss again. Corresponding to the framework shown in Fig. 5, the model is fine tuned using only the classification part in the first stage. Then, in the second stage, both the classification part and the supervised contrastive learning part are used. After the two-stage joint fine tuning, only the classification part of the framework is reserved to complete inference work.

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

1) *Datasets Description*: In our experiments, three public remote-sensing datasets are utilized to evaluate the ViT-CL: Aerial Image Dataset (AID) [32], Northwestern Polytechnical University Dataset (NWPU-RESISC45) [37], and UC-Merced Land Use Dataset (UCM) [12], the detail information of the three datasets are displayed in Table II. Among them, the NWPU-RESISC45 dataset and UCM dataset are more challenging than the AID dataset.

2) *Hardware and Software Environment*: All subsequent experiments are conducted on a personal computer, and the detailed computing environment is shown in Table III.

TABLE III
PARAMETER STATISTICS FOR THE SMALL BASE AND LARGE VARIANTS OF THE VISION TRANSFORMER

Operation system	Ubuntu 16.04 Server
Memory	128G
Framework	Pytorch 1.11
GPU	4 × NVIDIA TITAN XP
CPU	6 × Intel(R) Core(TM) i7-6850 CPU@3.60GHz

3) *Parameter Optimization Setup*: For comparison purposes, the overall accuracy (OA) is employed to evaluate the performance of different classic methods, which indicates the percentage of correctly classified images in the total number of images. When training or fine tuning models, 50% and 80% of the UCM dataset, 20% and 50% of the AID dataset, and 10% and 20% of the NWPU-RESISC45 dataset are randomly selected for training, respectively. One thing to note is that when using two-stage ViT-CL, the division of the train set (referring to the samples assigned to the set rather than the sample proportion) needs to be consistent in both stages. The pretrained ViT model used as the backbone of our ViT-CL is selected as described in Section III-A, a B/16 version pretrained on Imagenet-21 k with AugReg strategies. It can be downloaded from https://storage.googleapis.com/vit_models/augreg/.

In the optimization stage, adaptive moment estimation (Adam) is introduced to update the parameters of all methods and the learning rate (LR) is set to 0.0001. Also, stepLR is used to control the LR, whose `step_size` is set to 20 and `gamma` is set to 0.9. That is to say, *LR* is multiplied by a factor of 0.9 every 20 epochs. All the methods are fine tuned 100 epochs and for each epoch, `batch_size` is set to 128 limited by the memory of the GPU. Besides, the input images are resized to 256×256 pixels. In addition, with respect to the parameters in joint loss, λ is set to 0.2 through many experiments, and temperature τ is set to 0.07 as recommended in most comparative learning papers [40], [41], [52].

B. Comparison With Some Classic Methods

we compare our method with five classic CNN-based methods, one traditional ViT model, and two ViT-based improved models, which are as follows:

- 1) Fine-tune ResNet-50 [17];
- 2) Fine-tune AlexNet [14];
- 3) Fine-tune VGGNet-16 [15];
- 4) Fine-tune GoogLeNet [16];
- 5) Fine-tune MobileNe_V2 [53];
- 6) V16_21k [43];
- 7) SCViT [45];
- 8) ET-GSNet [47].

Table IV shows detailed comparisons between ViT-CL and other models on three datasets, AID, NWPU-RESISC45, and UCM. It can be seen that the proposed two-stage joint fine-tuning method ViT-CL has an obviously higher OA, compared

with both these classical CNN-based models and the improved ViT-based models.

Furthermore, the confusion matrices (CMs) of the ViT models' prediction on the two more challenging datasets AID and NWPU-RESISC45 are calculated to prove the improvement of the ViT-CL model. The two CMs are shown in Fig. 6; (a) is for AID with a training ratio of 50% and (b) is for NWPU-RESISC45 with a training ratio 20%. It can be seen from the CMs that the proposed method performs well on the two datasets.

In the dataset AID, where the total number of categories is 30, the class number owning a greater than 90% accuracy is as high as 29. And among them, 25 have an accuracy greater than 95%. Besides, even the worst accuracy can reach 88%. In detail, the model achieves excellent results in categories with high intraclass diversity, such as "Airport" (97%), "Commercial" (97%), "Railway Station" (98%), "Church" (99%), "Farmland" (99%), and "Mountain" (100%) (some results is shown in Table V). Also, it performs well in categories with high interclass similarity. And these highly similar pairs of categories include "BareLand" (99%) and "Desert" (99%), "Park" (94%) and "Resort" (92%), and "Playground" (98%) and "Stadium" (99%) (some predict results is shown in Table VI).

In dataset NWPU-RESISC45, there are 40 categories, whose accuracy is over 90% out of 45, and 26 categories are above 95%. The five categories with the lowest accuracies are "Palace" (76%), "Commercial Area" (88%), "Dense Residential" (88%), "Medium Residential" (89%), and "Church" (89%). Except for the category "Palace" in which the model performs worst (76%), the accuracy is close to 90%. For the two categories "Church" and "Railway Station," which have high intraclass diversity in this dataset, the proposed model achieves accuracies of 89% and 93%, respectively (some predicted results of these categories are shown in Table VII). And for category pairs with a high interclass similarity: "Freeway" with "Runway," "Industrial Area" with "Railway Station," and "Railway Station" with "stadium," the accuracies achieved by the model are 92%, 95%, 90%, 95%, 95%, and 97%, respectively (some predict results is shown in Table VIII).

Moreover, from images in Tables V and VII, it can be seen that images belonging to the same scene can appear very different, and ViT-CL has the ability to capture diversity. When it comes to Tables VI and VIII, images of different scenes may look very similar or contain the same objects, ViT-CL also has the ability to distinguish between these scenarios. The aforementioned results fully show that ViT-CL can well distinguish both intraclass diversity and interclass similarity.

C. Ablation Study and Analysis

This article also conducted experiments on the variants of the model to illustrate the effectiveness of the two-stage joint optimization, including the following.

- 1) Fine-tune B/16_21 k: Fine tune the pretrained ViT model once, utilizing and only utilizing CE loss for classification.
- 2) One-stage ViT-CL: Fine tune the pretrained ViT model once utilizing joint loss and retain the classification part for classification.

TABLE IV
CLASSIFICATION ACCURACIES OF DIFFERENT METHODS ON THE AID, NWPU-RESISC45, AND UCM DATASETS (%)

Method Type	Method	Parameters	AID		NWPU-RESISC45		UCM	
			20%	50%	10%	20%	50%	80%
♡	Fine-tune ResNet-50	25.56M	88.17	92.94	85.94	88.47	95.33	98.33
	Fine-tune AlexNet	60.97M	85.05	90.42	79.23	83.95	93.62	96.90
	Fine-tune VGGNet-16	138.36M	83.99	93.08	83.99	87.88	88.56	93.08
	Fine-tune GoogLeNet	7M	90.03	93.08	86.31	88.77	96.10	96.90
	Fine-tune MobileNet_V2	3.5M	90.88	92.89	87.17	89.72	97.00	97.62
◇	V16_21k	-	94.97	-	92.60	-	-	98.14
	SCViT	-	95.56	96.98	92.72	94.66	98.90	99.57
	ET-GSNet	-	95.58	96.88	92.72	94.50	-	99.29
◇	ViT-CL	86.57M	95.60	97.42	92.85	94.69	99.14	99.76

♡ CNN-based model.
◇ ViT based model.

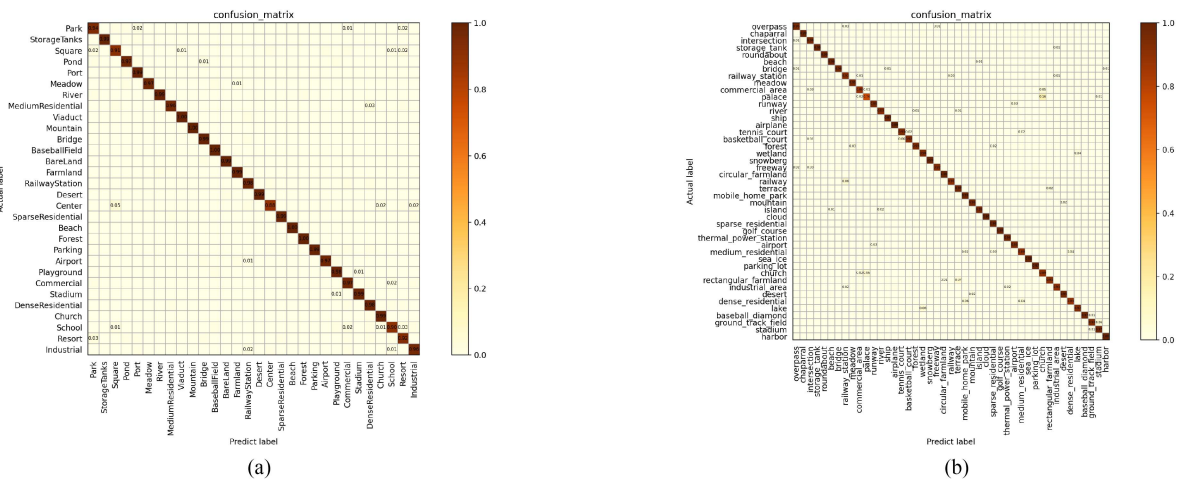


Fig. 6. CMs of ViT-CL on AID dataset and NWPU-RESISC45 dataset. (a) CM of ViT-CL on AID dataset with train percent 50%. (b) CM of ViT-CL on NWPU-RESISC45 dataset with train percent 20%.

TABLE V
PART PREDICT RESULTS OF DATASET AID, CATEGORIES ARE CHOSEN FOR THEIR BIG INTRACLASS DIVERSITY

AID (train ratio 50%)					
Airport	Commercial	Railway station	Church	Farmland	Mountain
Airport	Commercial	Railway station	Church	Farmland	Mountain
Airport	Commercial	Railway station	* Square	Farmland	Mountain

The labels in the second row of the table are the true labels of the images, and the labels below the images are the predicted labels. The predicted labels with * indicate wrong predictions.

TABLE VI
PART PREDICT RESULTS OF DATASET NWPU-RESISC45, CATEGORIES ARE CHOSEN FOR THEIR BIG INTRACLASS DIVERSITY




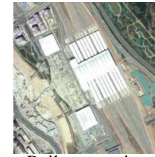
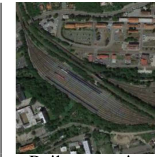




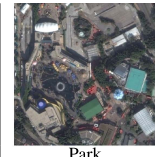

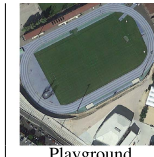





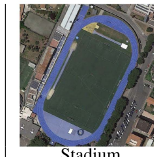












NWPU-RESISC45 (train ratio 20%)					
Church			Railway station		
					
Church	Church	Church	Railway station	Railway station	Railway station

TABLE VII
PART PREDICT RESULTS OF DATASET AID, CATEGORIES ARE CHOSEN FOR THEIR HIGH INTERCLASS SIMILARITY

AID (train ratio 50%)					
					
BareLand BareLand	BareLand BareLand	Park Park	Park * Resort	Playground Playground	Playground * Stadium
					
Desert Desert	Desert * BareLand	Resort Resort	Resort * Park	Stadium Stadium	Stadium * Playground

Of the two lines of tags below the images, the first acts as the true label and the second acts as the predicted label. The predict labels with * indicate wrong predictions.

TABLE VIII
SOME PREDICT RESULT OF DATASET NWPU-RESISC45, CATEGORIES ARE CHOSEN FOR THEIR HIGH INTERCLASS SIMILARITY

NWPU-RESISC45 (train ratio 20%)					
					
Freeway Freeway	Freeway Freeway	Industrial area Industrial area	Industrial area * Railway station	Railway station Railway station	Railway station Railway station
					
Runway Runway	Runway Runway	Railway station Railway station	Railway station Railway station	stadium stadium	stadium stadium

Of the two lines of tags below the images, the first acts as the true label and the second acts as the predicted label. The predict labels with * indicate wrong predictions.

TABLE IX
CLASSIFICATION ACCURACIES OF THREE ViT BASED METHOD ON THE AID, NWPU-RESISC45, AND UCM DATASETS (%)

Method	AID		NWPU-RESISC45		UCM	
	20%	50%	10%	20%	50%	80%
Fine-tune B/16_21k	94.45	96.56	91.34	93.37	98.47	99.52
one-stage ViT-CL	94.47	96.52	91.33	93.39	98.19	99.04
ViT-CL	95.60	97.42	92.85	94.69	99.14	99.76

3) ViT-CL: Our proposed two-stage joint fine-tune model.

Table IX shows detailed comparisons for the three ViT-based methods. First, compared with fine-tuning B/16_21 k, one-stage ViT-CL additional introduces supervised contrast loss. Their OAs illustrate that merely adding supervised contrast loss does not work and may disturb the backpropagation process to obtain better results. Second, comparing the last two models, one-stage ViT-CL and the proposed ViT-CL, the difference between them is when to execute the fine tuning by joint loss. It can be seen

TABLE X
CLASSIFICATION ACCURACIES OF THREE ViT-BASED METHOD ON THE AID, NWPU-RESISC45, AND UCM DATASETS (%)

Method	Index	AID		NWPU-RESISC45		UCM	
		20%	50%	10%	20%	50%	80%
Fine-tune B/16_21k	min	72.60	80.67	63.97	77.50	90.00	95.00
	max	100.00	100.00	99.05	99.46	100.00	100.00
	avg	94.01	96.31	91.35	93.34	98.48	99.52
	var	7.02	4.84	6.83	5.20	2.68	1.47
ViT-CL	min	83.19	88.46	74.13	75.71	96.00	95.00
	max	100.00	100.00	99.52	100.00	100.00	100.00
	mean	95.39	97.31	92.85	94.69	99.14	99.76
	std	4.68	3.09	5.38	4.38	1.32	1.06

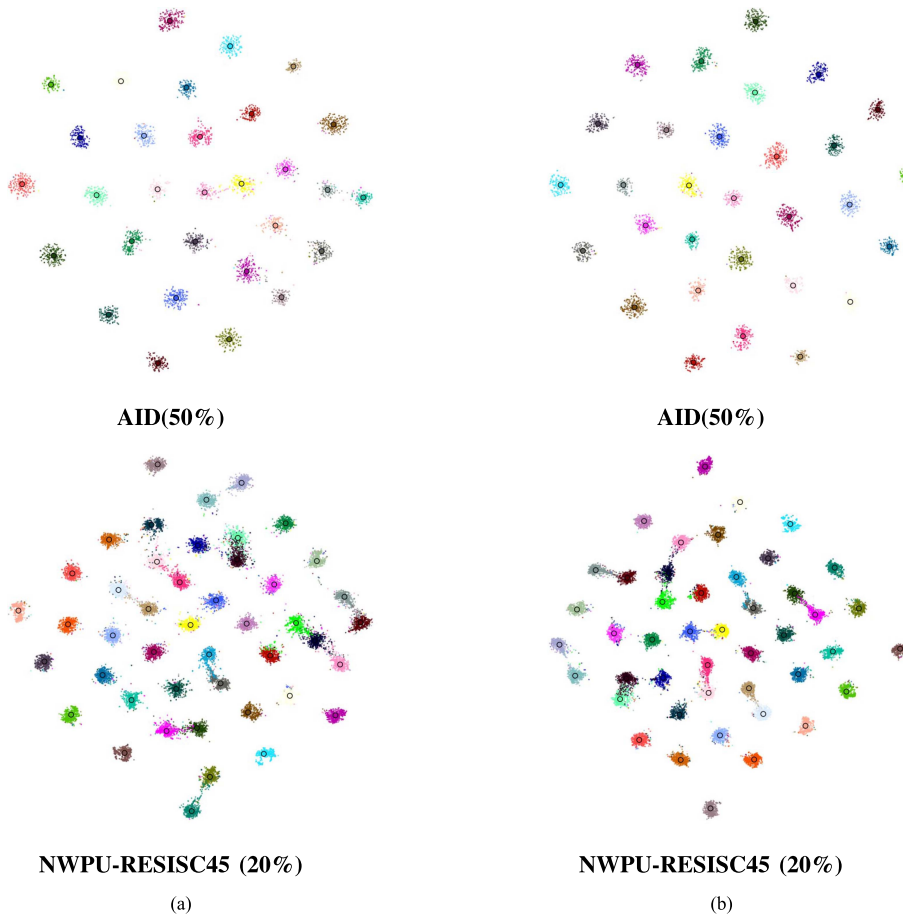


Fig. 7. 2-D visualization of feature representations extracted by Fine-tuned B/16_21 k and ViT-CL on the AID and NWPU-RESISC45 dataset using t-SNE. (a) Fine-tuned B/16_21 k. (b) ViT-CL.

that the two-stage optimization, which introduces supervised contrast loss after the pretrained model has achieved a good result on the target dataset by initial fine tuning, can make the joining loss effective. Finally, as for fine tuning B/16_21 k and ViT-CL, the main difference is that the latter is optimized by the two-stage joint framework, while the former is optimized only once by CE loss. Their results show that our proposed framework can improve the OAs by more than 1% on the two more challenging datasets AID and NWPU-RESISC45. And the lower proportion of the training set, the more significant the improvement.

Furthermore, this article statistic the distributions of category accuracy obtained by fine-tuning B/16_12 K and ViT-CL on different datasets (i.e., diagonal elements of the CM) to show how much the two-stage joint fine-tuning framework improving feature representations' expression ability compared with the CE loss. The maximum, minimum, *average*, and *variance* of each category's accuracy are statistical, respectively. The results are shown in Table X. It can be seen from Table X that the average classification accuracies of the ViT-CL on different three datasets are all higher than those of Fine-tune B/16_21 k, so do the minimum accuracies except the one on

dataset NWPU-RESISC45 with train ratio 20%. Together with the lower variances of classification accuracies of the ViT-CL, all of these indicate that the introduction of contrast loss can make similar samples from the same scene more clustered, while the confusion degree among different scenes becomes lower.

More intuitively, the t-distributed stochastic neighbor embedding (t-SNE) algorithm [54] can reduce the dimension of the feature representations generated by different models so that feature projections can be visualized in a 2-D space. The 2-D visualization images of feature representations extracted by Fine-tuning B/16_21 k and ViT-CL on the two dataset AID (train ratio 50%) and NWPU-RESISC45 (train ratio 20%) are shown in Fig. 7, where (a) is for Fine-tuned B/16_21 k and (b) is for ViT-CL. From Fig. 7, it can be found that the feature structure is clear no matter whether the two-stage joint operation is adopted, which demonstrates the effectiveness of the backbone ViT model. Furthermore, compared with the feature clusters extracted by Fine-tuned B/16_21 k, the feature clusters extracted by ViT-CL are closer together, and their boundaries of them are clearer. This fact confirms the usefulness of our framework.

V. CONCLUSION

In this work, a two-stage end-to-end framework named ViT-CL is proposed. The framework combines the ViT model with supervised contrastive learning and gives full play to the advantages of the two so that it can further improve the accuracy of scene classification. The backbone ViT of this framework can capture long-range dependencies among patches via a self-attention mechanism. And the proposed joint loss function composed of cross entropy loss and supervised contrast loss can help the model learn more robust and discriminating semantic features. Besides, to avoid time-consuming parameter tuning, a two-stage fine tuning is employed to ensure the joint loss function can show its best performance. ViT-CL has been evaluated on three public remote-sensing image datasets, and the experimental results demonstrate the effectiveness in improving the overall accuracy of scene classification, compared to some classical CNN-based methods and improved ViT-based models. Moreover, with the ablation experiment, how the two-stage joint fine-tuning framework improves the performance of scene classification is discussed and it concluded that both “two-stage” and “joint” are necessary. In the future, we will employ unsupervised contrast learning or data enhancement strategies to build a scenario classification framework with lower time consumption and better performance.

REFERENCES

- [1] X. Wu, D. Hong, Z. Huang, and J. Chanussot, “Infrared small object detection using deep interactive U-Net,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Nov. 2022, Art. no. 6517805, doi: [10.1109/LGRS.2022.3218688](https://doi.org/10.1109/LGRS.2022.3218688).
- [2] J. Yao et al., “Semi-active convolutional neural networks for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5537915, doi: [10.1109/TGRS.2022.3206208](https://doi.org/10.1109/TGRS.2022.3206208).
- [3] C. Toth and G. Józkó, “Remote sensing platforms and sensors: A survey,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 115, pp. 22–36, 2016.
- [4] U. Maulik and D. Chakraborty, “Remote sensing image classification: A survey of support-vector-machine-based advanced techniques,” *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 33–52, Mar. 2017.
- [5] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, “Graph convolutional networks for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.
- [6] H. Zhao et al., “GCFNet: Global collaborative fusion network for multispectral and panchromatic image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5632814, doi: [10.1109/TGRS.2022.3215020](https://doi.org/10.1109/TGRS.2022.3215020).
- [7] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the Fisher kernel for large-scale image classification,” in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.
- [8] R. M. Anwer, F. S. Khan, J. Van De Weijer, M. Molinier, and J. Laaksonen, “Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 138, pp. 74–85, 2018.
- [9] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, “ORSIm detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5146–5158, Jul. 2019.
- [10] M. J. Swain and D. H. Ballard, “Color indexing,” *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.
- [11] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.
- [12] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2010, pp. 270–279.
- [13] 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.
- [14] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, “CNN-RNN: A unified framework for multi-label image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2285–2294.
- [15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*.
- [16] C. Szegedy et al., “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
- [17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [18] X. Wu, D. Hong, and J. Chanussot, “UIU-Net: U-Net in U-Net for infrared small object detection,” *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, Dec. 2022.
- [19] S. Chaib, H. Liu, Y. Gu, and H. Yao, “Deep feature fusion for VHR remote sensing scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.
- [20] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, “Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4355–4369, May 2021.
- [21] R. Minetto, M. P. Segundo, and S. Sarkar, “Hydra: An ensemble of convolutional neural networks for geospatial land classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6530–6541, Sep. 2019.
- [22] J. Xie, N. He, L. Fang, and A. Plaza, “Scale-free convolutional neural network for remote sensing scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916–6928, Sep. 2019.
- [23] H. Sun, S. Li, X. Zheng, and X. Lu, “Remote sensing scene classification by gated bidirectional network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.
- [24] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, “When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.
- [25] Y. Liu, C. Y. Suen, Y. Liu, and L. Ding, “Scene classification using hierarchical Wasserstein CNN,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2494–2509, May 2019.
- [26] C. Shi, X. Zhang, J. Sun, and L. Wang, “A lightweight convolutional neural network based on group-wise hybrid attention for remote sensing scene classification,” *Remote Sens.*, vol. 14, no. 1, 2021, Art. no. 161.
- [27] D. Hong et al., “More diverse means better: Multimodal deep learning meets remote-sensing imagery classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.
- [28] L. Bai, Q. Liu, C. Li, Z. Ye, M. Hui, and X. Jia, “Remote sensing image scene classification using multiscale feature fusion covariance network with octave convolution,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5620214, doi: [10.1109/TGRS.2022.3160492](https://doi.org/10.1109/TGRS.2022.3160492).

- [29] S. Liu et al., “A shallow-to-deep feature fusion network for VHR remote sensing image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5410213, doi: [10.1109/TGRS.2022.3179288](https://doi.org/10.1109/TGRS.2022.3179288).
- [30] F. Zhang, B. Du, and L. Zhang, “Scene classification via a gradient boosting random convolutional network framework,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Mar. 2016.
- [31] 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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1150–1162, Jan. 2022, doi: [10.1109/JSTARS.2022.3141826](https://doi.org/10.1109/JSTARS.2022.3141826).
- [32] W. Luo, H. Li, G. Liu, and L. Zeng, “Semantic annotation of satellite images using author–genre–topic model,” *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1356–1368, Feb. 2014.
- [33] G. Cheng, X. Xie, J. Han, L. Guo, and G. Xia, “Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020, doi: [10.1109/JSTARS.2020.3005403](https://doi.org/10.1109/JSTARS.2020.3005403).
- [34] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, “Multimodal GANs: Toward crossmodal hyperspectral–multispectral image segmentation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5103–5113, Jun. 2021.
- [35] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [36] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” 2020, *arXiv:2010.11929*.
- [37] 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.
- [38] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
- [39] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, “Big self-supervised models are strong semi-supervised learners,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 22243–22255.
- [40] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
- [41] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” 2020, *arXiv:2003.04297*.
- [42] P. Khosla et al., “Supervised contrastive learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 18661–18673.
- [43] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, “Vision transformers for remote sensing image classification,” *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 516.
- [44] L. Bashmal, Y. Bazi, and M. Al Rahhal, “Deep vision transformers for remote sensing scene classification,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2815–2818.
- [45] 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, Mar. 2022, Art. no. 4409512, doi: [10.1109/TGRS.2022.3157671](https://doi.org/10.1109/TGRS.2022.3157671).
- [46] P. Deng, K. Xu, and H. Huang, “When CNNs meet vision transformer: A joint framework for remote sensing scene classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Sep. 2021, Art. no. 8020305, doi: [10.1109/LGRS.2021.3109061](https://doi.org/10.1109/LGRS.2021.3109061).
- [47] K. Xu, P. Deng, and H. Huang, “Vision transformer: An excellent teacher for guiding small networks in remote sensing image scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5618715, doi: [10.1109/LGRS.2021.3109061](https://doi.org/10.1109/LGRS.2021.3109061).
- [48] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, “How to train your ViT? Data, augmentation, and regularization in vision transformers,” 2021, *arXiv:2106.10270*.
- [49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
- [50] W. Liu, Y. Wen, Z. Yu, and M. Yang, “Large-margin softmax loss for convolutional neural networks,” 2016, *arXiv:1612.02295*.
- [51] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, “Early convolutions help transformers see better,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 30392–30400.
- [52] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.

- [53] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
- [54] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.



Meiqiao Bi received the M.S. degree in computer technology from the Hebei University of Technology, Tianjin, China, in 2016.

She is currently a visiting student with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her research interests include remote sensing scene classification, computer vision, machine learning, and image processing.



Minghua Wang (Member, IEEE) received the B.S. degree in automation from the Harbin Institute of Technology (HIT), Harbin, China, in 2016, and the Ph.D. degree in control science and engineering from HIT, in 2021.

She was a visiting Ph.D. student with the Université Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, France, in 2019–2020. She is currently a Postdoc with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her research interests include hyper-

spectral image denoising, anomaly detection, tensor learning, and low-rank representation.



Zhi Li received the B.S. degree in remote sensing science and technology from the China University of Geosciences, Wuhan, China, in 2021. He is currently working toward the Ph.D. degree in cartography and geographic information systems with the University of Chinese Academy of Sciences, Beijing, China.

His research interests include remote sensing scene classification and hyperspectral image processing.



Danfeng Hong (Senior Member, IEEE) received the M.Sc. degree (summa cum laude) in computer vision from the College of Information Engineering, Qingdao University, Qingdao, China, in 2015, the Dr.-Ing degree (summa cum laude) from the Signal Processing in Earth Observation, Technical University of Munich, Munich, Germany, in 2019.

He is currently a Professor with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences (CAS), Beijing, China. Before joining the CAS, he has been a Research Scientist and led a Spectral Vision Working Group, Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany. He was also an Adjunct Scientist with GIPSA-lab, Grenoble INP, CNRS, Université Grenoble Alpes, Grenoble, France. His research interests include signal/image processing, hyperspectral remote sensing, machine/deep learning, artificial intelligence, and their applications in Earth Vision.

Dr. Hong is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), an Editorial Board Member for *Remote Sensing*, and an Editorial Advisory Board Member of the *ISPRS Journal of Photogrammetry and Remote Sensing*. He was the recipient of the Best Reviewer Award of the IEEE TGRS in 2021 and 2022, the Best Reviewer Award of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING in 2022, the Jose Bioucas Dias Award for recognizing the outstanding paper at WHISPERS in 2021, the Remote Sensing Young Investigator Award in 2022, the IEEE GRSS Early Career Award in 2022, and a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters) in 2022.