Annals of Emerging Technologies in Computing (AETiC)
Vol. 8, No. 4, 2024

Research Article

Variance Consistency Learning: Enhancing Cross-Modal Knowledge Distillation for Remote Sensing Image Classification

Huaxiang Song*, Yong Zhou, Wanbo Liu, Di Zhao, Qun Liu and Jinling Liu

School of Geography Science and Tourism, Hunan University of Arts and Science, China
cn11028719@huas.edu.cn; zy720625@huas.edu.cn; liuwanbo0201@huas.edu.cn; zhaodi@huas.edu.cn; liuqun@huas.edu.cn; liujl@huas.edu.cn
*Correspondence: cn11028719@huas.edu.cn

Received: 1thAugust 2024; Accepted: 29th September 2024; Published: 1st October 2024

Abstract: Vision Transformers (ViTs) have demonstrated exceptional accuracy in classifying remote sensing images (RSIs). However, existing knowledge distillation (KD) methods for transferring representations from a large ViT to a more compact Convolutional Neural Network (CNN) have proven ineffective. This limitation significantly hampers the remarkable generalization capability of ViTs during deployment due to their substantial size. Contrary to common beliefs, we argue that domain discrepancies along with the RSI inherent natures constrain the effectiveness and efficiency of cross-modal knowledge transfer. Consequently, we propose a novel Variance Consistency Learning (VCL) strategy to enhance the efficiency of the cross-modal KD process, implemented through a plug-and-plug module within a ViTteachingCNN pipeline. We evaluated our student model, termed VCL-Net, on three RSI datasets. The results reveal that VCL-Net exhibits superior accuracy and a more compact size compared to 33 other state-of-the-art methods published in the past three years. Specifically, VCL-Net surpasses other KD-based methods with a maximum improvement in accuracy of 22% across different datasets. Furthermore, the visualization analysis of model activations reveals that VCL-Net has learned long-range dependencies of features from the ViT teacher. Moreover, the ablation experiments suggest that our method has reduced the time costs in the KD process by at least 75%. Therefore, our study offers a more effective and efficient approach for cross-modal knowledge transfer when addressing domain discrepancies.

Keywords: ***Cross-Modal; Deep Learning; Knowledge Distillation; Remote Sensing Image Classification***

1. Introduction

Remote Sensing Images (RSIs) provide a crucial method for Earth observation, facilitating various applications such as environmental monitoring [1], agriculture [2], land management [3], and surveying [4]. With advancements in imaging techniques, RSIs have evolved into a form of big data, encompassing multiple spatial and temporal dimensions. Consequently, only computer algorithms can effectively perform RSI recognition tasks. Classification is a fundamental component of these algorithms for RSI understanding, as advancements in classification often drive improvements in subsequent tasks like detection and segmentation. A decade ago, shallow machine learning models were central to RSI classification, requiring extensive feature mining experiments but often resulting in suboptimal accuracy. With the advent of deep learning, Convolutional Neural Networks (CNNs) have dominated RSI classification tasks over the past decade due to their superior accuracy and automatic feature extraction capabilities [5].

Convolutional Neural Networks (CNNs) possess a progressively expanding visual field, which enhances their ability to generalize local patterns effectively [6]. However, this expansion can lead to the loss of dependencies among local features, a phenomenon that contradicts human cognitive habits. The

Huaxiang Song, Yong Zhou, Wanbo Liu, Di Zhao, Qun Liu and Jinling Liu, "Variance Consistency Learning: Enhancing Cross-Modal Knowledge Distillation for Remote Sensing Image Classification", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 56-76, Vol. 8, No. 4, 1st October 2024, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2024.04.003, Available: http://aetic.theiaer.org/archive/v8/v8n4/p3.html.

Vision Transformer (ViT) [7], a novel architecture capable of capturing long-range feature dependencies, has emerged as a promising solution in computer vision. While ViTs can compete with CNNs, they often require a significantly larger number of parameters to achieve comparable accuracy. This poses a challenge for the field of remote sensing, which frequently relies on mobile or embedded devices. Consequently, the substantial size of ViTs significantly limits their application and deployment in RSI tasks.

Model compression holds significant potential for addressing the trade-off between accuracy and efficiency. Bucilă *et al.* [8] introduced an innovative method for transferring knowledge from a complex model (the teacher) to a more compact one (the student) using prediction logits. This concept was later expanded as knowledge distillation (KD). However, the logit-based KD process typically requires a substantial number of training epochs, often in the tens of thousands, to mitigate accuracy loss [10–11]. Consequently, researchers [12–13] proposed feature alignment, which aligns features of intermediate layers. Nevertheless, feature-based approaches necessitate additional function modules for both teacher and student models, leading to a higher parameter count than logit-based methods. Therefore, logit-based KD methods retain their advantages, provided the efficiency of knowledge transfer is improved.

Currently, effective logit-based KD methods primarily focus on knowledge transfer between models with the same architecture, such as a CNN teacher and a CNN student [14–15]. However, given the unique advantages of both CNNs and ViTs [16], the importance of cross-modal KD becomes evident. Since 2021, researchers have introduced innovative concepts for cross-modal KD, leading to significant advancements [17–19]. Nonetheless, most cross-modal KD methods have exhibited substantial accuracy losses [20–22]. Therefore, there remains considerable potential for improvement in current cross-modal KD strategies [23].

Recently, researchers have proposed several KD methods for RSI classification. However, existing studies still face significant limitations. Most KD techniques either achieve compactness at the expense of competitive accuracy [24–27] or deliver acceptable accuracy with a considerably larger volume [28–30]. In other words, the majority of these methods have not successfully balanced accuracy and efficiency. The authors believe that this dilemma arises from several underlying factors.

Figure 1. Comparative Analysis of Feature Recognition in Natural Images and Remote Sensing Images

As illustrated in Figure 1, the two natural images on the left can be easily differentiated using only partial visual features of the animals, such as their heads. In contrast, the four RSI samples on the right necessitate a comprehensive combination of local features for classification. For example, the round roof cannot be used to distinguish between the center and church scenes. Likewise, the rectangular roofs of buildings are not unique to commercial or industrial classes. Consequently, the noisy backgrounds and greater similarity between categories constitute domain gaps when compared to natural images.

Currently, researchers typically employ CNNs or ViTs developed on ImageNet-1K, a large-scale dataset containing a million natural samples, for RSI classification. However, most of these methods simply replicate the training procedures designed for natural images, resulting in suboptimal models due to unaddressed domain discrepancies [31].

The KD process is a function approximation that employs a student model to approximate the teacher function. Intrinsically, various KD techniques exhibit different capabilities for reducing deviations between teacher and student predictions. However, no KD algorithm can completely eliminate the variance originating from data distribution in training. As depicted in Figure 2, the round spots in black represent different data distributions with varying variances, denoted as $D(x)$, and the teacher and student functions are represented as $f(x)$ in red and blue, respectively. In the left sub-chart, smaller variances in the training data guide the function curves of the teacher and student to be more consistent. In other words, smaller variances facilitate the student's approximation to the teacher, assuming that the

KD algorithm is effective. Conversely, as depicted in the right sub-chart, larger variances result in the student having larger margins with the teacher. This is because their function curves are more likely to contain larger variances.

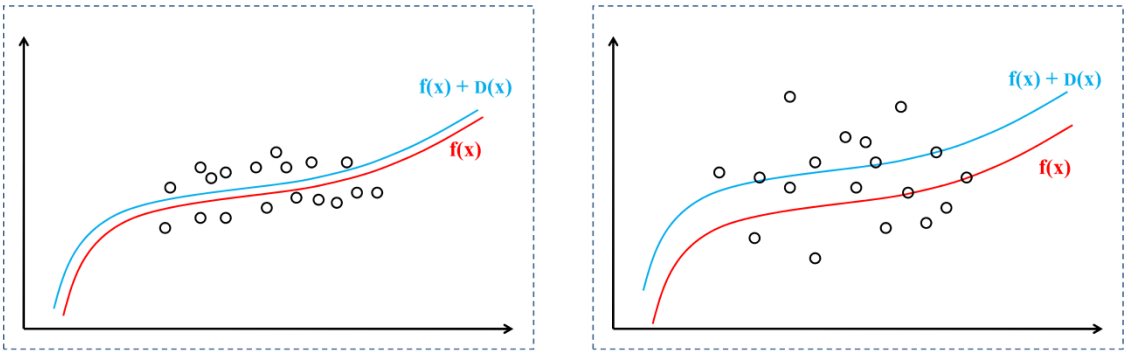

Figure 2. Comparative Analysis of Variance Impact on Student-Teacher Approximation in KD Techniques

At present, logit-based KD algorithms primarily focus on minimizing the deviation between the function curves of the teacher and student models [32–33]. To the best of our knowledge, existing studies in the literature have entirely overlooked the issue of variance in the KD process. Objectively, the problem of variance can be mitigated when there is a sufficient volume of training data, such as a dataset with a million samples. However, RSI datasets typically only contain tens of thousands of data points, as RSI tasks are often case-specific. Moreover, the visual features of RSIs generally exhibit a larger inter-class dissimilarity compared to natural images. Therefore, we argue that these inherent characteristics of RSIs need to be adequately addressed when applying KD techniques.

In response to the aforementioned issues, we propose a novel approach for cross-modal knowledge transfer. Contrary to popular belief, we argue that the inefficiency in the logit-based KD process arises from disparities in data distribution within RSI datasets. Consequently, we introduce a novel variance consistency learning (VCL) strategy aimed at providing efficient and precise KD solutions for RSI classification. Specifically, we have incorporated our VCL algorithm into a plug-and-play module that facilitates DA and regularization techniques in a manageable manner, resulting in constrained yet acceptable variance levels within RSIs. The effectiveness of our method was evaluated on three benchmark RSI datasets. Experimental results demonstrate that our VCL-based model, termed VCL-Net, outperforms 33 other advanced methods from the past three years, exhibiting significant improvements in both accuracy and efficiency. Notably, compared to other KD methods, VCL-Net achieves substantial accuracy enhancements of up to 22%. The primary contributions of this research are as follows:

1) We introduce a cross-modal approach for RSI classification that utilizes a ViT to teach a CNN through distillation. Our VCL-Net is more lightweight and exhibits superior accuracy and efficiency compared to other state-of-the-art methods in the literature.
2) VCL-Net exhibits remarkable improvements in accuracy when compared to other KD methods reported in the past three years. These results suggest that our VCL strategy is a more effective and efficient method for cross-modal distillation.
3) We initially emphasized the crucial role of variances arising from the inherent nature of RSIs in the KD process. As a promising solution, our VCL module, with its plug-and-play attribute, can be integrated into any KD process.

2. Related Works

Since 2018, researchers have put forward a series of KD methods to create compact yet precise classifiers for RSIs. For instance, Chen *et al.* [26] have shown the effectiveness of logit-based KD in classifying RSIs. Xing *et al.* [27] have introduced a collaborative KD method, which brings in the novel concept of mutually supervised learning during training. Hu *et al.* [28] have designed a functional module to boost distillation effectiveness. However, these methods primarily excel in compactness. On the other hand, despite the larger size of student models, Li *et al.* [29] and Zhao *et al.* [30] have put forward hint-based KD methods that have shown improvements in accuracy.

In the specialized field of cross-modal KD, a limited number of innovative approaches have been proposed by researchers. For instance, Xu *et al.* [24] presented a ViT-teaching-CNN method, employing a strategy that allows the ViT to cease instruction partway through the process. In a different approach, Wang *et al.* [25] suggested a self-distillation technique for ViT, utilizing a contrastive learning pipeline. Nabi *et al.* [34] put forward a synchronized training method via a CNN-teaching-ViT framework. Zhao *et al.* [35] developed functional modules that can distill knowledge extracted by a multi-sample contrastive network to enhance a CNN or ViT. Despite the creativity of these methods, there still remains significant potential for improvement in both the accuracy and compactness of the models.

RSIs typically have noisier backgrounds compared to natural images. Consequently, a neural architecture search (NAS) for CNNs may yield more efficient and compact models than those developed on ImageNet-1K. For instance, Ao *et al.* [36] presented their NAS method, which employs a two-phase evolutionary process. Broni-Bediako *et al.* [37] suggested a NAS method that utilizes a symbolic linear generative encoding strategy. Shen *et al.* [38] put forward a NAS method that incorporates a multistage network progressive fusion pipeline. However, despite some of the resulting models being compact, most of these methods only achieved below-average accuracy.

In a similar vein to NAS methods, researchers have proposed other lightweight methods based on CNN or ViT, utilizing self-designed architectures. For instance, Chen *et al.* [39] proposed a multi-branch local attention method that employs an enhancement strategy for ResNet with embedded attention modules. Huang *et al.* [40] introduced a stochastic depth method that integrates convolutional blocks with their coordinate attention. Shi *et al.* [41] presented a self-designed CNN that incorporates a unique self-compensating convolution structure. Xu *et al.* [42] suggested a CNN structure-based method that utilizes Lie group encoding for image decomposition. Bai *et al.* [43] put forward a multi-scale feature fusion approach that employs octave convolution for processing RSI multi-frequency and multi-scale features. Zhang *et al.* [44] proposed a Laplacian-CNN method that leverages Laplacian operators to capture high-frequency features. Huang *et al.* [45] suggested a lightweight transformer-based method that uses multi-level group convolution modules in conjunction with transformer blocks. However, most of these methods do not demonstrate competitive performance when evaluated based on accuracy.

Domain discrepancies in RSIs have hindered the effectiveness of models developed on ImageNet-1K. Despite having more parameters than CNNs, most ViT-based methods reported in the literature are not highly competitive for RSI classification. For instance, Bazi *et al.* [46] demonstrated the potential effectiveness of ViT for RSI classification. Wang *et al.* [47] posited that ViT models, such as Swin-ViT, could be more effective when pre-trained on large-scale datasets comprising one million RSIs. Lv *et al.* [48] proposed a progressive aggregation strategy for ViTs, aiming to enhance the representation abilities of spatial channel features.

Furthermore, domain gaps often hinder many innovative methods from being competitive, even though their pipelines frequently consist of multiple models. For instance, Shen *et al.* [49] and Xu *et al.* [50] proposed two analogous multi-CNN methods that cascade two CNNs in parallel to capture global and local features. Similarly, Tang *et al.* [51] and Wang *et al.* [52] proposed two CNN cooperative methods that utilize unique loss functions as supervision indicators for two parallel CNNs. In contrast, researchers have also proposed other multi-model concepts based on ViTs. For example, Zhang *et al.* [53] proposed a ViT-based method that uses the features from a CNN's final pooling layer as patch embeddings. Wang *et al.* [54] proposed another ViT-based method that employs the features extracted by CNNs as additional embedded tokens for another ViT. Deng *et al.* [55] and Zhao *et al.* [56] proposed two similar methods that utilize a CNN and a ViT in parallel as cooperative components. Ma *et al.* [57] proposed an intriguing concept that uses a dual-branch module to mine homogeneous and heterogeneous patches for two parallel ViTs. Cheng and Lei [58] proposed an ensemble method that employs multiple lightweight CNNs cascaded before Hidden Markov Models as individual classifiers. However, few of these multi-model methods have achieved a competitive balance between accuracy and model size.

3. Methodologies

3.1. Variance Consistency Learning

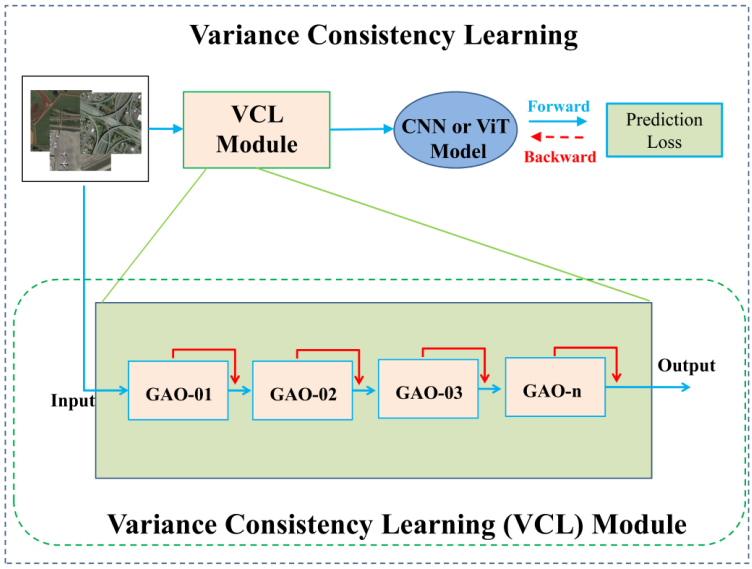

Figure 3. Framework and Structure of the VCL Strategy and Module

Figure 3 illustrates the framework of our VCL strategy and the structure of the VCL module. Within our learning framework, the VCL module processes original RSI samples as inputs and outputs transformed samples for model training. Specifically, we have designed seven gated activation operators (GAOs) within the VCL module, each comprising three crucial components: a stochastic probability generator, a DA or regularization function, and a conditional branch. During operation, the generator samples a stochastic probability along with the input samples. Within the branch, if the sampled probability does not exceed a predetermined threshold, the function transforms the input samples. Otherwise, the input and output samples remain identical. The seven GAOs, sequentially arranged within the VCL module, include color jitter, horizontal or vertical flip, rotation, random erasing, random resize crop, and CutMix.

3.2. Proposed KD Method's Framework

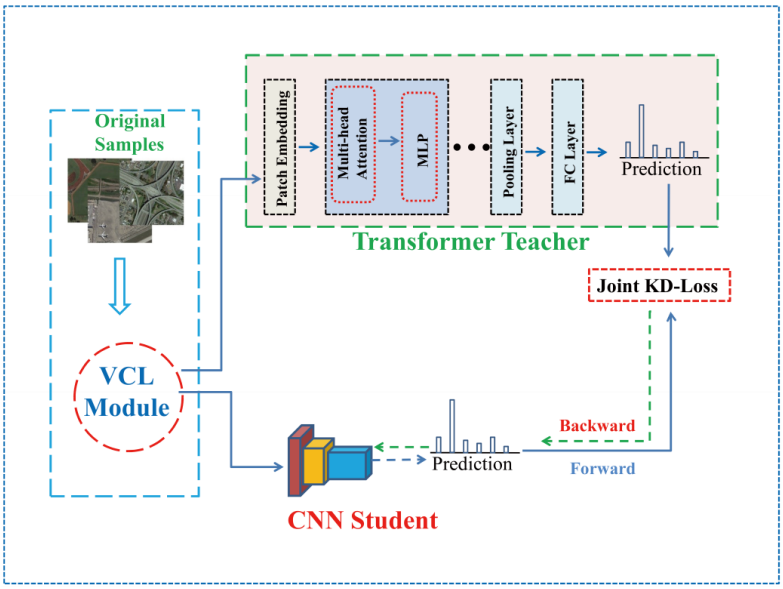

Figure 4. Framework of the proposed KD method

Figure 4 illustrates the framework of the proposed KD approach. Initially, the original RSI samples are processed using the VCL module. Subsequently, the transformed samples are input into the ViT teacher and CNN student for forward computation. As shown at the top of Figure 4, the ViT teacher

consists of multiple transformer blocks, each containing two essential components: multi-head attention and multi-layer perceptron (MLP) layers. After forward computation, the KD loss function processes the prediction logits from both the teacher and student models to produce a joint KD loss. The parameters of the student model are then updated via backpropagation based on this joint KD loss.

The training pipeline spans a total of 600 epochs, with only the student CNN being trainable. A learning rate (LR) of 0.0002 is employed throughout the training process across all RSI datasets. Notably, the probability settings vary only for a fine-grained RSI dataset due to its differences in object granularity.

3.3. Model Structures

Our study used the Next-ViT-Small (N-ViT-S) as the teacher model, with reference [59] providing a detailed description of its architecture. Unlike the traditional ViT [7], N-ViT-S introduces a hybrid-attention structure that integrates cascaded convolution operations with self-attention mechanisms. Table 1 demonstrates that N-ViT-S, despite using fewer parameters, achieves higher accuracy than the conventional ViT-base model. In this table, 'FLOPs' measures floating-point operations in billions (G), and the 'Top-1' column indicates the accuracy of these models on the ImageNet-1K dataset.

In assessing the efficacy of our proposed method, we employed the EfficientNet-B0 as the student model, with its comprehensive architecture detailed in reference [60]. This model, unlike other CNNs such as ResNets, incorporates built-in channel attention structures. Table 1 illustrates that EfficientNet-B0, despite its compact size, achieves satisfactory accuracy.

Table 1. Comparative Analysis of Model Performances on the ImageNet-1K Dataset

Model	Param (M)	FLOPs (G)	Top-1 (%)
ViT-Base	86.6	17.6	81.1
N-ViT-S	31.7	5.8	82.5
EfficientNet-B0	5.3	0.4	77.7

3.4. KD Loss

Let's consider an RSI dataset, symbolized as $S = \{x_i, y_i\}$, where x_i and y_i denote each RSI sample and its corresponding label within the set S, respectively. During forward computation, a classifier, when given x_i as input, will generate a prediction logit not only for the intended category but also for non-target classes. In the context of modern deep learning models, each input x_i is typically normalized to a tensor with values within the range of 0 to 1. Consequently, a classifier can fundamentally be perceived as a function, symbolized as f, which takes tensors as input and yields logit vectors as output. The function f can be characterized as follows:

$$y_i = f(x_i) \tag{1}$$

Assume that c represents the category number within S. Consequently, each y_i in Equation (1) is a vector encompassing c prediction logits, which can be represented as $y_i \in \mathbb{R}^{1\times c}$.

At present, deep learning models often possess an enormous number of parameters, sometimes even reaching into the trillions. This reality makes deployment in remote sensing particularly challenging, especially in cases like the ViT-Base model, which has close to 90 million parameters. Bucila et al. [8] pioneered the concept of model compression, a process that employs a smaller (student) model to mimic a larger, more robust (teacher) model. Expanding on this foundation, Hinton et al. [9] refined this technique of knowledge transfer, giving it the name KD.

Utilizing the logits defined in Equation (1), y_i undergoes a transformation into probabilities that correspond to each category, symbolized as p_i, via a softmax function. This transformation process can be articulated as follows:

$$p_i = softmax(y_i) = \frac{\exp(y_i)}{\sum_{i=1}^{c}\exp(y_i)} \tag{2}$$

Commonly, logit-based KD methods utilize the objective function of Kullback-Leibler (KL) divergence for the loss computation. Throughout the distillation process, this function accepts probabilities from both the teacher and student models as inputs. The loss, symbolized as $\mathcal{L}_{KD}$, can be articulated as follows:

$$\mathcal{L}_{KD} = KL(P_t \parallel P_s) = \sum_{i=1}^{c}(P_{t,i} \times log\frac{P_{t,i}}{P_{s,i}}) \tag{3}$$

In this context, $P_{t,i}$ and $P_{s,i}$ represent the probabilities derived from the teacher and student models, respectively.

In a robust model, the prediction probability for the intended categories often escalates to nearly 98%, while it could plummet to as low as 0.1% or even lower for the non-intended classes. As a result, the logits computed by equation (2) associated with the intended categories will overshadow those of the non-intended ones, particularly when the teacher model demonstrates high precision. To mitigate this concern, we commonly introduced a hyperparameter, denoted as t, to modulate the model's prediction logits. They integrated t into equation (2), thereby modifying the data distribution of the logits. The softening process can be delineated as follows:

$$p_i^t = softmax\left(\frac{y_i}{t}\right) \tag{4}$$

Given these conditions, the softened loss, symbolized as $\mathcal{L}_{\mathrm{KD}}^t$, can be reformulated as follows:

$$\mathcal{L}_{\mathrm{KD}}^t = \sum_{i=1}^{c}(t^2 \times P_t^t \ log\frac{P_t^t}{P_s^t}) \tag{5}$$

The distillation process, when solely reliant on $\mathcal{L}_{\mathrm{KD}}^t$, can be quite time-consuming in reducing the accuracy gaps. To accelerate the KD process, the training loss, which is exclusively computed based on the student model's predictions, like cross-entropy loss, is often combined with $\mathcal{L}_{\mathrm{KD}}^t$. Consequently, a conventional KD training loss, represented as $\mathcal{L}_{\mathrm{KD-training}}$, usually consists of two elements, which can be articulated as follows:

$$\mathcal{L}_{\mathrm{KD-training}} = -\sum_{\mathrm{i=1}}^{\mathrm{c}}(y_i \log P_s) + \sum_{i=1}^{c}(t^2 \times P_t^t \ log\frac{P_t^t}{P_s^t}) \tag{6}$$

Nonetheless, the loss in Equation (6) necessitates an extensive process, potentially spanning tens of thousands of training epochs, before a student model attains stratified accuracy [10–11]. This can be attributed to the reduced information entropy resulting from the exclusive use of the KL divergence.

Huang et al. [61] proposed an alternative loss function, termed DIST, as a potentially more efficient solution. Specifically, the DIST loss utilizes the Pearson distance to modify the loss computation defined in Equation (5). Let us denote the Pearson correlation coefficient and the Pearson distance as ρ and D_P, respectively. Then, D_P can be expressed as follows:

$$D_P = 1 - \rho(V_t, V_s) = 1 - \frac{\sum_{\mathrm{i=1}}^{\mathrm{c}}(V_t - \overline{V}_t)(V_s - \overline{V}_s)}{\sqrt{\sum_{\mathrm{i=1}}^{\mathrm{c}}(V_t - \overline{V}_t)^2 \sum_{\mathrm{i=1}}^{\mathrm{c}}(V_s - \overline{V}_s)^2}} \tag{7}$$

In Equation (7), it uses V_t and V_s to denote the probability vectors of the teacher and student models, respectively.

DIST, utilizing the Pearson distance, initially presents its inter-class loss, symbolized as $\mathcal{L}_{inter}$. This loss is computed via equation (7), taking V_t and V_s as inputs. Given that the training batch size is represented as N, $\mathcal{L}_{inter}$ can be articulated as follows:

$$\mathcal{L}_{inter} = \frac{1}{N}\sum_{i=1}^{N} D_P(V_t, V_s) \tag{8}$$

In addition, DIST presents an intra-class loss, symbolized as $\mathcal{L}_{intra}$. This loss is calculated using equation (7), with the transposes of V_t and V_s at the N and c dimensions serving as inputs. In this context, c signifies the count of categories within a dataset. Consequently, $\mathcal{L}_{intra}$ can be articulated as follows:

$$\mathcal{L}_{intra} = \frac{1}{c}\sum_{j=1}^{c} D_P({V_t}^T, {V_s}^T) \tag{9}$$

In this study, we initially utilize DIST to substitute the softened loss, as defined in Equation (5). Following this, we retain the cross-entropy loss of the student, as outlined in Equation (6), without any changes. Additionally, we establish the temperature hyperparameter at a value of 2 when calculating probabilities. Consequently, the KD loss in our knowledge transfer process should be redefined as follows:

$$\mathcal{L}_{\mathrm{KD-training}} = -\sum_{\mathrm{i=1}}^{\mathrm{c}}(y_i \log \mathrm{P_s}) + \mathcal{L}_{inter} + \mathcal{L}_{intra} \tag{10}$$

The CutMix method employs a hyperparameter, denoted as α, to determine the modified label subsequent to the cutting and mixing of samples. This hyperparameter is characterized as the proportion of the patch area associated with Class B to the total area of a specific sample from Class A. We can represent the CutMix labels as $Label_{CM}$, the Class A labels as $Label_A$, and the Class B labels as $Label_B$. The formulation of the CutMix labels can be expressed as follows:

$$Label_{CM} = (1 - \alpha) \times Label_A + \alpha \times Label_B \quad (11)$$

3.5. Distillation Algorithm

Algorithm 1. Distillation procedure using pseudo-code

Definitions: The training and testing RSI subsets are denoted as $S_{train} = \{(x_i, y_i)\}$ and $S_{test} = \{(x_i, y_i)\}$, respectively. The teacher and student models and the CutMix algorithm are symbolized by f_T, f_S, and f_{CM}, respectively. The transformations in DA are represented by DA_{Tr}. The notation f_{CE} is used to denote the cross-entropy loss function. The notations P_T and P_S are used to denote the predicted probabilities of teacher and student models, respectively.
Input: Images and labels from the training or testing subsets.
Output: the accuracy (Acc) results of the student classifier.

1 **For** Epoch = 1, 2, . . . , 600 **Do**
2 **For** iteration = 1 to $\left(\frac{length(S_{train})}{64} + 1\right)$ **Do**
3 Sample a batch of samples from S_{train}, and input them to the functions of f_T and f_S.
4 Predict probabilities using the equation: $P_T = f_T(\, f_{CM}(DA_{Tr}(x_i))\,)$
5 Predict probabilities using the equation: $P_S = f_S(\, f_{CM}(DA_{Tr}(x_i))\,)$.
6 Calculate the loss using the equation: $Loss = f_{CE}(P_S,\ f_{CM}(y_i)) + \mathcal{L}_{inter}(P_T, P_S) + \mathcal{L}_{intra}(P_T, P_S)$.
7 Update parameters through back propagation.
8 **End For**
9 Calculate the student classifier's accuracy using the equation: $Acc = (f_S(x_i) == y_i)$, where $x_i, y_i \in S_{test}$, and save the Acc result.
10 **End For**
11 **Return** the Acc results

Algorithm 1 provides a detailed outline of the distillation procedures for the proposed method. The entire process spans 600 training epochs, as stated in line 1. For each epoch, lines 2 and 3 illustrate that a batch of 64 images and their corresponding labels are fed into the teacher and student models after undergoing DA techniques. The prediction probabilities for the samples are then calculated and used in the distillation loss function, as shown in lines 4, 5, and 6. Gradients are subsequently computed to facilitate the update of the student model's parameters. As depicted in lines 8, 9, 10, and 11, the accuracy of the student classifier is assessed at the conclusion of each epoch, and a record of the accuracy is maintained.

Regarding the additional hyperparameter configurations, the initial LR is established at 2×10^{-4} and is governed by the cosine decay algorithm, with a lower limit for the LR set at 2×10^{-5}. The Adam-W optimizer is utilized with a weight decay parameter of 1×10^{-6}. For the weights in the batch normalization layer, the decay parameter is assigned a value of zero. Moreover, a consistent resolution of 256^2 is preserved throughout both the training and testing phases across all datasets.

During the training phase of the N-ViT-S teacher model, we utilized an algorithm specifically designed for inherent RSI characteristics. Detailed information can be found in reference [23]. We made a significant adjustment to the algorithm: we used a smaller initial LR of 5×10−5 and extended the training duration to 600 epochs.

3.6. Dataset and Division

We use two widely recognized benchmarks [63], the Aerial Image Dataset 30 (AID30) and the North Western Polytechnic University 45 (NWPU45), for effectiveness evaluation. The AID30 includes 30 separate categories, with a total of 10,000 images, each uniformly presenting a resolution of 600^2. By comparison, the NWPU45 contains 45 distinct categories with a total of 31,500 images, each consistently exhibiting a resolution of 256^2.

In order to evaluate the effectiveness of the method on low-resolution RSIs, we also utilized the Aircraft Fine Grain Recognition 50 (AFGR50) dataset [54] as a benchmark. The AFGR50 dataset encompasses 50 distinct categories, collectively containing 12,500 images, each uniformly presenting a resolution of 128^2.

Both the NWPU45 and AFGR50 maintain a consistent number of samples across categories, with 700 and 250 samples per category, respectively. In contrast, the AID30 is imbalanced, with the number of samples per class varying between 220 and 420. Randomly selected samples from each category across these three datasets are displayed in Figures 5, 6, and 7.

Figure 5. Representative Samples per Category of the AID30 Dataset

Figure 6. Representative Samples per Category of the NWPU45 Dataset

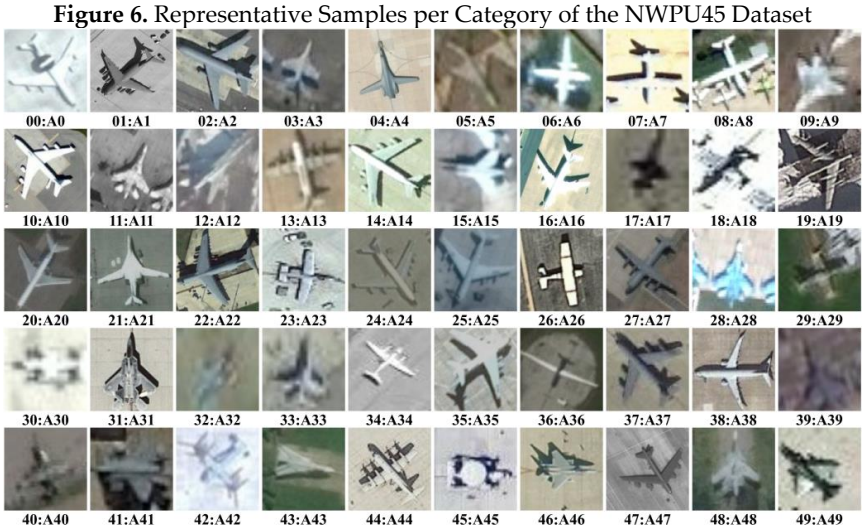

Figure 7. Representative Samples per Category of the AFGR50 Dataset

To ensure an equitable comparison, we adhered to the training ratio (TR) delineated in the extant literature: AID30 at 20% and 50%; NWPU45 at 10% and 20%; and AFGR50 at 10%, 20%, and 30%. For each TR, we randomly selected samples from the whole dataset to form the training subsets, while the remaining samples were designated as testing subsets.

3.7. Performance Evaluation Metrics

We employed the overall accuracy (OA) and the confusion matrix as evaluation metrics, consistent with existing literature. The symbol N_c denotes the total count of samples correctly classified, while N_t signifies the total count of classified samples. Hence, OA can be expressed as follows:

$$OA = \frac{N_c}{N_t} \tag{12}$$

The confusion matrix exhibits the classification outcomes for all categories within a dataset. This figure is an organized tabular representation that offers comprehensive details about the number of samples that have been correctly and incorrectly classified per category.

4. Experimental Results

4.1. OA Results

We conducted a comparative analysis of effectiveness using OA as the criterion. The results of this analysis are displayed in Tables 2, 3, and 4. This comparison encompasses 32 advanced methods published within the last three years. The tables sequentially present the results for AID30, NWPU45, and AFGR50. In these tables, the column titled 'Algorithm Uniqueness' offers an overview of the unique pipelines used by various methods. The 'Params' column contains either the original data as mentioned in the relevant literature or evaluations based on the foundational structures of the corresponding models. The use of 'None' in the tables signifies the absence of comprehensive details in the related literature.

4.1.1. OA results for AID30

Table 2. OA (%) Comparison among Different Methods Using the AID30 Dataset

Methods	Algorithm Uniqueness	Params (M)	AID30 (%)	
			TR-20%	TR-50%
RE-EfficientNet [23]	Single EfficientNet-B3	12.0	97.11 ± 0.06	98.15 ± 0.10
ET-GS-Net [24]	ViT-teaching-CNN	11.7	95.58 ± 0.18	96.88 ± 0.19
LaST-Net [25]	ViT self-distillation	> 28.0	83.23	87.34
TST-Net [26]	CNN-teaching-CNN	1.0	85.50	None
CKD-Net [27]		None	None	None
VSDNet [28]	Hint-based KD	>8.0	96.73 ± 0.15	97.95 ± 0.10
DKD-Net [29]		4.4	95.09	96.94
ESD-MBENet [30]		23.9	96.39 ± 0.21	98.40 ± 0.23
CT-ViT [34]	CNN-teaching-ViT	86.9	96.74 ± 0.13	None
EMSC-Net [35]		>88.6	96.02 ± 0.18	97.35 ± 0.17
TPENAS-Net [36]	NAS	1.7	None	None
SLGE-Net [37]		5.1	96.10 ± 0.18 (TR-60%)	
DARTS-Net [38]		3.8	95.65 (TR-60%)	
MBLA-Net [39]	Attention CNN	>25.6	95.60 ± 0.17	97.14 ± 0.03
LRSCM-Net [40]		7.6	95.41	97.28
SCCNN [41]	Self-designed CNN	0.5	93.15 ± 0.25	97.31 ± 0.10
LGRIN [42]		4.6	94.74 ± 0.23	97.65 ± 0.25
MF2C-Net [43]		33.2	95.54 ± 0.17	97.02 ± 0.28
LH-Net [44]		>46.8	93.30 ± 0.10	97.81 ± 0.13
LT-Net [45]	Single Transformer	8.2	94.98 ± 0.08	None
ViT-Base [46]		88.6	94.97 ± 0.01	None
Swin-ViT-Tiny [47]		28.3	96.55 ± 0.03	98.10 ± 0.06
SCViT [48]		>88.6	95.56±0.17	96.98±0.16
ACGL-Net [49]	Dual-CNNs	33.6	94.44 ± 0.09	96.10 ± 0.10
GLDBS-Net [50]		>23.4	95.45 ± 0.19	97.01 ± 0.22
AC-Net [51]		>276.6	93.33 ± 0.29	95.38 ± 0.29
T-CNN [52]		15.9	94.55 ± 0.27	96.72 ± 0.23
TRS-Net [53]	CNN-and-ViT	46.3	95.54 ± 0.18	98.48 ± 0.06
P2FEViT [54]		>94.9	94.72 ± 0.04	95.85 ± 0.15
CT-Net [55]		>107.8	96.25 ± 0.10	97.70 ± 0.11
L2RCF-Net [56]		46.7	97.00 ± 0.17	97.80 ± 0.22
HHTL-Net [57]	Dual-ViTs	>177.2	96.52 ± 0.13	96.88 ± 0.21
CNN–HMM Ensemble [58]	Four-CNN ensemble	19	93.93 ± 0.15	97.81 ± 0.04
ViT Teacher (This work)	Single N-ViT-S	28.4	97.65 ± 0.07	98.56 ± 0.13
VCL-Net (This work)	Single Efficient-B0	5.3	**97.10** ± 0.09	**98.22** ± 0.02

As demonstrated in Table 2, our teacher model exhibits superior OA values in comparison to other methods. We posit that the extensive attention operators in the N-ViT-S architecture render it more sensitive than CNNs. Consequently, our training strategy, which is based on the inherent characteristics of RSIs, evidently outperforms other Vision ViT methods. The OA of our student model indicates that our

cross-modal method can effectively transfer knowledge from the ViT teacher. Moreover, the OA margin at the 20% TR is slightly larger than that at the 50% TR. It is reasonable to assume that a smaller training subset may contain larger variances in data distribution. Therefore, we contend that the results of the OA margins are justifiable.

In comparison to other methods, we observed that only RE-EfficientNet [23] exhibits a competitive OA value at the 20% TR, despite having 2.3 times the parameters of VCL-Net. However, for the 50% TR, ESD-MBENet [30] and TRS-Net [53] show OA improvements of approximately 0.2% when using VCL-Net as a baseline. We attribute these improvements to three factors as follows:

Firstly, AID30 is an imbalanced dataset, with the number of samples in the most confusing categories significantly below the average. For instance, the 'church' class only has 240 samples, while the average number is 333. Consequently, we posit that the OA improvements may include accuracy fluctuations resulting from the random division of training subsets.

Secondly, ESD-MBENet and TRS-Net possess significantly more parameters than our VCL-Net. In deep learning, a larger model volume typically correlates with better generalization capability, albeit at the expense of model efficiency.

Lastly, and more importantly, CNNs excel at local feature extraction, while ViTs are superior at handling long-range dependencies of features. Therefore, our VCL-Net may not be able to transfer all the 'dark knowledge' from the ViT teacher due to structural differences.

In comparison to other KD methods, we observed that VCL-Net demonstrates significant improvements in OA, ranging from 1.0% to 10.0% at the 20% TR. For the 50% TR, the advantages of VCL-Net are largely consistent, with the exception of ESD-MBENet. These results suggest that VCL-Net is more effective than other KD methods when benchmarked against AID30.

When compared to other multi-model approaches, we found that the vast majority have not achieved competitive accuracy, despite their extensive parameters serving as expectation indicators. This comparison further underscores that VCL-Net is a more efficient classifier than its counterparts.

4.1.2. OA results for NWPU45

Table 3. OA (%) Comparison among Different Methods Using the NWPU45 Dataset

Methods	Algorithm Uniqueness	Params (M)	NWPU45 (%)	
			TR-10%	TR-20%
RE-EfficientNet [23]	Single EfficientNet-B3	12.0	94.60 ± 0.05	96.15 ± 0.03
ET-GS-Net [24]	ViT-teaching-CNN	11.7	92.72 ± 0.28	94.50 ± 0.18
LaST-Net [25]	ViT self-distillation	> 28.0	72.58	73.67
TST-Net [26]	CNN-teaching-CNN	1.0	80.00(TR-50%)	
CKD-Net [27]		None	91.6 (TR is not clear)	
VSD-Net [28]	Hint-based KD	>8.0	93.24 ± 0.11	95.67 ± 0.11
DKD-Net [29]		4.4	93.72	95.76
ESD-MBENet [30]		23.9	93.05 ± 0.18	95.36 ± 0.14
CT-ViT [34]	CNN-teaching-ViT	86.9	93.88 ± 0.07	
EMSC-Net [35]		>88.6	93.58 ± 0.22	95.37 ± 0.07
TPENAS-Net [36]	NAS	1.7	None	90.38
SLGE-Net [37]		5.1	96.56 ± 0.13 (TR-80%))	
DARTS-Net [38]		3.8	95.32 (TR-60%)	
MBLA-Net [39]	Attention CNN	>25.6	92.32 ± 0.15	94.66 ± 0.11
LRSCM-Net [40]		7.6	92.18	94.74
SCCNN [41]	Self-designed CNN	0.5	92.02 ± 0.50	94.39 ± 0.16
LGRIN [42]		4.6	91.95 ± 0.15	94.43 ± 0.16
MF2C-Net [43]		33.2	92.07 ± 0.22	93.85 ± 0.27
LH-Net [44]		>46.8	89.89 ± 0.15	92.53 ± 0.13
LT-Net [45]	Single Transformer	8.2	92.21 ± 0.11	None
ViT-Base [46]		88.6	92.60 ± 0.10	None
Swin-ViT-Tiny [47]		28.3	93.02 ± 0.12	94.51 ± 0.05
SCViT [48]		>88.6	92.72±0.04	94.66±0.10
ACGL-Net [49]	Dual-CNNs	33.6	None	None
GLDBS-Net [50]		>23.4	92.24 ± 0.21	94.46 ± 0.15
AC-Net [51]		>276.6	91.09 ± 0.13	92.42 ± 0.16
T-CNN [52]		15.9	90.25 ± 0.14	93.05 ± 0.12
TRS-Net [53]	CNN-and-ViT	46.3	93.06 ± 0.11	95.56 ± 0.20

P2FEViT [54]		>94.9	94.97 ± 0.13	95.74 ± 0.19
CT-Net [55]		>107.8	92.24 ± 0.21	94.46 ± 0.15
L2RCF-Net [56]		46.7	94.58 ± 0.16	95.60 ± 0.12
HHTL-Net [57]	Dual-ViTs	>177.2	92.07 ± 0.44	94.21 ± 0.09
CNN–HMM Ensemble [58]	Four-CNN ensemble	19	93.43 ± 0.25	95.51 ± 0.21
ViT Teacher (This work)	Single N-ViT-S	28.4	94.84 ± 0.12	96.30 ±0.13
VCL-Net (This work)	Single Efficient-B0	5.3	**94.55** ± 0.01	**96.23** ± 0.09

As depicted in Table 3, our ViT teacher model achieves superior OA values on the NWPU45 dataset, except that P2FEViT [54] has a 0.1% improved OA value at the 10% TR. It is reasonable to note that P2FEViT has at least 3.4 times the parameters of our teacher model. VCL-Net exhibits approximately a 0.3% OA margin with the teacher at the 10% TR, but this OA gap narrows to 0.1% at the 20% TR. These results align with the OA improvements on AID30 when the training samples become sufficient.

In comparison to 33 other advanced approaches, we observed that only RE-EfficientNet presents competitive OA values with VCL-Net at both the 10% and 20% TRs. Given that RE-EfficientNet and P2FEViT clearly possess more parameters, the results demonstrate that VCL-Net maintains a more balanced approach in terms of efficiency and effectiveness.

When juxtaposed with other KD techniques, it is evident that VCL-Net exhibits substantial enhancements in OA, with a range of 2.0% to 22.0% at the 10% TR. The benefits of VCL-Net remain steady at the 20% TR. These findings further underscore the superior efficacy of VCL-Net over other KD strategies when evaluated using the NWPU45 benchmark.

Our evaluation of various multi-model methodologies revealed that most did not reach satisfactory accuracy levels. Furthermore, when tested against the complex NWPU45 dataset, both ESD-MBENet and TRS-Net failed to demonstrate competitive OA values. This comparative analysis further accentuates the superior efficiency and robustness of VCL-Net as a classifier over its counterparts.

4.1.3. OA results for AFGR50

Table 4. OA (%) Comparison among Different Methods Using the AFGR50 Dataset

Methods	Algorithm Uniqueness	Params (M)	AFGR50 (%)		
			TR-10%	TR-20%	TR-30%
P2FEViT [54]	CNN and ViT	>94.9	89.24 ± 0.10	95.22 ± 0.13	97.27 ± 0.15
ViT Teacher (This work)	Single N-ViT-S	28.4	91.10 ± 0.51	96.50 ± 0.22	97.25 ± 0.07
VCL-Net (This work)	Single Efficient-B0	5.3	**90.68** ± 0.18	**96.11** ± 0.09	**97.16** ± 0.13

In RSIs, the task of fine-grained object recognition is widely performed. As a result, we evaluated the generalization capability of VCL-Net using the AFGR50 dataset. The OA outcomes of three methods applied to the AFGR50 dataset are presented in Table 4. Currently, the utilization of this dataset as a performance benchmark in public studies is minimal, given its recent release in March 2023.

Table 4 reveals that our teacher-student model surpasses P2FEViT in terms of OA on the 10% and 20% TRs. However, when evaluated using a TR of 30%, P2FEViT manages to reduce the OA gap. Given that the AFGR50 dataset comprises 12,500 samples, it is logical to assume that an increase in training samples can boost model performance. Hence, these findings suggest that VCL-Net exhibits greater effectiveness and resilience when dealing with subsets comprising varying volumes.

Furthermore, the OA values derived from the NWPU45 and AFGR50 datasets further substantiate the superior efficiency and robustness of VCL-Net across diverse RSI datasets, despite P2FEViT possessing 17.9 times more parameters than VCL-Net.

4.1.4. Overview of OA Comparisons

The consistent enhancement in the OA values of VCL-Net across various RSI datasets underscores the efficacy of our cross-modal method for knowledge transfer. When evaluated on the basis of accuracy, VCL-Net outperforms other KD-based approaches significantly. In comparison to other multi-model methods, VCL-Net not only offers superior accuracy but also maintains a lightweight architecture. Furthermore, VCL-Net exhibits commendable robustness across diverse RSI datasets.

4.2 Confusion Matrixes

Figures 8, 9, and 10 display confusion matrices that examine the distribution of errors across different categories. An accuracy of 100% is denoted by 1.0. Categories marked in red are particularly challenging, while those in blue are susceptible to misclassification. For clarity, the categories with an OA exceeding 98% are excluded.

4.2.1. Confusion Results for AID30.

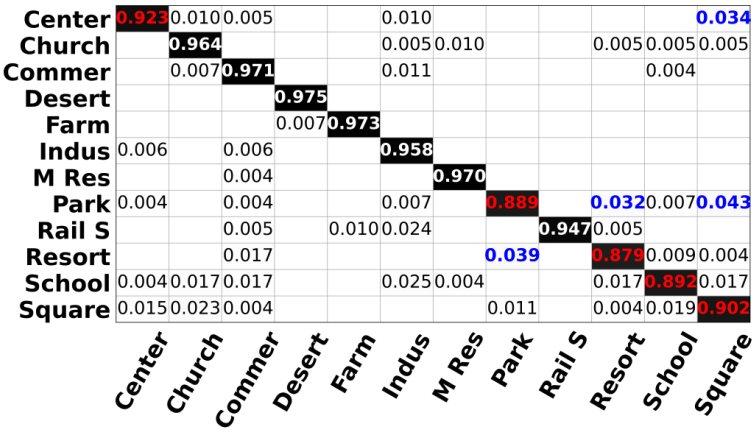

Figure 8. Confusion Matrix for AID30 at the 20% TR

Figure 8 demonstrates that within the AID30 dataset, five categories, namely center, park, resort, school, and square, pose significant challenges in terms of differentiation. These categories display OAs falling below 93%. Delving deeper, three categories—park, resort, and square—exhibit a misclassification ratio exceeding 0.3% with another class. Notably, the results reveal a high degree of similarity between the square category and both the center and park categories, leading to confusion.

The AID30 dataset exhibits an uneven distribution of sample numbers across categories, with an average count of 333. Notably, the categories that pose greater challenges within AID30 typically have fewer samples than this average. For example, the sample counts for the center, resort, and school categories are 260, 290, and 300, respectively. Consequently, this imbalance in AID30 could result in a significant bias in OA due to random divisions of training subsets.

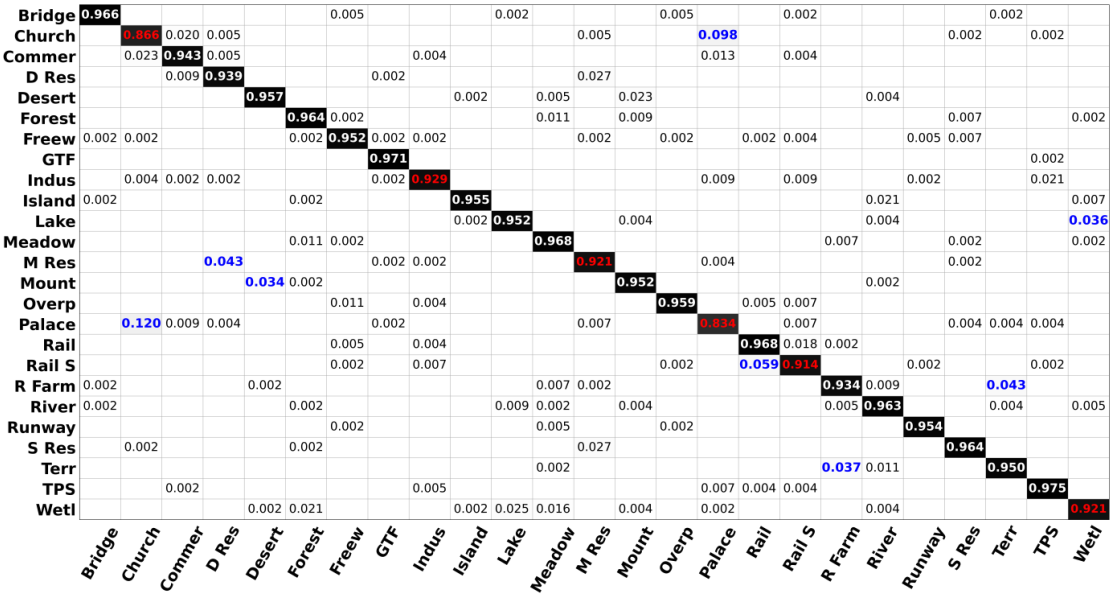

Figure 9. Confusion Matrix for NWPU45 at the 20% TR

4.2.2. Confusion Results for NWPU45

Figure 9 reveals that within the NWPU45 dataset, six categories display OAs of less than 93.0%. Among these, the categories of church and palace prove especially difficult to differentiate, with their OAs

dipping below 90%. Moreover, these two categories, church and palace, exhibit a high degree of confusion with each other, resulting in elevated misclassification ratios of 12% and 10%, respectively. Conversely, the remaining categories demonstrate a more distributed confusion pattern with relatively smaller ratio values.

4.2.3. Confusion Results for AFGR50

Figure 10 reveals that within the AFGR50 dataset, nine categories display OAs of less than 93.0%. Among these, the categories of A24, A25, A39 and A46 prove especially difficult to differentiate, with their OAs dipping below 90%. Moreover, these two categories, A24 and A44, exhibit elevated misclassification ratios of up 6% to 10%, respectively. These observations highlight that fine-grained RSIs have very different features compared building scenes.

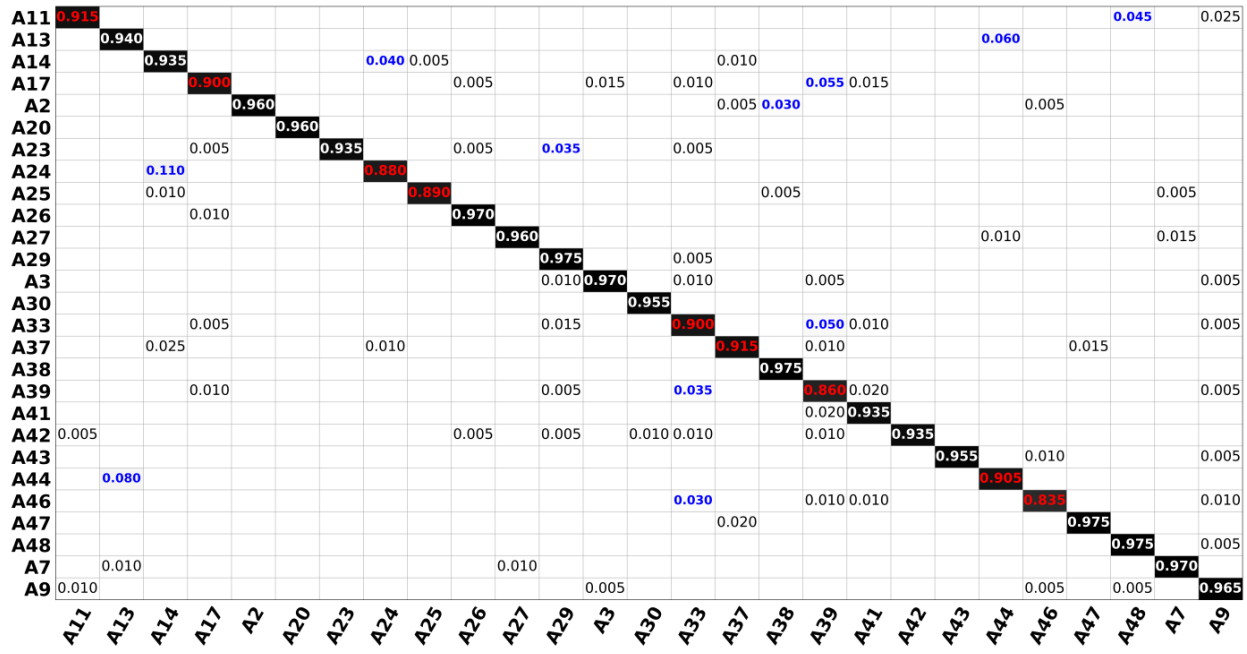

Figure 10. Confusion Matrix for AFGR50 at the 20% TR

Upon juxtaposing the outcomes of various methodologies, it is observed that the categories causing the most confusion essentially remain unchanged. Additionally, both TRS-Net and ESD-MBENet, despite demonstrating marginally superior OAs on AID30, consistently exhibit lower OAs across categories on NWPU45. Importantly, each category in NWPU45 possesses an identical sample size of 700, rendering NWPU45 a more appropriate benchmark for performance assessment compared to AID30. These findings suggest that certain multi-model methods excel only in specific training subsets. In contrast, the perplexing outcomes on both AID30 and NWPU45 further attest to the efficacy and robustness of VCL-Net.

4.4. Visualization and Analysis

To shed light on the model's activation mappings and validate the effectiveness of its features, we employ two separate techniques. The first approach involves the use of Gradient-Weighted Class Activation Mapping (Grad-CAM) [64], which provides visual explanations for the model's predictions. Following this, we leverage t-Distributed Stochastic Neighbor Embedding [65], commonly known as t-SNE, to examine the strength of the model's features.

4.4.1. Grad-CAM results

Figure 11 demonstrates the CAM results, which include seven samples from the ambiguous categories of AID30 and NWPU45. Specifically, the samples of the center, park, resort, school, and square are part of AID30, while the remaining two are from NWPU45. In the figure, the original images are displayed in the first row, and their corresponding CAMs are presented in rows two to four. Among all the CAMs, the first row is associated with EfficientNet-B0, which employs the training strategy mentioned

in reference [23]. The second and final rows are attributed to our N-ViT-S teacher and VCL-Net, respectively.

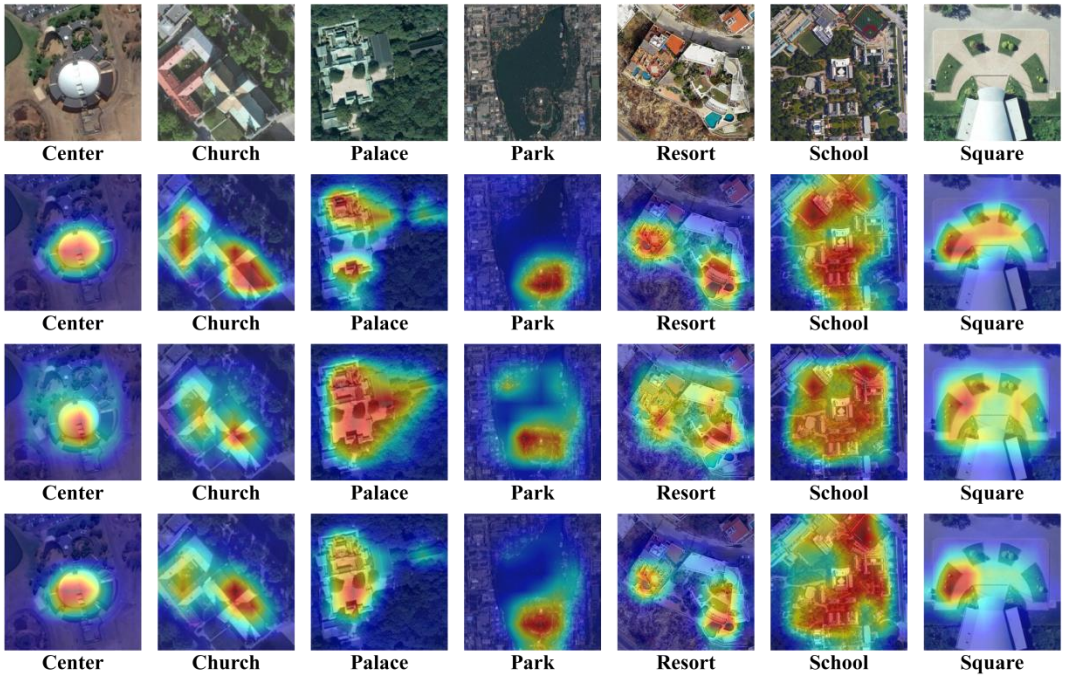

Figure 11. Analysis of Representative RSI Samples Using Grad-CAM

Rows 2, 3, and 4 highlight the brighter areas in the CAMs, signifying the main activation regions in the original images. These highlighted zones are intimately associated with earthbound objects that act as key visual elements, contributing to the semantic labels of the categories. For instance, the center's circular roof and the school's functional structure serve as such features.

As shown in rows 2 and 3, the varying bright regions indicate that N-ViT-S possesses distinct feature representations compared to the pure CNN. Specifically, the ViT teacher's activation focuses more on the long-range dependencies of features. This is particularly evident in the CAMs of the palace, resort, and school categories, which include multiple building objects in the scenes.

In contrast, the CAMs of VCL-Net exhibit a balance between the activation patterns of the ViT and pure CNN. Specifically, the activations of VCL-Net demonstrate more feature dependencies compared to pure CNN. Conversely, the bright regions of VCL-Net are more concentrated compared to the teacher. These CAMs validate that VCL-Net has effectively identified crucial local features in RSIs and has learned more feature dependencies through the process of distillation.

4.4.2. t-SNE Results

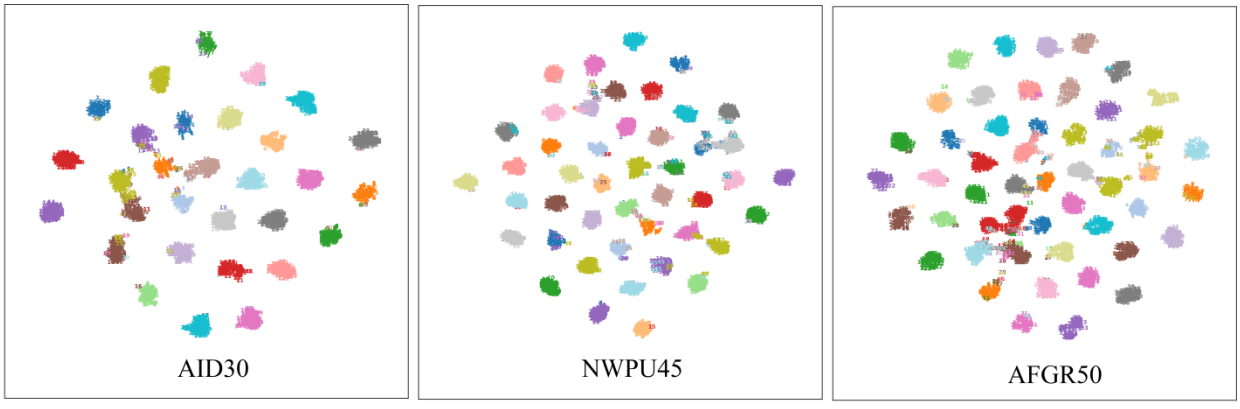

Figure 12. t-SNE Visualization on the AID30, NWPU45, and AFGR50 Datasets

Figure 12 presents the t-SNE results for the AID30, NWPU45, and AFGR50 datasets, using a two-dimensional projection to represent the high-dimensional data distribution across different categories.

This approach allows for the distinction of spatial distances among samples and provides a straightforward method for assessing a model's feature effectiveness by distinguishing between categories.

Figure 12 demonstrates that all categories within the three datasets are clearly differentiated, with the exception of several pairs of categories that show slight overlap. This observation aligns with those identified in the confusion matrix diagrams. For instance, in the AID30 dataset, the overlapping pairs include the categories of center with church and resort with school. Similarly, in the NWPU45 dataset, the overlapping pairs are railway with railway station, lake with wetland, and church with palace. Despite these overlaps, the extent of separation among categories is sufficient for differentiating various classes.

In summary, these t-SNE results indicate that the deep features derived by VCL-Net are effective.

4.5. Computational Efficiency Analysis

In this section, we evaluate the inference speeds of various models using a dataset of 25,200 RSI samples. The testing resolution follows Algorithm 1.

As indicated in Table 5, VCL-Net, which employs an Efficient-B0 backbone, exhibits the fastest inference speed among all methods. By comparison, VCL-Net requires only about 28.8% of the inferring times and 16.7% of the parameters of the ViT teacher. Similarly, VCL-Net is also a more lightweight classifier with superior accuracy when compared to RE-EfficientNet.

Table 5. Comparison of Inferring Speeds for Various Models

Model	Params(M)	FLOPs(G)	Inferring time(second)
ResNet-50	11.7	1.8	63.5 ± 0.17
RE-EfficientNet [32]	12.0	1.8	59.9 ± 0.24
N-ViT-S Teacher	31.7	5.8	105 ± 0.03
VCL-Net	5.3	0.4	30.2 ± 0.20

4.6. Ablation Experiments

In this segment of the study, we conduct a range of ablation tests to confirm the effectiveness and necessity of the methods we propose. During these tests, all hyperparameters used in training comply with Algorithm 1, barring the parameter subjected to ablation.

In the initial set of experiments, we implement a training procedure devoid of KD. To begin with, we examine the effect of DA methods on accuracy when the training approach is simply replicated in the domain of natural images, termed "duplication". In other words, we assign a value of 1.0 to the probabilities within all GAOs. Following this, we assess the influence of our proposed DA strategy on accuracy. As depicted in Table 6, the findings suggest that both DA combinations yield noticeably lower OA values in the absence of the ViT teacher.

Table 6. Evaluating the OA (%) of VCL-Net with Different DA Approaches.

Model	KD	DA Strategy		AID30	NWPU45
		Duplication	Ours	TR20%	TR10%
VCL-Net	×	✓	×	95.34 ± 0.12	91.80 ± 0.29
	×	×	✓	94.60 ± 0.04	90.75 ± 0.10
	✓	×	✓	**97.10** ± 0.09	**94.55** ± 0.01

Additionally, we assess the impact of our DA configurations within the KD procedure. Specifically, we first substitute the settings in Table 1 with the duplication, followed by conducting the KD process as outlined in Algorithm 1. As indicated in Table 7, VCL-Net fails to attain satisfactory OA outcomes despite undergoing training for six times more epochs, ranging from 600 to 2400.

Hence, the results of the ablation study underscore the critical role of our strategy. It enables VCL-Net to attain exceptional accuracy during the cross-modal knowledge transfer process while reducing the time costs by at least a factor of six.

Table 7. Assessing the OA (%) of VCL-Net under Different Training Methods.

Model	DA Strategy	Training Epochs		AID30	NWPU45
		600	2400	TR10%	TR20%
VCL-Net	Duplication	×	✓	96.61 ± 0.06	94.40 ± 0.06
	Ours	✓	×	**97.10** ± 0.09	**94.55** ± 0.01

5. Conclusions

In this study, we introduce an innovative cross-modal knowledge transfer method designed to create efficient and precise classifiers for RSI classification. This method utilizes a ViT-teaching-CNN pipeline and adeptly mitigates the domain differences between RSIs and natural images. It incorporates novel yet simple principles to better adapt to the intrinsic properties of RSIs, thereby markedly improving efficiency and resilience during the distillation stage.

The advantages of our approach stem mainly from two factors. Firstly, we contend that the discrepancies in data distribution within RSI datasets have significantly hindered the efficacy and efficiency of the logit-based KD process. Secondly, we devise a more potent algorithm that incorporates a blend of manageable DA and regularizations to tackle the intrinsic attributes of RSIs. Specifically, the discrepancies come from cluttered backgrounds and substantial resemblances across categories.

Our distillation model, referred to as VCL-Net, was evaluated on three standard RSI datasets. The findings revealed that VCL-Net demonstrated superior precision and resilience in comparison to 33 other cutting-edge techniques published in the last three years. Specifically, VCL-Net achieved a maximum accuracy improvement of 22% across various RSI datasets when contrasted with other KD techniques documented in the literature. Additionally, the Grad-CAM outcomes suggest that VCL-Net has acquired long-range dependencies from the ViT instructor through distillation. Furthermore, the ablation studies confirm that our approach has significantly cut down the time costs of knowledge transfer by at least 75% compared to simply replicating strategies in the natural image domain. Hence, we illustrate that cross-modal knowledge transfer can be more effective and efficient when domain differences are appropriately managed.

Our research, while promising, is in its nascent stages and acknowledges certain limitations that necessitate future enhancements. Initially, we have not conducted an exhaustive grid search across all hyperparameters, indicating potential avenues for refining our methodology. Additionally, we have yet to fully leverage the unique attributes of RSIs in crafting bespoke and more effective distillation techniques. We aspire to address these issues in our forthcoming endeavors.

Acknowledgement

Hunan National Science Fund provided research support for this study (grant number 2024JJ7314).

References

[1] Mangana B. Rampheri, Timothy Dube, Farai Dondofema and Tatenda Dalu, "Progress in the remote sensing of groundwater-dependent ecosystems in semi-arid environments", *Physics and Chemistry of the Earth, Parts A/B/C*, Print ISSN: 14747065, Vol. 130, pp. 103359, June 2023, Published by Elsevier Ltd, DOI: 10.1016/j.pce.2023.103359, Available: https://linkinghub.elsevier.com/retrieve/pii/S1474706523000037.

[2] Siwei Zhang, Jun Ma, Xiaohu Zhang and Cui Guo, "Atmospheric remote sensing for anthropogenic methane emissions: applications and research opportunities", *Science of the Total Environment*, Print ISSN: 00489697, Vol. 893, pp. 164701, October 2023, Published by Elsevier Ltd, DOI: 10.1016/j.scitotenv.2023.164701, Available: https://linkinghub.elsevier.com/retrieve/pii/S0048969723033247.

[3] Suraj Sawant, Rahul Dev Garg, Vishal Meshram and Shrayank Mistry, "Sen-2 lulc: land use land cover dataset for deep learning approaches", *Data in Brief*, Print ISSN: 23523409, Vol. 51, pp. 109724, December 2023, DOI: 10.1016/j.dib.2023.109724, Available: https://linkinghub.elsevier.com/retrieve/pii/S2352340923007953.

[4] Hui Xiang, Chunmei Zhou, Cuidong and Huaxiang Song, "High-quality agricultural development in the central china: empirical analysis based on the dongting lake area", *Geomatica*, Print ISSN: 1195-1036, Vol. 76, No. 1, pp. 100010, 9 July 2024, Published by Elsevier Ltd., DOI: 10.1016/j.geomat.2024.100010, Available: https://linkinghub.elsevier.com/retrieve/pii/S1195103624000107.

[5] Huaxiang Song, "FST-efficientnetv2: exceptional image classification for remote sensing", *Computer Systems Science and Engineering*, Print ISSN: 0267-6192, Vol. 46, No. 3, pp. 3959–3978, 2023, Published by Tech Science Press, DOI: 10.32604/csse.2023.038429, Available: https://www.techscience.com/csse/v46n3/52217.

[6] Huaxiang Song, "A consistent mistake in remote sensing images' classification literature", *Intelligent Automation & Soft Computing*, Print ISSN: 1079-8587, Vol. 37, No. 2, pp. 1381–1398, 2023, Published by Tech Science Press, DOI: 10.32604/iasc.2023.039315, Available: https://www.techscience.com/iasc/v37n2/53269.

[7] Huaxiang Song, Yuxuan Yuan, Zhiwei Ouyang, Yu Yang and Hui Xiang, "Quantitative regularization in robust vision transformer for remote sensing image classification", *The Photogrammetric Record*, Print ISSN: 0031-868X,

1477-9730, Vol. 39, No. 186, pp. 340–372, June 2024, Published by John Wiley & Sons Ltd., DOI: 10.1111/phor.12489, Available: https://onlinelibrary.wiley.com/doi/10.1111/phor.12489.

[8] Cristian Buciluǎ, Rich Caruana and Alexandru Niculescu-Mizil, "Model Compression", *in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining 2006 (KDD '06)*, 20-23 August 2006, Philadelphia, Pennsylvania, USA, ISBN: 978-1-59593-339-3, pp. 535-541, Published by Association for Computing Machinery, DOI: 10.1145/1150402.1150464, Available: https://dl.acm.org/doi/10.1145/1150402.1150464.

[9] Huaxiang Song, Chai Wei and Zhou Yong, "Efficient knowledge distillation for remote sensing image classification: a CNN-based approach", *International Journal of Web Information Systems*, Print ISSN: 1744-0084, Vol. 20, No. 2, pp. 129–158, 1 January 2024, Published by Emerald Publishing Limited, DOI: 10.1108/IJWIS-10-2023-0192, Available: https://www.emerald.com/insight/content/doi/10.1108/IJWIS-10-2023-0192/full/html.

[10] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi and Andrew Gordon Wilson, "Does Knowledge Distillation Really Work?", In *Proceedings of the Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*, 6-14 December 2021, Red Hook, NY, USA, ISBN: 978-1-71384-539-3, Published by Curran Associates Inc., Available: https://dl.acm.org/doi/abs/10.5555/3540261.3540790.

[11] Lucas Beyer, Xiaohua Zhai, Amelie Royer, Larisa Markeeva, Rohan Anil *et al.*, "Knowledge Distillation: A Good Teacher Is Patient and Consistent", In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18-24 June 2022, New Orleans, LA, USA, ISBN: 978-1-66546-946-3, pp. 10915–10924, Published by IEEE, DOI: 10.1109/CVPR52688.2022.01065, Available: https://ieeexplore.ieee.org/document/9879513/.

[12] Wonpyo Park, Dongju Kim, Yan Lu and Minsu Cho, "Relational Knowledge Distillation", In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15-20 June 2019, Long Beach, CA, USA, ISBN: 978-1-72813-293-8, pp. 3962–3971, Published by IEEE, DOI: 10.1109/CVPR.2019.00409, Available: https://ieeexplore.ieee.org/document/8954416/.

[13] Yixia Chen, Mingwei Lin, Zhu He, Kemal Polat, Adi Alhudhaif *et al.*, "Consistency- and dependence-guided knowledge distillation for object detection in remote sensing images", *Expert Systems with Applications*, Print ISSN: 09574174, Vol. 229, pp. 120519, November 2023, Published by Elsevier Ltd., DOI: 10.1016/j.eswa.2023.120519, Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417423010217.

[14] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan *et al.*, "Focal and Global Knowledge Distillation for Detectors", In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18-24 June 2022, New Orleans, LA, USA, ISBN: 978-1-6654-6946-3, pp. 4633–4642, Published by IEEE, DOI: 10.1109/CVPR52688.2022.00460, Available: https://ieeexplore.ieee.org/document/9879869.

[15] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu *et al.*, "Cross-Image Relational Knowledge Distillation for Semantic Segmentation", In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18-24 June 2022, New Orleans, LA, USA, ISBN: 978-1-6654-6946-3, pp. 12309–12318, Published by IEEE, DOI: 10.1109/CVPR52688.2022.01200, Available: https://ieeexplore.ieee.org/document/9879845.

[16] Jinsu Yoo, Taehoon Kim, Sihaeng Lee, Seung Hwan Kim, Honglak Lee *et al.*, "Enriched CNN-Transformer Feature Aggregation Networks for Super-Resolution", In *Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 02-07 January 2023, Waikoloa, HI, USA, ISBN: 978-1-6654-9346-8, pp. 4945–4954, Published by IEEE, DOI: 10.1109/WACV56688.2023.00493, Available: https://ieeexplore.ieee.org/document/10030797.

[17] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles *et al.*, "Training Data-Efficient Image Transformers & Distillation through Attention", In *Proceedings of the 38th International Conference on Machine Learning*, 8-24 July 2021, Virtual, ISSN: 2640-3498, pp. 10347-10357, Published by PMLR, DOI: 10.48550/arXiv.2012.12877, Available: https://proceedings.mlr.press/v139/touvron21a.html.

[18] Sucheng Ren, Zhengqi Gao, Tianyu Hua, Zihui Xue, Yonglong Tian *et al.*, "Co-Advise: Cross Inductive Bias Distillation", In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18-24 June 2022, New Orleans, LA, USA, ISBN: 978-1-6654-6946-3, pp. 16752–16761, Published by IEEE, DOI: 10.1109/CVPR52688.2022.01627, Available: https://ieeexplore.ieee.org/document/9879858.

[19] Yu Wang, Zhenfeng Shao, Tao Lu, Lifeng Liu, Xiao Huang *et al.*, "A lightweight distillation cnn-transformer architecture for remote sensing image super-resolution", International Journal of Digital Earth, Print ISSN: 1753-8947, 1753-8955, Vol. 16, No. 1, pp. 3560–3579, October 2023, Published by Taylor & Francis Group, DOI: 10.1080/17538947.2023.2252393, Available: https://www.tandfonline.com/doi/full/10.1080/17538947.2023.2252393.

[20] Yufan Liu, Jiajiong Cao, Bing Li, Weiming Hu, Jingting Ding *et al.*, "Cross-Architecture Knowledge Distillation", *In Proceedings of the Proceedings of the Asian Conference on Computer Vision (ACCV)*, 4–8 December 2022, Macao, China, ISBN: 978-3-031-26347-7, pp. 3396–3411, Published by Springer, DOI: 10.1007/978-3-031-26348-4_11, Available: https://link.springer.com/chapter/10.1007/978-3-031-26348-4_11.

[21] Borui Zhao, Renjie Song and Jiajun Liang, "Cumulative Spatial Knowledge Distillation for Vision Transformers", *In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 01-06 October 2023, Paris, France, ISBN: 979-8-3503-0718-4, pp. 6146–6155, Published by IEEE, DOI: 10.1109/ICCV51070.2023.00565, Available: https://ieeexplore.ieee.org/document/10377169.

[22] Alex Andonian, Shixing Chen and Raffay Hamid, "Robust Cross-Modal Representation Learning with Progressive Self-Distillation", *In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18-24 June 2022, New Orleans, LA, USA, ISBN: 978-1-6654-6946-3, pp. 16409–16420, Published by IEEE, DOI: 10.1109/CVPR52688.2022.01594, Available: https://ieeexplore.ieee.org/document/9879136.

[23] Huaxiang Song and Yong Zhou, "Simple is best: a single-cnn method for classifying remote sensing images", *Networks and Heterogeneous Media*, Print ISSN: 1556-1801, Vol. 18, No. 4, 2023, pp. 1600–1629, Published by AIMS Press, DOI: 10.3934/nhm.2023070, Available: http://www.aimspress.com/article/doi/10.3934/nhm.2023070.

[24] Kejie Xu, Peifang Deng and Hong Huang, "Vision transformer: an excellent teacher for guiding small networks in remote sensing image scene classification", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Online: 1558-0644, Vol. 60, pp. 1–15, 2022, Published by IEEE, DOI: 10.1109/TGRS.2022.3152566, Available: https://ieeexplore.ieee.org/document/9716120.

[25] Xuying Wang, Jiawei Zhu, Zhengliang Yan, Zhaoyang Zhang, Yunsheng Zhang *et al.*, "LaST: label-free self-distillation contrastive learning with transformer architecture for remote sensing image scene classification", *IEEE Geoscience and Remote Sensing Letters*, Print ISSN: 1545-598X, Online ISSN: 1558-0571, Vol. 19, pp. 1–5, 2022, Published by IEEE, DOI: 10.1109/LGRS.2022.3185088, Available: https://ieeexplore.ieee.org/document/9802117.

[26] Guanzhou Chen, Xiaodong Zhang, Xiaoliang Tan, Yufeng Cheng, Fan Dai *et al.*, "Training small networks for scene classification of remote sensing images via knowledge distillation", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 10, No. 5, pp. 719, May 2018, Published by MDPI, DOI: 10.3390/rs10050719, Available: http://www.mdpi.com/2072-4292/10/5/719.

[27] Shiyi Xing, Jinsheng Xing, Jianguo Ju, Qingshan Hou and Xiurui Ding, "Collaborative consistent knowledge distillation framework for remote sensing image scene classification network", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 14, No. 20, pp. 5186, October 2022, Published by MDPI, DOI: 10.3390/rs14205186, Available: https://www.mdpi.com/2072-4292/14/20/5186.

[28] Yutao Hu, Xin Huang, Xiaoyan Luo, Jungong Han, Xianbin Cao *et al.*, "Variational self-distillation for remote sensing scene classification", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Online ISSN: 1558-0644, Vol. 60, pp. 1–13, 2022, Published by IEEE, DOI: 10.1109/TGRS.2022.3194549, Available: https://ieeexplore.ieee.org/document/9844008/.

[29] Daxiang Li, Yixuan Nan and Ying Liu, "Remote sensing image scene classification model based on dual knowledge distillation", *IEEE Geoscience and Remote Sensing Letters*, Print ISSN: 1545-598X, Online ISSN: 1558-0571, Vol. 19, pp. 1–5, 2022, Published by IEEE, DOI: 10.1109/LGRS.2022.3208904, Available: https://ieeexplore.ieee.org/document/9900370.

[30] Qi Zhao, Yujing Ma, Shuchang Lyu and Lijiang Chen, "Embedded self-distillation in compact multibranch ensemble network for remote sensing scene classification", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Online ISSN: 1558-0644, Vol. 60, pp. 1–15, 2022, Published by IEEE, DOI: 10.1109/TGRS.2021.3126770, Available: https://ieeexplore.ieee.org/document/9606819.

[31] Huaxiang Song, "MBC-net: long-range enhanced feature fusion for classifying remote sensing images", *International Journal of Intelligent Computing and Cybernetics*, Print ISSN: 1756-378X, Vol. 17, No. 1, pp. 181–209, 1 January 2024, Published by Emerald Publishing Limited, DOI: 10.1108/IJICC-07-2023-0198, Available: https://www.emerald.com/insight/content/doi/10.1108/IJICC-07-2023-0198/full/html.

[32] Huaxiang Song, Yafang Li, Xiaowen Li, Yuxuan Zhang, Yangyan Zhu *et al.*, "ERKT-net: implementing efficient and robust knowledge distillation for remote sensing image classification", *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, Print ISSN: 2410-0218, Vol. 11, No. 3, July 2024, Published by EAI, DOI: 10.4108/eetinis.v11i3.4748, Available: https://publications.eai.eu/index.php/inis/article/view/4748.

[33] Huaxiang Song, Yuxuan Yuan, Zhiwei Ouyang, Yu Yang and Hui Xiang, "Efficient knowledge distillation for hybrid models: a vision transformer-convolutional neural network to convolutional neural network approach for classifying remote sensing images", *IET Cyber-Systems and Robotics*, Print ISSN: 2631-6315, Online ISSN: 2631-6315, Vol. 6, No. 3, pp. e12120, September 2024, Published by John Wiley & Sons Ltd., DOI: 10.1049/csy2.12120, Available: https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/csy2.12120.

[34] Mostaan Nabi, Luca Maggiolo, Gabriele Moser and Sebastiano B. Serpico, "A CNN-Transformer Knowledge Distillation for Remote Sensing Scene Classification", In *Proceedings of the IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 17-22 July 2022, Kuala Lumpur, Malaysia, ISBN: 978-1-66542-792-0, pp. 663–666, Published by IEEE, DOI: 10.1109/IGARSS46834.2022.9884099, Available: https://ieeexplore.ieee.org/document/9884099/.

[35] Yibo Zhao, Jianjun Liu, Jinlong Yang and Zebin Wu, "EMSCNet: efficient multisample contrastive network for remote sensing image scene classification", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Online ISSN: 1558-0644, Vol. 61, pp. 1–14, 2023, Published by IEEE, DOI: 10.1109/TGRS.2023.3262840, Available: https://ieeexplore.ieee.org/document/10086539/.

[36] Lei Ao, Kaiyuan Feng, Kai Sheng, Hongyu Zhao, Xin He *et al.*, "TPENAS: a two-phase evolutionary neural architecture search for remote sensing image classification", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 15, No. 8,

pp. 2212, April 2023, Published by MDPI, DOI: 10.3390/rs15082212, Available: https://www.mdpi.com/2072-4292/15/8/2212.

[37] Clifford Broni-Bediako, Yuki Murata, Luiz H. B. Mormille and Masayasu Atsumi, "Searching for cnn architectures for remote sensing scene classification", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Online ISSN: 1558-0644, Vol. 60, pp. 1–13, 2022, Published by IEEE, DOI: 10.1109/TGRS.2021.3097938, Available: https://ieeexplore.ieee.org/document/9497513/.

[38] Junge Shen, Bin Cao, Chi Zhang, Ruxin Wang and Qi Wang, "Remote sensing scene classification based on attention-enabled progressively searching", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Online ISSN: 1558-0644, Vol. 60, pp. 1–13, 2022, Published by IEEE, DOI: 10.1109/TGRS.2022.3186588, Available: https://ieeexplore.ieee.org/document/9807377/.

[39] Sibao Chen, Qingsong Wei, Wenzhong Wang, Jin Tang, Bin Luo *et al.*, "Remote sensing scene classification via multi-branch local attention network", *IEEE Transactions on Image Processing*, Print ISSN: 1057-7149, Online ISSN: 1941-0042, Vol. 31, pp. 99–109, 2022, Published by IEEE, DOI: 10.1109/TIP.2021.3127851, Available: https://ieeexplore.ieee.org/document/9619948/.

[40] Xinyu Wang, Haixia Xu, Liming Yuan and Xianbin Wen, "A lightweight and stochastic depth residual attention network for remote sensing scene classification", *IET Image Processing*, Print ISSN: 1751-9659, Online ISSN: 1751-9667, Vol. 17, No. 11, pp. 3106–3126, September 2023, Published by John Wiley & Sons Ltd., DOI: 10.1049/ipr2.12836, Available: https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.12836.

[41] Cuiping Shi, Xinlei Zhang, Jingwei Sun and Liguo Wang, "Remote sensing scene image classification based on self-compensating convolution neural network", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 14, No. 3, pp. 545, January 2022, Published by MDPI, DOI: 10.3390/rs14030545, Available: https://www.mdpi.com/2072-4292/14/3/545.

[42] Chengjun Xu, Guobin Zhu and Jingqian Shu, "A lightweight and robust Lie group-convolutional neural networks joint representation for remote sensing scene classification", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Online ISSN: 1558-0644, Vol. 60, pp. 1–15, 2022, Published by IEEE, DOI: 10.1109/TGRS.2020.3048024, Available: https://ieeexplore.ieee.org/document/9325064/.

[43] Lin Bai, Qingxin Liu, Cuiling Li, Zhen Ye, Meng Hui *et al.*, "Remote sensing image scene classification using multiscale feature fusion covariance network with octave convolution", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Online ISSN: 1558-0644, Vol. 60, pp. 1–14, 2022, Published by IEEE, DOI: 10.1109/TGRS.2022.3160492, Available: https://ieeexplore.ieee.org/document/9737532/.

[44] Wenhua Zhang, Licheng Jiao, Fang Liu, Jia Liu and Zhen Cui, "LHNet: laplacian convolutional block for remote sensing image scene classification", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Online ISSN: 1558-0644, Vol. 60, pp. 1–13, 2022, Published by IEEE, DOI: 10.1109/TGRS.2022.3192321, Available: https://ieeexplore.ieee.org/document/9832932/.

[45] Xinyan Huang, Fang Liu, Yuanhao Cui, Puhua Chen, Lingling Li *et al.*, "Faster and better: a lightweight transformer network for remote sensing scene classification", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 15, No. 14, pp. 3645, July 2023, Published by MDPI, DOI: 10.3390/rs15143645, Available: https://www.mdpi.com/2072-4292/15/14/3645.

[46] Yakoub Bazi, Laila Bashmal, Mohamad M. Al Rahhal, Reham Al Dayil and Naif Al Ajlan, "Vision transformers for remote sensing image classification", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 13, No. 3, pp. 516, February 2021, Published by MDPI, DOI: 10.3390/rs13030516, Available: https://www.mdpi.com/2072-4292/13/3/516.

[47] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia and Dacheng Tao, "An empirical study of remote sensing pretraining", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, 1558-0644, Vol. 61, pp. 1–20, 2023, Published by IEEE, DOI: 10.1109/TGRS.2022.3176603, Available: https://ieeexplore.ieee.org/document/9782149/.

[48] Pengyuan Lv, Wenjun Wu, Yanfei Zhong, Fang Du and Liangpei Zhang, "SCViT: a spatial-channel feature preserving vision transformer for remote sensing image scene classification", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Online ISSN: 1558-0644, Vol. 60, pp. 1–12, 2022, Published by IEEE, DOI: 10.1109/TGRS.2022.3157671, Available: https://ieeexplore.ieee.org/document/9729845/.

[49] Junge Shen, Tianwei Yu, Haopeng Yang, Ruxin Wang and Qi Wang, "An attention cascade global–local network for remote sensing scene classification", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 14, No. 9, pp. 2042, April 2022, Published by MDPI, DOI: 10.3390/rs14092042, Available: https://www.mdpi.com/2072-4292/14/9/2042.

[50] Kejie Xu, Hong Huang and Peifang Deng, "Remote sensing image scene classification based on global–local dual-branch structure model", *IEEE Geoscience and Remote Sensing Letters*, Print ISSN: 1545-598X, Online ISSN: 1558-0571, Vol. 19, pp. 1–5, 2022, Published by IEEE, DOI: 10.1109/LGRS.2021.3075712, Available: https://ieeexplore.ieee.org/document/9425547/.

[51] Xu Tang, Qiushuo Ma, Xiangrong Zhang, Fang Liu, Jingjing Ma *et al.*, "Attention consistent network for remote sensing scene classification", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Print ISSN: 1939-1404, Online ISSN: 2151-1535, Vol. 14, pp. 2030–2045, 2021, Published by IEEE, DOI: 10.1109/JSTARS.2021.3051569, Available: https://ieeexplore.ieee.org/document/9324913/.

[52] Weiquan Wang, Yushi Chen and Pedram Ghamisi, "Transferring cnn with adaptive learning for remote sensing scene classification", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Online ISSN: 1558-0644, Vol. 60, pp. 1–18, 2022, Published by IEEE, DOI: 10.1109/TGRS.2022.3190934, Available: https://ieeexplore.ieee.org/document/9829875/.

[53] Jianrong Zhang, Hongwei Zhao and Jiao Li, "TRS: transformers for remote sensing scene classification", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 13, No. 20, pp. 4143, October 2021, Published by MDPI, DOI: 10.3390/rs13204143, Available: https://www.mdpi.com/2072-4292/13/20/4143.

[54] Guanqun Wang, He Chen, Liang Chen, Yin Zhuang, Shanghang Zhang *et al.*, "P2FEViT: plug-and-play cnn feature embedded hybrid vision transformer for remote sensing image classification", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 15, No. 7, pp. 1773, March 2023, Published by MDPI, DOI: 10.3390/rs15071773, Available: https://www.mdpi.com/2072-4292/15/7/1773.

[55] Peifang Deng, Kejie Xu and Hong Huang, "When cnns meet vision transformer: a joint framework for remote sensing scene classification", *IEEE Geoscience and Remote Sensing Letters*, Print ISSN: 1545-598X, Online ISSN: 1558-0571, Vol. 19, pp. 1–5, 2022, Published by IEEE, DOI: 10.1109/LGRS.2021.3109061, Available: https://ieeexplore.ieee.org/document/9531646/.

[56] Maofan Zhao, Qingyan Meng, Linlin Zhang, Xinli Hu and Lorenzo Bruzzone, "Local and long-range collaborative learning for remote sensing scene classification", *IEEE Transactions on Geoscience and Remote Sensing*, Print ISSN: 0196-2892, Online ISSN: 1558-0644, Vol. 61, pp. 1–15, 2023, Published by IEEE, DOI: 10.1109/TGRS.2023.3265346, Available: https://ieeexplore.ieee.org/document/10093899/.

[57] Jingjing Ma, Mingteng Li, Xu Tang, Xiangrong Zhang, Fang Liu *et al.*, "Homo–heterogenous transformer learning framework for rs scene classification", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Print ISSN: 1939-1404, Online ISSN: 2151-1535, Vol. 15, pp. 2223–2239, 2022, Published by IEEE, DOI: 10.1109/JSTARS.2022.3155665, Available: https://ieeexplore.ieee.org/document/9726930/.

[58] Xiang Cheng and Hong Lei, "Remote sensing scene image classification based on mmscnn–hmm with stacking ensemble model", *Remote Sensing*, Print ISSN: 2072-4292, Vol. 14, No. 17, pp. 4423, September 2022, Published by MDPI, DOI: 10.3390/rs14174423, Available: https://www.mdpi.com/2072-4292/14/17/4423.

[59] Jiashi Li, Xin Xia, Wei Li, Huixia Li, Xing Wang *et al.*, "Next-ViT: Next Generation Vision Transformer for Efficient Deployment in Realistic Industrial Scenarios", *arXiv*, 2022, DOI: 10.48550/arXiv.2207.05501, Available: https://arxiv.org/abs/2207.05501.

[60] Adekanmi Adeyinka Adegun, Serestina Viriri and Jules-Raymond Tapamo, "Review of deep learning methods for remote sensing satellite images classification: experimental survey and comparative analysis", *Journal of Big Data*, Print ISSN: 2196-1115, Vol. 10, No. 1, pp. 93, June 2023, Published by Springer, DOI: 10.1186/s40537-023-00772-x, Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00772-x.

[61] Tao Huang, Shan You, Fei Wang, Chen Qian and Chang Xu, "Knowledge Distillation from a Stronger Teacher", In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2024, Red Hook, NY, USA, ISBN: 978-1-71387-108-8, pp. 33716-33727, DOI: 10.5555/3600270.3602713, Published by Curran Associates Inc., Available: https://dl.acm.org/doi/abs/10.5555/3600270.3602713.

[62] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo *et al.*, "CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features", In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 27 October-02 November 2019, Seoul, South Korea, ISBN: 978-1-72814-803-8, pp. 6022–6031, DOI: 10.1109/ICCV.2019.00612, Published by IEEE, Available: https://ieeexplore.ieee.org/document/9008296/.

[63] Ivica Dimitrovski, Ivan Kitanovski, Dragi Kocev and Nikola Simidjievski, "Current trends in deep learning for earth observation: an open-source benchmark arena for image classification", *ISPRS Journal of Photogrammetry and Remote Sensing*, Print ISSN: 09242716, Vol. 197, pp. 18–35, March 2023, Published by Elsevier Ltd., DOI: 10.1016/j.isprsjprs.2023.01.014, Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271623000205.

[64] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh *et al.*, "Grad-cam: visual explanations from deep networks via gradient-based localization", *International Journal of Computer Vision*, Print ISSN: 0920-5691, 1573-1405, Vol. 128, No. 2, pp. 336–359, February 2020, Published by Springer, DOI: 10.1007/s11263-019-01228-7, Available: http://link.springer.com/10.1007/s11263-019-01228-7.

[65] Laurens van der Maaten and Geoffrey Hinton, " Visualizing Data using t-SNE", *Journal of Machine Learning Research*, Print ISSN 1532-4435, Vol. 9, No. 86, pp. 2579–2605, 2008, Published by MIT Press, Available: http://jmlr.org/papers/v9/vandermaaten08a.html.