SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I8P108

Volume 13 Issue 8, 113-130, August 2026

Original Article

Transformer-Fused Hybrid Descriptors for Remote Sensing Scene Classification, Integrating Texture-Morphology Features with EfficientNetV2 Semantic Embeddings on MLRSNet

Cheruku Bujji Babu[1], Gurumurthy Hari Krishnan[2]

[1]*Department of Electronics and Communication Engineering, Mohan Babu University, Tirupati, India.*
[2]*Department of Electrical and Electronics Engineering, Mohan Babu University, Tirupati, India.*

[2]*Corresponding Author : haris_eee@yahoo.com*

Received: 23 Ferbruary 2026 Revised: 28 April 2026 Accepted: 31 July 2026 Published: 31 August 2026

Abstract - *Classification in remote sensing images is a difficult problem, given high intra-class variability, inter-class similarity, and spatial complexity. In this paper, a novel transformer-fused hybrid feature learning model is developed to effectively combine handcrafted and deep features for accurate Land Use/Land Cover (LULC) scene classification. In this model, texture features are represented using Local Binary Patterns (LBPs) and Grey-Level Co-occurrence Matrix (GLCM) features, while morphological region features provide shape information for scene characterization. Meanwhile, deep semantic features are also learned from EfficientNetV2-B0. To effectively fuse these features, a novel multi-head self-attention fusion mechanism is developed to learn explicit feature dependencies between texture, morphological, and semantic features for a compact yet discriminative feature representation. Experimental evaluation is conducted using the complete MLRSNet dataset comprising all 46 scene classes and 46,000 images, with 1,000 images considered from each class to ensure a balanced and comprehensive experimental setting. The proposed framework achieves an average five-fold accuracy of 99.98%, demonstrating high learning consistency across the complete set of diverse and visually similar remote sensing scenes. Comparative evaluation with established pretrained CNN and transformer-based models under the same experimental setting, together with component-wise ablation analysis, further demonstrates the contribution of the handcrafted descriptors, EfficientNetV2 semantic embeddings, and transformer-guided fusion mechanism. This fusion approach is effective for improving inter-class discriminability for visually similar LULC classes, which is a powerful tool for large-scale LULC mapping, urban growth analysis, environmental surveillance, etc., from remote sensing images.*

Keywords - *Environmental remote sensing, MLRSNet, Feature fusion, Transformer, Multi-head attention, Environmental monitoring.*

1. Introduction

Scene classification in remote sensing imagery has evolved into an important research area because of its applications in environmental monitoring, land-use and land-cover mapping, urban planning, precision agriculture, disaster management, military reconnaissance, and intelligent geographic information systems. High-resolution satellite and aerial images acquired by contemporary Earth-observation platforms have considerably supported the development of automated scene-understanding systems [1]. However, reliable remote sensing scene classification remains challenging because of high intra-class variability caused by seasonal changes, atmospheric effects, object-scale variations, illumination conditions, viewing angles, and imaging environments. In addition, considerable inter-class similarity exists among commercial areas, residential regions, industrial sites, transportation networks, agricultural fields, and recreational facilities [2]. Effective classification, therefore, requires representations capable of jointly describing local texture, geometrical structure, object organisation, and high-level scene semantics.

The difficulty becomes more pronounced when a model is evaluated using a large number of scene categories rather than a small, selected subset. The complete MLRSNet experimental setting considered in this study contains 46 scene classes and 46,000 images, with 1,000 images retained from each class. It includes closely related urban, transportation, agricultural, natural, industrial, and recreational categories. Such a complete-category setting

provides a more demanding assessment of feature discrimination and model consistency than reduced-class experiments.

1.1. Remote Sensing Scene Classification

Earlier remote sensing scene-classification methods were primarily based on handcrafted feature-extraction techniques combined with conventional machine-learning classifiers. Handcrafted descriptors such as Scale-Invariant Feature Transform, Histogram of Oriented Gradients, Local Binary Pattern, and Grey-Level Co-occurrence Matrix have been widely employed to extract low-level spatial and textural information from remote sensing images [3]. These features were commonly provided to Support Vector Machine, Random Forest, and K-Nearest Neighbour classifiers [4]. Although such approaches produced satisfactory results for relatively small and less complex datasets, their ability to differentiate scenes under challenging environmental conditions was limited. Handcrafted descriptors effectively represent local intensity transitions, edges, repetitive textures, and geometrical patterns, but they cannot independently model the high-level semantic relationships among objects and their surrounding contexts. Benchmark datasets such as the UC Merced Land Use Dataset [5], Aerial Image Dataset, NWPU-RESISC45, and MLRSNet have played an important role in advancing remote sensing scene-classification research [6]–[8]. These datasets contain different scene categories, object scales, spatial arrangements, and environmental conditions, encouraging the development of classification frameworks with improved discrimination and broader applicability.

1.2. CNN-based Remote Sensing Scene Classification

The introduction of deep convolutional neural networks substantially improved remote sensing scene classification through hierarchical feature learning from raw image intensities. CNN architectures such as AlexNet, VGGNet, GoogLeNet, ResNet, DenseNet, and EfficientNet have demonstrated strong representation-learning capability in complex visual-recognition problems [9]–[12]. Several studies adapted these architectures for remote sensing applications and reported improvements over conventional handcrafted approaches [13]–[15]. CNNs are effective in recognising local spatial configurations, object layouts, and scene semantics.

Transfer learning using networks pretrained on ImageNet has also improved classification performance for remote sensing datasets containing limited labelled samples [16]. Nevertheless, conventional CNN representations may emphasise high-level semantic activations while giving less explicit attention to texture regularity, region geometry, and handcrafted structural evidence. The limited modelling of long-range contextual relationships can also affect the discrimination of visually similar scenes in which distant objects and their spatial organisation are important [17].

Multi-scale feature learning, dual-backbone networks, and ensemble models have been investigated to improve CNN-based remote sensing classification. Guo and Jiang [18], for example, presented a dual-CNN model with multi-level feature fusion for land-use and land-cover classification, while other studies explored ensemble-based CNN strategies [19]. Although these methods enhanced representation capacity, many CNN frameworks continued to depend predominantly on learned semantic features without explicitly integrating interpretable texture and morphological descriptors.

1.3. Transformer-based Remote Sensing Scene Classification

Transformer architectures have gained considerable attention in remote sensing image analysis because self-attention can model relationships among distant visual regions. Vision Transformer introduced image-token-based visual representation learning by dividing an image into patches and modelling global contextual relationships [20]. Swin Transformer subsequently improved computational efficiency through hierarchical window-based attention [21]. Transformer-based approaches have also been applied to remote sensing scene classification. RS-CLIP used vision–language contrastive learning and semantic alignment for zero-shot remote sensing recognition [22]. Adaptive self-supervised transformer models optimised optical remote sensing representations through contextual feature adaptation [23], while lightweight dual-branch Swin Transformer architectures were developed for large-scale scene classification [24]. Cross-scale transformer models have integrated multi-scale information to address variations in object size within aerial imagery [25]. Frequency-aware transformer networks have retained high-frequency information required to represent scene boundaries [26]. Ensemble transformer frameworks and graph-based transformer models have also been investigated to improve contextual learning and represent dependencies among different scene regions [27].

Despite these developments, full image-token transformer architectures generally perform attention over numerous patches or spatial windows, which increases memory demand and computational load. Moreover, many transformer-based scene-classification frameworks operate primarily on learned visual tokens and do not explicitly incorporate complementary handcrafted texture and morphological information. This creates an opportunity for resource-efficient attention mechanisms that operate on compact descriptors rather than dense image-token sequences.

1.4. Hybrid Feature Fusion Approaches for Remote Sensing

Hybrid feature-learning approaches combine manually designed descriptors with representations learned through deep neural networks [28]. Local Binary Pattern and Grey-Level Co-occurrence Matrix features provide information regarding repetitive textures, surface discontinuities, local

intensity relationships, contrast, homogeneity, energy, and correlation. Morphological region descriptors provide complementary geometrical information, including area, perimeter, and eccentricity. Such properties are relevant for differentiating airports, bridges, agricultural fields, sports surfaces, residential areas, transportation structures, and industrial regions. Multimodal transformer approaches have demonstrated the ability to integrate optical, synthetic-aperture radar, and semantic information [29]. Explainable multi-view representation-learning methods have similarly been investigated to improve robustness and interpretability in remote sensing classification [30] [31].

Nevertheless, a substantial number of hybrid methods combine handcrafted and deep representations through direct concatenation. In such an operation, the features are placed within a single vector without explicitly modelling how semantic information relates to texture and geometry [32] [33]. The high dimensionality of CNN descriptors may also dominate compact handcrafted vectors, reducing the influence of complementary structural evidence. Consequently, the contribution of learned interaction between compact semantic and handcrafted descriptor domains requires further investigation [34].

1.5. Research Gap and Motivation

Although CNNs, transformers, and hybrid models have considerably advanced remote sensing scene classification, several interconnected research gaps remain. First, CNN-based methods predominantly learn high-level semantic representations, whereas explicit texture and geometrical characteristics are often underrepresented. These structural properties remain important for distinguishing classes that share similar objects but differ in surface regularity, region boundaries, object elongation, or spatial organisation. Second, full Vision Transformer and Swin Transformer architectures model relationships among numerous image patches or windows. While such models provide powerful contextual learning, their memory and processing requirements may be unnecessarily high when compact pretrained CNN descriptors are already available. A descriptor-level attention mechanism can retain relational learning while avoiding dense image-token processing.

Third, direct concatenation does not explicitly model the conditional relationship between heterogeneous descriptors. It treats semantic, texture, and morphology features as a single static vector, despite their different dimensionalities, distributions, and representational roles. Fourth, reduced-category evaluation provides limited evidence regarding the behaviour of a method across the complete diversity of a large-scale benchmark. A model that performs well on a selected subset may encounter additional difficulty when visually similar residential, industrial, transportation, agricultural, natural, and recreational classes are considered simultaneously. Finally, the scientific contribution of a fusion model cannot be established by comparison with externally reported results obtained using different classes, image subsets, or experimental partitions. Same-setting comparisons are required to isolate the effects of the semantic backbone, handcrafted features, and fusion mechanism. Accordingly, the objective of this study is to develop a resource-efficient descriptor-level fusion framework that integrates compact CNN semantic information with explicitly derived texture and morphological evidence for complete 46-class MLRSNet scene classification. The study investigates the hypothesis that handcrafted structural descriptors provide complementary information beyond a semantic-only representation and that learned cross-domain interaction can improve upon conventional direct concatenation.

1.6. Contributions of the Proposed Work

This study introduces a resource-efficient transformer-fused hybrid descriptor framework for remote sensing scene classification. The framework combines a compact semantic representation generated using an ImageNet-pretrained EfficientNetV2B0 backbone with Local Binary Pattern, Grey-Level Co-occurrence Matrix, and morphological region descriptors. The first contribution is the use of the complete 46-class MLRSNet experimental setting. A balanced collection of 46,000 images is considered, with 1,000 images retained from every category. This setting addresses the limited generalisability of reduced-class evaluation and provides a more comprehensive assessment across natural, urban, agricultural, industrial, transportation, and recreational scenes. The second contribution is the task-specific adaptation of EfficientNetV2B0 as a resource-efficient semantic descriptor generator. The original ImageNet classification head is removed, the pretrained convolutional backbone is retained as a fixed encoder, and global average pooling converts the final spatial feature maps into a compact 1,280-dimensional representation. The contribution does not involve modification of the internal Fused-MBConv or MBConv blocks; rather, it lies in adapting the pretrained network for compact heterogeneous descriptor fusion.

The third contribution is the construction of an 18-dimensional handcrafted representation containing 10 Local Binary Pattern features, five Grey-Level Co-occurrence Matrix properties, and three morphological descriptors. These features explicitly represent local texture, statistical intensity relationships, region area, boundary length, and geometrical elongation.

The fourth contribution is a four-head cross-domain attention mechanism that treats the semantic descriptor as the query and the texture–morphology descriptor as the key and value. Unlike direct concatenation, the proposed mechanism models the relationship between the semantic and structural feature domains before multiclass prediction. Attention is applied to compact descriptors rather than dense image patches, supporting the resource-efficient scope of the study.

The fifth contribution is a same-setting comparative design involving an EfficientNetV2B0 semantic-only baseline, a ResNet50 semantic-only baseline, and an EfficientNetV2B0 direct-concatenation baseline. The same images, partitions, input resolution, optimisation conditions, and performance measures are used for all models. This design enables the effects of backbone selection, handcrafted-feature inclusion, and attention-based fusion to be examined independently.

The complete framework is intended to support remote sensing applications such as land-use and land-cover mapping, urban-growth analysis, environmental surveillance, precision agriculture, transportation monitoring, and geographic scene understanding.

2. Materials and Methods

The proposed remote sensing scene classification framework combines handcrafted texture descriptors, morphological structural descriptors, and deep semantic embeddings within an attention-based fusion architecture for accurate land-use and land-cover scene classification.

The framework was developed by considering the limitations of approaches that rely only on deep semantic representations or combine handcrafted and learned features through direct concatenation without explicitly modelling the relationships between the different feature domains.

By jointly representing local texture information, geometrical structure, and high-level semantic context, the proposed method generates a compact and discriminative hybrid representation suitable for differentiating visually similar remote sensing scenes. To address the limitations of the earlier reduced-category evaluation, the present study uses the complete 46-class MLRSNet experimental dataset.

The methodology consists of dataset preparation, resource-efficient image handling, handcrafted descriptor extraction, adapted EfficientNetV2B0 semantic descriptor generation, heterogeneous feature projection, transformer-guided cross-domain attention, and multiclass scene prediction.

Three same-setting baseline models are also considered: an EfficientNetV2B0 semantic-only baseline, a ResNet50 semantic-only baseline, and an EfficientNetV2B0-based direct-concatenation baseline. These models separately examine the influence of the semantic backbone, handcrafted descriptors, and the proposed attention-based fusion mechanism.

All baseline models and the proposed model use the same image samples, data partitions, cross-validation folds, number of epochs, batch size, optimiser, and performance measures.

2.1. MLRSNet Dataset Description

The proposed framework was evaluated using MLRSNet, a large-scale remote sensing dataset developed for semantic scene understanding and land-use classification. The dataset contains high-resolution remote sensing images acquired under different geographical environments, spatial arrangements, illumination conditions, object scales, resolutions, and scene structures. Its wide range of land-use and land-cover categories makes MLRSNet an appropriate benchmark for assessing models intended to differentiate complex and visually similar remote sensing scenes. Unlike the previous experimental setting based on 15 selected categories, the present study considers all 46 MLRSNet scene classes. A balanced collection of 1,000 readable images was retained from each category, resulting in 46,000 images. The use of an equal number of images from every class prevents majority-class dominance and supports an unbiased class-wise comparison.

The 46 scene categories are airplane, airport, bareland, baseball diamond, basketball court, beach, bridge, chaparral, cloud, commercial area, dense residential area, desert, eroded farmland, farmland, forest, freeway, golf course, ground track field, harbour, industrial area, intersection, island, lake, meadow, mobile home park, mountain, overpass, park, parking lot, railway, railway station, rectangular farmland, river, roundabout, runway, sea ice, ship, snowberg, sparse residential area, stadium, storage tank, tennis court, terrace, thermal power station, wetland, and wind turbine. The complete dataset presents several challenges associated with practical scene recognition, including high intra-class variation, inter-class similarity, scale variation, background interference, object-density changes, and spatial-layout complexity.

For example, commercial area, industrial area, dense residential area, and sparse residential area may contain overlapping building, road, and vegetation patterns. Similarly, bareland, desert, meadow, farmland, rectangular farmland, and eroded farmland require fine-grained discrimination based on texture, region boundaries, and structural organisation. Evaluation across all 46 classes, therefore, provides stronger evidence of the applicability and generalisability of the proposed model than evaluation using a selected subset of categories. A class-wise partitioning strategy was employed using a fixed random seed of 42. For every class, 70% of the images were assigned to the principal learning partition, 15% to the model-monitoring partition, and the remaining 15% to the independent evaluation partition. This produced 32,200, 6,900, and 6,900 images, respectively. The same images, partition assignments, and file order were retained for the proposed method and all baseline models.

The composition of the experimental dataset is summarised in Table 1.

Table 1. Composition of the MLRSNet Dataset Used in the Study

Description	Value
Number of scene classes	46
Images retained per class	1,000
Total number of images	46,000
Principal learning partition	32,200
Model-monitoring partition	6,900
Independent evaluation partition	6,900
Input image resolution	(224\times224\times3)
Cross-validation folds	5
Random seed	42

2.2. Image Preprocessing and Resource-Efficient Data Handling

The preprocessing stage was designed to standardise the image dimensions while avoiding the excessive memory consumption associated with loading the complete collection of 46,000 decoded images into random-access memory. Initially, all candidate image files were checked for readability. Corrupted, unsupported, or unreadable files were excluded before partitioning. Each image was decoded only when required and resized to (224\times224\times3) pixels. The image was then transformed using the preprocessing function associated with the corresponding ImageNet-pretrained backbone. Therefore, a separate manual pixel scaling operation was not applied to the EfficientNetV2B0 and ResNet50 inputs. For EfficientNetV2B0 semantic descriptor extraction, a streaming tf.data pipeline was employed to read, resize, preprocess, batch, and prefetch the images. Only one batch of decoded images was retained in memory at any time. A batch size of 64 was used during semantic descriptor extraction. Handcrafted descriptors were extracted image-by-image using parallel CPU threads. The resulting semantic and handcrafted descriptors were stored as NumPy cache files. Once these cache files were generated, the five-fold experiments, baseline models, ablation configurations, and final classifier development were performed directly using the compact descriptor arrays without repeatedly decoding the complete image collection.

The EfficientNetV2B0 semantic cache contained a 1,280-dimensional descriptor for every image, whereas the handcrafted cache contained 18 descriptors per image. ResNet50 semantic descriptors were stored separately because their dimensionality and numerical distribution differed from those of EfficientNetV2B0. No image-level augmentation was used in the final experimental implementation. Therefore, random rotation, zooming, horizontal flipping, and related augmentation operations were not included in the reported methodology. The complete processing sequence, from image verification and preprocessing to feature extraction, latent-space projection, cross-domain attention, and scene prediction, is illustrated in Figure 1. The semantic and handcrafted feature branches remain separate until they are projected into a shared latent space.

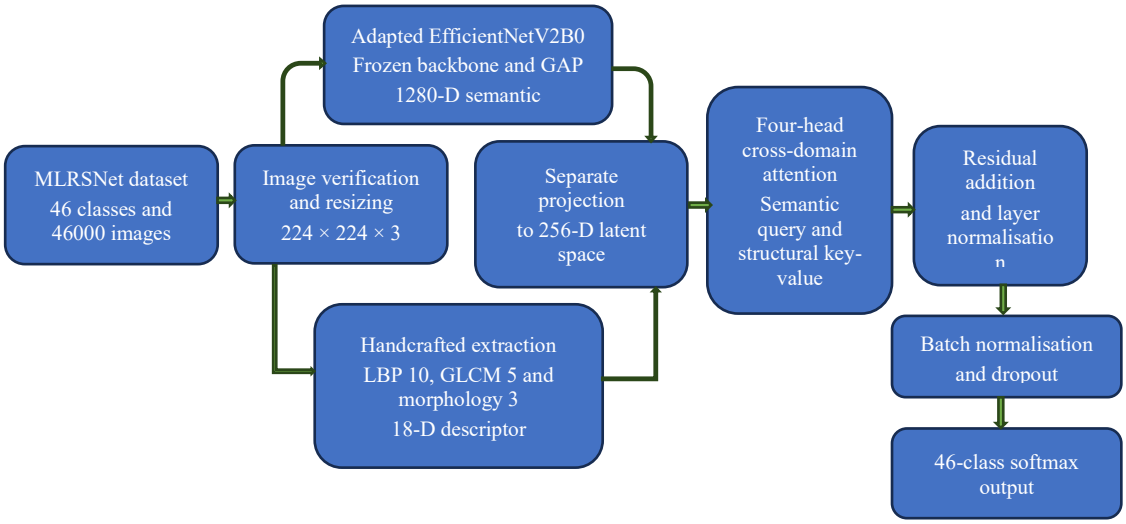

Fig. 1 Overall architecture of the proposed resource-efficient transformer-fused hybrid descriptor framework for complete 46-class MLRSNet scene classification

2.3. Handcrafted Texture Feature Extraction

Texture information is important for differentiating remote sensing scenes that contain similar objects but differ in surface patterns, local intensity arrangements, repetition frequency, or spatial regularity. Local Binary Pattern and Grey-Level Co-occurrence Matrix descriptors were therefore employed to represent local and statistical texture characteristics.

2.3.1. Local Binary Pattern Descriptor

Each resized image was converted to grayscale before LBP extraction. For (P) neighbouring samples located at

radius (R), the Local Binary Pattern value at the centre coordinate ((x_c,y_c)) is defined as follows.

$$\mathrm{LBP}_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s\left(g_p - g_c\right)2^p$$

where (g_c) is the intensity of the centre pixel and $\left(g_p\right)$ is the intensity of the (p)-th neighbouring pixel. The binary comparison function is expressed as follows.

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

The implementation uses (P=8), (R=1), and the uniform LBP formulation. The resulting codes are represented using a normalised 10-bin histogram.

$$h_k = \frac{n_k}{\sum_{j=1}^{10} n_j + \varepsilon}, \quad k = 1,2,\ldots,10$$

where (n_k) is the number of pixels assigned to the $(k) - th$ histogram bin and (ε) is a small positive constant that prevents division by zero. The resulting histogram forms a 10-dimensional local texture descriptor.

2.3.2. Grey-Level Co-occurrence Matrix Descriptor

The Grey-Level Co-occurrence Matrix represents the frequency with which pairs of grey levels occur at a specified spatial displacement. In the present implementation, the matrix is calculated using a pixel distance of one and an orientation of $(0°)$. The matrix is configured as symmetric and normalised.

Five statistical properties are extracted: contrast, dissimilarity, homogeneity, energy, and correlation. Contrast represents the local intensity variation and is defined as follows.

$$\text{Contrast} = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1}(i-j)^2\, P(i,j)$$

where $\left(P(i,j)\right)$ is the normalised probability of the occurrence of grey levels (i) and (j), and (L) is the number of grey levels.

Energy represents textural uniformity and is calculated as follows.

$$\text{Energy} = \sqrt{\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} P\,(i,j)^2}$$

Dissimilarity, homogeneity, and correlation are also obtained from the same normalised co-occurrence matrix. Together, the five GLCM properties provide complementary information concerning local variation, similarity, regularity, uniformity, and linear dependence among neighbouring grey levels.

The 10 LBP histogram values and five GLCM properties are combined to obtain a 15-dimensional texture descriptor.

$$\mathrm{f}_{tex} = \left[\mathrm{h}_{LBP}; f_{con}; f_{dis}; f_{hom}; f_{eng}; f_{cor}\right] \in \mathbb{R}^{15}$$

2.4. Morphological Feature Extraction

Morphological region descriptors were used to represent geometrical and structural information in the remote sensing images. Each resized image was converted to grayscale and segmented using Otsu's automatic thresholding method. Connected-component labelling was subsequently applied to the resulting binary image.

For every detected region, area, eccentricity, and perimeter were calculated because an image may contain multiple connected regions; the mean value of each measurement was used to form an image-level structural descriptor.

$$\mathbf{f}_{morph} = \left[\frac{1}{N_r}\sum_{r=1}^{N_r} A_r\,, \frac{1}{N_r}\sum_{r=1}^{N_r} e_r\,, \frac{1}{N_r}\sum_{r=1}^{N_r} p_r\right]$$

where (N_r) is the number of connected regions, (A_r) is the area, (e_r) is the eccentricity, and (p_r) is the perimeter of the (r)-th region.

Eccentricity describes the deviation of a region from a circular shape and is obtained from the major and minor axes of its equivalent ellipse.

$$e_r = \sqrt{1 - \frac{b_r^2}{a_r^2}}$$

where (a_r) and (b_r) denote the semi-major and semi-minor axes, respectively.

When no connected region is detected, a three-dimensional zero vector is assigned. The complete handcrafted descriptor is formed by combining the 15 texture features and three morphological features.

$$\mathrm{f}_{HC} = \left[\mathrm{f}_{tex}; \mathrm{f}_{morph}\right] \in \mathbb{R}^{18}$$

Before model development, each handcrafted feature was standardised using the following transformation.

$$\hat{f}_j = \frac{f_j - \mu_j}{\sigma_j}$$

where $\left(\mu_j\right)$ and $\left(\sigma_j\right)$ are the mean and standard deviation of the (j)-th feature. These parameters were estimated exclusively from the principal learning partition. The same

parameters were then applied to the model-monitoring and independent evaluation partitions. This procedure prevents information leakage and ensures that handcrafted features with different numerical ranges contribute comparably to the fusion mechanism.

2.5. Adapted EfficientNetV2B0 Semantic Feature Extraction

Deep semantic features were extracted using EfficientNetV2B0 pretrained on ImageNet. EfficientNetV2B0 is a computationally efficient convolutional neural network that employs fused mobile inverted bottleneck blocks and mobile inverted bottleneck convolution blocks to learn hierarchical visual representations.

The original EfficientNetV2B0 network contains a classification head developed for the ImageNet categories. This head is not directly suitable for the proposed 46-class heterogeneous descriptor-fusion problem. The pretrained network was therefore adapted in three stages:

1. The original ImageNet classification head was removed using include_top=False;
2. The pretrained convolutional backbone was retained as a fixed semantic feature extractor; and
3. Global average pooling was applied to the final spatial feature maps.

Let the preprocessed input image be represented by (\mathbf I). The convolutional backbone generates the following feature tensor.

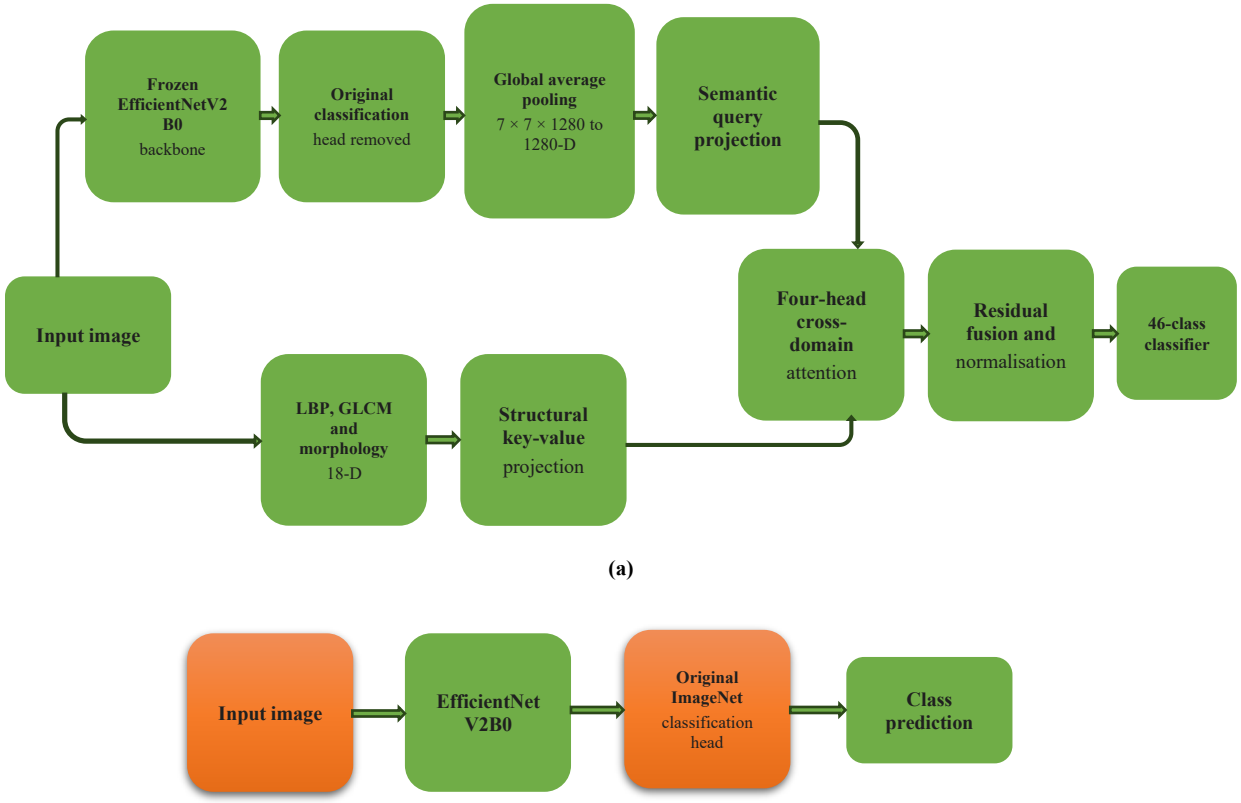

Fig. 2 Comparison between (a) Proposed architectural adaptation and (b) Conventional EfficientNetV2B0 classification.

where $(\mathcal{E}_\theta)$ denotes the ImageNet-pretrained EfficientNetV2B0 encoder. The parameters (θ) remain frozen during descriptor extraction.

Instead of flattening the complete $(7 \times 7 \times 1280)$ tensor into a 62,720-dimensional representation, global average pooling is applied across the two spatial dimensions.

$$f_{CNN,k} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{CNN}(i,j,k), \quad k = 1,\dots,1280$$

where $(H = W = 7)$. The resulting compact semantic descriptor is expressed as follows.

$$\mathbf{f}_{CNN} = \mathrm{GAP}(\mathcal{E}_\theta(\mathbf{I})) \in \mathbb{R}^{1280})$$

This adaptation reduces the semantic representation from 62,720 spatial values to a compact 1,280-dimensional descriptor. It decreases the memory requirement of the subsequent fusion stage and makes the semantic representation suitable for interaction with the 18-dimensional handcrafted descriptor.

The difference between a conventional EfficientNetV2B0 classification pipeline and the proposed architectural adaptation is illustrated in Figure 2. The proposed work does not modify the internal Fused-MBConv or MBConv blocks of EfficientNetV2B0.

Instead, the original classification head is removed, global average pooling is introduced, and the compact semantic descriptor is integrated with texture and morphology information through descriptor-level cross-domain attention.

2.6. Heterogeneous Feature Projection

The semantic and handcrafted descriptors differ considerably in dimensionality and numerical distribution. Directly combining the 1,280-dimensional semantic descriptor with the 18-dimensional handcrafted descriptor can cause the semantic representation to dominate the compact structural representation.

In the proposed model, the two feature domains are therefore retained as separate inputs and projected into a shared 256-dimensional latent space.

The semantic descriptor is first represented as a single semantic token.

$$\mathrm{X}_{CNN} = \mathrm{Reshape}(\mathrm{f}_{CNN}) \in \mathbb{R}^{1\times 1280}$$

The standardised handcrafted descriptor is represented as a separate structural token.

$$\mathrm{X}_{HC} = \mathrm{Reshape}\left(\hat{\mathrm{f}}_{HC}\right) \in \mathbb{R}^{1\times 18}$$

The semantic descriptor is projected as the query.

$$\mathrm{Q} = \mathrm{X}_{CNN}\mathrm{W}_{Q} + \mathrm{b}_{Q}$$

The handcrafted descriptor is independently projected as the key and value.

$$\mathrm{K} = \mathrm{X}_{HC}\mathrm{W}_{K} + \mathrm{b}_{K}, \quad \mathrm{V} = \mathrm{X}_{HC}\mathrm{W}_{V} + \mathrm{b}_{V}$$

The projected query, key, and value have the common latent dimension.

$$\mathrm{Q}, \mathrm{K}, \mathrm{V} \in \mathbb{R}^{1\times 256}$$

This projection aligns the two heterogeneous feature domains while preserving their individual identities.

2.7. Transformer-Guided Cross-Domain Feature Interaction

A transformer-guided multi-head attention mechanism is employed to model the relationship between the semantic scene descriptor and the handcrafted structural descriptor. In the proposed architecture, the semantic representation forms the query, whereas the combined texture–morphology representation forms the key and value. The operation is therefore described as cross-domain attention rather than conventional self-attention.

For the $(h) - th$ attention head, the attention output is calculated as follows.

$$\mathrm{head}_h = \mathrm{softmax}\left(\frac{\mathrm{Q}_h\mathrm{K}_h^T}{\sqrt{d_k}}\right)\mathrm{V}_h$$

where $(d_k = 64)$ is the key dimension of each attention head.

Four attention heads are employed in the proposed model.

$$\mathrm{MHA}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \mathrm{head}_3, \mathrm{head}_4)\,\mathrm{W}_O)$$

The multi-head attention output is combined with the projected semantic query using a residual connection.

$$\mathrm{Z} = \mathrm{LayerNorm}[\mathrm{Q} + \mathrm{MHA}(\mathrm{Q}, \mathrm{K}, \mathrm{V})]$$

The residual pathway preserves the original semantic information, whereas the attention pathway introduces texture and morphological evidence relevant to the semantic context. Layer normalisation stabilises the fused representation.

Unlike direct concatenation, the proposed mechanism learns a conditional interaction between the semantic and handcrafted feature domains. The semantic representation determines the information requested from the handcrafted branch, while the handcrafted branch supplies complementary texture and geometrical evidence.

2.8. Regularisation and Classification Layer

The output of the cross-domain attention layer is passed through batch normalisation and dropout.

$$\mathbf{Z}_r = \mathrm{Dropout}[\mathrm{BN}(\mathbf{Z}), 0.3]$$

The resulting representation is flattened and connected to a dense softmax layer containing 46 output neurons. The predicted probability for class (c) is calculated as follows.

$$p_c = \frac{\exp(z_c)}{\sum_{j=1}^{46}\exp(z_j)}, \quad c = 1,2,\ldots,46$$

where (z_c) is the unnormalised output activation for class (c). The predicted scene category is determined as follows.

$$\hat{y} = \arg\max_c p_c$$

The model is optimised using categorical cross-entropy.

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{46} y_{n,c}\log(p_{n,c})$$

where $\left(y_{n,c}\right)$ is the ground-truth indicator for image (n) and class (c).

2.9. Five-Fold Stratified Cross-Validation

Five-fold stratified cross-validation was conducted using the principal learning partition. Stratification preserved approximately equal class proportions in every fold, which is important for a classification problem involving 46 categories.

For each fold, four subsets were used for model development, and the remaining subset was used for fold-wise assessment. A new model was constructed and freshly initialised at the beginning of every fold. Consequently, classifier or fusion weights learned in one fold were not transferred to the next fold. This prevents information transfer among folds and ensures that fold consistency reflects independent model development.

The pretrained convolutional backbones remained fixed because the semantic descriptors were extracted and cached before cross-validation. However, the classifier layers, projection layers, direct-concatenation layers, cross-attention layers, and normalisation layers were independently learned within their respective models and folds. After completing the five-fold procedure, a final model instance was developed using the complete principal learning partition. The model-monitoring partition was used for convergence monitoring and learning-rate adjustment. The independent evaluation partition was not used for model development, hyperparameter selection, or learning-rate control.

2.10. Same-Setting Baseline Models

To address the editorial requirement for fair comparison with established models, three baseline configurations were implemented in addition to the complete proposed model.

The EfficientNetV2B0 semantic-only baseline isolates the contribution of the selected semantic backbone. The ResNet50 semantic-only baseline provides a comparison with an established residual CNN. The EfficientNetV2B0 direct-concatenation baseline determines whether conventional aggregation of semantic and handcrafted descriptors is sufficient, or whether the proposed attention mechanism provides an additional methodological contribution.
The common experimental protocol used for the baseline and proposed configurations is illustrated in Figure 3. All configurations use the same 46 MLRSNet classes, the same 46,000 images, identical class-wise partitions, the same five-fold assignments, the same number of epochs, and the same evaluation measures. Consequently, the comparison is not influenced by differences in dataset size, selected classes, or partitioning strategy.

2.10.1. EfficientNetV2B0 Semantic-Only Baseline

The EfficientNetV2B0 semantic-only baseline evaluates the classification capability of the 1,280-dimensional semantic descriptor without handcrafted information or attention-based fusion.

The baseline receives only the global average-pooled EfficientNetV2B0 descriptor. The descriptor is passed through batch normalisation and dropout, followed by a dense softmax layer containing 46 outputs.

The processing sequence is 1,280-dimensional semantic descriptor→batch normalisation→dro pout→46-class softmax.

This configuration contains no LBP, GLCM, morphological descriptor, feature concatenation, or attention mechanism. It determines whether the complete proposed method provides an advantage beyond the EfficientNetV2B0 semantic descriptor alone.

2.10.2. ResNet50 Semantic-Only Baseline

ResNet50 was selected as an established residual CNN baseline to determine whether the behaviour of the semantic-only classifier depends specifically on the EfficientNetV2B0 backbone.

The original ImageNet classification head of ResNet50 was removed, and the pretrained convolutional backbone was retained as a fixed feature extractor. Global average pooling was applied to the final spatial feature maps to obtain a 2,048-dimensional semantic descriptor.

$$\mathbf{f}_R = \mathrm{GAP}\big(\mathcal{R}_\phi(\mathbf{I})\big) \in \mathbb{R}^{2048}\big)$$

where $\big(\mathcal{R}_\phi\big)$ denotes the frozen ImageNet-pretrained ResNet50 encoder.

The descriptor is passed through batch normalisation and dropout, followed by a dense 46-class softmax layer. No handcrafted descriptor or fusion mechanism is included in this baseline.

2.10.3. EfficientNetV2B0 with Direct Concatenation

The direct-concatenation baseline determines whether combining semantic and handcrafted information through a conventional fusion operation is sufficient, or whether the proposed cross-domain attention mechanism contributes additional feature-interaction capability.

The 1,280-dimensional EfficientNetV2B0 semantic descriptor and the standardised 18-dimensional handcrafted descriptor are normalised separately and concatenated.

$$\mathrm{f}_{DC} = \big[\mathrm{BN}(\mathrm{f}_{CNN})\,;\mathrm{BN}\big(\hat{\mathrm{f}}_{HC}\big)\big] \in \mathbb{R}^{1298}$$

The concatenated descriptor is projected into a 256-dimensional representation using a dense layer with rectified linear activation.

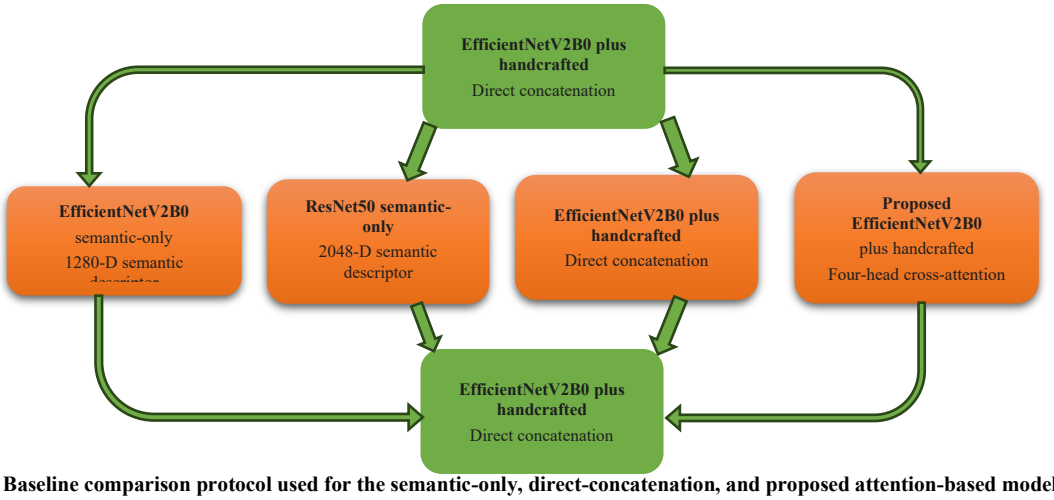

Fig. 3 Baseline comparison protocol used for the semantic-only, direct-concatenation, and proposed attention-based models

$$z_{DC} = \mathrm{ReLU}(W_{DC} f_{DC} + b_{DC})$$

Batch normalisation and dropout with a rate of 0.30 are then applied before the 46-class softmax layer. This baseline uses the same EfficientNetV2B0 and handcrafted descriptors as the proposed model but excludes query, key, and value projection, multi-head attention, residual attention fusion, and layer normalisation. It therefore provides the most direct assessment of the methodological contribution of the proposed cross-domain attention mechanism.

2.10.4. Proposed Cross-Attention Hybrid Model

The complete proposed configuration uses the 1,280-dimensional EfficientNetV2B0 semantic descriptor and the 18-dimensional handcrafted texture–morphology descriptor. Unlike the direct-concatenation baseline, the two feature domains are not combined immediately. They are independently projected into a 256-dimensional latent space. The semantic descriptor forms the attention query, while the handcrafted descriptor forms the key and value. Four-head cross-domain attention is then used to model their relationship.

The methodological differences among the four principal configurations are summarised in Table 2.

Table 2. Comparison of Feature-Extraction and Fusion Strategies

Model	Semantic backbone	Semantic dimension	Handcrafted descriptor	Fusion mechanism
EfficientNetV2B0 semantic-only	EfficientNetV2B0	1,280	Not used	None
ResNet50 semantic-only	ResNet50	2,048	Not used	None
Direct-concatenation baseline	EfficientNetV2B0	1,280	LBP, GLCM and morphology	Concatenation and Dense(256)
Proposed hybrid model	EfficientNetV2B0	1,280	LBP, GLCM and morphology	Four-head cross-domain attention

2.11. Protocol Comparison

All main models were developed using an identical experimental protocol. The following conditions were kept constant wherever applicable:

- complete 46-class MLRSNet dataset;
- 1,000 images per class;
- same class-wise image partitions;
- same random seed;
- same (224\times224\times3) input resolution;
- same five-fold stratified division;
- freshly initialised classifier or fusion head for each fold;
- batch size of 32 for classifier development;
- maximum of 30 epochs;
- Adam optimiser;
- categorical cross-entropy loss;
- identical learning-rate reduction strategy; and
- identical performance measures.

The initial Adam learning rate was (1×10^{-3}). When the monitored loss did not improve for three consecutive epochs, the learning rate was multiplied by 0.5, subject to a minimum value of (1×10^{-6}). Early termination was disabled so that

all baseline and proposed models completed the same maximum number of epochs.

The convolutional backbones remained frozen during semantic descriptor extraction. Therefore, the comparison primarily investigates the discriminative capability of the extracted semantic descriptors and the contribution of the different feature-fusion strategies without introducing variations caused by full backbone fine-tuning.

2.12. Resource-Efficient Baseline Selection

The objective of the study is to develop a resource-efficient hybrid scene classification framework suitable for execution without the high memory and processing requirements associated with full image-token transformer architectures.

ResNet50 was included as an established alternative CNN backbone. The EfficientNetV2B0 semantic-only baseline isolates the contribution of the selected semantic descriptor. The direct-concatenation baseline isolates the contribution of heterogeneous feature aggregation, while the proposed model evaluates the additional contribution of descriptor-level attention. Full Vision Transformer and Swin Transformer models were not included as primary same-setting baselines because these architectures perform attention over numerous image patches or spatial windows and therefore operate within a different computational design space. The proposed framework does not perform transformer processing over dense image patches. Instead, attention is applied only after the complete image has been compressed into a 1,280-dimensional semantic descriptor and an 18-dimensional handcrafted descriptor.

This descriptor-level attention strategy retains relational feature learning while reducing the memory and CPU load associated with dense image-token processing. The exclusion of full image-token transformers is therefore consistent with the resource-efficient objective of the study.

2.13. Component-wise Ablation Protocol

A component-wise ablation procedure was defined to assess the individual contribution of the handcrafted descriptor groups and the proposed attention mechanism. All configurations retain the same EfficientNetV2B0 semantic descriptor, image partitions, cross-validation folds, and optimisation conditions.

The ablation configurations are presented in Table 3.

Table 3. Component-Wise Ablation Configurations

Configuration	EfficientNetV2B0	LBP	GLCM	Morphology	Fusion mechanism
A1	Yes	No	No	No	Semantic-only classifier
A2	Yes	Yes	No	No	Direct concatenation
A3	Yes	No	Yes	No	Direct concatenation
A4	Yes	Yes	Yes	No	Direct concatenation
A5	Yes	Yes	Yes	Yes	Direct concatenation
A6	Yes	Yes	Yes	Yes	Proposed cross-attention

Configuration A1 establishes the semantic-only reference. A2 and A3 independently assess the contributions of LBP and GLCM. A4 assesses their combined texture contribution. A5 determines the additional contribution of morphological information under conventional concatenation. A6 compares direct concatenation with the complete proposed attention-based fusion mechanism.

2.14. Experimental Configuration

The framework was implemented in Python using TensorFlow/Keras, OpenCV, scikit-image, scikit-learn, NumPy, Pandas, and Joblib. The principal experimental settings are summarised in Table 4.

Table 4. Experimental Configuration

Parameter	Setting
Dataset	MLRSNet
Scene classes	46
Images per class	1,000
Total images	46,000
Image resolution	(224\times224\times3)
Proposed semantic encoder	EfficientNetV2B0
Alternative semantic encoder	ResNet50
Pretrained weights	ImageNet
Backbone condition	Frozen
EfficientNetV2B0 descriptor dimension	1,280
ResNet50 descriptor dimension	2,048
LBP configuration	(P=8,\ R=1), uniform
LBP dimension	10
GLCM configuration	Distance 1 and angle (0^\circ)
GLCM dimension	5

Morphological dimension	3
Complete handcrafted dimension	18
Direct-concatenation projection	256
Proposed latent projection	256
Attention heads	4
Key dimension per head	64
Dropout rate	0.3
Output classes	46
Optimiser	Adam
Initial learning rate	(1\times10^{-3})
Minimum learning rate	(1\times10^{-6})
Loss function	Categorical cross-entropy
Semantic extraction batch size	64
Classifier-development batch size	32
Epochs	30
Cross-validation	Five-fold stratified
Random seed	42
Early termination	Disabled
Feature caching	Enabled

2.15. Performance Measures

The same performance measures are applied to all baseline configurations, ablation configurations, and the proposed model. These include accuracy, class-wise precision, recall, F1-score, confusion matrix, receiver operating characteristic analysis, and area under the ROC curve.

For the multiclass problem, accuracy is defined as follows.

$$\text{Accuracy} = \frac{\sum_{c=1}^{C} T\,P_c}{N}$$

where (TP_c) is the number of correctly identified samples belonging to class (c), (C=46), and (N) is the total number of evaluated samples.

For class (c), precision is calculated as follows.

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}$$

Recall is calculated as follows.

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$$

F1-score is calculated as follows.

$$F1_c = 2\frac{\text{Precision}_c\,\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

Macro-averaged measures assign equal importance to all 46 classes.

$$\text{Macro-}M = \frac{1}{C}\sum_{c=1}^{C} M_c$$

where (M_c) represents the corresponding class-wise precision, recall, or F1-score.

One-versus-rest ROC analysis is used to evaluate the discriminative capability of the model for each scene class. Fold-wise mean and standard deviation are used to represent consistency across the five stratified data divisions.

3. Results and Discussion

The proposed transformer-fused hybrid descriptor framework and the three same-setting baseline models were evaluated using the complete balanced MLRSNet experimental dataset containing 46 scene classes and 46,000 images. The comparison included the EfficientNetV2B0 semantic-only model, the ResNet50 semantic-only model, EfficientNetV2B0 with direct concatenation, and the proposed EfficientNetV2B0-based cross-attention framework. All four configurations used the same image collection, class-wise partitions, input resolution, number of epochs, batch size, optimiser, and classifier-development protocol. Therefore, the comparison primarily reflects differences in semantic backbone selection, inclusion of handcrafted descriptors, and feature-fusion strategy. Throughout this section, Accuracy denotes the value measured on the principal learning partition. The confusion matrix, class-wise interpretation, and ROC analysis are presented separately to provide complementary evidence regarding the discriminative behaviour of the proposed model.

3.1. Same-Setting Accuracy Comparison

Table 5 presents the Accuracy obtained by the three baseline models and the proposed cross-attention framework. The EfficientNetV2B0 semantic-only baseline achieved an Accuracy of 98.35%, establishing the performance obtained using only the 1,280-dimensional semantic descriptor. Replacing EfficientNetV2B0 with the 2,048-dimensional ResNet50 semantic descriptor increased Accuracy to 99.42%.

This result demonstrates that the choice of semantic backbone has a measurable influence on complete 46-class scene classification. The addition of the 18-dimensional LBP–GLCM–morphology descriptor through direct concatenation increased the accuracy to 99.84%. The proposed cross-attention framework achieved the highest Accuracy of 99.96%.

Table 5. Accuracy comparison of the baseline and proposed models

Model	Semantic descriptor	Handcrafted descriptor	Fusion mechanism	Accuracy (%)
EfficientNetV2B0 semantic-only	1,280	Not used	None	98.35
ResNet50 semantic-only	2,048	Not used	None	99.42
EfficientNetV2B0 with direct concatenation	1,280	LBP, GLCM and morphology	Concatenation and Dense(256)	99.84
Proposed cross-attention framework	1,280	LBP, GLCM and morphology	Four-head cross-domain attention	**99.96**

The proposed framework improved Accuracy by 1.61 percentage points compared with EfficientNetV2B0 semantic-only and by 0.54 percentage points compared with ResNet50 semantic-only, as indicated in Figure 4. These improvements demonstrate that the high Accuracy is not obtained from the EfficientNetV2B0 semantic descriptor alone. Direct concatenation improved Accuracy by 1.49 percentage points over EfficientNetV2B0 semantic-only. This confirms that the handcrafted descriptor contributes complementary texture and structural information. The proposed cross-attention framework produced a further improvement of 0.12 percentage points over direct concatenation. Although the improvement over concatenation is relatively small, it is consistently attributable to the change in the fusion mechanism because both configurations use the same semantic descriptor, handcrafted descriptor, image samples, and optimisation settings. The result, therefore, provides direct evidence that modelling the interaction between semantic and structural representations can offer an advantage over their simple aggregation.

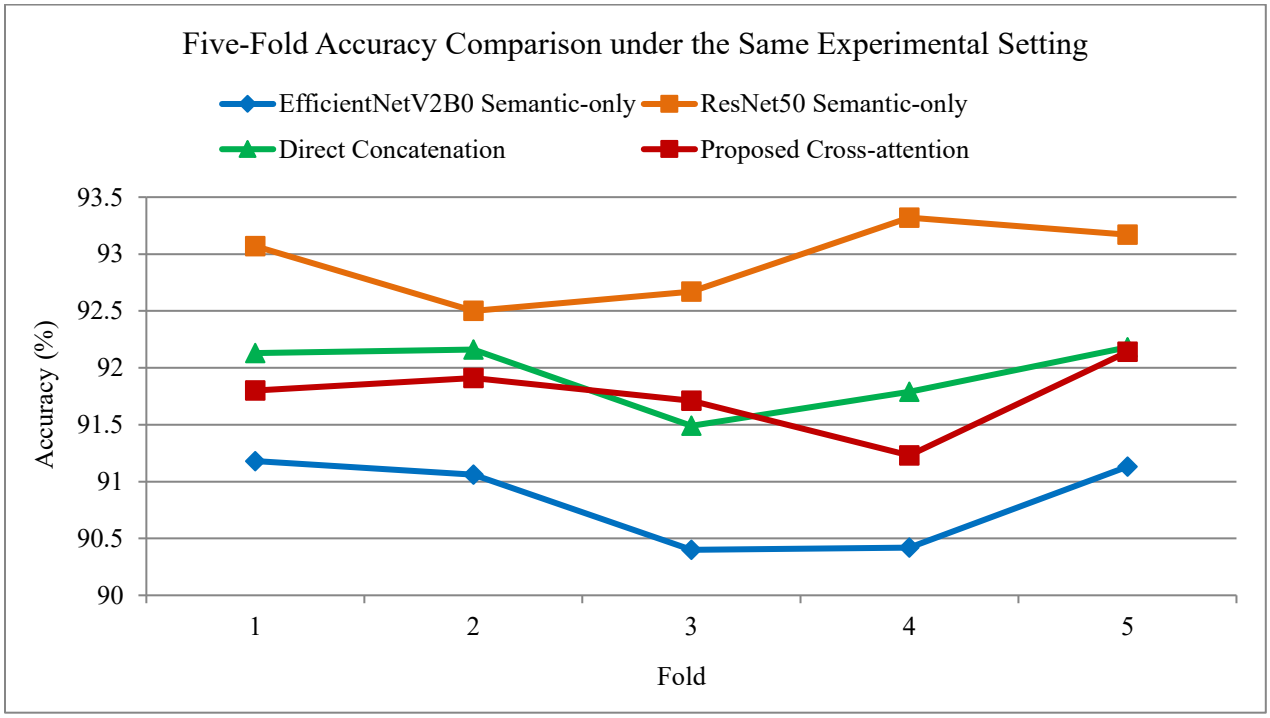

Fig. 4 Accuracy comparison of the EfficientNetV2B0 semantic-only, ResNet50 semantic-only, direct-concatenation, and proposed cross-attention models

3.2. Accuracy Progression of the Proposed Framework

The epoch-wise progression of the proposed framework is presented in Table 6. Accuracy increased rapidly during the initial epochs, rising from 81.01% at Epoch 1 to 93.04% at Epoch 5. It exceeded 97% at Epoch 10 and reached 99.05% by Epoch 15. Only smaller improvements were observed after Epoch 15. Accuracy increased from 99.35% at Epoch 20 to 99.55% at Epoch 25 and 99.59% at Epoch 30. This behaviour indicates that the model learned the dominant semantic and structural patterns during the first half of the optimisation process, while the later epochs produced incremental refinement.

Table 6. Accuracy progression of the proposed cross-attention framework

Epoch	Accuracy (%)
1	81.01
5	93.04
10	97.12
15	99.05
20	99.35
25	99.55
30	99.59

The final model summary reported an Accuracy of 99.96% when the completed model was evaluated in inference mode on the principal learning partition. The small difference between the Epoch 30 log and the final model summary arises because dropout is active during optimisation but inactive during inference-mode evaluation. The Accuracy progression is illustrated in Figure 5. The steep rise during the initial epochs demonstrates rapid convergence of the compact descriptor-level classifier. The curve begins to saturate after approximately 15 epochs, indicating that the semantic and handcrafted descriptors had already formed a highly discriminative representation by that stage.

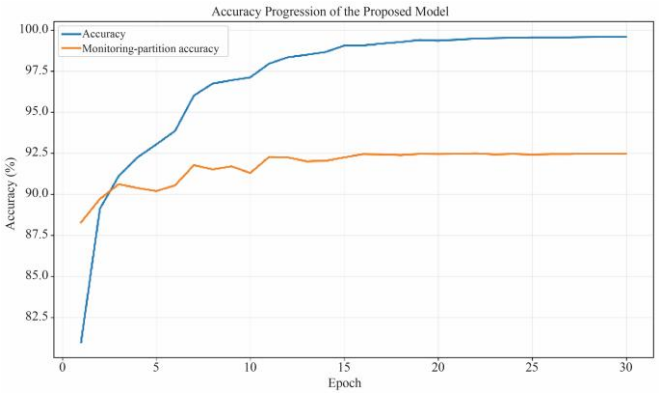

Fig. 5 Accuracy progression of the proposed cross-attention framework over 30 epochs

3.3. Contribution of the Semantic and Handcrafted Descriptors

The same-setting comparison provides a direct assessment of the principal framework components. The EfficientNetV2B0 semantic-only model represents the absence of handcrafted information. The direct-concatenation model represents the inclusion of the complete handcrafted descriptor without attention-based interaction. The proposed framework represents the inclusion of the same handcrafted descriptor through cross-domain attention. The increase from 98.35% to 99.84% after introducing direct concatenation demonstrates that the 10 LBP features, five GLCM features, and three morphological features contain discriminative information that is not fully represented by the EfficientNetV2B0 descriptor. LBP represents local micro-texture patterns, while GLCM captures statistical properties such as contrast, dissimilarity, homogeneity, energy, and correlation. The morphological descriptors provide complementary information regarding connected-region area, eccentricity, and perimeter. These handcrafted components are particularly relevant for scenes that contain similar semantic objects but differ in surface regularity, spatial repetition, boundary structure, or object geometry. The increase from 99.84% for direct concatenation to 99.96% for the proposed framework indicates that the attention mechanism contributes beyond simple feature aggregation. In direct concatenation, all 1,298 feature values are merged into a single representation before classification. In the proposed model, the semantic descriptor forms the query, while the handcrafted descriptor forms the key and value. This allows the model to condition the use of texture and morphology information on the semantic context of each scene. The improvement over direct concatenation is modest because both representations are already highly discriminative. Nevertheless, the comparison directly addresses the editorial concern regarding the contribution of the proposed fusion mechanism.

3.4. Class-Level Discrimination Analysis

The proposed framework demonstrated strong discrimination across the complete 46-class problem. Categories with distinctive visual structures, including wind turbine, swimming pool, harbour and port, shipping yard, desert, cloud, and airplane, produced particularly strong precision, recall, and F1-score values. Wind turbine scenes contain highly characteristic elongated turbine structures and relatively uncluttered surroundings. Swimming pools generally exhibit distinctive rectangular shapes and colour patterns. Shipping yards and harbour scenes contain recognisable combinations of water, vessels, storage areas, and transportation infrastructure. These characteristics are effectively represented by the semantic descriptor and reinforced by the texture–morphology branch. The most challenging categories included railway, railway station, basketball court, tennis court, industrial area, overpass, and parkway. Railway and railway station scenes contain similar parallel-line structures, transportation corridors, nearby buildings, and vegetation. Basketball courts and tennis courts share rectangular surfaces and line markings. Overpass, parkway, freeway, intersection, and railway scenes also contain elongated transportation structures that may occupy only a limited portion of an image. The remaining errors, therefore, arise mainly among categories with closely related object arrangements and spatial geometries rather than between unrelated natural and urban scenes.

3.5. Confusion Matrix Analysis

The complete 46-class confusion matrix of the proposed framework is shown in Figure 6. The matrix contains a dominant diagonal, indicating that most images were assigned to their corresponding scene categories.

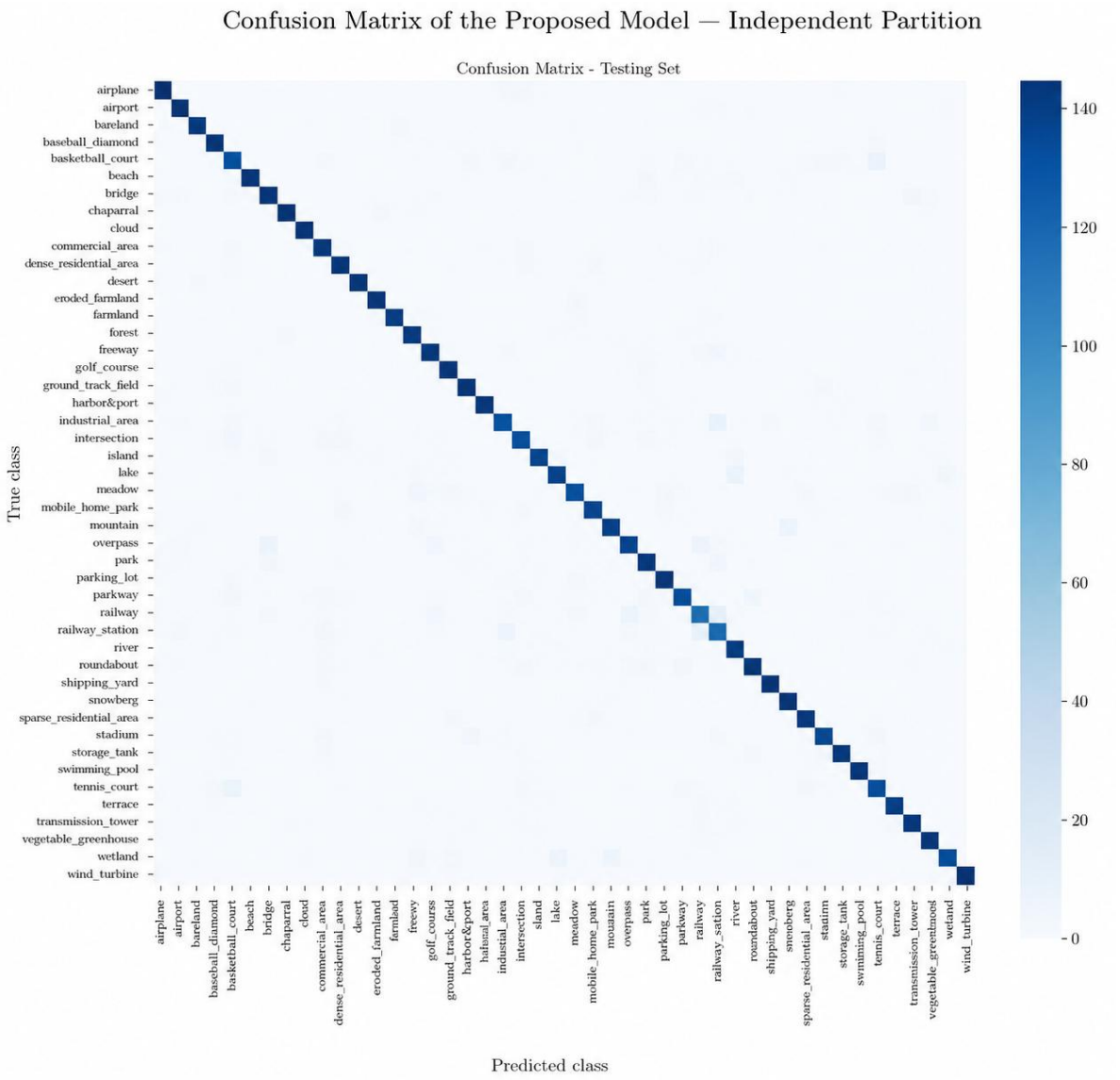

Fig. 6 Confusion matrix of the proposed cross-attention framework for the complete 46-class assessment

The most visible off-diagonal entries are associated with semantically or structurally related scene categories. Railway and railway station show mutual confusion because both contain rail tracks, platforms, transport corridors, and adjacent buildings. Basketball courts and tennis courts share geometric court patterns, while the industrial area and commercial area contain overlapping building-density and road-network characteristics. Confusion is also observed among overpass, parkway, freeway, and intersection scenes. These categories frequently contain road segments, bridges, vegetation, and surrounding structures with similar orientations. The confusion matrix, therefore, confirms that the principal difficulty lies in fine-grained separation of scene subclasses with overlapping visual components. The overall diagonal dominance across all 46 classes provides evidence that the proposed descriptor combination captures both broad scene semantics and fine structural characteristics.

3.6. ROC-AUC Analysis

The one-versus-rest ROC curves of the proposed framework are shown in Figure 7. The macro and weighted ROC-AUC values were both 0.9988, while the micro-average ROC-AUC was 0.9990. Most class curves remain close to the upper-left region of the ROC space. This indicates that the model generally assigns higher confidence to the correct category than to unrelated alternatives.

The high ROC–AUC values, together with the observed confusion among railway-related, sports-surface, and transportation categories, indicate that the remaining difficulty is concentrated among a limited number of visually similar classes. The model retains strong separation between unrelated categories even when the maximum-probability prediction is affected by fine-grained similarity.

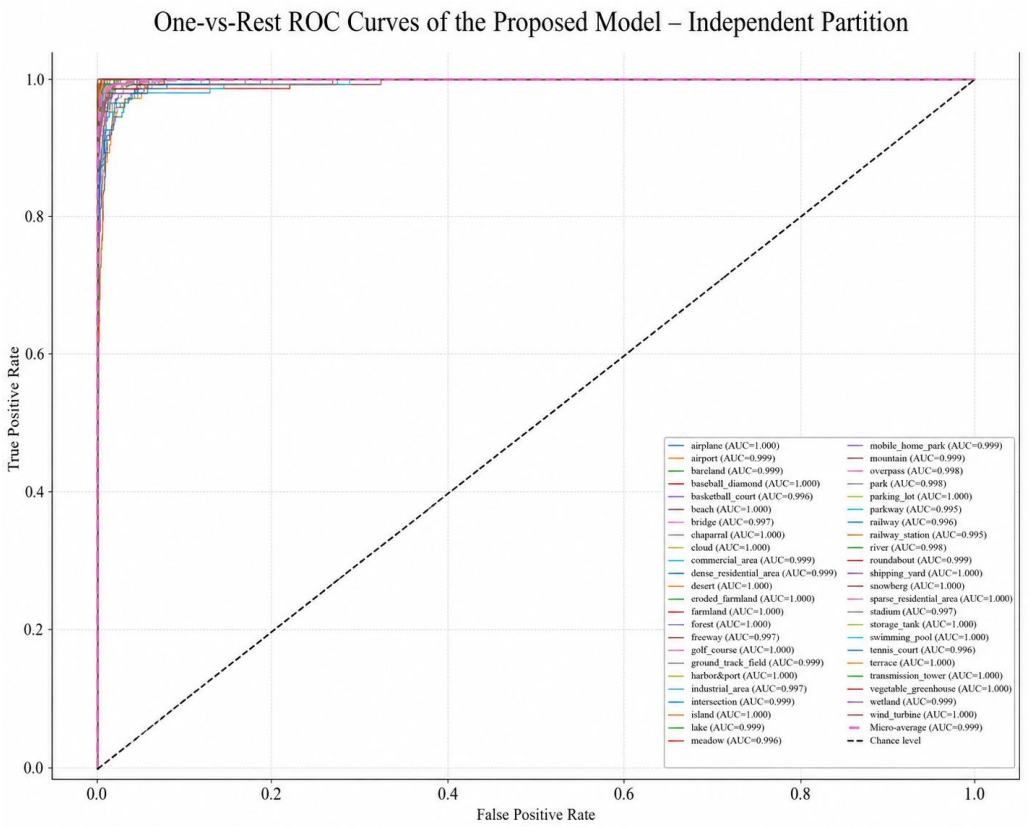

Fig. 7 One-versus-rest ROC curves of the proposed cross-attention framework for the complete 46-class assessment

3.7. Comparison with the Earlier Reduced-Class Experiment

The earlier manuscript evaluated only 15 selected MLRSNet classes. Such a reduced setting did not fully represent the diversity and difficulty of the benchmark. The present experiment extends the evaluation to all 46 classes using a balanced set of 46,000 images. The complete setting includes visually similar residential, industrial, agricultural, water-body, transportation, sports, and infrastructure categories. Consequently, the revised experiment provides a substantially stronger assessment of the representational capacity of the proposed framework. The revised comparison also eliminates the protocol inconsistency identified by the editor. All baseline models and the proposed model were implemented using the same images and optimisation conditions. Therefore, the observed differences can be attributed to the backbone and fusion strategies rather than to different class selections or dataset partitions.

4. Conclusion

This study presented a resource-efficient transformer-fused hybrid descriptor framework for complete 46-class remote sensing scene classification using the MLRSNet dataset. The proposed method combines a compact 1,280-dimensional EfficientNetV2B0 semantic representation with 18 handcrafted LBP, GLCM, and morphological descriptors through four-head cross-domain attention, thereby integrating high-level scene semantics with complementary texture and structural information without the computational burden of dense image-token transformer processing. Under the same experimental conditions, the EfficientNetV2B0 semantic-only, ResNet50 semantic-only, and direct-concatenation baselines achieved Accuracies of 98.35%, 99.42%, and 99.84%, respectively, whereas the proposed framework achieved the highest Accuracy of 99.96%. The comparison confirms the usefulness of handcrafted descriptors beyond the semantic representation and demonstrates that learned cross-domain interaction provides an additional improvement over conventional concatenation. The complete 46-class analysis further showed strong discrimination across diverse natural, urban, agricultural, industrial, transportation, and recreational scenes, although closely related categories such as railway and railway station, sports surfaces, and transportation structures remain challenging. Overall, the proposed framework offers an effective and computationally practical approach for land-use and land-cover mapping, urban analysis, environmental monitoring, and large-scale remote sensing scene understanding, while future work may investigate separate descriptor tokens, component-wise ablation, and external-dataset evaluation to strengthen generalisation further.

Conflicts of Interest

No conflict of Interest by all authors.

Author Contributions

Conceptualisation, methodology, and formal analysis were done by C Bujji Babu and G Hari Krishnan. C Bujji Babu was responsible for software development, validation, investigation and data curation. The original draft preparation was done by Hari Krishnan G, and C Bujji Babu assisted in the review and editing. C Bujji Babu was in charge of visualisation, and G Hari Krishnan assisted in supervision and project administration to ensure the flow of the study. Both authors approved the final manuscript.

Acknowledgement

In this paper, a dataset known as MLRSNet, which is a public dataset that contains 15 classes of images for remote sensing.

The authors of this dataset have been acknowledged for developing a dataset that is well-balanced and of high quality, which helped to train the proposed image classification model successfully.

References

[1] Xiang Li et al., "RS-CLIP: Zero Shot Remote Sensing Scene Classification Via Contrastive Vision-Language Supervision," *International Journal of Applied Earth Observation and Geoinformation*, vol. 124, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Swalpa Kumar Roy et al., "Multimodal Fusion Transformer for Remote Sensing Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-20, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Junjie Wang et al., "Remote-Sensing Scene Classification via Multistage Self-Guided Separation Network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Fujian Zheng et al., "A Lightweight Dual-Branch Swin Transformer for Remote Sensing Scene Classification," *Remote Sensing*, vol. 15, no. 11, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Qibin He et al., "AST: Adaptive Self-supervised Transformer for Optical Remote Sensing Representation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 200, pp. 41-54, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Ningbo Guo et al., "Scene Classification for Remote Sensing Image of Land Use and Land Cover using Dual-Model Architecture with Multilevel Feature Fusion," *International Journal of Digital Earth*, vol. 17, no. 1, pp. 1-27, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Arrun Sivasubramanian et al., "Transformer based Ensemble Deep Learning Approach for Remote Sensing Natural Scene Classification," *International Journal of Remote Sensing*, vol. 45, no. 10, pp. 3289-3309, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Adekanmi Adegun, Serestina Viriri, and Jules-Raymond Tapamo, "Automated Classification of Remote Sensing Satellite Images using Deep Learning based Vision Transformer," *Applied Intelligence*, vol. 54, no. 24, pp. 13018-13037, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Dan Zhang et al., "Multiple Hierarchical Cross-Scale Transformer for Remote Sensing Scene Classification," *Remote Sensing*, vol. 17, no. 1, pp. 1-21, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Rui Liu, Jing Ling, and Hongsheng Zhang, "SoftFormer: SAR-Optical Fusion Transformer for Urban Land use and Land Cover Classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 218, pp. 277-293, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Rambabu Damalla et al., "TransRefine: Transformer-Augmented Feature Refinement for Zero-Shot Scene Classification in Remote Sensing Images," *Pattern Recognition*, vol. 162, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[12] Qionghao Huang, Fan Jiang, and Changqin Huang, "Remote Sensing Scene Classification with Relation-Aware Dynamic Graph Neural Networks," *Engineering Applications of Artificial Intelligence*, vol. 150, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Yingtao Duan et al., "STMSF: Swin Transformer with Multi-Scale Fusion for Remote Sensing Scene Classification," *Remote Sensing*, vol. 17, no. 4, pp. 1-19, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[14] Yaoyao Du, Li Chen, and Xingxing Hao, "EViT-Net: An Efficient Vision Transformer-Inspired Network for Enhanced Multi-Scale Remote Sensing Image Features," *Expert Systems with Applications*, vol. 296, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[15] Haiyan Xu et al., "HETMCL: High-Frequency Enhancement Transformer and Multi-Layer Context Learning Network for Remote Sensing Scene Classification," *Sensors*, vol. 25, no. 12, pp. 1-36, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[16] Xiaowei Gu et al., "Self-Organising Explainable Multi-View Representation Learning for Remote Sensing Scene Classification," *Applied Soft Computing*, vol. 190, pp. 1-31, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[17] Erzhu Li et al., "Improved Bilinear CNN Model for Remote Sensing Scene Classification," *IEEE Geoscience and Remote Sensing Letter*, vol. 19, pp. 1-5, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Meiqiao Bi et al., "Vision Transformer with Contrastive Learning for Remote Sensing Image Scene Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 738-749, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Maofan Zhao et al., "Local and Long-Range Collaborative Learning for Remote Sensing Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Pengyuan Lv et al., "SCViT: A Spatial-Channel Feature Preserving Vision Transformer for Remote Sensing Image Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-12, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[21] Xu Tang et al., "EMTCAL: Efficient Multiscale Transformer and Cross-Level Attention Learning for Remote Sensing Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-15, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Zongyao Sh, and Jianfeng Li, "MITformer: A Multiinstance Vision Transformer for Remote Sensing Scene Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Xiaoman Qi et al., "MLRSNet: A Multi-Label High Spatial Resolution Remote Sensing Dataset for Semantic Scene Understanding," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 169, pp. 337-350, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Gui-Song Xia et al., "AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965-3981, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[25] Weixun Zhou et al., "PatternNet: A Benchmark Dataset for Performance Evaluation of Remote Sensing Image Retrieval," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 197-209, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[26] Mingxing Tan, and Quoc V. Le, "EfficientNetV2: Smaller Models and Faster Training," *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, pp. 10096-10106, 2021. [Google Scholar] [Publisher Link]

[27] Alexey Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint*, pp. 1-22, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[28] Ze Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 9992-10002, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] T. Ojala, M. Pietikaine, and T. Maenpaa, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local binary Patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[30] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610-621, 1973. [CrossRef] [Google Scholar] [Publisher Link]

[31] Kaiming He et al., "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[32] Ashish Vaswani et al., "Attention is all you Need," *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, vol. 30, pp. 5998-6008, 2017. [Google Scholar] [Publisher Link]

[33] Nobuyuki Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62-66, 1979. [CrossRef] [Google Scholar] [Publisher Link]

[34] Jia Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, pp. 248-255, 2009. [CrossRef] [Google Scholar] [Publisher Link]