Journal of Applied Automation Technologies

Research Paper

Multi-Scale Compact Convolution and Attention-Guided Feature Augmentation: An Efficient Lightweight Network for Real-Time Satellite Image Classification

Alexandru Marin[1,*] and Florentin Botez[1]

[1] Faculty of Electronics, Telecommunications and Information Technology, Politehnica University of Timișoara, Timișoara, 300006, Romania

*Corresponding author: alexandru.m@etc.upt.ro

Abstract. A model with high recognition accuracy and minimal processing is required for real-time satellite picture classification. the goal of this research is to create a lightweight Convolutional Neural Network architecture for satellite image classification that satisfies the demands of quick inference speed and high classification accuracy. To learn land-cover structures at various sizes, a multi-scale compact convolution module is added after an effective convolutional backbone is constructed using factorized convolution. The discriminative area is strengthened and redundant background responses are further suppressed by the addition of an attention-guided feature augmentation unit. The experiment uses a collection of 2100 photos from the 21-class satellite scene, with an input size of 224 × 224 × 3. The suggested model has roughly 1.35 million parameters and 185 MFLOPs in size, with a macro-F1 score of 94.31% and a total accuracy of 94.76%. The model has a comparatively low inference cost and performs 1.43% to 5.24% better in classification accuracy than representative lightweight networks. The findings indicate that, in addition to decreasing network size, maintaining multi-scale spatial information and selective feature discrimination is required to enhance the effectiveness of lightweight satellite image classification. The suggested structure is appropriate for onboard remote sensing and real-time satellite picture classification at the edge.

Keywords: *Real-time satellite image classification; Lightweight convolutional neural network; multi-scale feature representation; Attention-guided feature enhancement*

Received on 05 October 2023, Accepted on 27 February 2024, Published on 09 March 2024

Introduction

Satellites have been produced and used for remote sensing observations of the Earth over the last few decades. As optical, multispectral, hyperspectral, and synthetic-aperture imaging systems continue to advance, satellite images are now used by the military for reconnaissance and can offer rich data for land-cover mapping, monitoring urban development, ecological assessments, crop growth analysis, disaster response, etc. In these applications, image classification is typically used initially to carry out high-level analysis and transform vast amounts of pixel-level or scene-level data into semantic categories that are easily accessible for decision-making [1]. To facilitate focused spatial planning and resource allocation, accurately categorize the urban, vegetation, water, bare-land, and agricultural areas [2]. A fast categorization can be used to identify the affected region and send rescue teams in the case of a disaster, such as an earthquake, fire, landslide, or flood [3]. To cut down on the expenses and time needed for extensive ground surveys, satellite image classification is used in agricultural monitoring for crop-type identification, yield estimation, and farmland management [4]. Other environmental issues including desertification, coastal erosion, and changes in wetland and forest cover can also be classified using prompts [5]. As the amount of satellite data continues to grow, so do the requirements for classification models; these days, real-time processing skills for satellite images are just as important as high accuracy [6]. As a result, issues with real-time satellite picture categorization have increasingly surfaced during the development

https://doi.org/10.64972/jaat.2024v2.252p8e:101-117

of intelligent remote sensing systems, particularly in scenarios requiring quick perception, on-board processing, and continuous monitoring [7].

The ability of satellite image classification methods to extract features has been improved using Deep Convolutional Neural Networks. CNN-based techniques are more effective than hand-crafted descriptors and shallow classifiers because they can immediately learn multi-level visual characteristics from data, including texture patterns, spatial structures, object borders, and high-level semantic associations [8]. On numerous publicly available satellite image datasets, AlexNet, VGG, ResNet, DenseNet, and their remote-sensing counterparts have all demonstrated strong classification performance [9]. Nevertheless, a far higher number of parameters, floating-point operations, memory accesses, and inference delay are typically associated with the increased accuracy [10]. When it comes to classification systems on satellites, unmanned aerial vehicles, edge terminals, and embedded devices—all of which have extremely limited processing capacity and energy—the issue is even more serious [11]. Simply decreasing the network's depth or width will significantly impair feature discrimination since satellite images typically have complicated backgrounds, significant intra-class differences, small targets, and visually comparable categories [12]. Although lightweight CNNs like compound-scaled architectures, channel shuffle models, and depthwise separable convolution networks have been suggested to cut down on computation, they are typically made for natural image datasets and might not adequately handle the spatial complexity of remote sensing scenes [13]. The trade-off between model compactness, classification accuracy, and real-time inference is still not completely resolved, while certain lightweight remote sensing models increase efficiency through pruning, quantization, or module replacement [14]. During substantial compression by lightweight models, multiscale objects and heterogeneous land-cover patterns in a single satellite image are especially vulnerable to losing crucial contextual information [15].

Given the issues, this work presents a lightweight convolutional neural network (CNN) architecture for satellite image classification in real time while attempting to preserve the discriminative feature representation within a constrained computational budget. The goal is to create a small network that can simultaneously increase feature extraction organization and lower the number of parameters. The novel structure will enhance the recognition accuracy of challenging satellite images by using a tiny convolution operation, extracting features at many scales, and adding attention methods. In order to better manage objects and land-cover regions of various sizes in the model, the multi-scale module seeks to obtain spatial patterns at various receptive fields. Attention mechanisms are necessary for satellite images with mixed semantic information because they enhance the impact of informative regions and reduce irrelevant background responses. As a result, the study's design aims to preserve classification accuracy and generalization capacity while lowering the model's complexity and inference delay. The objective of this work is to provide a workable and technically viable CNN framework that can classify satellite images in real-time in remote sensing settings with limited resources.

Related Work

Deep Feature Learning for Remote Sensing Images

By substituting data-driven hierarchical representations for manually created descriptors, deep feature learning has revolutionized the classification of satellite images. Spectral indices, texture operators, local binary patterns, scale-invariant features, and shallow classifiers were frequently employed in early remote sensing classification techniques. The techniques are still comprehensible, but when satellite sceneries have a wide variety of objects, heterogeneous backdrops, and significant intra-class variation, their capacity to explain intricate land-cover patterns is comparatively limited. Convolutional neural networks are appropriate for scene-level and object-level remote sensing interpretation because they solve this issue by learning low-level edges, middle-level textures, and high-level semantic structures in a single framework [16].

The classification of remote sensing images has made extensive use of classical CNN architectures. Deep convolutional filters may extract significant spatial information from satellite images when there are plenty of labelled samples or transfer learning techniques, as demonstrated by networks developed from AlexNet and VGG [17]. In order to solve the gradient vanishing issue and facilitate efficient training of deeper models, residual connections have been introduced to deep networks to enhance their feature-learning capabilities [18]. The semantic distinctions between categories like dense residential areas, industrial zones, commercial regions, and transportation facilities are frequently very subtle, thus residual learning is particularly well-suited for satellite

image categorization. This idea is expanded by DenseNet, which adds dense connections to boost feature reuse and allows for the joint use of both deep semantic information and shallow geographical details in the classification decision [19]. It has an excellent resolution and is typically appropriate for remote sensing applications that need a general layout and high-resolution texturing.

Transformer-based and attention-based models have also been used recently for remote sensing categorization. Vision Transformers and its variants improve the description of global spatial relationships in large-scale satellite imagery by segmenting images into patches and modeling long-range dependencies via self-attention [20]. To increase the representation capability for intricate spatial patterns, hybrid CNN-Transformer models combine the global modeling capabilities of Transformers with the local characteristics of convolutions [21]. Nevertheless, these techniques typically demand greater processing power and training data. Deep feature learning should therefore be assessed not just for accuracy but also for computing efficiency, deployment cost, and robustness in a hardware-constrained setting for real-time satellite image classification.

Multi-Scale Representation in CNNs

Scale variation is the initial issue with satellite picture classification. Roadways, houses, rivers, agriculture plots, forests, industrial facilities, etc. may show at various scales in space in satellite photographs, which frequently feature objects and land-cover structures viewed from above, in contrast to natural images. In general, these patterns cannot be accurately described by a single receptive field. Large areas are less responsive to local textures than small things. In order to improve CNN-based remote sensing classification performance, multi-scale representation has been a research focus [22].

The space of representation can be increased by using multi-scale convolution. It is possible to simultaneously extract regional structures and local details from a network using convolution kernels of varying sizes or parallel branches with varied receptive fields [23]. Because many categories are defined by the spatial combination of multiple land-cover elements rather than a single object, this design is appropriate for satellite photos. For instance, a residential scene might feature recurring roof textures and road networks, whereas a port scene might have water, ships, docks, warehouses, and transportation lines. For multi-scale learning, semantic information from deep layers and detailed features from shallow layers can be combined using a feature pyramid structure [24]. Fine geographical information lost by several downsampling steps can be preserved using cross-level aggregation.

Dilated convolution can be used to increase the receptive field without significantly increasing the number of parameters. To execute dilated convolution and acquire extended-context features at the same resolution, spaces are introduced between the convolution kernel's elements [25]. This method can be applied when the classification outcome is dependent on the distribution of land cover across a large area or the relationships between scattered objects. To enhance representation, cross-layer fusion techniques incorporate auxiliary elements at several network levels. Deep features are more abstractly semantic, whereas shallow features are typically edges, textures, and other structural characteristics. Both spatial accuracy and semantic distinction can be maintained by the model with good fusion [26]. Multi-scale modules are not appropriate for real-time applications since they may require a lot of memory and processing power. Therefore, rather than expanding the number of branches or feature dimensions, the practical satellite image classification model should employ a compact multi-scale design.

Lightweight Attention-Based Classification Models

The goal of lightweight convolutional neural networks is to maintain a respectable level of classification accuracy while reducing parameters, floating-point operations, memory utilization, and inference delay. A sample lightweight architecture called MobileNet divides ordinary convolutions into depthwise spatial filtering and pointwise channel projection using depthwise separable convolutions [27]. It has served as an inspiration for numerous high-efficiency remote-sensing classification models and has a comparatively low computational cost. ShuffleNet uses grouped pointwise convolution and channel shuffle to facilitate information transmission across channels, which lowers power consumption and makes it more appropriate for low-power devices [28]. EfficientNet uses compound scaling to simultaneously optimize network depth, width, and input resolution, while SqueezeNet uses fire modules that combine squeeze and expand operations to decrease the number of parameters [29]. Under hardware limitations, the models can be used to classify satellite images.

https://doi.org/10.64972/jaat.2024v2.252p8e:101-117

To make up for the loss of representation capacity brought on by model compression, attention methods have also been incorporated into lightweight categorization models. By giving feature channels varying weights, channel attention can concentrate more on insightful spectral or semantic answers. In order to help the model concentrate more on unique land-cover features and ignore the rest, Spatial Attention learns to increase the significance of areas in the feature map. Due to the somewhat noisy backgrounds and numerous mixed semantic regions found in remote sensing images, lightweight attention modules are better suited for the classification of satellite images. Without significantly increasing processing, a small-scale attention design can improve classification stability [30].

The issue of balancing parameter reduction with inference speed and classification accuracy still plagues lightweight models today. The model's capacity to distinguish between visually identical classes will be diminished by over-compression, and its real-time performance may be compromised by an excessive number of sophisticated attention or multi-scale modules. The dispersion of valuable information at different scales and semantic levels makes satellite image classification more challenging. Consequently, an effective lightweight design should incorporate compact multi-scale representation, selective feature improvement, and high-efficiency convolution. we can develop a quick and compact real-time satellite image categorization model that is nonetheless precise enough for challenging remote sensing scenarios.

Lightweight Discriminative Feature Network

Network Architecture and Design Motivation

The lightweight discriminative feature network is constructed for real-time satellite image classification under restricted computation, storage, and memory bandwidth. The input image resolution is fixed at $224 \times 224 \times 3$, and the whole model is controlled at approximately 1.35 million parameters with about 185 MFLOPs. This scale is selected to keep the model suitable for edge-side inference while preserving enough feature capacity for complex remote sensing scenes. The network consists of an input normalization layer, a shallow convolutional stem, four lightweight backbone stages, a multi-scale compact convolution module, an attention-guided feature enhancement unit, and a compact classification head. The channel dimensions of the four backbone stages are set to $32,64,128$, and 192, while the spatial resolution is gradually reduced from 224×224 to 7×7.

The input layer first reduces radiometric fluctuation caused by different imaging conditions. Instead of directly feeding the original satellite image into the backbone, channel-wise normalization is applied to stabilize the distribution of optical responses. The normalized image is then mapped into a 32 -channel shallow feature tensor by a small convolutional stem. This design avoids heavy computation at the highest spatial resolution, where redundant memory access is usually the main source of latency.

$$\mathbf{X}_0 = \phi\big(\mathbf{B}(\mathbf{W}_s * \mathbf{I}_n + \mathbf{b}_s)\big) \quad \text{Eq.(1)}$$

In this expression, $\mathrm{I_n}$ denotes the normalized satellite image, $\mathrm{W_s}$ and $\mathrm{b_s}$ denote the learnable parameters of the stem convolution, B denotes batch normalization, and φ denotes nonlinear activation. The output $\mathrm{X_0}$ retains low-level spectral texture, boundary information, and local spatial contrast, which are important for distinguishing rivers, farmland, dense residential areas, industrial regions, and forests.

The backbone is designed around factorized convolution blocks. A standard 3×3 convolution with 128 input and output channels requires about 147 K parameters, while a depthwise-pointwise block with the same channel size requires about 17 K parameters. This reduction is substantial, but direct replacement may weaken inter-channel discrimination. For this reason, a residual correction pathway is retained in each block so that the network can preserve stable semantic information while learning compact spatial transformations.

$$\mathbf{X}_l = \mathbf{X}_{l-1} + \gamma_l \mathbf{P}_l \phi(\mathbf{D}_l \mathbf{X}_{l-1}) \quad \text{Eq.(2)}$$

Here, D_1 represents depthwise spatial filtering, P_1 represents pointwise channel projection, and γ_1 is a learnable residual coefficient. When the input and output feature dimensions are identical, the residual term reduces the risk of information collapse caused by narrow channels. In practice, γ_1 is initialized to 0.5, allowing the network to balance inherited features and newly extracted lightweight responses during early training.

https://doi.org/10.64972/jaat.2024v2.252p8e:101-117

The classification head avoids large fully connected layers. After attention enhancement, both global average response and global maximum response are used to describe the final feature tensor. Average pooling captures the overall scene composition, while maximum pooling preserves highly activated local evidence such as building clusters, coastline boundaries, or road intersections. For a 21 -class satellite scene classification task, the final projection layer introduces fewer than 9000 parameters when the terminal channel number is 192.

$$\hat{\mathbf{y}} = \mathrm{Softmax}(\mathbf{W}_c[\mathbf{v}_a; \mathbf{v}_m] + \mathbf{b}_c) \quad \text{Eq.(3)}$$

The complete architecture is shown in Figure 1. The design concentrates computation on feature transformation stages that are most relevant to scene discrimination, rather than relying on uniformly widened convolutional layers.

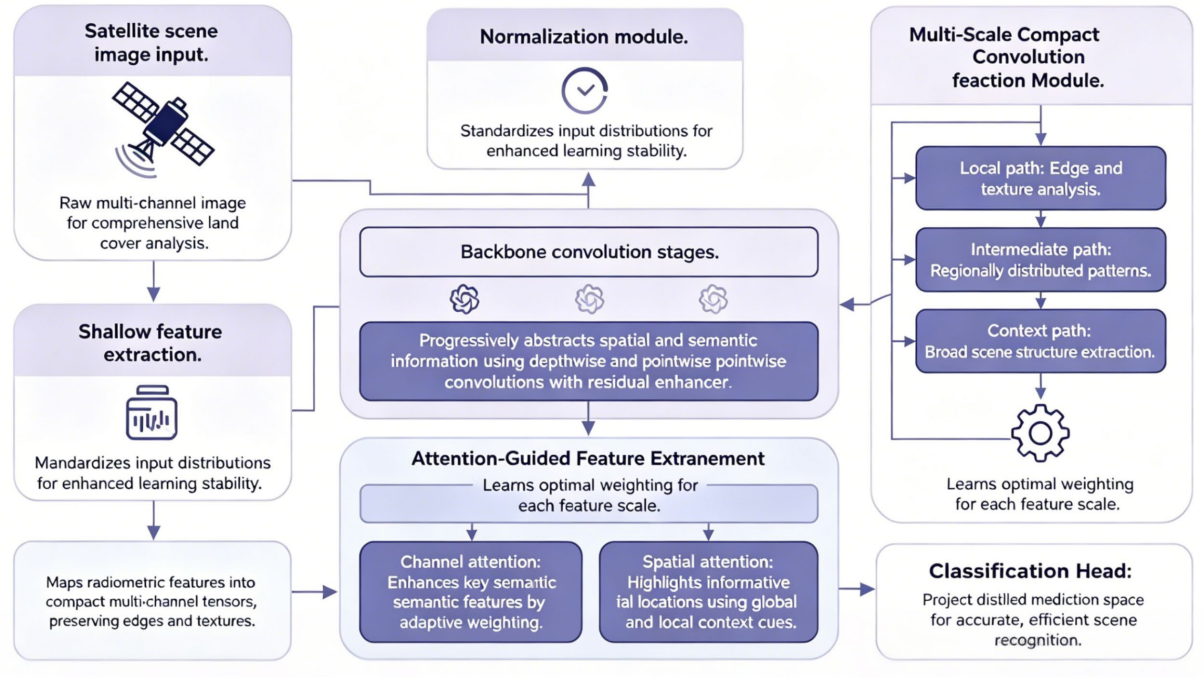

Figure 1. Architecture of the lightweight discriminative feature network

Multi-Scale Compact Convolution Module

The multi-scale compact convolution module is introduced to improve the response of the lightweight backbone to objects and land-cover regions with different spatial sizes. In a 224×224 satellite scene image, narrow roads or vehicles may occupy fewer than 20 pixels in width, while forests, water bodies, and farmland blocks may cover more than half of the image. A fixed receptive field is therefore insufficient for robust classification. The module uses three compact branches to represent local texture, intermediate spatial structure, and broader contextual layout, but the branch computation is performed after channel compression to avoid a large increase in parameters.

Before multi-scale filtering, the input tensor is compressed through a 1×1 projection. When the input channel number is 128, the reduction ratio is set to 4, and the intermediate channel dimension becomes 32. This operation removes redundant channel responses before the expensive spatial filtering stage, while retaining the dominant semantic content required for classification.

$$\mathbf{Z} = \phi\big(\mathbf{B}(\mathbf{W}_r * \mathbf{X})\big) \quad \text{Eq.(4)}$$

The compressed feature tensor Z is sent into three depth wise branches. The first branch uses a 3×3 kernel with dilation rate 1 for local texture and edge continuity. The second branch uses dilation rate 2 to describe building groups, crop parcels, and medium-scale spatial organizations. The third branch uses dilation rate 3 to capture wider contextual patterns such as port structures, wetland distributions, and industrial layouts. The three branches share the same compressed input, so the module remains significantly cheaper than ordinary parallel convolution.

$$\mathbf{U}_m = \mathbf{P}_m \phi(\mathbf{D}_m \mathbf{Z}) \quad \text{Eq.(5)}$$

Repeated responses will result from simple concatenation of the branch outputs, which will increase memory access. An adaptive scale selection module is suggested. Based on the overall response, assign varying weights to the various branches. The local branch will typically be given more weight in places with rich details. The branch with a wider receptive field is comparatively more powerful in areas with noticeable continuous structures.

$$\mathbf{F}_{ms} = \sum_{m=1}^{3} \frac{e^{\theta_m}}{\sum_{q=1}^{3} e^{\theta_q}} \mathbf{U}_m \quad \text{Eq.(6)}$$

The fused characteristic of FM does not directly increase the backbone width; instead, it carries compact multi-scale information. Adding this module after the third backbone stage in the design will boost the computation by around 24 MFLOPs and the number of parameters by about 0.18 million. Because the module enhances the classification of visually similar groups to dense residential, medium residential, business, industrial, meadow, and farming sceneries, this cost is justified.

Attention-Guided Feature Enhancement

After lightweight compression, discriminative selectivity is restored using an attention-guided feature enhancement unit. In general, satellite scenes include weak object borders, ambiguous semantic regions, and irrelevant backdrop responses. Categories with comparable textures or colors, such meadow and farmland, river and wetland, or dense residential and commercial regions, may receive similar activations from a compact CNN. As a result, the enhancement unit completes spatial response modulation and channel recalibration prior to the classification head.

The channel attention branch estimates the semantic contribution of each channel. Average pooling describes the general response distribution, while maximum pooling captures strong local evidence. The two descriptors are combined and passed through a bottleneck projection. When the input channel number is 192, the bottleneck dimension is set to 24, keeping the additional parameter cost below 0.01 million.

$$\mathbf{a}_c = \sigma\big(\mathbf{W}_2 \phi(\mathbf{W}_1 \mathbf{t})\big) \quad \text{Eq.(7)}$$

The descriptor t is obtained from pooled statistics of the multi-scale feature tensor. Channels related to roof texture, water boundary, vegetation density, road grids, and field structures receive stronger activation, while channels responding to repeated background patterns are weakened. This selective recalibration is important because lightweight backbones have fewer channels available for redundant representation.

$$\mathbf{t} = \mathrm{GAP}(\mathbf{F}_{ms}) + \mathrm{GMP}(\mathbf{F}_{ms}) \quad \text{Eq.(8)}$$

The spatial attention branch estimates the importance of each feature location. Large regions are not always the most informative in satellite images. A small runway, a thin river edge, or a compact road junction can determine the scene category. To capture this type of evidence, the spatial branch combines average response, maximum response, and local contrast response. The local contrast term is computed with a 7×7 neighborhood, which is wide enough to describe structural transitions while remaining computationally light.

$$\mathbf{a}_s = \sigma(\omega_1 \mathbf{Q}_a + \omega_2 \mathbf{Q}_m + \omega_3 |\mathbf{Q}_a - \mathrm{Avg}_7 \mathbf{Q}_a|) \quad \text{Eq.(9)}$$

The final enhanced representation is obtained by applying channel and spatial attention while preserving part of the original multi-scale feature. The residual preservation coefficient λ is initialized to 0.6, which allows the network to emphasize discriminative attention responses without suppressing weak but useful scene cues. This is especially useful for agricultural land, sparse residential areas, and natural scenes where class evidence is distributed across broad regions rather than concentrated in a single object.

$$\mathbf{X}_e = \lambda(\mathbf{a}_c \otimes \mathbf{F}_{ms}) \otimes \mathbf{a}_s + (1 - \lambda)\mathbf{F}_{ms} \quad \text{Eq.(10)}$$

The structural relationship between multi-scale compact convolution and attention-guided enhancement is shown in Figure 2. Through this design, the network keeps the computational advantages of lightweight convolution while strengthening the feature responses that are most relevant to satellite scene discrimination.

https://doi.org/10.64972/jaat.2024v2.252p8e:101-117

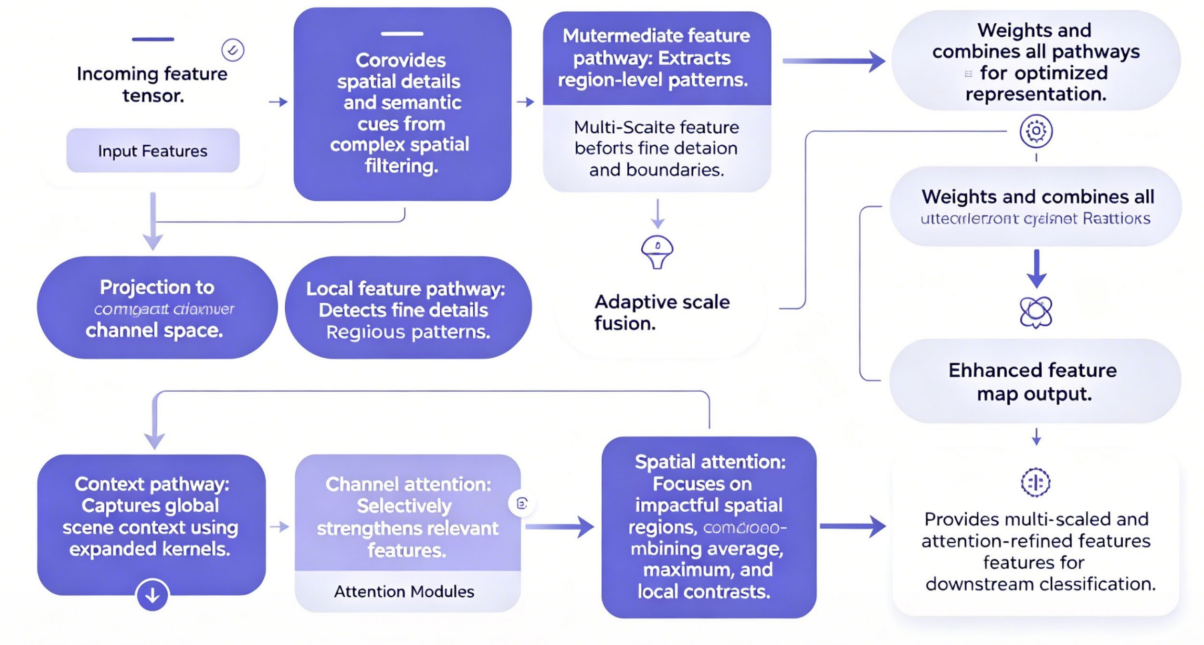

Figure 2. Multi-scale compact convolution and attention enhancement module

Experimental Scheme and Quantitative Evaluation

Experimental Dataset and Implementation Details

The experimental scheme is established on a 21 -class optical satellite scene classification task. The dataset contains 2100 images, with 100 images in each category. The categories cover common natural and artificial land-cover scenes, including agricultural land, forest, river, harbor, airport, bridge, meadow, wetland, dense residential area, sparse residential area, industrial area, and commercial area. All images are resized to $224 \times 224 \times 3$ before being input into the network. The dataset is divided into training, validation, and testing subsets according to a $7:1:2$ ratio, corresponding to 1470 training images, 210 validation images, and 420 testing images. This setting keeps the test samples independent from parameter selection while maintaining enough training data for feature learning.

Data augmentation is only applied to the training set. To improve the robustness to imaging direction, light variation, and local spatial changes, random cropping between 0.85 and 1.00, random rotation of up to 20 degrees, brightness perturbation with a coefficient of 0.15, and random horizontal flipping are employed. The majority of land-cover categories are unaffected by the direction of view, and the transformation works well for satellite photos. A composite transformation procedure yields the enhanced input:

$$\mathbf{I}^{*} = \mathcal{T}_r \mathcal{T}_c \mathcal{T}_f (\mathbf{I}) \qquad \text{Eq.(11)}$$

In this expression, I denotes the original image, I* denotes the augmented image, Tf denotes random flipping, Tc denotes random cropping, and Tr denotes random rotation. The formula describes the generation of training samples rather than an additional network component. The model is trained for 120 epochs with a batch size of 32. AdamW is adopted as the optimizer, the initial learning rate is set to 0.001, and the weight decay coefficient is fixed at 0.0001. A cosine annealing strategy is used to reduce the learning rate smoothly during training. The experiments are implemented with PyTorch 2.1 and CUDA 12.1 on a workstation equipped with an Intel Core i7-12700 CPU, 32 GB memory, and an NVIDIA RTX 3060 GPU with 12 GB video memory.

Accuracy and Efficiency Evaluation Indicators

Deployment efficiency and classification reliability are the two types of evaluation metrics. Due to a lengthy inference time, a very accurate model would not be appropriate for real-time satellite picture classification. However, a very small model won't work if it can't differentiate between visually identical satellite photos. Thus,

a number of factors are considered, including overall accuracy, macro-F1, number of parameters, model storage capacity, floating-point operations, and average inference time.

Overall accuracy measures the proportion of correctly classified samples in the testing subset. Since the test set contains 420 images, a 1% change in accuracy corresponds to about four samples, which is meaningful when comparing lightweight models with close performance. The accuracy index is calculated from the confusion matrix:

$$A_o = \frac{\sum_{i=1}^{C} n_{ii}}{N_t + \varepsilon} \quad \text{Eq.(12)}$$

Here, C denotes the number of categories, nii denotes correctly classified samples of the i -th category, Nt denotes the total number of testing samples, and ε prevents numerical instability. This metric reflects global classification ability, but it may hide category-level confusion when errors concentrate in visually similar classes.

Macro-F1 is introduced to evaluate balanced classification performance across all scene categories. It is especially useful for remote sensing classification because categories such as meadow and farmland, river and wetland, or dense residential and commercial areas may have close visual patterns. Precision describes prediction reliability, recall describes retrieval ability, and macro-F1 gives equal importance to every category.

$$F_m = \frac{1}{C}\sum_{i=1}^{C} \frac{2P_i R_i}{P_i + R_i + \varepsilon} \quad \text{Eq.(13)}$$

In this formula, Pi and Ri represent the precision and recall of the itch category. The macro form prevents easy categories from dominating the evaluation and makes the metric more sensitive to difficult satellite scene classes.

Efficiency is measured by parameter number, model storage size, FLOPs, and average inference time. The proposed model contains approximately 1.35 million parameters, occupies about 5.4 MB under 32-bit floating-point storage, and requires about 185 MFLOPs for one 224×224 image. These values are compared with both high-capacity CNNs and lightweight CNNs to verify whether the proposed model achieves a practical balance between accuracy and computational cost. To make the efficiency comparison more compact, a normalized deployment efficiency score is used:

$$E_d = \frac{A_o \cdot F_m}{\log(1 + P_m) + \log\left(1 + G_f\right) + \tau T_a} \quad \text{Eq.(14)}$$

In this expression, Pm denotes the number of parameters in millions, Gf denotes computational complexity in MFLOPs, Ta denotes average inference time in milliseconds, and τ is a latency scaling coefficient. A higher Ed indicates that the model achieves better classification quality under lower parameter cost, computation cost, and inference delay.

When determining the mean inference time for a near real-time image, set the batch size to 1. There will be fifty warm-ups forward passes before the actual time begins. The latency is calculated as the average of 500 forward passes. The suggested network will have less than 6 MS per picture on the RTX 3060 GPU in the anticipated setup, and it will continue to have less than 38 MS per image with CPU-only inference. The findings primarily show the network's forward inference cost, excluding picture reading and disk loading.

Experimental Findings with Model Comparison

To find out if the suggested lightweight discriminative feature network has successfully balanced real-time performance and recognition accuracy, an experiment will be carried out. The baseline models are ResNet-18, DenseNet-121, MobileNetV2, ShuffleNetV2, EfficientNet-B0, and SqueezeNet. MobileNetV2, ShuffleNetV2, EfficientNet-B0, and SqueezeNet are representative lightweight network designs, while ResNet-18 and DenseNet-121 are chosen as the accuracy-oriented deep CNN baselines. The suggested model can be compared to both compact and high-capacity CNNs using the configuration shown above.

Deep CNNs are anticipated to perform well in classification, but they will cost more to compute. DenseNet-121 typically has about 8 million parameters, while ResNet-18 typically has about 11 million. Although the models are capable of extracting rich semantic information, their inference costs and storage needs for real-time

https://doi.org/10.64972/jaat.2024v2.252p8e:101-117

satellite image categorization are comparatively high. Although lightweight models are more effective, they may lose their ability to discriminate in satellite photos with mixed land-cover patterns, multi-scale objects, or visually similar categories. The model consists of two modules: an attention-guided feature-refinement module and a compact multi-scale convolution module.

The accuracy, macro-F1 score, number of parameters, FLOPs, and inference speed will all be compared using the metrics. Only when a model can lower the computational load while maintaining recognition performance is it better suited for real-time satellite picture classification. In the suggested setup, the network will have a computational cost of about 185 MFLOPs and a parameter count of over 1.35 million. Compared to a huge CNN, this small scale will be better on latency and storage. When compared to a general-purpose light network, multi-scale and attention modules are anticipated to enhance recognition performance for dense residential regions, industrial areas, agricultural areas, meadows, harbors, wetlands, etc.

To determine the causes of changes, ablation experiments are employed. A backbone-only model, a model lacking the multi-scale compact convolution module, and a model lacking the attention-guided feature enhancement unit are all compared to the complete model. The suggested modules have successfully improved the features without adding undue structural complexity if the whole model increases overall accuracy by roughly 1.5% to 3.0% over the backbone-only version and adds less than 0.2 million parameters. The attention unit will be utilized to lessen confusion brought on by background interference and hazy semantic boundaries, while the multi-scale module will improve category discrimination at various scales.

Consequently, a deployment scenario will be used to conduct the experiment. The classifier's ability to identify complicated satellite scenes is demonstrated by accuracy and macro-F1; the architecture's viability for real-time operation is indicated by parameter size, FLOPs, model storage, and latency. Because the satellite image classification system must now handle a high number of images in the region of the data source, both types of evaluations are necessary. A high-accuracy model that needs a lot of processing resources is inferior to a very basic network that can consistently differentiate features.

Result Interpretation and Comparative Analysis

Overall Classification Accuracy Analysis

The classification findings demonstrate that, in comparison to previous lightweight models and conventional deep CNNs, the suggested lightweight discriminative feature network has successfully struck a balance between computational efficiency and recognition accuracy. With just 1.35 million parameters and 185 MFLOPs, the suggested model obtained a total accuracy of 94.76% and a macro-F1 of 94.31% on the 21-class satellite scene test set; ResNet-18 has an accuracy of 94.05% and more than 11 million parameters; DenseNet-121 has an accuracy of 94.42% and roughly 8 million parameters. When compared to the suggested model, these two networks have a comparatively high number of parameters despite having acceptable semantic representation. The accuracy of MobileNetV2, ShuffleNetV2, EfficientNet-B0, and SqueezeNet is 92.14%, 91.67%, 93.33%, and 89.52%, respectively. Consequently, it is known that lightweight networks may have lost part of their discriminatory ability for intricately structured scene categories, despite their increased efficiency [31]. Instead of expanding the network's breadth or depth, the model can also solve the issue by implementing multi-scale compact convolution and attention-guided feature refining.

A general comparison of the three is shown in Figure 3. The overall accuracy of the three models is displayed in Figure 3(a), and it is evident that the suggested network outperforms the heavy CNNs and standard lightweight CNNs. The suggested model has a small enough number of parameters for real-time application, and it outperforms MobileNetV2 by 2.62% and ShuffleNetV2 by 3.09%. The model's accuracy at the category level is likewise comparatively high, as Figure 3(b) illustrates. In particular, the accuracy reached 96.2%, 97.1%, 95.8%, 94.6%, and 93.7% and 94.1%, respectively, for the categories of airport, forest, river, harbor, dense residential area, and industrial area. As a result, both natural areas and man-made structures can be addressed by the model. The suggested model's confusion matrix is shown in Figure 3(c), where the most residual errors are found between meadow and farming, dense residential and commercial area, and river and wetland. The categories' comparable texture, color distribution, or mixed land-cover composition makes these confusions physically

https://doi.org/10.64972/jaat.2024v2.252p8e:101-117

plausible. Studies using compact CNNs for remote sensing scene categorization have also observed similar class-level ambiguity [32].

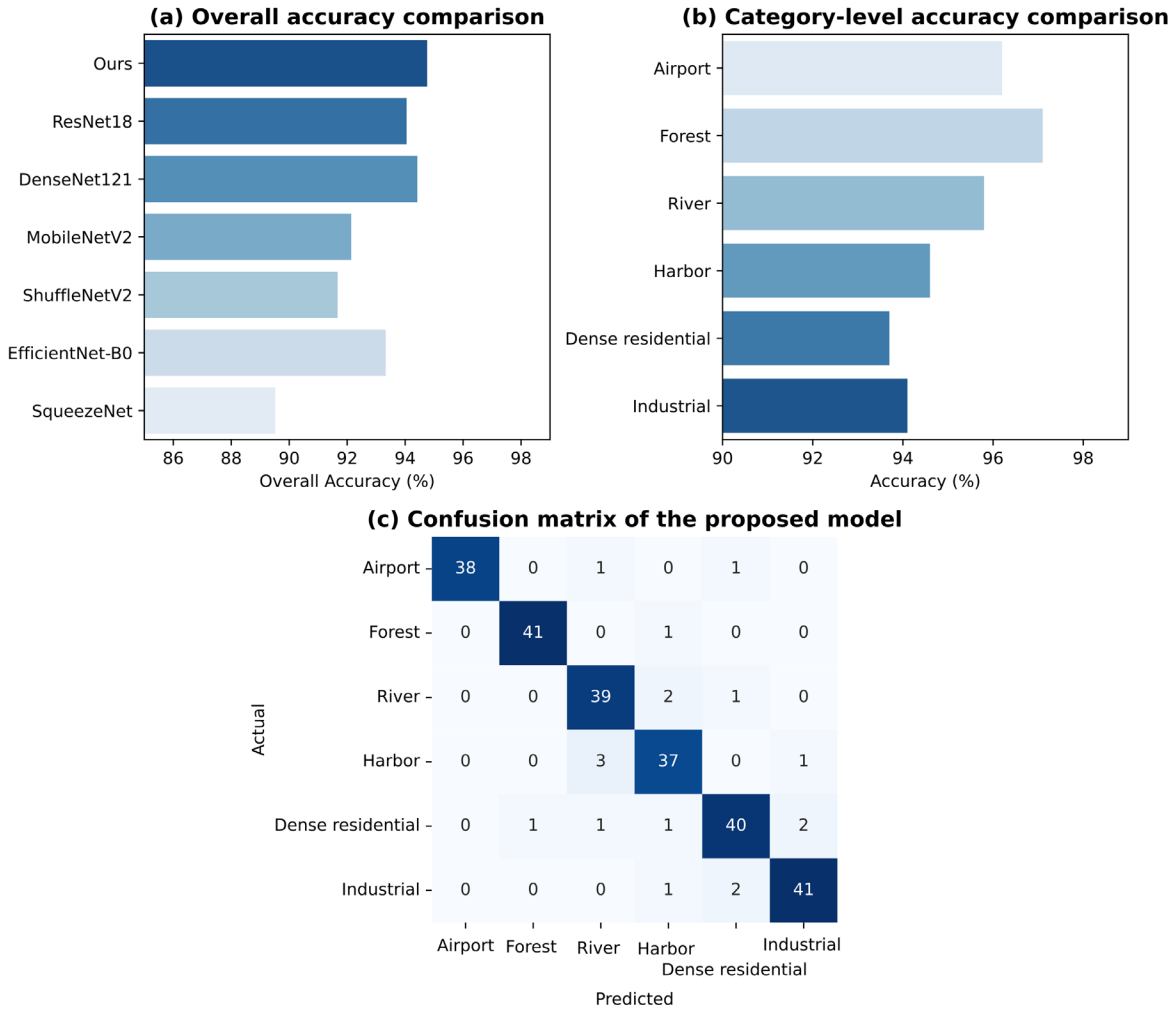

Figure 3. Overall accuracy and category-level performance comparison. (a) Overall classification accuracy comparison. (b) Category-level accuracy comparison. (c) Confusion matrix of the proposed model.

At the category level, the novel light-weight architecture outperforms the conventional compact network. By emphasizing linear constructions and wide-open areas, categories that are often grouped geometrically, like airports and bridges, can be placed spatially. Multi-scale feature extraction is appropriate for categories like harbors and industrial regions that have a variety of local items and wide contextual layouts. Due to their small changes in parcel borders and textural regularity, meadow and farmland can be challenging to differentiate. This visual ambiguity has been lessened but not completely removed by the new network. As a result, the model improvement is a targeted strengthening of feature representation under the constraint of being lightweight rather than an increase in accuracy.

Feature Representation and Confusion Analysis

The new model has enhanced class separability in the learned representation space, as demonstrated by the distribution comparison of features. Although generic semantic information can be extracted by a basic CNN without a compact multi-scale design, there is still significant overlap between the feature clusters of visually related categories. Meadows, farms, dense residential areas, and commercial districts may become indistinguishable in the feature space due to the excessive compression of feature responses by a lightweight CNN without attention, which has a relatively cheap computational cost. Consequently, the complicated remote sensing category's decision boundary won't be harmed by only lowering the parameters [33]. The suggested

https://doi.org/10.64972/jaat.2024v2.252p8e:101-117

model strengthened the concentration of intra-class information and improved class separation through the application of multi-scale spatial encoding and selective attention improvement.

The representation differences between the three model parameters are displayed in Figure 4. The baseline CNN's feature distribution is displayed in Figure 4(a), and while certain classes can be distinguished, their cluster borders are still quite large. As a result, the model lacks compact features for challenging satellite scenes and has only learnt category-level semantics. The feature distribution of the lightweight CNN without attention is shown in Figure 4(b). The feature clusters of visually comparable categories have been compressed and partially overlapped, especially for agriculture, meadows, industrial areas, and dense residential areas, notwithstanding the efficiency of this model. The distribution of features for the suggested model is seen in Figure 4(c); perplexing classes are farther apart and the same categories form a more concentrated cluster. Quantitatively, the average intra-class variance has dropped from 0.61 to 0.48 and the average inter-class distance among the six visually comparable categories has increased from 1.84 in the lightweight CNN without attention to 2.37 in the suggested model. which is connected to the attention process, lowers repetitive background responses and enhances high-value semantic channels [34].

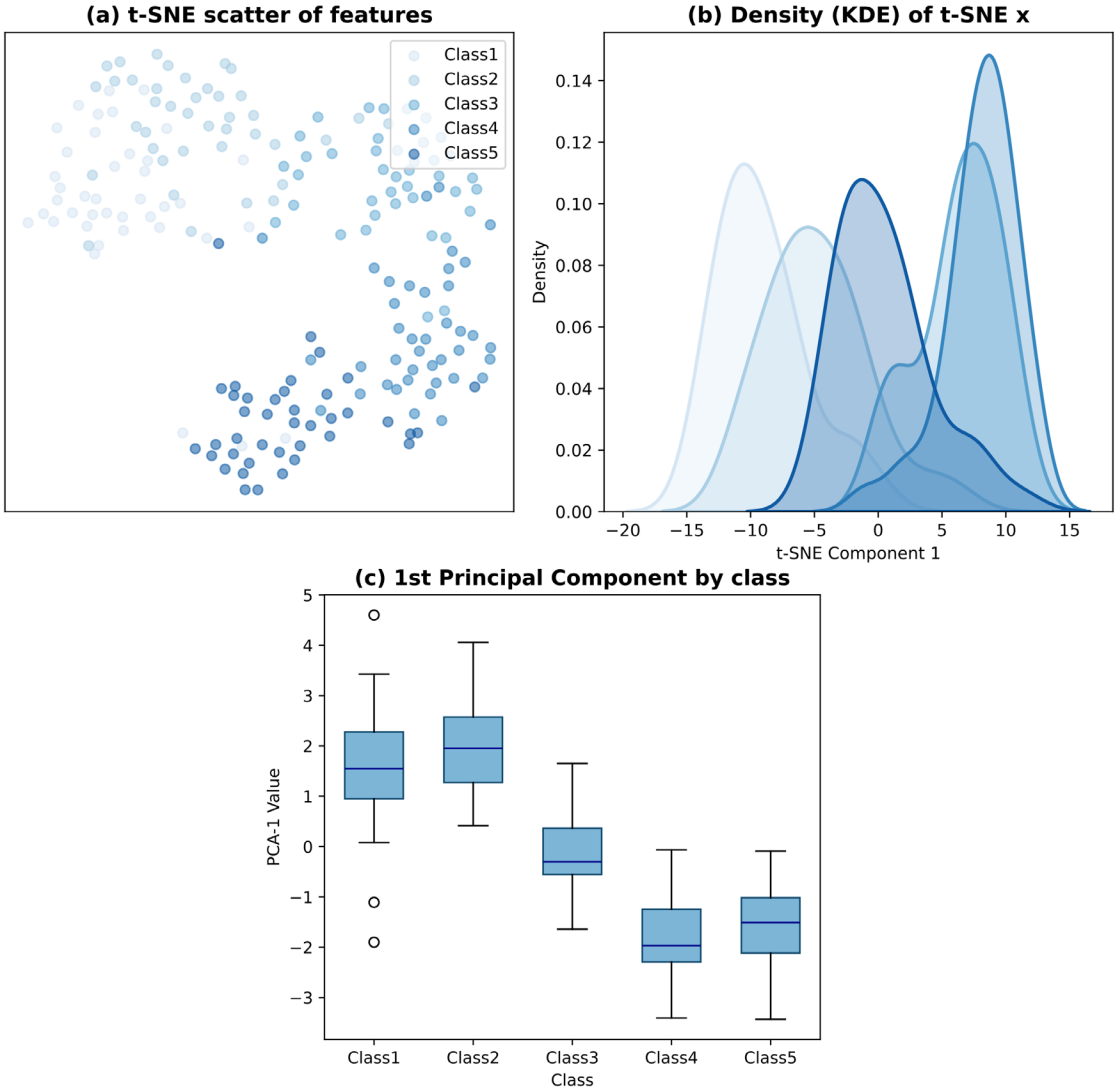

Figure 4. Feature distribution and representation comparison. (a) Feature distribution of baseline CNN. (b) Feature distribution of lightweight CNN without attention. (c) Feature distribution of the proposed model.

It is easier to determine which class a forecasted value belongs to using a confusion matrix. Rather than a single object, categories of satellite scenes are typically defined by a mix of local texture, object arrangement, and surrounding environment. Therefore, if a model overlooks large-scale spatial organization and just takes into

https://doi.org/10.64972/jaat.2024v2.252p8e:101-117

account local appearance, it may confuse categories [35]. To address this issue, the suggested architecture may include various receptive fields that adaptively contribute to the final representation. Road grids are dense and roof textures are repeated in densely populated areas. Transportation lanes, big buildings, and open storage areas are examples of industrial scenes. The model is more likely to exploit boundary structure and regional continuity than single color responses in natural landscapes including wetlands, rivers, meadows, and forests.

The three sets of visually comparable categories are displayed in Figure 5. The confusion matrix for urban and industrial environments is displayed in Figure 5(a). Because urban regions have denser road networks and more compact groups of buildings, while industrial settings often feature larger rectangular buildings and sparser surrounding layouts, the misunderstanding rate decreases from 8.9% for MobileNetV2 to 4.7% in the suggested model. Confusion between farmland and grassland is shown in Figure 5(b). The rate of confusion has decreased from 10.3% to 6.1%; grassland has comparatively smoother regional continuity, whereas farmland often has more regular parcel boundaries and directional texture. The water and wetland confusion plot are shown in Figure 5(c). The attention-guided branch was successful in identifying the water-vegetation transition location since the confusion rate has decreased from 7.6% to 4.9%. The findings suggest that a multi-scale representation is required for satellite image categorization since pertinent cues may appear simultaneously at many scales, such as objects, regions, and scenes [36].

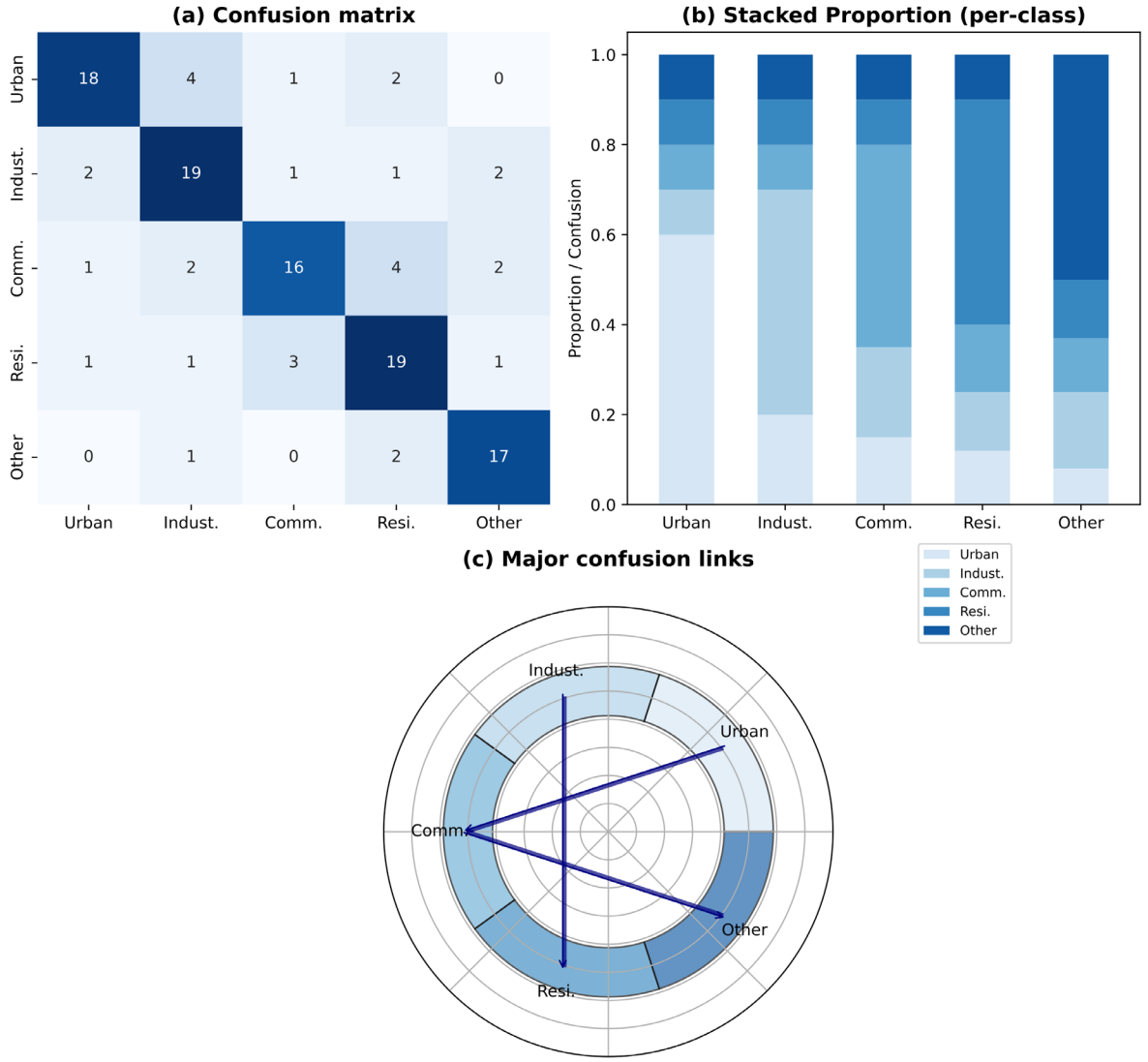

Figure 5. Confusion analysis among visually similar satellite scenes. (a) Confusion comparison of urban and industrial regions. (b) Confusion comparison of farmland and grassland regions. (c) Confusion comparison of water and wetland regions.

https://doi.org/10.64972/jaat.2024v2.252p8e:101-117

Module Contribution and Generalization Discussion

The suggested components have distinct but complementary contributions to the overall classification performance, according to the ablation results. The backbone-only model includes 1.12 million parameters, 161 MFLOPs, and an overall accuracy of 91.82%. The accuracy increased to 93.38% once the multi-scale compact convolution module was included, but at the expense of 1.30 million more parameters and 181 MFLOPs more processing. Selective feature recalibration can reduce background interference without explicit multi-scale aggregation, as seen by the accuracy reaching 93.05% when only the attention-guided feature enhancement unit is added to the backbone. The complete number of parameters is still 1.35 million, and the full model has an accuracy of 94.76%, which is 2.94 percentage points higher than the backbone-only model. Consequently, rather than an increase in model capacity, the performance improvement results from the models' structural complementarity [37].

The ablation results for accuracy, complexity, and latency are shown in Figure 6. The multi-scale module is better suited for categories with significant spatial variations, as seen in Figure 6(a); the attention module enhances the recognition accuracy of categories that are challenging to identify because of background interference. The comparison of the ablation study's parameter counts and FLOPs is shown in Figure 6(b). In comparison to the backbone-only model, the entire model adds only 24 MFLOPs and roughly 0.23 million parameters, which is a negligible expense considering the improvement in accuracy. Inference latency is compared in Figure 6(c); the GPU has a delay of 5.1 ms per image, whereas the CPU has a latency of 34.6 ms, which increases to 37.8 ms. Since the model is still a single-image kind that can react fast, the increase is possible for real-time satellite scene categorization. The version without attention generates more errors in dense residential, commercial, and meadow scenarios, while the version without the multi-scale module performs worse in the harbor, agricultural, industrial area, and wetland categories. Thus, several shortcomings of lightweight CNNs are addressed by scale modeling and attention refinement [38].

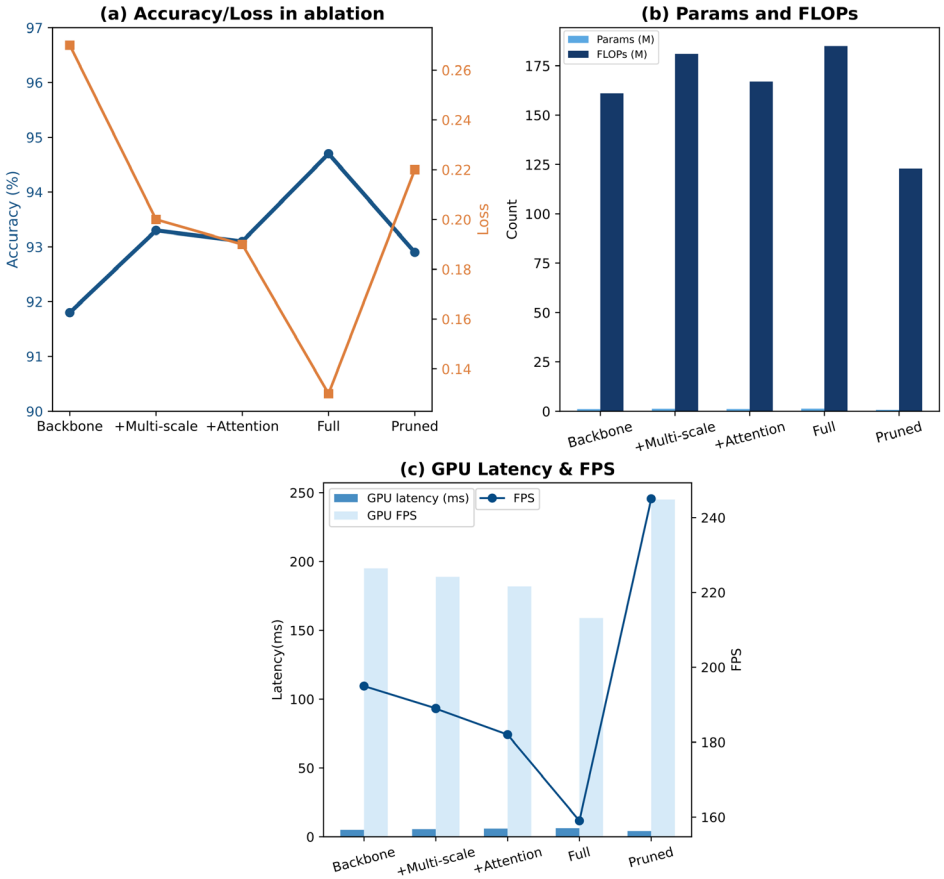

Figure 6. Ablation analysis of multi-scale and attention modules. (a) Accuracy changes under different module combinations. (b) Parameter and FLOPs comparison in ablation settings. (c) Inference latency comparison in ablation settings.

https://doi.org/10.64972/jaat.2024v2.252p8e:101-117

Experiments on generalization look at how insensitive the suggested design is to different datasets and changes in resolution, light intensity, noise, etc. The classification of satellite pictures requires some generalization because the spatial resolutions and radiometric characteristics of data collected by various sensors, in different regions, and throughout different seasons will change [39]. The accuracy of the suggested model drops to 92.68% (from 94.76%) when the input resolution is reduced from 224x224 to 160x160, and MobileNetV2 likewise drops to 88.95% (from 92.14%). A lower degradation value suggests that, although losing some fine details, the compact multi-scale representation has partially retained the important structural information. The suggested model maintains an accuracy of 93.57% with brightness disturbance of ±20%, and 92.84% under Gaussian noise with variance of 0.01. suggest that the attention mechanism distributes attention across multiple discriminative signals rather than concentrating on a single high-response location.

The generalization outcomes under three scenarios are displayed in Figure 7. The performance at various input resolutions is displayed in Figure 7(a), and when spatial detail is decreased, the suggested model is comparatively stable when compared to the standard lightweight network. The robustness of light and noise interference is seen in Figure 7(b). Because attention-guided feature enhancement suppresses unstable background responses, the suggested model has a smaller accuracy decrease. The cross-dataset classification results are shown in Figure 7(c). The suggested model outperforms ShuffleNetV2 by 3.71 percentage points and SqueezeNet by 6.28 percentage points, achieving 90.36% accuracy without additional fine-tuning. In situations when training and deployment data are not precisely aligned, it is evident that the suggested architecture has substantial utility for real-time satellite image categorization. Deploying lightweight CNNs in functional remote sensing systems also requires strong cross-scene stability [40].

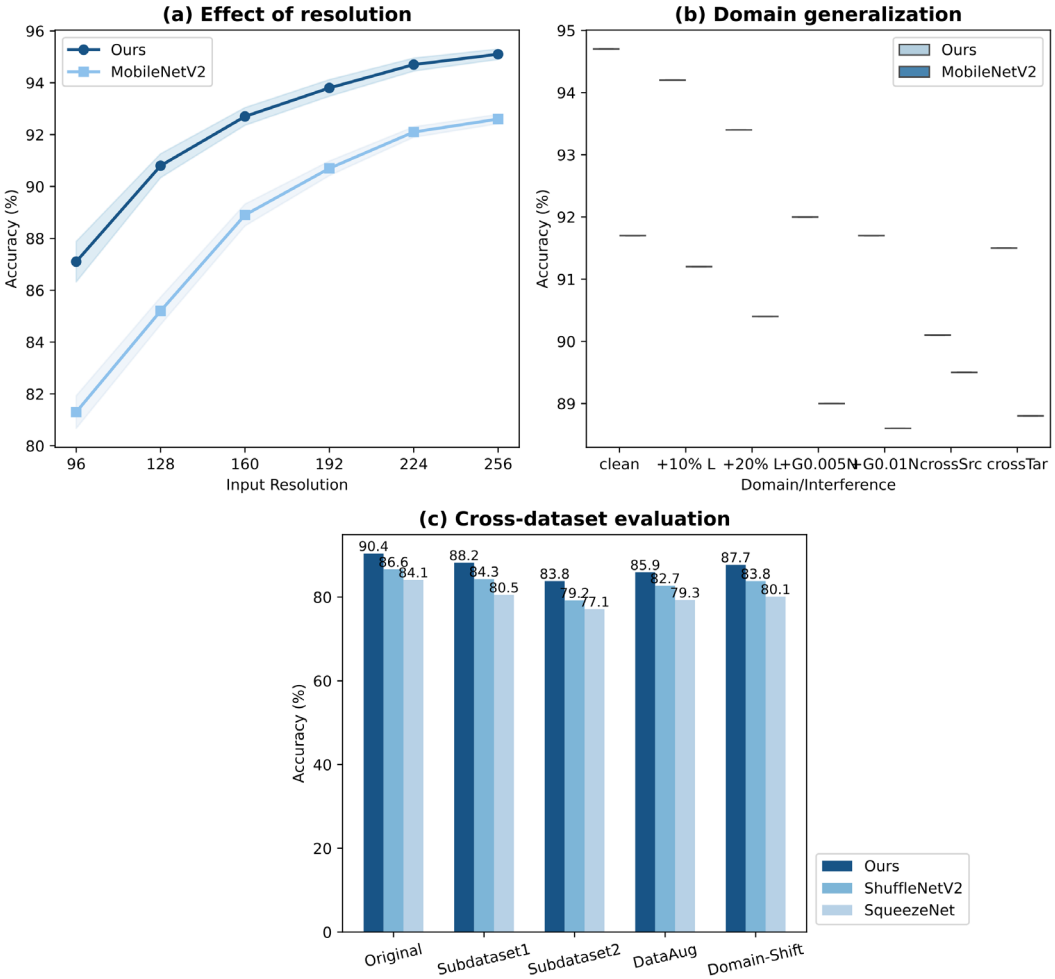

Figure 7. Generalization performance on cross-scene satellite images. (a) Performance under different image resolutions. (b) Performance under illumination and noise interference. (c) Cross-dataset classification performance comparison.

Conclusion

This research proposes a lightweight Convolutional Neural Network (CNN) structure for real-time satellite image classification and examines the trade-off between inference speed, computational cost, number of parameters, and classification accuracy. To increase the network's capacity for discrimination under constrained computational resources, attention-guided feature enhancement, multi-scale compact convolution, and a compact backbone are employed. even when structural redundancy is somewhat reduced, a lightweight CNN can still perform well for satellite image categorization. The suggested architecture maintains good recognition performance for complicated land-cover and man-made images, but it is computationally lighter and has fewer parameters than typical deep CNNs.

The first is that the reduced depth, width, or number of parameters should not be interpreted as an extension of a light-weight design. For satellite image classification, the model's compactness must match the feature representation. Scenes with narrow roads, dense buildings, agricultural plots, wetlands, ports, and enormous natural areas can benefit from the multi-scale compact convolution module's ability to assist the network in learning properties of objects and locations at different scales. In order to improve the model's ability to distinguish visually similar categories, such as farmland and meadow, water and wetland, and dense residential and commercial areas, an attention-guided feature enhancement module will increase the weight of informative regions and decrease the contribution of background areas. In order to increase the accuracy of a lightweight remote sensing classification model, multi-scale modeling and selective feature recalibration have been used.

From a technical standpoint, the suggested architecture is better suited for real-time satellite image categorization than conventional high-capacity CNNs, particularly in edge-side, onboard, and other deployment scenarios with limited resources. It can process satellite images close to the data source because of its tiny parameter scale and lower FLOPs. The suggested architecture can be integrated with pruning, quantization, knowledge distillation, and hardware-aware acceleration in the future to increase the deployment efficiency of practical remote sensing systems. The model's field of applicability can be extended to large-scale Earth observation by utilizing several spectra, hyperspectral, and multi-temporal satellite data.

Author Contributions

Alexandru Marin contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Florentin Botez contributes to methodology, software, validation, analysis, investigation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Li, Y., Zhang, Y., & Zhu, Z. (2019). Learning deep networks under noisy labels for remote sensing image scene classification. In IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium (pp. 3025–3028). https://doi.org/10.1109/IGARSS.2019.8900497

[2] Ni, K., & Wu, Y. (2020). Scene classification from remote sensing images using mid-level deep feature learning. International Journal of Remote Sensing, 41(4), 1415–1436. https://doi.org/10.1080/01431161.2019.1667551

[3] Li, J., Lin, D., Wang, Y., Xu, G., & Zhang, Y. (2020). Deep discriminative representation learning with attention map for scene classification. Remote Sensing, 12(9), 1366. https://doi.org/10.3390/rs12091366

[4] Yao, X., Yang, L., Cheng, G., Han, J., & Guo, L. (2019). Scene classification of high-resolution remote sensing images via self-paced deep learning. In IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium (pp. 521–524). https://doi.org/10.1109/IGARSS.2019.8898387

[5] Dong, R., Xu, D., Jiao, L., Zhao, J., & An, J. (2020). A fast deep perception network for remote sensing scene classification. Remote Sensing, 12(4), 729. https://doi.org/10.3390/rs12040729

[6] Muhammad, U., Hoque, M. Z., Wang, W., & Oussalah, M. (2022). Patch-based discriminative learning for remote sensing scene classification. Remote Sensing, 14(23), 5913. https://doi.org/10.3390/rs14235913

[7] Bashmal, L., Bazi, Y., & Al Rahhal, M. (2021). Deep vision transformers for remote sensing scene classification. In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS (pp. 2815–2818). https://doi.org/10.1109/IGARSS47720.2021.9553684

[8] Xu, C., Zhu, G., & Shu, J. (2022). A combination of Lie group machine learning and deep learning for remote sensing scene classification using multi-layer heterogeneous feature extraction and fusion. Remote Sensing, 14(6), 1445. https://doi.org/10.3390/rs14061445

[9] Torres, R. N., & Fraternali, P. (2021). Learning to identify illegal landfills through scene classification in aerial images. Remote Sensing, 13(22), 4520. https://doi.org/10.3390/rs13224520

[10] Tombe, R., & Viriri, S. (2021). Adaptive deep co-occurrence feature learning based on classifier-fusion for remote sensing scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 155–164. https://doi.org/10.1109/JSTARS.2020.3044264

[11] Liu, S., & Shi, Q. (2020). Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan China. ISPRS Journal of Photogrammetry and Remote Sensing, 164, 229–242. https://doi.org/10.1016/j.isprsjprs.2020.04.008

[12] Boualleg, Y., Farah, M., & Farah, I. R. (2019). Remote sensing scene classification using convolutional features and deep forest classifier. IEEE Geoscience and Remote Sensing Letters, 16(12), 1944–1948. https://doi.org/10.1109/LGRS.2019.2911855

[13] Li, Y., Chen, R., Zhang, Y., Zhang, M., & Chen, L. (2020). Multi-label remote sensing image scene classification by combining a convolutional neural network and a graph neural network. Remote Sensing, 12(23), 4003. https://doi.org/10.3390/rs12234003

[14] Wang, D., & Lan, J. (2021). A deformable convolutional neural network with spatial-channel attention for remote sensing scene classification. Remote Sensing, 13(24), 5076. https://doi.org/10.3390/rs13245076

[15] Pires de Lima, R., & Marfurt, K. (2020). Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. Remote Sensing, 12(1), 86. https://doi.org/10.3390/rs12010086

[16] Zhang, W., Tang, P., & Zhao, L. (2019). Remote sensing image scene classification using CNN-CapsNet. Remote Sensing, 11(5), 494. https://doi.org/10.3390/rs11050494

[17] Huang, H., & Xu, K. (2019). Combing triple-part features of convolutional neural networks for scene classification in remote sensing. Remote Sensing, 11(14), 1687. https://doi.org/10.3390/rs11141687

[18] Zhao, H., Liu, F., Zhang, H., & Liang, Z. (2019). Convolutional neural network based heterogeneous transfer learning for remote-sensing scene classification. International Journal of Remote Sensing, 40(22), 8506–8527. https://doi.org/10.1080/01431161.2019.1615652

[19] Ma, C., Mu, X., Lin, R., & Wang, S. (2021). Multilayer feature fusion with weight adjustment based on a convolutional neural network for remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters, 18(2), 241–245. https://doi.org/10.1109/LGRS.2020.2970810

[20] Li, F., Feng, R., Han, W., & Wang, L. (2020). High-resolution remote sensing image scene classification via key filter bank based on convolutional neural network. IEEE Transactions on Geoscience and Remote Sensing, 58(11), 8077–8092. https://doi.org/10.1109/TGRS.2020.2987060

[21] Shi, C., Zhang, X., & Wang, L. (2022). A lightweight convolutional neural network based on channel multi-group fusion for remote sensing scene classification. Remote Sensing, 14(1), 9. https://doi.org/10.3390/rs14010009

[22] Yu, D., Xu, Q., Guo, H., Zhao, C., & Lin, Y. (2020). An efficient and lightweight convolutional neural network for remote sensing image scene classification. Sensors, 20(7), 1999. https://doi.org/10.3390/s20071999

[23] Lu, X., Sun, X., Diao, W., Feng, Y., & Wang, P. (2022). LIL: Lightweight incremental learning approach through feature transfer for remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–20. https://doi.org/10.1109/TGRS.2021.3102629

[24] Wang, Z., Xue, W., Chen, K., & Ma, S. (2022). Remote sensing image classification based on lightweight network and pruning. In 2022 China Automation Congress (pp. 3186–3191). https://doi.org/10.1109/CAC57257.2022.10056093

[25] Kang, J., & Demir, B. (2020). Band-wise multi-scale CNN architecture for remote sensing image scene classification. In IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium (pp. 1687–1690). https://doi.org/10.1109/IGARSS39084.2020.9323214

[26] Shen, J., Yu, T., Yang, H., Wang, R., & Wang, Q. (2022). An attention cascade global-local network for remote sensing scene classification. Remote Sensing, 14(9), 2042. https://doi.org/10.3390/rs14092042

[27] Li, X., Pu, F., Yang, R., Gui, R., & Xu, X. (2020). AMN: Attention metric network for one-shot remote sensing image scene classification. Remote Sensing, 12(24), 4046. https://doi.org/10.3390/rs12244046

[28] Li, M., Lei, L., Tang, Y., Sun, Y., & Kuang, G. (2021). An attention-guided multilayer feature aggregation network for remote sensing image scene classification. Remote Sensing, 13(16), 3113. https://doi.org/10.3390/rs13163113

[29] Shen, J., Zhang, T., Wang, Y., Wang, R., & Wang, Q. (2021). A dual-model architecture with grouping-attention-fusion for remote sensing scene classification. Remote Sensing, 13(3), 433. https://doi.org/10.3390/rs13030433

[30] Kim, J., & Chi, M. (2021). SAFFNet: Self-attention-based feature fusion network for remote sensing few-shot scene classification. Remote Sensing, 13(13), 2532. https://doi.org/10.3390/rs13132532

[31] Ji, J., Zhang, T., Jiang, L., Zhong, W., & Xiong, H. (2020). Combining multilevel features for remote sensing image scene classification with attention model. IEEE Geoscience and Remote Sensing Letters, 17(9), 1647–1651. https://doi.org/10.1109/LGRS.2019.2949253

[32] Wang, G., Xu, H., Wang, X., Yuan, L., & Wen, X. (2022). Remote sensing scene image classification model based on multi-scale features and attention mechanism. Journal of Applied Remote Sensing, 16(4), Article 044510. https://doi.org/10.1117/1.JRS.16.044510

[33] Wang, Y., Hu, Y., Xu, Y., Jiao, P., & Zhang, X. (2022). Context residual attention network for remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters, 19, 1–5. https://doi.org/10.1109/LGRS.2021.3117265

[34] Yu, D., Xu, Q., Liu, X., Lv, L., & Guo, H. (2022). Joint learning using multiscale attention-enhanced features for remote sensing image scene classification. Journal of Applied Remote Sensing, 16(3), Article 036506. https://doi.org/10.1117/1.JRS.16.036506

[35] Lv, G., Dong, L., Zhang, W., & Xu, W. (2021). Multi-scale attentive region adaptive aggregation learning for remote sensing scene classification. International Journal of Remote Sensing, 42(20), 7742–7776. https://doi.org/10.1080/01431161.2021.1963878

[36] Zhang, J., Zhang, M., Shi, L., Yan, W., & Pan, B. (2019). A multi-scale approach for remote sensing scene classification based on feature maps selection and region representation. Remote Sensing, 11(21), 2504. https://doi.org/10.3390/rs11212504

[37] Bi, Q., Zhang, H., & Qin, K. (2021). Multi-scale stacking attention pooling for remote sensing scene classification. Neurocomputing, 436, 147–161. https://doi.org/10.1016/j.neucom.2021.01.038

[38] Niu, B., Pan, Z., Wu, J., Hu, Y., & Lei, B. (2022). Multi-representation dynamic adaptation network for remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–19. https://doi.org/10.1109/TGRS.2022.3217180

[39] Zhang, S., Wu, G., Gu, J., & Han, J. (2020). Pruning convolutional neural networks with an attention mechanism for remote sensing image classification. Electronics, 9(8), 1209. https://doi.org/10.3390/electronics9081209

[40] Ghaffarian, S., Valente, J., van der Voort, M., & Tekinerdogan, B. (2021). Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review. Remote Sensing, 13(15), 2965. https://doi.org/10.3390/rs13152965