

Article

Frequency Domain and Gradient-Spatial Multi-Scale Swin KANsformer for Remote Sensing Scene Classification

Xiaozhang Zhu ¹, Junqing Huang ² and Haihui Wang ^{1,*} 

¹ School of Mathematical Sciences, Beihang University, Beijing 102206, China; zzhuxz@buaa.edu.cn

² Department of Mathematics, Ghent University, 9000 Ghent, Belgium; junqing.huang@ugent.be

* Correspondence: whhmath@buaa.edu.cn

Highlights

What are the main findings?

- We propose the FG-Swin KANsformer, utilizing frequency domain and gradient prior information and combining the nonlinear modeling capability of the Swin Transformer with the fusion of KAN to enhance the utilization of original image information and feature extraction capabilities.
- The FG-Swin KANsformer achieves exceptional performance in remote sensing scene classification on multiple benchmark datasets, with a classification accuracy superior to numerous mainstream models based on CNN and Transformer.

What are the implications of the main findings?

- This study utilizes the original image and global semantic information, integrating detailed features and multi-scale spatial relationships, providing a reliable model for remote sensing scene classification tasks.
- The model is scalable and offers a new feature extraction modeling approach based on the Transformer architecture for other remote sensing tasks.

Abstract

Transformer-based deep learning techniques have recently shown outstanding potential in remote sensing scene classification (RSSC), benefiting from their ability to capture global semantic relationships and contextual dependencies. However, effectively utilizing the raw image and global semantic information while simultaneously taking into account detailed features and multi-scale spatial relationships remains a major challenge. Therefore, this paper proposes a novel FG-Swin KANsformer model that integrates frequency domain and gradient prior information from raw images with the Kolmogorov–Arnold Network (KAN) to enhance nonlinear feature modeling. The FG-Swin KANsformer consists of three key components: the Discrete Cosine Transform (DCT) module, the gradient-spatial feature extraction (GSFE) module, and the Swin Transformer module integrated with KAN. In the feature embedding phase, the DCT module extracts frequency domain features, while the GSFE module uses multi-scale convolutions and Sobel operators to extract spatial structures and gradient information at different scales, thereby enhancing the utilization of the original image's frequency domain and gradient prior information. In the Swin Transformer feature modeling phase, the conventional multilayer perceptron (MLP) in Swin Transformer Blocks is replaced by KAN, which decomposes complex multivariate functions into univariate compositions, thereby improving nonlinear representation capacity and enhancing feature discrimination. The thorough experiments on three distinct public remote sensing (RS) datasets demonstrate that FG-Swin KANsformer exhibits outstanding performance.



Academic Editor: Salah Bourennane

Received: 19 November 2025

Revised: 28 January 2026

Accepted: 2 February 2026

Published: 5 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: remote sensing scene classification (RSSC); feature extraction; Transformer; Kolmogorov–Arnold Network (KAN)

1. Introduction

With the continuous advancement of RS and satellite imaging technologies, modern high-resolution satellite imagery is able to capture ground details and object characteristics with greater clarity, enabling RS images to contain a wide variety of target objects. Consequently, it becomes essential to fully exploit the information embedded in raw images and acquire global semantic representations in order to achieve more effective feature extraction, and subsequently to make more accurate category judgments for RS images [1]. RSSC automatically categorizes RS images into specific semantic labels based on the geographic environmental information contained in the imagery. Owing to its strong capability in describing geographical features, RSSC has become a research hotspot in fields such as urban planning, environmental monitoring, and land use analysis [2–8]. Compared to natural image classification tasks, RS imagery exhibits more complex backgrounds and significant scale variations among ground objects. Furthermore, as shown in Figure 1, such images often suffer from large intraclass diversity and high interclass similarity, making RSSC tasks more challenging [9].

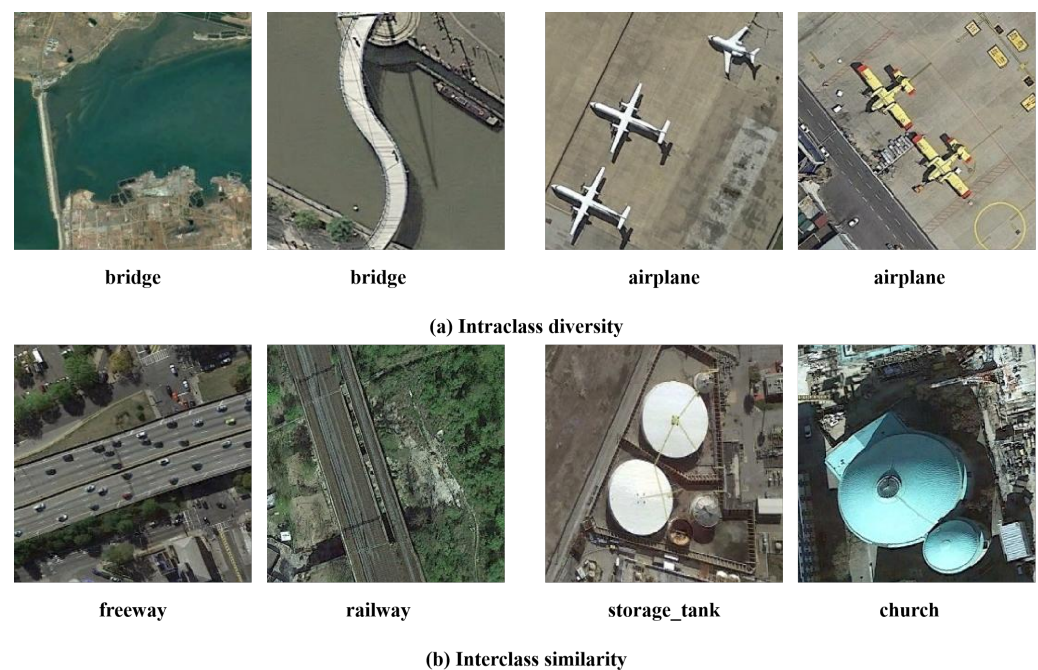


Figure 1. Image selected from the NWPU dataset.

In RS image analysis, scene classification typically involves identifying and classifying various land use types and surface cover types [10]. And the primary objective of RSSC is to systematically categorize different regions into meaningful land cover types based on geographic information in the image data. Unlike traditional single-object detection or classification tasks, RSSC emphasizes the comprehensive analysis of entire regions, focusing on distinguishing complex geographic scenes with similar spatial or semantic characteristics, rather than simply recognizing individual objects. Therefore, the accuracy and comprehensiveness of RSSC depend not only on the precise identification of each target but also on a thorough understanding of the spatial relationships between the targets and their global semantic information, which is crucial for improving classification performance.

The standard pipeline for RSSC typically consists of two steps: first, extracting the discriminative features of the original image; second, effectively classifying the scene categories based on the extracted features. In the aforementioned steps, identifying the discriminative features is the core component of scene classification and is generally achieved using either traditional handcrafted feature extraction methods [11–17] or deep learning methods [18–25]. Early handcrafted feature extraction techniques are capable of characterizing the statistical properties and geometric structures of RS images to some extent. However, due to their limited feature representation capability, they struggle to effectively cope with the complexity, large scale variations, and diversity inherent in RS scenes, which consequently leads to suboptimal classification performance. With the continuous rise of computer vision technologies, intelligent classification models represented by CNNs and Transformers have gradually become the mainstream. These methods achieve significant success in natural image classification tasks and are widely applied and validated in the field of RSSC. Nevertheless, unlike natural images, RS images are typically acquired from vertical perspectives by satellite or aerial platforms, and are characterized by complex backgrounds, significant scale variations, and high interclass similarity. These characteristics greatly increase the difficulty of extracting effective and distinctive features from RS images [26]. CNNs demonstrate strong performance in local feature modeling, whereas Transformer-type models rely on self-attention mechanisms, excelling in effectively establishing long-range connections and modeling global semantic relationships. Although they achieve outstanding results in various computer vision tasks, existing methods still have the following limitations:

1. Insufficient utilization of frequency domain information—Traditional Transformer-based models typically perform patch embedding directly from raw pixel inputs, without fully exploiting the potential of frequency domain features in modeling high-frequency texture details and low-frequency global structures. As a result, frequency domain information is not sufficiently leveraged.
2. Inadequate modeling of gradient and structural priors—Most existing studies focus on capturing contextual semantics using self-attention mechanisms or convolutional operations, while neglecting structural priors such as gradients and edges, making the model difficult to recognize in scenarios with blurred class boundaries or dense small targets, affecting classification accuracy.
3. Limited expressive capability of feedforward networks—The MLPs exhibit restricted nonlinear mapping capabilities and are insufficient for modeling the complex and highly variable visual patterns present in RS imagery. In scenarios involving multi-scale objects and high interclass similarity, this limitation often leads to insufficient feature extraction.

Therefore, we propose a novel FG-Swin KANsformer model that integrates frequency domain extraction via the DCT module, gradient extraction through the GSFE module, and the KAN module. The proposed model simultaneously enhances three crucial aspects—input feature construction, global semantic representation, and deep nonlinear modeling—thereby significantly improving scene classification performance in RS imagery. In the input feature construction stage, the DCT Extractor is employed to separate the high-frequency and low-frequency components of the target, and the DCT Adaptive Fusion is further introduced to dynamically adjust their weights, enabling an adaptive balance between texture details and the overall structure. Meanwhile, the GSFE module extracts gradient information and spatial structural features at different scales through multi-scale convolution and the Sobel operator and performs adaptive fusion with the original input after residual connection. This design not only highlights object boundaries, shapes, and texture characteristics, but also strengthens the spatial contextual representation within

scenes, thereby providing more discriminative feature inputs for the subsequent hierarchical modeling. In the deep feature modeling stage, we choose the Swin Transformer as the backbone and replace the MLP in the feedforward layer of Swin Transformer Blocks with KAN. By combining the window-based self-attention mechanism of Swin Transformer with the nonlinear feature representation capability of KAN, the architecture effectively enhances global semantic modeling and deep nonlinear representation while maintaining computational efficiency.

The major contributions of this article are as follows:

1. We design a new RSSC model named FG-Swin KANsformer. Based on the Swin Transformer, it introduces the frequency domain feature enhancement DCT module and the gradient-spatial structure enhancement GSFE module, while replacing the MLP in the Swin Transformer Blocks with KAN. This approach enhances both input feature enhancement and deep nonlinear modeling, significantly improving the overall performance of RSSC.
2. An effective-frequency spatial-gradient-aware feature construction mechanism is introduced to strengthen the quality of input representations. The DCT module decomposes input images into high- and low-frequency components, which are then adaptively fused to maintain a balance between global structural information and fine-grained texture features. Meanwhile, the GSFE module utilizes multi-scale convolutional filters and the Sobel operator to extract discriminative gradient and spatial structural features, leading to more expressive feature embeddings that capture edges, shapes, and contextual dependencies.
3. To enhance nonlinear modeling capability, the KAN-augmented feedforward layer is constructed by replacing the standard MLP in Swin Transformer Blocks with KAN. This modification enhances the FG-Swin KANsformer's nonlinear function approximation ability and strengthens the representation of complex feature representation, leading to more effective deep semantic modeling and higher classification accuracy.

The structure of the article is as follows: The systematic review of related research progress is provided in Section 2. The overall framework and core module designs of the proposed FG-Swin KANsformer model are presented in Section 3. In Section 4, the superior performance of FG-Swin KANsformer in RSSC tasks is demonstrated through extensive comparative experiments and validation analyses. Ablation studies are conducted and visualization-based interpretations are offered in Section 5. Finally, the conclusion is presented, and the potential areas for improvement are also discussed in Section 6.

2. Related Work

2.1. Transformer-Based Models for RSSC

The core objective of RSSC is to categorize RS images into the corresponding correct categories based on visual content. In the context of the rapid development of new network architectures and hardware, deep learning-based methods have gradually replaced traditional handcrafted feature extraction strategies, owing to their adaptive learning capability and hierarchical feature modeling ability. Among them, CNN and Transformer-based approaches [27–35] have shown particularly remarkable performance. Their superiority depends on their capability to autonomously extract hierarchical feature representations from unprocessed data, thereby effectively identifying complex patterns and semantic information. And, in particular, with the widespread adoption of Transformer architectures, deep learning-based RSSC approaches have exhibited remarkable effectiveness in global and local semantic modeling, multi-scale feature fusion, and cross-category discrimination, significantly enhancing the ability to distinguish highly similar scenes.

The Transformer architecture was originally designed for handling tasks within the field of natural language processing, where its core idea was to model and transform input sequences through an encoder–decoder architecture [36]. Later, Dosovitskiy et al. directly extended the Transformer to computer vision and introduced the Vision Transformer (ViT), which represented a new paradigm for leveraging Transformers in image understanding [37]. For the input image $X \in \mathbb{R}^{H \times W \times C}$, the ViT model first partitions the image into a collection of distinct patches. And each patch is then reshaped into a fixed-size token, and absolute positional encodings are added to each token to retain spatial relationships, forming the final input sequence. This sequence is subsequently processed by L layers of a Transformer encoder, each consisting of the multi-head self-attention (MHSA) mechanism and the MLP. The update process for the l -th layer is:

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

$$z^l = \text{MLP}\left(\text{LN}\left(z^{l'}\right)\right) + z^{l'}, \quad l = 1, 2, \dots, \quad (2)$$

where z^{l-1} corresponds to the output features of the preceding layer, $z^{l'}$ refers to the intermediate features after multi-head attention, z^l denotes the ultimate output of this layer, and LN denotes Layer Normalization. In the MHSA module, the input sequence is projected into the query, key, and value matrices $Q, K, V \in \mathbb{R}^{N \times d}$, and the attention is computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (3)$$

where in $d = D/h$, D denotes the input feature dimension, h denotes the number of attention heads, and N denotes the sequence length. Through parallel execution of MHSA and integration of their outputs, this enables the capture of the global dependency relationships between different patches in the sequence. Unlike traditional CNNs, which rely on local convolution kernels, ViT employs the MHSA mechanism to directly model long-range relationships across different regions of an image, thereby exhibiting stronger global modeling capability.

With the introduction of ViT, Transformer-based architectures have achieved significant breakthroughs in image processing, fully demonstrating their superior capability in processing visual data. Numerous Transformer variants have since been proposed for computer vision tasks [38–40]. Among them, Swin Transformer emerges as a more efficient and scalable visual Transformer architecture. By introducing a hierarchical representation and a shifted window-based local self-attention mechanism, Swin Transformer greatly reduces computational complexity through localized computation, thereby alleviating the quadratic growth in computational cost with respect to input resolution observed in ViT, and improving the efficiency when processing high-resolution images [41]. At the same time, Swin Transformer preserves the capability of modeling long-distance contextual relationships and naturally generates features at multiple spatial resolutions, making it more adaptable within dense prediction scenarios.

To resolve the problem of information isolation resulting from non-overlapping window partitioning, Swin Transformer adopts the shifted window strategy in successive layers by shifting the window partition by $(M/2, M/2)$, thereby enabling cross-window connections and enhancing inter-region interactions. This mechanism not only preserves the computational efficiency of local self-attention but also effectively expands the receptive field, thus facilitating global modeling. In each Swin Transformer Block, local self-attention is computed within non-overlapping windows. Given the query Q , key K , and value V pro-

jected from the input features, the window-based multi-head self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + B\right)V, \quad (4)$$

where d is the head dimension, $B \in \mathbb{R}^{M^2 \times M^2}$ denotes the learnable relative position bias within each window, and M denotes the window size.

In addition, Swin Transformer employs a hierarchical representation. At the end of each stage, a patch merging operation is applied to reduce spatial resolution while expanding channel dimensionality:

$$X^{(l)} \in \mathbb{R}^{\frac{H}{2^l} \times \frac{W}{2^l} \times C_l}, \quad l = 1, 2, \dots, L. \quad (5)$$

In this formulation, $X^{(l)}$ represents the feature map produced of the l -th layer, H and W specify the input image's height and width respectively, and C_l indicates the channel count of feature map at the l -th layer. This design enables Swin Transformer to maintain high computational efficiency when processing high-resolution images while capturing multi-scale image features. Owing to its efficient computation and powerful feature extraction capabilities, Swin Transformer has achieved remarkable success across multiple domains. By incorporating a shifted window mechanism and a multi-stage architectural design, Swin Transformer addresses the limitations of ViT in handling large-scale visual data, establishing itself as a highly impactful framework within contemporary computer vision.

2.2. Kolmogorov–Arnold Networks (KAN)

The MLP constitutes a classical feedforward model and serves as a core building block in deep learning [42–45]. Moreover, the universal approximation theorem [44] asserts that the MLP is theoretically able to approximate arbitrary continuous mappings. However, the MLP still faces several limitations. For instance, it inherently lacks structured mechanisms for modeling sequential dependencies or spatial relationships, leading to its reduced effectiveness in tasks where such information is critical. In addition, due to the lack of follow-up analytical interpretability tools, the MLP often exhibits poor interpretability. Due to these limitations, Guo et al. replace the MLP feedforward part of ViT with an inverted residual structure based on convolution, introducing local inductive bias of convolution [46]. Liu et al. propose the KAN, which differs fundamentally from the MLP [47]. Unlike the MLP, whose theoretical foundation is rooted in the universal approximation theorem, the theoretical foundation of KAN stems from the Kolmogorov–Arnold representation theorem, showing that any multivariate continuous mapping can be expressed as a sum of several univariate components [48,49]:

$$f(x_1, x_2, \dots, x_n) = \sum_{q=1}^{2n+1} \phi_q \left(\sum_{p=1}^n \psi_{qp}(x_p) \right), \quad (6)$$

where both ψ_{qp} and ϕ_q are univariate functions parameterized using spline functions and x_p denotes the input features. The inner functions ψ_{qp} transform each input feature into intermediate representations, which are then aggregated by the outer functions ϕ_q to generate the final output. This structured formulation enables the KAN to effectively capture compositional structures through combinations of univariate functions, providing a robust framework for function approximation.

Compared with the MLP, although both architectures employ fully connected structures, they differ significantly in their structural design and processing mechanisms. In the MLP, the fixed activation function is applied at each node, whereas the KAN applies learnable activation functions on the edges, which are automatically adjusted based on the

training data [47]. Specifically, the KAN replaces the traditional linear weight matrix with learnable one-dimensional functions parameterized by spline functions on each connection edge, as illustrated in Figure 2. Since the KAN removes nonlinear operations from the node level, each node only performs a simple linear summation of the input signals without requiring conventional activation functions for nonlinear transformation. This design considerably reduces the computational complexity at the node level while enhancing the model's flexibility and performance for specific tasks.

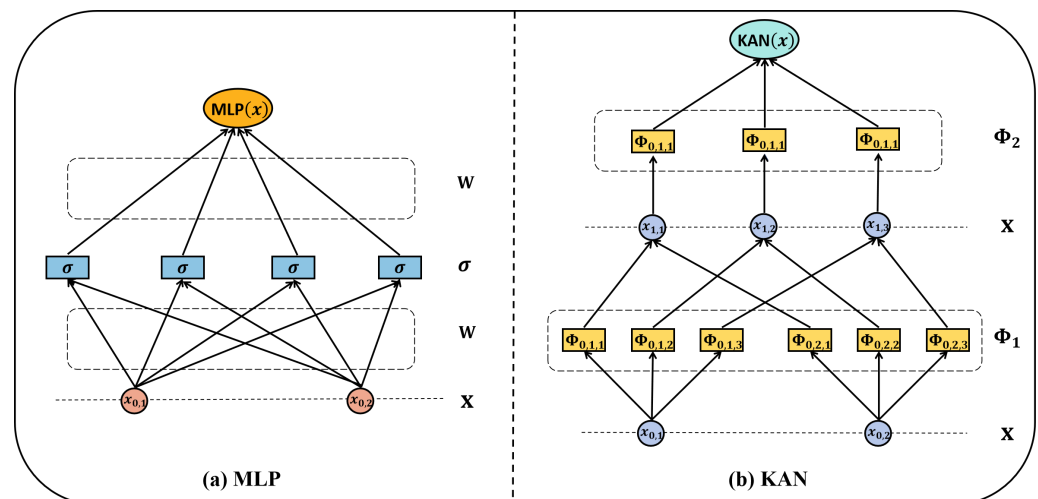


Figure 2. Two types of feedforward neural networks.

Beyond KAN's architectural characteristics, the nonlinear approximation capability offered by spline-based activation functions makes KAN particularly suitable for remote sensing scene classification (RSSC). Remote sensing imagery often exhibits heterogeneous spatial textures, multi-scale object distributions, and high interclass similarity under complex backgrounds, which impose stronger requirements on the expressive capacity of the model. By allowing each connection edge to learn adaptive nonlinear mappings, KAN is able to capture fine-grained variations and irregular spatial patterns that are difficult to represent with fixed activation functions in conventional MLPs. This task-driven suitability demonstrates that replacing MLP with KAN in our model is not merely a structural substitution, but a rational design tailored to the inherent characteristics of RSSC tasks.

2.3. Feature Enhancement Method Based on Frequency Domain and Gradients in Image Analysis

Early vision tasks extensively adopted DCT to break images down into various spectral domain components for image compression and denoising [50,51]. In recent years, despite the remarkable progress achieved by deep learning methods, visual tasks in natural image domains still encounter several challenges. Natural scene images often contain high-dimensional redundant features and complex background noise, which may lead to information redundancy and insufficient discriminative capability in learned representations. Consequently, frequency domain concepts have been reintroduced into deep learning frameworks to remedy the shortcomings of convolutional operations in capturing long-distance dependencies and global statistical structures. One class of approaches performs linear or gated operations directly in the frequency domain, followed by reconstruction or discrimination in the pixel domain—for example, spectral masking and band-pass filtering commonly used in image restoration and denoising. Another class embeds frequency domain operators into feature extraction backbones, utilizing Fast Fourier Transform (FFT)/DCT as separable global token mixers or as approximate substitutes for attention mechanisms to reduce computational complexity [52]. Moreover, some studies

incorporate frequency domain priors into multi-branch architectures alongside spatial convolution or attention mechanisms, where adaptive weighting is used to dynamically select between different frequency bands in order to balance texture detail preservation and structural stability [53,54].

In addition to incorporating frequency domain prior information, spatial gradient computation is also one of the fundamental operations within the domain of image analysis, and, in particular, performs exceptionally well in applications such as boundary detection and feature extraction. And the traditional gradient-based operators have gradually been integrated with neural network frameworks, leading to broader applicability and further extensions. For example, Ravivarma et al. employed the Sobel operator to extract boundary information in visual analysis applications [55]. Wang et al. proposed the LGO, which, unlike traditional Sobel filters, is capable of adaptively learning gradient-spatial information, thereby improving its adaptability to complex textures [56]. The advancement of deep learning enables spatial gradient computation to be jointly optimized and automatically learned within neural network architectures, significantly improving both the efficiency and accuracy of image processing.

As discussed earlier, RS images often exhibit complex targets, large scale variations, and high interclass similarity, which pose persistent challenges for existing image classification methods during feature extraction. To tackle these issues, we introduce the FG-Swin KANsformer, which incorporates the frequency domain DCT module and the gradient-based GSFE module during the initial feature formation stage. The DCT module effectively derives different frequency domain components from images, thereby greatly strengthening the framework's capacity to capture. Meanwhile, the GSFE module adaptively fuses multi-scale gradient features, strengthening the feature extraction process of RS images. The dual enhancement across the frequency and spatial domains strengthens the framework's capability to differentiate targets with varying scales and to handle variable background scenarios.

3. Methods

This part provides an in-depth explanation of the overall structure and specific modules of the FG-Swin KANsformer model used for RSSC. We begin with a systematic overview of the overall framework of FG-Swin KANsformer. We then present the design and implementation of the DCT and GSFE modules used in the input feature construction stage, which enhance the model's classification capability by extracting discriminative features from the frequency and spatial domains, respectively. Finally, we delve into the design principles of the KAN-integrated Swin Transformer, where the unique MHSA mechanism of Swin Transformer is combined with the nonlinear modeling capability of KAN. The below subsections provide detailed explanations of each component within the FG-Swin KANsformer architecture.

3.1. Overview

In the field of RSSC, we propose a novel model named FG-Swin KANsformer, whose overall framework is illustrated in Figure 3. Specifically, the input image initially undergoes processing through the DCT module, generating corresponding high- and low-frequency enhanced representations. These representations are then fused with the original image through the residual-weighted strategy, and finally result in enhanced image characteristics. And the fused feature map is further refined by the GSFE module, which combines multi-scale convolution with a Sobel gradient detector to capture edge-aware features, while the attention mechanism emphasizes key regions. After enhancement by the DCT and GSFE modules, the features are fed into the KAN-integrated Swin Transformer, which

leverages its unique window attention mechanism to efficiently model local as well as global dependencies in RS imagery. Within each Swin Transformer Block, the traditional MLP is replaced by KAN to further improve the model's capability in capturing discriminative features. Finally, the model produces classification outputs based on the learned feature representations. The detailed design of each component in FG-Swin KANsformer is presented one by one below.

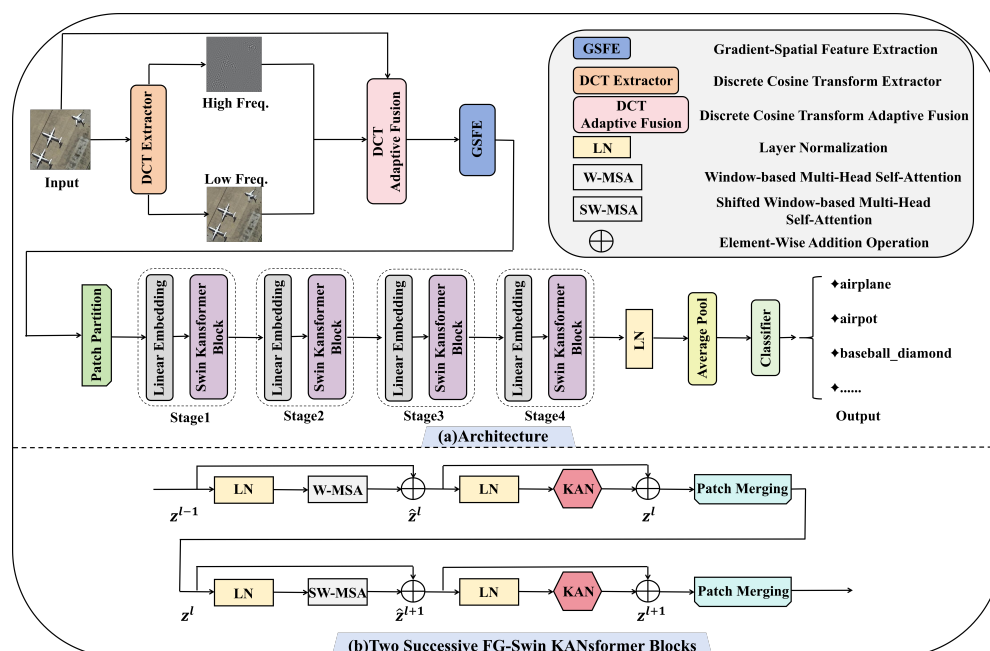


Figure 3. Overview of our proposed FG-Swin KANsformer. (a) Architecture of FG-Swin KANsformer. (b) Structure of two successive FG-Swin KANsformer Blocks.

3.2. Feature Enhancement Module

In RS image analysis, the details, edges, and the multi-scale properties of the image are of great significance. Therefore, we design two feature enhancement modules during the input feature construction phase: the frequency domain decomposition DCT module and the gradient-space feature extraction GSFE module. These two modules enhance the image's feature representation from the perspectives of frequency domain and spatial gradient information, effectively improving the FG-Swin KANsformer's capacity to capture the details and boundary features of RS images, thereby boosting classification performance.

3.2.1. DCT Module

The detailed features and background structure of an image are located in different frequency components, and frequency domain information can provide more clues about details and textures. To capture these important frequency details, we introduce a frequency domain decomposition module based on DCT in the model. As shown in Figure 4, the DCT Extractor module first extracts high-frequency and low-frequency features from the input image, and then the DCT Adaptive Fusion module applies an adaptive fusion mechanism to enhance the image's detail and texture representation. Specifically, DCT is applied to each channel of the input image individually, transforming the image from the spatial domain to the frequency domain:

$$I_D(u, v, c) = \alpha(u)\alpha(v) \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} I(x, y, c) \cos\left(\frac{\pi(2x+1)u}{2H}\right) \cos\left(\frac{\pi(2y+1)v}{2W}\right), \quad (7)$$

where $I(x, y, c)$ denotes the pixel values of the c -th color channel of the input image, $I_D(x, y, c)$ represents the transformed frequency domain representation, H and W denote the image height and width, x and y are spatial indices, u and v are frequency indices, and $\alpha(u)$ and $\alpha(v)$ are normalization coefficients. Subsequently, high-pass and low-pass filters are applied to the frequency domain image to extract the high-frequency component $I_{\text{high}}(u, v, c)$ and the low-frequency component $I_{\text{low}}(u, v, c)$, respectively. The high-frequency component primarily preserves details and edge structures, while the low-frequency component reflects the global structure and background information of the image. The computations of the different frequency domain components are defined as follows:

$$I_{\text{high}}(u, v, c) = I_D(u, v, c) \cdot H_{\text{high}}(u, v), \quad (8)$$

$$I_{\text{low}}(u, v, c) = I_D(u, v, c) \cdot H_{\text{low}}(u, v), \quad (9)$$

where $H_{\text{high}}(u, v)$ and $H_{\text{low}}(u, v)$ represent the high-pass and low-pass filters, respectively. After extracting the different frequency domain components, the Inverse Discrete Cosine Transform (IDCT) is applied to transform the frequency domain representations back into the spatial domain. The resulting features are then further refined by the DCT Adaptive Fusion module, which enhances the image representation through an adaptive fusion strategy. Specifically, the original image, high-frequency component, and low-frequency component are fused in a learnable manner to obtain the final output.

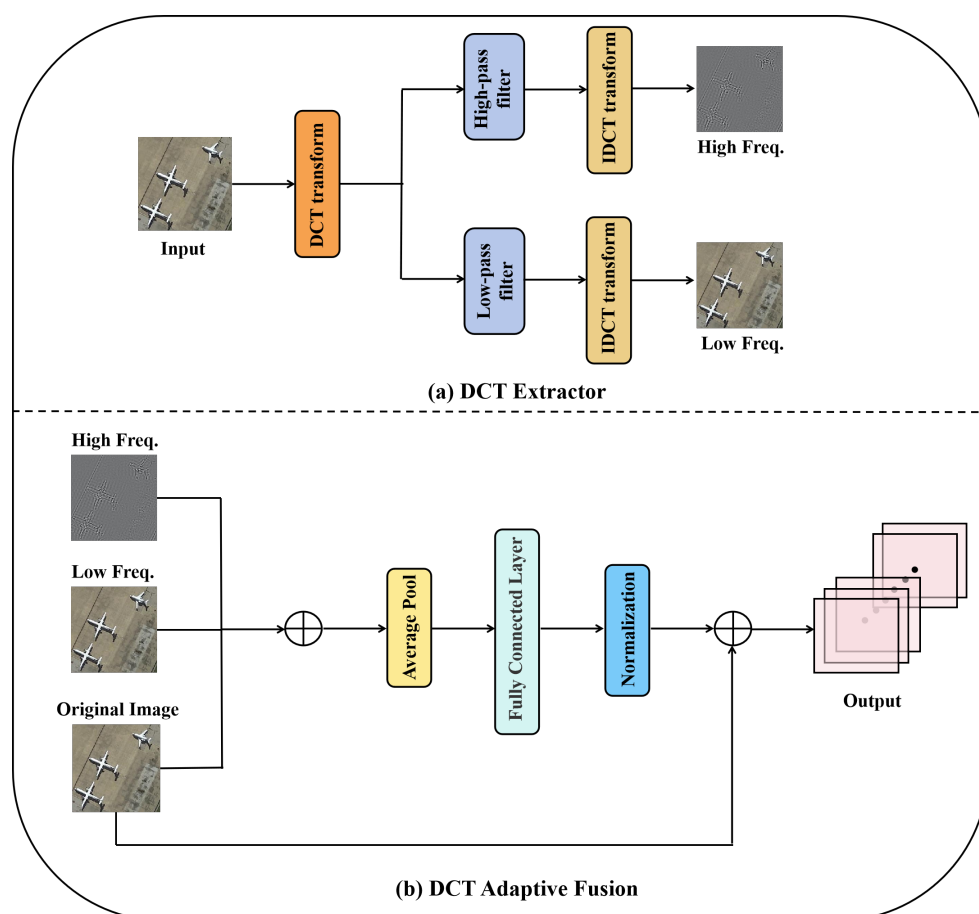


Figure 4. Detailed structure diagram of DCT module.

3.2.2. GSFE Module

In RS images, edge and detail regions often carry critical semantic information. To enhance the model's attention to these regions, we design the GSFE module, as shown in Figure 5. This module first computes gradient information using the Sobel operator to extract edge features from the image, and then combines these edge features with multi-scale convolutional representations. Finally, the extracted edge features and multi-scale convolution features are adaptively fused through a learnable weighting mechanism, thereby strengthening the representation of important regions in the image.

Concretely, the GSFE module first extracts multi-scale features from the image through convolution operations with different receptive fields. To capture local information at various scales, convolution kernels of different sizes are adopted to extract fine-grained features across multiple spatial levels. Specifically, given an input image I , multi-scale features F_1 , F_3 , and F_5 are obtained by applying convolution operations with kernel sizes of 1×1 , 3×3 , and 5×5 , respectively. Meanwhile, the input I is processed by the Sobel operator to compute gradient information and extract edge features. The gradient computation of the Sobel operator is defined as:

$$\text{grad}_x = I * \text{Sobel}_x, \quad (10)$$

$$\text{grad}_y = I * \text{Sobel}_y, \quad (11)$$

where Sobel_x and Sobel_y represent the horizontal and vertical convolution kernels, respectively. Through gradient computation, the structural edge characteristics of the image are obtained, which helps the model capture fine-grained details. The computed gradient features are subsequently processed by a lightweight convolutional network consisting of two convolutional layers with ReLU activation applied afterward, resulting in the enhanced gradient feature representation G . Finally, the multi-scale convolutional features and the gradient feature are concatenated to obtain the fused feature F_{concat} , expressed as:

$$F_{\text{concat}} = \text{concat}(F_1, F_3, F_5, G). \quad (12)$$

Then, the fusion-weighted convolution block generates the weight coefficient A through a lightweight two-layer convolutional network. And these weights are applied to the concatenated features via element-wise multiplication before being fed into the convolutional fusion module:

$$F_{\text{fused}} = \text{Conv}(A \odot F_{\text{concat}}), \quad (13)$$

and the weighted feature map F_{fused} is residual-connected with the original input I , resulting in the final output feature map F_{output} :

$$F_{\text{output}} = F_{\text{fused}} + F_{\text{concat}}. \quad (14)$$

Thus, the GSFE module can substantially reinforce the feature representation of edges and detailed areas, consequently strengthening the model's capability to recognize salient regions in RS images.

3.3. KAN Module in Swin Transformer Blocks

To further strengthen the model's nonlinear expression capability, the MLP in the Swin Transformer Blocks is replaced with KAN. Compared with traditional MLP, the KAN module leverages B-spline basis functions and piecewise polynomial linear layers to achieve more flexible and fine-grained feature transformations. To illustrate the mechanism

of the KAN module in detail, we refer to the architecture diagram in Figure 6. As depicted, the KAN module stacks multiple KAN linear layers, each of which is implemented as follows: First, the input data x undergoes a nonlinear transformation through a basic activation function. The activated result is subsequently fed into a linear layer to produce the baseline linear output, defined as:

$$X_{\text{Base Output}} = W_{\text{base}} \cdot \text{Activation}(x), \quad (15)$$

where W_{base} represents the parameter matrix associated with the base linear layer, and $\text{Activation}(x)$ is the result of applying the base activation function to the input x . At the same time, the input x passes through B-spline basis functions for spline interpolation, which perform spline transformations to input fine-grained structural information from the data. The B-spline basis function $B(x)$ provides higher expressiveness than standard activation functions, and is expressed as:

$$B(x) = [B_1(x), B_2(x), \dots, B_k(x)], \quad (16)$$

where $B_i(x)$ is the i -th B-spline basis function. This transformation provides a richer feature representation and increases the sensitivity to the fine details of the data. After the B-spline transformation, the data is passed into the Spline Linear Layer, where it is linearly combined with the piecewise polynomial weights W_{spline} , thus mapping the data processed by B-splines to a higher-dimensional feature space, leading to the piecewise polynomial output. The piecewise polynomial output is given by:

$$X_{\text{Spline Output}} = W_{\text{spline}} \cdot B(x). \quad (17)$$

Finally, by combining the base linear output and the piecewise polynomial output, the final output features are obtained:

$$X_{\text{Output}} = \alpha \cdot X_{\text{Base Output}} + \beta \cdot X_{\text{Spline Output}}, \quad (18)$$

where α and β are learnable weights used to control the importance of the base linear output and the piecewise polynomial output in the final result. This combination can acquire linear and nonlinear representations of the data, and strengthens its ability to express complex features.

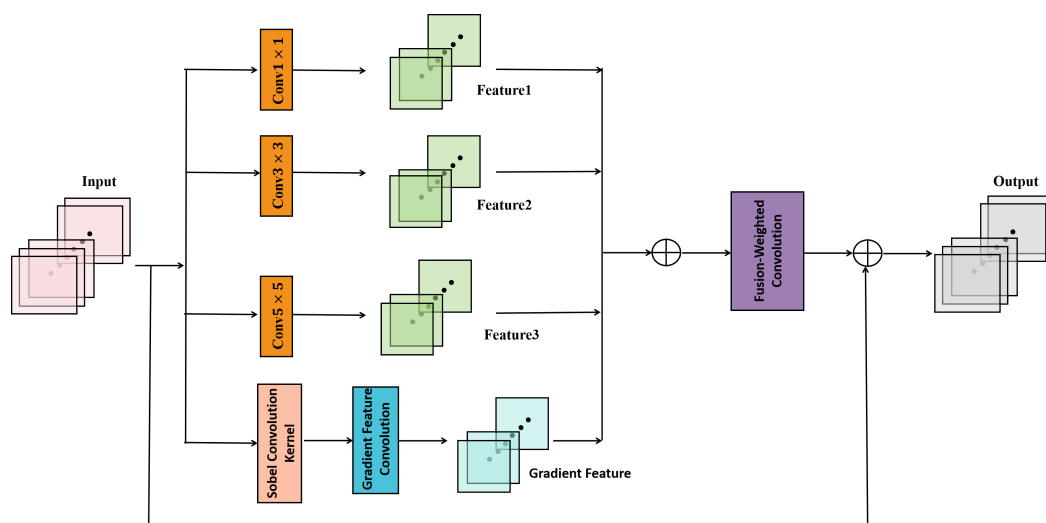


Figure 5. Detailed structure diagram of GSFE module.

Compared with the original KAN, we further optimize and refine the module in terms of memory efficiency and regularization strategy. Regarding memory efficiency, the origi-

nal KAN implementation requires expanding intermediate variables for each activation function, which leads to high memory overhead. Specifically, in the original design, each activation function computes and stores an independent intermediate tensor, which is not effectively reused and results in redundant memory consumption. To address this issue, we activate the input features using different B-spline basis functions instead of generating separate intermediate variables for each activation function. By linearly combining the outputs of multiple basis functions, this approach significantly reduces memory usage, avoids redundant computations, and thus improves both memory utilization and overall computational efficiency. In terms of regularization, the original KAN applies L_1 regularization to penalize network weights and performs nonlinear operations on all intermediate tensors. However, the approach implementation is incompatible with the computation of B-spline basis functions, which involve extensive linear combinations and cannot efficiently accommodate nonlinear regularization over intermediate tensors. Therefore, we apply L_1 regularization only to the network weights rather than directly to all intermediate tensors, which avoids unnecessary nonlinear operations while preserving the regularization capability of the model. Moreover, this strategy aligns better with conventional regularization techniques in neural networks and reduces unnecessary computational complexity during training, thereby improving training efficiency. Empirically, we do not observe degradation in convergence stability and generalization results after adopting L_1 regularization, indicating that the simplified regularization strategy still maintains convergence stability and generalization performance.

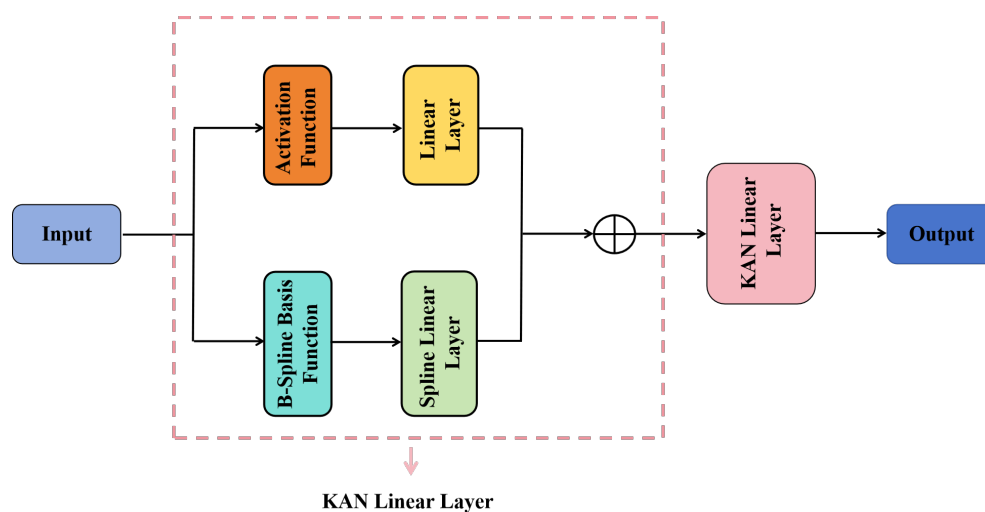


Figure 6. Detailed structure diagram of KAN module.

4. Results

The following text evaluates the RSSC ability of the FG-Swin KANsformer on three publicly available RS scene datasets. We first provide an explanation of the dataset used, the way the experimental parameters are set, and the key details during the implementation process. Then, through extensive experiments, we evaluate the performance of FG-Swin KANsformer with other advanced methods, and the results are reported accordingly.

4.1. Dataset

In order to comprehensively validate the superior performance of the FG-Swin KANsformer, we conduct systematic experiments on the following publicly available RS scene image datasets: the UC Merced Land Use Dataset (UCM) [57], the Aerial Image Dataset (AID) [26], and the NWPU-RESISC45 Dataset (NWPU) [58]. These datasets represent different scales, where UCM, AID, and NWPU correspond to small, medium, and large

datasets, respectively, and they also exhibit significant differences in category composition. Therefore, evaluating the model performance across these three datasets provides a more comprehensive assessment of its classification capability. The detailed information about each dataset is summarized in Table 1.

Table 1. Detailed description of each dataset.

Dataset	Total Number	Classes	Images per Class	Resolution	Image Size
UCM	2100	21	100	0.3 m	256×256
AID	10,000	30	220–420	0.5–0.8 m	600×600
NWPU	31,500	45	700	0.2–30 m	256×256

Below is a comprehensive overview of the publicly available dataset used in this article.

The UCM RS scene collection is the benchmark dataset for RSSC. It contains 100 images for each of the 21 categories, and each image has a resolution of 256×256 . In this work, two dataset partitioning strategies are adopted: 50%/50% and 80%/20% for training/testing splits. The dataset encompasses various scenes with rich texture and color information, making it an important benchmark for evaluating RSSC methods, and Figure 7 presents several representative examples.



Figure 7. Example diagram of the UCM dataset.

The AID is an aerial image dataset specifically designed for scene classification. Among the 30 RS scenarios in this dataset, there are 10,000 images with a spatial resolution of 600×600 . In this study, two dataset partitioning strategies are adopted: 20%/80% and 50%/50% for training/testing splits. And Figure 8 presents several representative examples from the dataset.

The NWPU dataset contains 700 images for each of the 45 categories, and each image measures 256×256 in resolution. And in this work, two dataset partitioning strategies are employed: 10%/90% and 20%/80% for training/testing splits. Relative to the two datasets mentioned earlier, NWPU is larger in scale and contains a wider variety of scene categories, and several sample images are displayed in Figure 9.

4.2. Training Settings

This study employs overall accuracy (OA) and confusion matrix (CM) as quantitative indicators for the FG-Swin KANsformer model. OA calculates the ratio of correctly classified scenarios to the total number of scenarios under evaluation, and the CM provides insights into potential misclassifications across different categories. By combining these

two metrics, model evaluation can be conducted on two levels: OA can evaluate the classification accuracy of the FG-Swin KANsformer, while CM enables more detailed analysis of interclass relationships and confusion patterns.

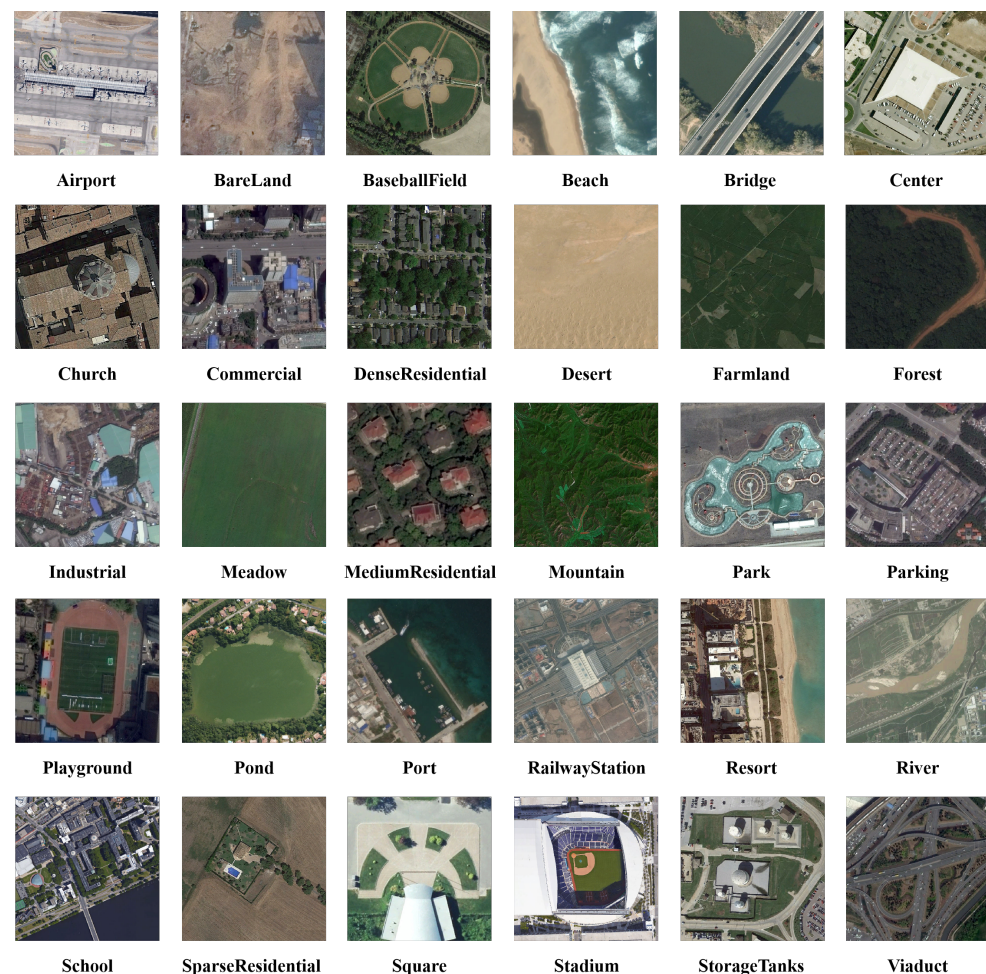


Figure 8. Example diagram of the AID dataset.

To accelerate computation, all experiments are conducted using an NVIDIA RTX 4060 GPU, and the model is trained for 100 epochs on each RS dataset with a batch size of 8. The cosine learning rate scheduler with linear warm up is adopted during training, and the Adam optimizer is used for parameter updates with an initial learning rate of 0.0001. All input images are resized to 224×224 . In addition, various data augmentation techniques are applied, including random horizontal flipping and random erasing. Each experiment is repeated five times and the final results are reported in terms of the mean and standard deviation.

4.3. Comparison Experiments

We design comparative experiments on the AID, NWPU, and UCM datasets by benchmarking FG-Swin KANsformer against the cutting-edge CNN-based and Transformer-based models. For the former, we select GoogLeNet [1], VGG [1], ResNet [59], EMSC-Net [60], GCSANet [61], MSANetwork [62], and MDRCD [63]. For Transformer-based methods, PiT-S [64], EMTCAL [65], CSAT [66], SCViT [32], LDBST [67], STMSF [68], ACTFormer [69], MSCN [70], and SceneFormer [71] are chosen for comparison. Through these comparisons, the superiority of FG-Swin KANsformer in fully exploiting frequency and gradient information from raw images, as well as the effectiveness of integrating Swin Transformer with KAN, can be further demonstrated. To ensure fairness, we directly adopt the performance results of the pretrained models reported in the original literature, which

represent their optimal performance. This strategy guarantees the reliability of experimental results and enables a more accurate comparison with current mainstream approaches.

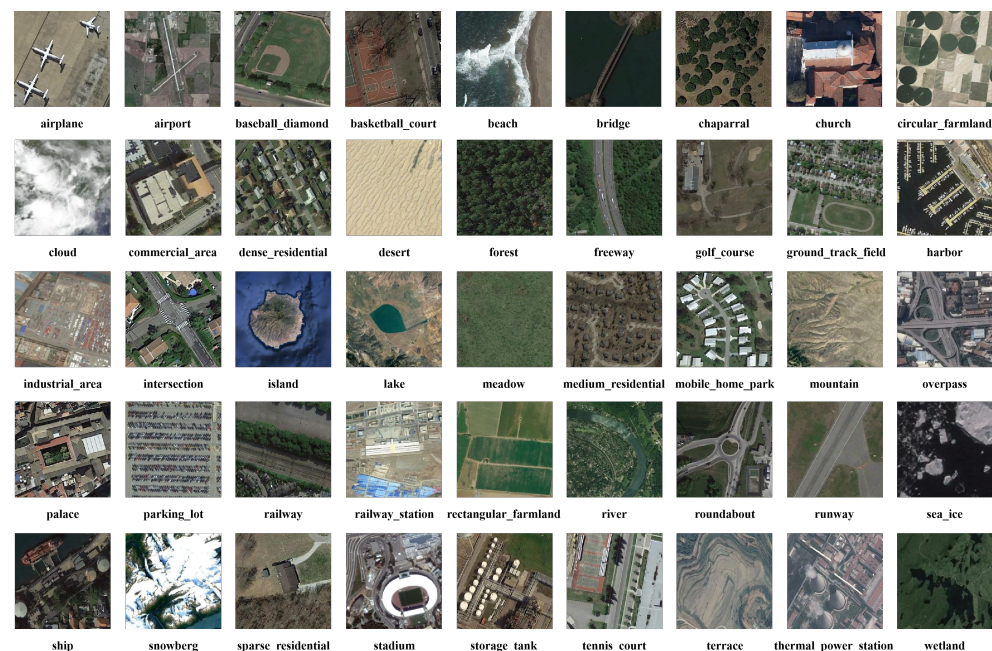


Figure 9. Example diagram of the NWPU dataset.

4.3.1. Experimental Results of UCM Dataset

For the UCM dataset, two experimental settings are adopted: one with 80% of the data used for training and 20% for testing, and the other with an equal split of 50% for both the training and testing sets. The detailed result outcomes can be found in Table 2, where the best performances are highlighted in bold. Evidently, under the 50% training split, FG-Swin KANsformer achieves an accuracy of 99.51%, outperforming all other compared models. As the training ratio rises to 80%, FG-Swin KANsformer reaches an accuracy of 99.65%, still demonstrating highly competitive performance. These results on the UCM dataset clearly verify the effectiveness of the FG-Swin KANsformer for RSSC tasks.

Table 2. Comparison of the OA(%) on the UCM dataset.

Method	50% Training Ratio	80% Training Ratio
GoogLeNet	92.70 ± 0.60	94.31 ± 0.89
VGG	94.05 ± 0.64	96.58 ± 0.41
ResNet	96.05 ± 0.65	97.75 ± 0.24
MSA-Network	97.80 ± 0.33	98.96 ± 0.21
PCNet	98.71 ± 0.22	99.25 ± 0.37
GCSANet	98.32 ± 0.71	99.31 ± 0.56
EMSCNet	98.70 ± 0.46	99.44 ± 0.16
MDRCN	98.57 ± 0.19	99.64 ± 0.12
PiT-S	95.83 ± 0.39	98.33 ± 0.50
EMTCAL	98.67 ± 0.16	99.57 ± 0.28
CSAT	95.72 ± 0.23	97.86 ± 0.16
SCViT	98.90 ± 0.19	99.57 ± 0.31
LDBST	98.76 ± 0.29	99.52 ± 0.24
STMSF	99.01 ± 0.31	99.58 ± 0.23
SceneFormer	-	99.00 ± 0.28
FG-Swin KANsformer (ours)	99.51 ± 0.13	99.65 ± 0.16

Bold denotes the best result.

As illustrated in Figure 10, we further present the categorization performance of FG-Swin KANsformer on the UCM dataset using the CM for more intuitive visualization. When the training ratio is configured to 80%, misclassification only occurs between the “intersection” and “medium residential” categories, while the remaining 19 categories all achieve 100% classification accuracy. Even when the training ratio is reduced to 50%, 18 out of the 21 categories still reach 100% accuracy, with only the “dense residential” category exhibiting an accuracy slightly below 98%. These observations demonstrate that the FG-Swin KANsformer can accurately distinguish the vast majority of categories in the UCM dataset, highlighting its excellent classification performance.

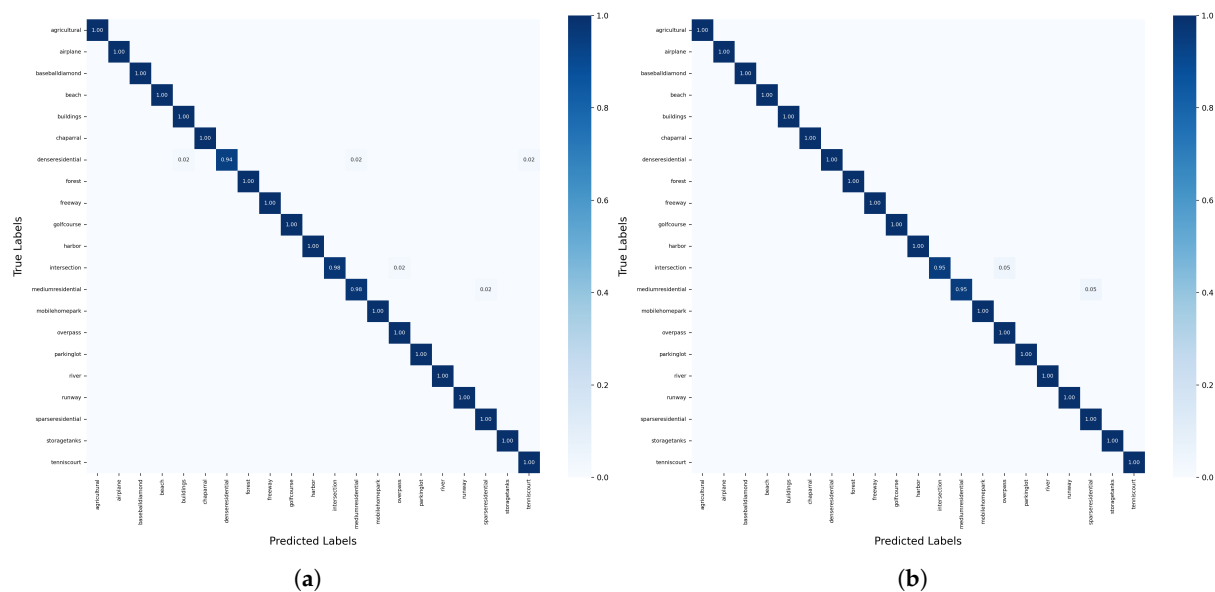


Figure 10. Confusion matrix for the UCM: (a) 50% training percentage; (b) 80% training percentage.

4.3.2. Experimental Results of AID

To further demonstrate the advantages of the FG-Swin KANsformer, evaluations are carried out on the AID, and the dataset is randomly partitioned into two training/testing ratios: 20%/80% and 50%/50%. As shown in Table 3, when using 20% of the data for training, FG-Swin KANsformer achieves an accuracy of 96.86%, outperforming STMSF by 0.71% and LDBST by 1.76%. When the training ratio is increased to 50%, the proposed model attains an accuracy of 97.93%, exceeding STMSF and LDBST by 0.42% and 1.09%, respectively.

In addition, the CMs illustrating the classification results of FG-Swin KANsformer on the AID are presented in Figure 11. When the training ratio is set to 50%, Figure 11b shows that only four categories, namely “resort”, “park”, “school”, and “square”, exhibit classification accuracies below 95%. Even with a reduced training ratio of 20%, Figure 11a indicates that 22 out of the 30 categories still achieve accuracies above 95%. These observations demonstrate that the FG-Swin KANsformer maintains high-quality classification performance even under significantly limited training data.

4.3.3. Experimental Results of NWPU Dataset

In contrast to the AID, the NWPU dataset is larger in scale and contains more complex scene categories, making the classification task more challenging. And two experimental settings are adopted by using 10% and 20% of the data for training. The overall accuracies of different methods on the NWPU dataset are reported in Table 4, and the conclusions are consistent with those obtained on the previous two datasets. When 10% of the data is used for training, FG-Swin KANsformer achieves an accuracy of 93.62%, outperforming

STMSF by 0.74% and SCViT by 0.9%. When the training ratio increases to 20%, the FG-Swin KANsformer reaches an accuracy of 95.27%, exceeding STMSF and SCViT by 0.32% and 0.61%, respectively. These outcomes illustrate the capability of the FG-Swin KANsformer on large-scale RS datasets. Overall, the experiments show that FG-Swin KANsformer consistently achieves the best classification performance on the UCM, AID, and NWPU datasets.

Table 3. Comparison of the OA(%) on the AID.

Method	20% Training Ratio	50% Training Ratio
GoogLeNet	83.44 ± 0.40	86.39 ± 0.55
VGG	92.76 ± 0.10	95.37 ± 0.09
ResNet	93.23 ± 0.12	94.12 ± 0.19
MSA-Network	93.53 ± 0.21	96.01 ± 0.43
PCNet	95.53 ± 0.16	96.76 ± 0.25
GCSANet	95.96 ± 0.38	97.53 ± 0.32
EMSCNet	95.13 ± 0.10	96.96 ± 0.10
MDRCN	93.64 ± 0.19	95.66 ± 0.18
PiT-S	90.51 ± 0.57	94.17 ± 0.36
EMTCAL	94.69 ± 0.14	96.41 ± 0.23
CSAT	92.55 ± 0.28	95.44 ± 0.17
SCViT	95.56 ± 0.17	96.98 ± 0.16
LDBST	95.10 ± 0.09	96.84 ± 0.20
STMSF	96.15 ± 0.16	97.51 ± 0.37
SceneFormer	96.14 ± 0.16	-
MSCN	95.86 ± 0.16	97.46 ± 0.12
ACTFormer	96.29 ± 0.19	97.56 ± 0.24
FG-Swin KANsformer (ours)	96.86 ± 0.16	97.93 ± 0.25

Bold denotes the best result.

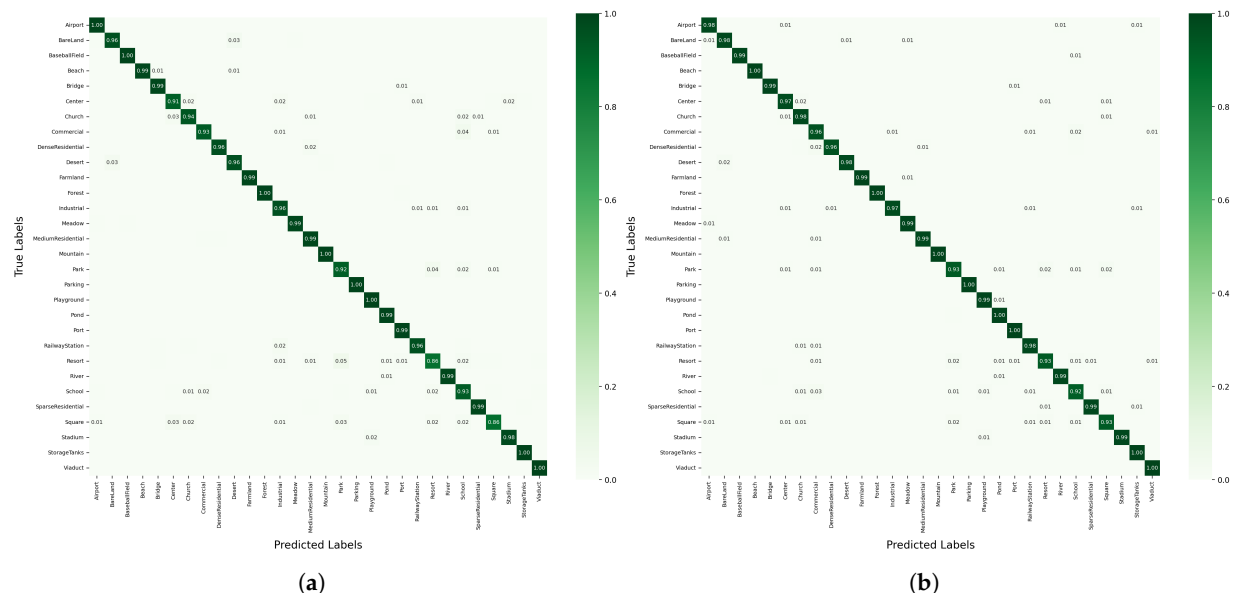


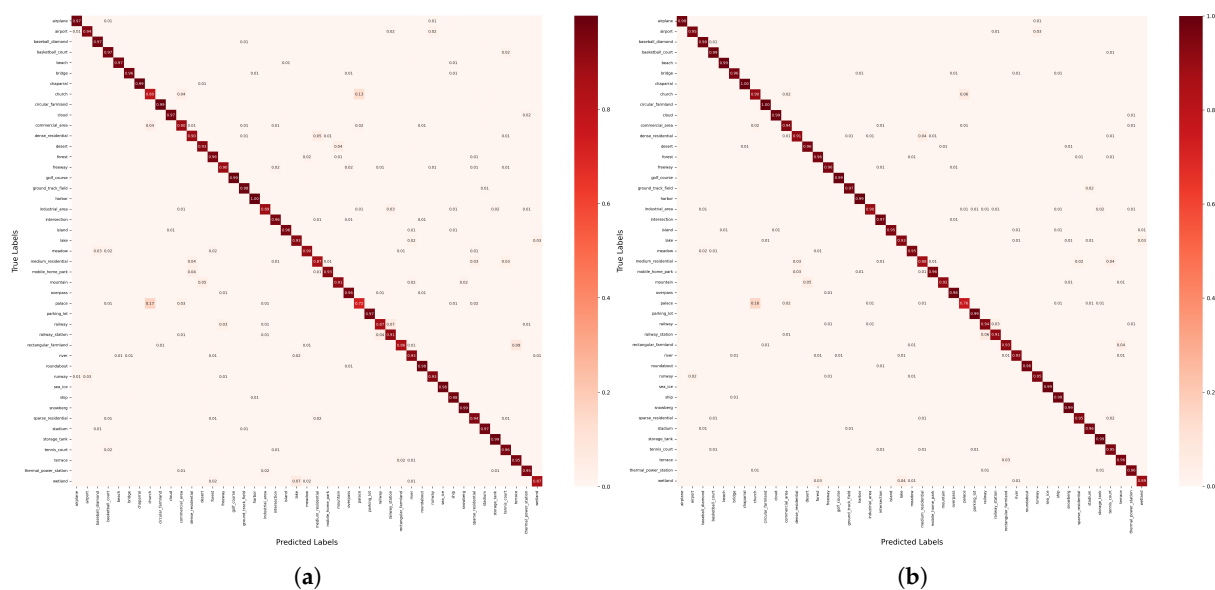
Figure 11. Confusion matrix for the AID: (a) 20% training percentage; (b) 50% training percentage.

In addition to OA, the classification performance of FG-Swin KANsformer on the NWPU dataset is further illustrated in Figure 12. When the training ratio is set to 20%, 42 out of the 45 categories achieve an accuracy higher than 90%. Even when the training ratio is reduced to 10%, Figure 12a indicates that 38 categories still maintain accuracies above 90%. These results confirm that the proposed FG-Swin KANsformer maintains strong classification performance even on the more complex and large-scale NWPU dataset.

Table 4. Comparison of the OA(%) on the NWPU dataset.

Method	10% Training Ratio	20% Training Ratio
GoogLeNet	76.19 ± 0.38	78.48 ± 0.26
VGG	87.14 ± 0.17	90.64 ± 0.14
ResNet	89.20 ± 0.14	92.12 ± 0.09
MSA-Network	90.38 ± 0.17	93.52 ± 0.21
PCNet	92.64 ± 0.13	94.59 ± 0.07
GCSANet	93.39 ± 0.39	93.95 ± 0.36
EMSCNet	92.16 ± 0.07	94.08 ± 0.20
MDRCN	91.59 ± 0.29	93.82 ± 0.17
PiT-S	85.85 ± 0.18	89.91 ± 0.19
EMTCAL	91.63 ± 0.19	93.65 ± 0.12
CSAT	89.70 ± 0.18	93.06 ± 0.16
SCViT	92.72 ± 0.04	94.66 ± 0.10
LDBST	90.83 ± 0.11	93.56 ± 0.07
STMSF	92.88 ± 0.16	94.95 ± 0.11
SceneFormer	92.51 ± 0.18	-
MSCN	92.64 ± 0.09	94.59 ± 0.11
ACTFormer	92.85 ± 0.14	94.76 ± 0.21
FG-Swin KANsformer (ours)	93.62 ± 0.15	95.27 ± 0.10

Bold denotes the best result.

**Figure 12.** Confusion matrix for the NWPU: (a) 10% training percentage; (b) 20% training percentage.

5. Discussion

We perform ablation experiments to evaluate the contribution of each module to the overall performance of the FG-Swin KANsformer. In addition, the Grad-CAM and the t-SNE are employed for visualization analysis to gain deeper insights into the feature extraction process and classification decision-making mechanism of the model in RSSC tasks.

5.1. Ablation Study

To thoroughly validate the enhancement effect of the FG-Swin KANsformer over the baseline model and examine the key roles of each module, we conduct ablation experiments using the Swin Transformer as the main network (denoted as ST). Specifically, we independently analyze the contributions of the DCT module, the GSFE module, and the

KAN-integrated Swin Transformer module. The specific experimental results can be seen in Table 5, where the training sets are respectively composed of 20% and 10% of the data from the AID and the NWPU dataset. The results indicate that when the Swin Transformer integrating DCT, GSFE, and KAN modules works collaboratively, the classification accuracy reaches the highest level. This confirms the effectiveness of the FG-Swin KANsformer as well as the individual contributions of its constituent modules.

Table 5. Results of different modules in FG-Swin KANsformer on the AID and the NWPU dataset.

ST	KAN	GSFE	DCT	AID-20%	NWPU-10%
✓				92.22 ± 0.14	88.60 ± 0.21
✓	✓			96.07 ± 0.18	93.02 ± 0.11
✓	✓	✓		96.61 ± 0.13	93.38 ± 0.14
✓	✓		✓	96.58 ± 0.19	93.40 ± 0.11
✓	✓	✓	✓	96.86 ± 0.16	93.62 ± 0.15

Bold denotes the best result.

5.2. Evaluation of Size of Models

Table 6 shows the number of parameters and FLOPs for different models, comparing FG-Swin KANsformer with the baseline model Swin Transformer, as well as two representative frontier models ViT and VGG-VD-16 [26]. FG-Swin KANsformer has a parameter of 69.37 M and a FLOP of 14.98 G, with an overall accuracy (OA) of 96.86% on the AID. In contrast, other models have larger parameters and computational complexity, but their accuracies do not surpass FG-Swin KANsformer. This result indicates that FG-Swin KANsformer has good classification accuracy while maintaining lower computational overhead.

Table 6. Parameters and FLOPs for different models.

Methods	Parameters	FLOPs	OA% (AID-20%)
ViT-Base	85.67 M	16.86 G	91.16 ± 0.41
VGG-VD-16	138.4 M	15.5 G	87.18 ± 0.29
SwinT-Base	86.71 M	15.17 G	92.22 ± 0.14
FG-Swin KANsformer(Ours)	69.37 M	14.98 G	96.86 ± 0.16

Bold denotes the best result.

5.3. Visualization Study

To visually demonstrate the superiority of FG-Swin KANsformer in feature extraction, the Grad-CAM [72] is employed for visual interpretation. This technique highlights the image areas that contribute most significantly to the model's predictions, enabling a more intuitive understanding of the areas attended to during classification. As shown in Figure 13, we select five sample types from the AID, namely bridge, center, church, stadium, and storage tank, for visual analysis. The comparison results indicate that, compared with the Swin Transformer, the proposed method achieves more accurate localization of key regions and demonstrates superior capability in extracting local features from RS images.

The t-SNE [73] is a dimension-based visualization technique that can effectively reflect the distribution characteristics and clustering patterns of various RS scenes in the embedding space. Therefore, we use this technology to visualize the classification results of the AID test set with a 50% training ratio and the UCM test set with an 50% training ratio. By applying t-SNE to the output features, we can intuitively compare the representation capabilities of FG-Swin KANsformer and Swin Transformer on the same datasets. As shown in Figure 14, compared with Swin Transformer, FG-Swin KANsformer exhibits a more compact sample distribution within each category and greater interclass separability. Furthermore, the boundaries between different categories appear more distinct, indicating

that FG-Swin KANsformer is able to learn more discriminative feature representations. This advantage contributes to improved performance in RSSC tasks.

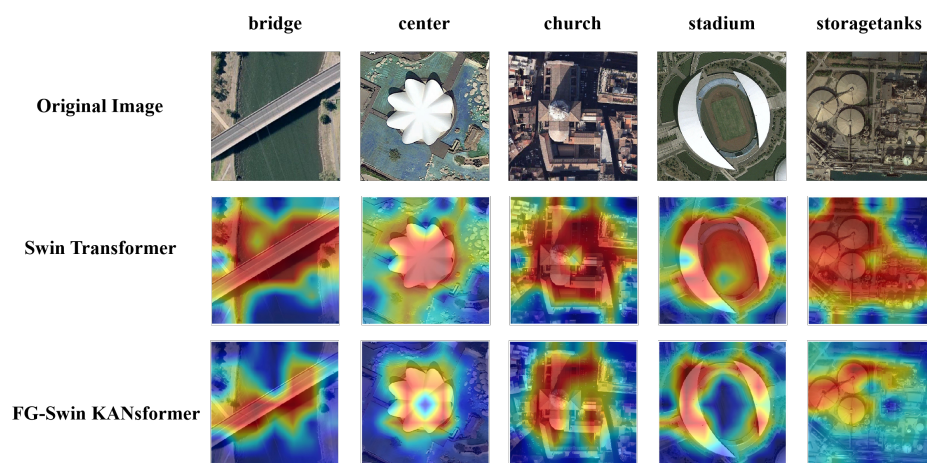


Figure 13. Grad-CAM visualization result. We compare the visualized feature map of FG Swin-KANsformer and the baseline model (Swin Transformer).

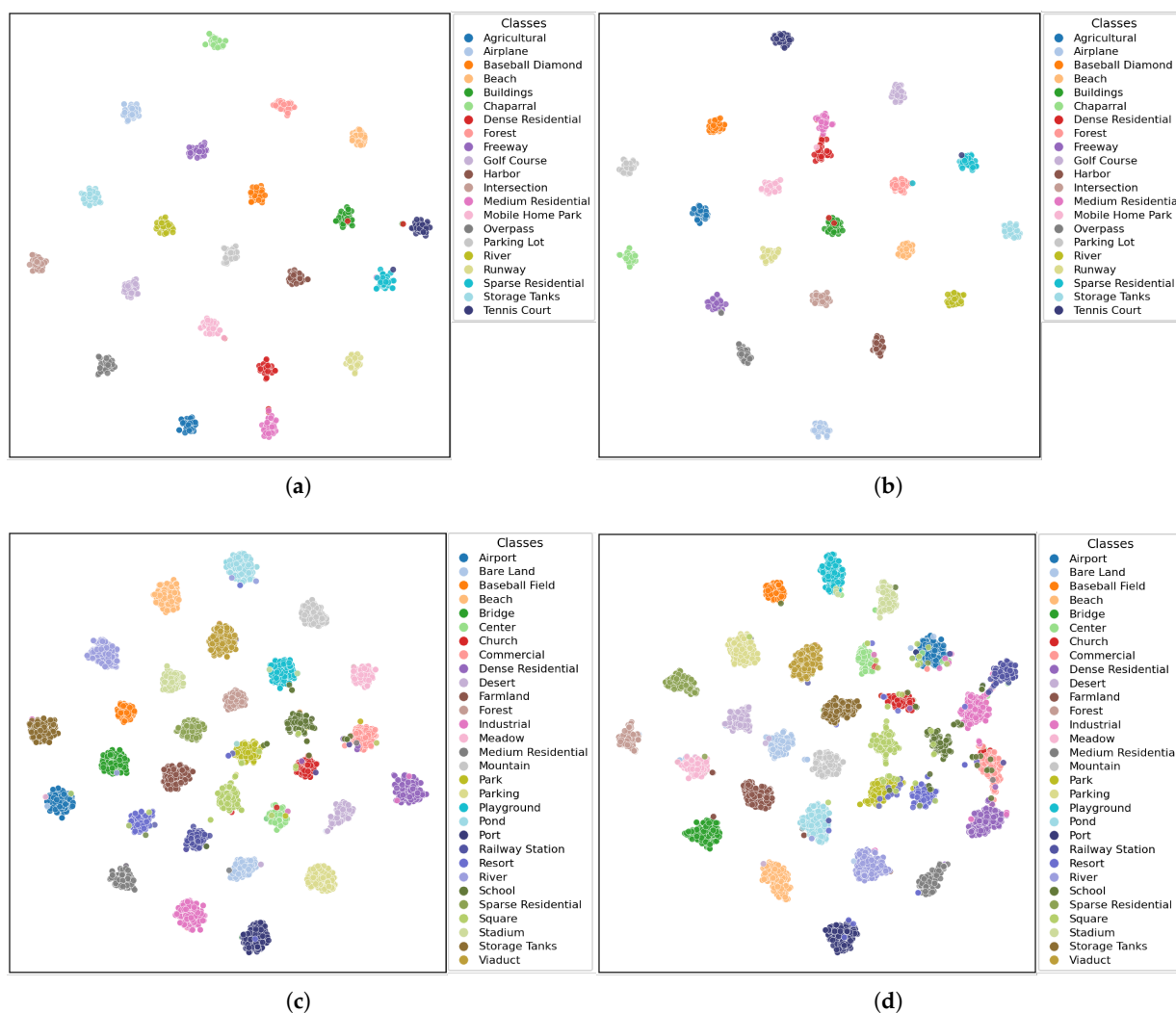


Figure 14. Visualization on the UCM dataset and AID using the t-SNE technique. (a) FG-Swin KANsformer on UCM dataset with 50% training ratio. (b) Swin Transformer on UCM dataset with 50% training ratio. (c) FG-Swin KANsformer on AID with 50% training ratio. (d) Swin Transformer on AID with 50% training ratio.

6. Conclusions

This article introduces a new universal RSSC method named FG-Swin KANsformer. To address key challenges such as complex backgrounds and large feature scale variations in RS, our model achieves collaborative enhancement from three perspectives: input feature construction, global semantic enhancement, and deep nonlinear modeling. In the input feature construction stage, the DCT module introduces frequency domain priors to balance texture details and global structural information. Meanwhile, the GSFE module enhances the ability to distinguish small targets and blurred boundaries through gradient and edge perception. In the deep semantic modeling stage, Swin Transformer integrated with KAN in the feedforward layer improves nonlinear function approximation and complex feature transformation, thereby boosting classification accuracy. Extensive experimental and visualization analyses demonstrate the superiority of FG-Swin KANsformer in multi-scale feature extraction and complex scene understanding, providing a new technical pathway for combining frequency domain perception with deep nonlinear modeling. Hence, our FG-Swin KANsformer offers a completely new and effective solution and approach for RSSC.

Despite exhibiting stable performance improvements, FG-Swin KANsformer still has certain limitations. Its application to ultra-high-resolution remote sensing images may lead to increased memory consumption and inference latency. Additionally, in scenarios where land cover types are highly similar or the background is extremely complex, a small number of misclassifications may still occur. And looking forward to the subsequent research work, it is possible to further extend this model by optimizing it for higher resolution and more complex RS scenarios. Additionally, we aim to explore its applicability to broader RS tasks in order to validate its generalization and universality. And we believe this research could advance Transformer-based methods toward wider applications in RS and foster the integration and development of RS technologies in fields such as environmental monitoring, smart city construction, and disaster early warning.

Author Contributions: Conceptualization, software, data curation, formal analysis, writing—original draft preparation, investigation, methodology, validation, X.Z.; software, formal analysis, investigation, methodology, validation, J.H.; writing—review and editing, formal analysis, investigation, methodology, validation, supervision, funding acquisition, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The relevant data supporting the conclusions of this study can be provided by the corresponding author upon reasonable request.

Acknowledgments: The authors express heartfelt gratitude to the anonymous reviewers and editors for their valuable suggestions and profound insights provided during the review process of this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE*. **2017**, *105*, 1865–1883. [\[CrossRef\]](#)
2. Han, W.; Zhang, X.H.; Wang, Y.; Wang, L.Z.; Huang, X.H.; Li, J.; Wang, S.; Chen, W.T.; Li, X.J.; Feng, R.Y.; et al. A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 87–113. [\[CrossRef\]](#)
3. Qiu, C.P.; Zhang, X.Y.; Tong, X.C.; Guan, N.Y.; Yi, X.D.; Ke, Y.; Zhu, J.J.; Yu, A.Z. Few-shot remote sensing image scene classification: Recent advances, new baselines, and future trends. *ISPRS J. Photogramm. Remote Sens.* **2024**, *209*, 368–382. [\[CrossRef\]](#)
4. Xu, K.J.; Huang, H.; Yuan, L.; Shi, G.Y. Multilayer feature fusion network for scene classification in remote sensing. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1894–1898. [\[CrossRef\]](#)

5. Wu, K.; Zhang, Y.Y.; Ru, L.X.; Dang, B.; Lao, J.W.; Yu, L.; Luo, J.W.; Zhu, Z.F.; Sun, Y.; Zhang, J.H.; et al. A semantic-enhanced multi-modal remote sensing foundation model for Earth observation. *Nat. Mach. Intell.* **2025**, *7*, 1235–1249. [\[CrossRef\]](#)
6. Zhu, Q.; Guo, X.; Deng, W.; Shi, S.; Guan, Q.F.; Zhong, Y.F.; Zhang, L.P.; Li, D. Land-use/land-cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 63–78. [\[CrossRef\]](#)
7. Kucharczyk, M.; Hugenholtz, C. Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities. *Remote Sens. Environ.* **2021**, *264*, 112577. [\[CrossRef\]](#)
8. Chen, Z.; Si, W.; Johnson, V.C.; Oke, S.A.; Wang, S.; Lv, X.; Tan, M.L.; Zhang, F.; Ma, X. Remote sensing research on plastics in marine and inland water: Development, opportunities and challenge. *J. Environ. Manag.* **2025**, *373*, 123815. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Cheng, G.; Xie, X.; Han, J.W.; Guo, L.; Xia, G.S. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [\[CrossRef\]](#)
10. Zhu, Q.; Sun, X.L.; Zhong, Y.F.; Zhang, L.P. High-resolution remote sensing image scene understanding: A review. In Proceedings of the 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 3061–3064.
11. Grigorescu, S.; Petkov, N.; Kruizinga, P. Comparison of texture features based on Gabor filters. *IEEE Trans. Image Process.* **2002**, *11*, 1160–1167. [\[CrossRef\]](#)
12. Tang, X.; Jiao, L.; Emery, W. SAR image content retrieval based on fuzzy similarity and relevance feedback. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 1824–1842. [\[CrossRef\]](#)
13. Mei, S.; Ji, J.; Hou, J.; Li, X.; Du, Q. Learning sensor-specific spatial-spectral features of hyperspectral images via convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4520–4533. [\[CrossRef\]](#)
14. Jiao, L.; Tang, X.; Hou, B.; Wang, S. SAR images retrieval based on semantic classification and region-based similarity measure for earth observation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3876–3891. [\[CrossRef\]](#)
15. Sedaghat, A.; Mohammadi, N. Uniform competency-based local feature extraction for remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 142–157. [\[CrossRef\]](#)
16. Tang, X.; Jiao, L.; Emery, W.J.; Liu, F.; Zhang, D. Two-stage reranking for remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5798–5817. [\[CrossRef\]](#)
17. Zhang, L.; Zhou, W.; Jiao, L. Wavelet support vector machine. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2004**, *34*, 34–39. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Peng, D.; Liu, X.; Zhang, Y.J.; Guan, H.; Li, Y.S.; Bruzzone, L. Deep learning change detection techniques for optical remote sensing imagery: Status, perspectives and challenges. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *136*, 104282. [\[CrossRef\]](#)
19. Goma, A.; Saad, O.M. Residual Channel-attention (RCA) network for remote sensing image scene classification. *Multimed. Tools Appl.* **2025**, *84*, 33837–33861. [\[CrossRef\]](#)
20. BAZI, Y.; Rahhal, M.M.; Alhichri, H.; Alajlan, N. Simple yet effective fine-tuning of deep CNNs using an auxiliary classification loss for remote sensing scene classification. *Remote Sens.* **2019**, *11*, 2908. [\[CrossRef\]](#)
21. Li, W.; Wang, Z.T.; Wang, Y.; Wu, J.Q.; Wang, J.; Jia, Y. Classification of high-spatial-resolution remote sensing scenes method using transfer learning and deep convolutional neural network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1986–1995. [\[CrossRef\]](#)
22. Wang, W.; Chen, Y.; Ghamisi, P. Transferring CNN with adaptive learning for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5533918. [\[CrossRef\]](#)
23. Wang, Q.; Liu, S.; Channusot, J.; Li, X. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 1155–1167. [\[CrossRef\]](#)
24. Lu, X.; Sun, H.; Zheng, X. A feature aggregation convolutional neural network for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7894–7906. [\[CrossRef\]](#)
25. Wang, X.; Wang, S.Y.; Ning, C.; Zhou, H.Y. Enhanced feature pyramid network with deep semantic embedding for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7918–7932. [\[CrossRef\]](#)
26. Xia, G.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [\[CrossRef\]](#)
27. Tang, X.; Zhang, X.R.; Liu, F.; Jiao, L. Unsupervised deep feature learning for remote sensing image retrieval. *Remote Sens.* **2018**, *10*, 1243. [\[CrossRef\]](#)
28. Wu, H.; Shi, C.P.; Wang, L.G.; Jin, Z. A cross-channel dense connection and multi-scale dual aggregated attention network for hyperspectral image classification. *Remote Sens.* **2023**, *15*, 2367. [\[CrossRef\]](#)
29. Shi, C.; Wu, H.; Wang, L. A feature complementary attention network based on adaptive knowledge filtering for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5527219. [\[CrossRef\]](#)
30. Roy, S.; Deria, A.; Hong, D.; Rasti, B.; Plaza, A.; Channusot, J. Multimodal fusion transformer for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5515620. [\[CrossRef\]](#)

31. Zhang, J.; Zhao, H.; Li, J. TRS: Transformers for remote sensing scene classification. *Remote Sens.* **2021**, *13*, 4143. [\[CrossRef\]](#)
32. Lv, P.; Wu, W.; Zhong, Y.F.; Du, F.; Zhang, L.P. SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4409512. [\[CrossRef\]](#)
33. Li, L.; Han, L.; Ye, Y.X.; Xiang, Y.M.; Zhang, T.Y. Deep learning in remote sensing image matching: A survey. *ISPRS J. Photogramm. Remote Sens.* **2025**, *225*, 88–112. [\[CrossRef\]](#)
34. Deng, P.; Xu, K.; Huang, H. When CNNs meet vision transformer: A joint framework for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8020305. [\[CrossRef\]](#)
35. Chen, X.; Ma, M.Y.; Li, Y.; Mei, S.H.; Han, Z.H.; Zhao, J. Hierarchical feature fusion of transformer with patch dilating for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4410516. [\[CrossRef\]](#)
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; p. 30.
37. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021; pp. 1–22.
38. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
39. Bazi, Y.; Bashmal, L.; Rahhal, M.M.; Dayii, R.; Ajlan, N. Vision transformers for remote sensing image classification. *Remote Sens.* **2021**, *13*, 516. [\[CrossRef\]](#)
40. Sha, Z.; Li, J. MITformer: A multiinstance vision transformer for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6510305. [\[CrossRef\]](#)
41. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, X.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
42. Haykin, S. *Neural Networks: A Comprehensive Foundation*; Prentice Hall PTR: Upper Saddle River, NJ, USA, 1994; p. 168.
43. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control. Signal.* **1989**, *2*, 303–314. [\[CrossRef\]](#)
44. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [\[CrossRef\]](#)
45. Lin, S.; Lyu, P.; Liu, D.; Tang, T.; Liang, X.; Song, A.; Chang, X.J. Mlp can be a good transformer learner. In Proceedings of the 2024 IEEE/CVF International Conference on Computer Vision, Seattle, WA, USA, 17–19 June 2024; pp. 19489–19498.
46. Guo, J.Y.; Han, K.; Wu, H.; Tang, Y.H.; Chen, X.H.; Wang, Y.H.; Xu, C. Cmt: Convolutional neural networks meet vision transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 12175–12185.
47. Liu, Z.; Wang, Y.X.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T.Y.; Tegmark, M. Kan: Kolmogorov-arnold networks. *arXiv* **2024**, arXiv:2404.19756. [\[PubMed\]](#)
48. Kolmogorov, A. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Transl. Am. Math. Soc.* **1963**, *2*, 55–59.
49. Braun, J.; Griebel, M. On a constructive proof of Kolmogorov’s superposition theorem. *Constr. Approx.* **2009**, *30*, 653–675. [\[CrossRef\]](#)
50. Ahmed, N.; Natarajan, T.; Rao, K. Discrete cosine transform. *IEEE Trans. Comput.* **2006**, *100*, 90–93. [\[CrossRef\]](#)
51. Mallat, S. Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans. Am. Math. Soc.* **1989**, *315*, 69–87.
52. Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; Catanzaro, B. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv* **2021**, arXiv:2111.13587.
53. Jin, L.; Song, Y.H.; Zhao, H.; Cao, J.Y.; Cheung, V.C.K.; Liao, W.H. Frequency-Aware Spatial-Temporal Attention Explainable Network for EEG Decoding. *IEEE J. Biomed. Health. Inf.* **2025**, *10*, 7175–7185. [\[CrossRef\]](#)
54. Huan, H.; Zhang, B. FDAENet: Frequency domain attention encoder-decoder network for road extraction of remote sensing images. *J. Appl. Remote Sens.* **2024**, *18*, 024510. [\[CrossRef\]](#)
55. Ravivarma, G.; Gavaskar, K.; Malathi, D.; Asha, K.G.; Ashok, B.; Aarthi, S. Implementation of Sobel operator based image edge detection on FPGA. *Mater. Today Proc.* **2021**, *45*, 2401–2407. [\[CrossRef\]](#)
56. Wang, C.; Yu, B.; Zhou, J. A learnable gradient operator for face presentation attack detection. *Pattern Recognit.* **2023**, *135*, 109146. [\[CrossRef\]](#)
57. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.

58. Cheng, G.; Li, Z.; Yao, X.; Guo, L.; Wei, Z. Remote sensing image scene classification using bag of convolutional features. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1735–1739. [[CrossRef](#)]
59. He, K.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
60. Zhao, Y.; Liu, J.; Yang, J.L.; Wu, Z. EMSCNet: Efficient multisample contrastive network for remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5605814. [[CrossRef](#)]
61. Chen, W.; Ouyang, S.; Tong, W.; Li, X.J.; Zheng, X.W.; Wang, L.Z. GCSANet: A global context spatial attention deep learning network for remote sensing scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1150–1162. [[CrossRef](#)]
62. Zhang, G.; Xu, W.; Zhao, W.; Huang, C.; Yk, E.N.; Chen, Y.; Su, J. A multiscale attention network for remote sensing scene images classification. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *14*, 9530–9545. [[CrossRef](#)]
63. Dai, W.; Shi, F.; Wang, X.; Xu, H.; Yuan, L.; Wen, X. A multi-scale dense residual correlation network for remote sensing scene classification. *Sci. Rep.* **2024**, *14*, 22197. [[CrossRef](#)]
64. Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; Oh, S. Rethinking spatial dimensions of vision transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 11936–11945.
65. Tang, X.; Li, M.; Ma, J.; Zhang, X.; Liu, F.; Jiao, L. EMTCAL: Efficient multiscale transformer and cross-level attention learning for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5626915. [[CrossRef](#)]
66. Guo, J.; Jia, N.; Bai, J. Transformer based on channel-spatial attention for accurate classification of scenes in remote sensing image. *Sci. Rep.* **2022**, *12*, 15473. [[CrossRef](#)]
67. Zheng, F.; Lin, S.; Zhou, W.; Huang, H. A lightweight dual-branch swin transformer for remote sensing scene classification. *Remote Sens.* **2023**, *15*, 2865. [[CrossRef](#)]
68. Duan, Y.; Song, C.; Zhang, Y.F.; Cheng, P.Y.; Mei, S.H. STMSF: Swin Transformer with Multi-Scale Fusion for Remote Sensing Scene Classification. *Remote Sens.* **2025**, *17*, 668. [[CrossRef](#)]
69. Xie, C.; Zhao, S.; Ye, S.; Fei, Y.Q.; Dai, X.Y.; Tan, Y.P. ACTFormer: A Transformer Network with Attention and Convolutional Synergy for Remote Sensing Scene Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 18674–18687. [[CrossRef](#)]
70. Ma, J.J.; Jiang, W.; Tang, X.; Zhang, X.R.; Liu, F.; Jiao, L. Multiscale sparse cross-attention network for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5605416. [[CrossRef](#)]
71. Tong, L.; Liu, J.; Du, B. SceneFormer: Neural architecture search of Transformers for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 3000415. [[CrossRef](#)]
72. Selvaraj, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Mlp can be a good transformer learner. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 618–626.
73. Maaten, L.; Hinton, H. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.