GeoSwin: Hierarchical Representation Learning and a Large-Scale Earth Observation Benchmark

Anonymous CVPR submission

Paper ID *****

1. Abstract

Land Use and Land Cover (LULC) classification is essential for environmental monitoring, urban planning, and resource management. However, existing benchmarks are often limited in scale, category diversity, and long-tailed distributions, which restrict generalization to real-world remote sensing scenarios characterized by seasonal variations, heterogeneous landscapes, and severe class imbalance. We introduce GeoLULC, a large-scale benchmark consisting of 73,988 satellite images at 224×224 resolution across 73 land-cover categories. The dataset emphasizes fine-grained intra-class variability and explicitly includes rare and visually complex categories to reduce bias toward dominant classes and to better reflect real-world distributions. We further present GeoSwin, a transformer-based architecture tailored for LULC classification. GeoSwin adopts a Swin Transformer backbone to capture hierarchical local–global dependencies and integrates multi-dimensional feature attention module to improve discrimination in heterogeneous scenes with high intra-class variance. Extensive experiments demonstrate that GeoSwin outperforms strong convolutional and transformer-based baselines on GeoLULC. Comprehensive ablation studies analyze the contribution of each architectural component. Cross-dataset evaluation on RSSCN7, EuroSAT, SIRI-WHU, UC-Merced, AID, and NWPU-RESISC45 further confirms the robustness and generalization capability of the proposed approach.

2. Introduction

Land use and land cover classification plays a crucial role in several applications, including environmental monitoring, urban planning, resource management, and climate change modeling [10, 57]. The emergence of multispectral and hyperspectral data from satellites such as Sentinel-2, Landsat-8, and MODIS, as well as UAVs and LiDAR, has revolutionized how we analyze and classify land cover. However, as remote sensing data become increasingly available, challenges such as spectral ambiguity among visually similar land-cover types, seasonal variations, class imbalance, and high computational costs persist. These challenges necessitate the development of more advanced models capable of capturing both spatial and spectral features with high accuracy and efficiency.

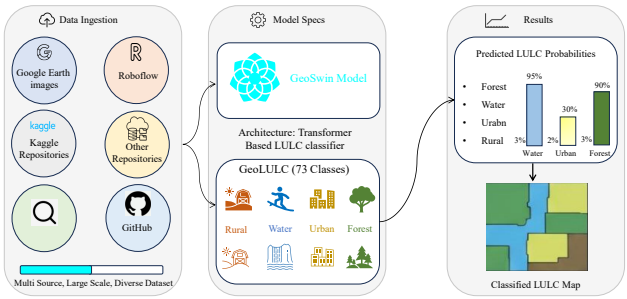

Figure 1. High-level overview of the proposed framework.

Traditional machine learning (ML) methods, including Support Vector Machines (SVMs), Random Forests (RFs), and K-Nearest Neighbors (KNNs), have long been used for LULC classification. As demonstrated by Saini et al. [36] and Alshari et al. [6], these methods have proven effective, especially in hybrid models. Ensemble approaches, such as those used at UC-Merced [16], have also shown strong performance. Despite their success, these ML models face significant limitations, such as the need for manual feature extraction, susceptibility to noise, and poor generalization across diverse regions. Furthermore, they often struggle to handle large-scale, high-dimensional data effectively.

In contrast, deep learning (DL) models, particularly convolutional neural networks (CNNs) and transformer-based architectures, have gained widespread popularity due to their ability to automatically learn robust features from raw data. Models like ResNet-50 and VGG-16 have been successful in LULC classification [15]. More recently, Vision Transformers (ViT) and TinyViT [51] have outperformed traditional CNNs, especially in capturing long-range dependencies and contextual information. Recent methods includes, methods includes Optimal Guidance Whale Optimization Algorithm [50], hybrid CNN-Transformer models

[4, 18], ExViT [59], and ForestViT [28]. Despite recent progress, several fundamental challenges remain in LULC classification. First, existing datasets often exhibit severe class imbalance, where dominant land-cover categories are overrepresented while rare or complex classes are underrepresented. Such skewed distributions result in biased models with poor performance on minority categories. Moreover, many benchmarks contain a limited number of classes and samples, restricting diversity and reducing their suitability for large-scale and cross-regional generalization. Second, conventional machine learning and deep learning approaches frequently rely on generic pre-trained backbones that are not optimized for the unique spectral spatial characteristics of remote sensing imagery. As a result, these models struggle to capture long-range dependencies and multi-scale contextual information in heterogeneous environments. Third, LULC classification requires distinguishing between visually similar categories with subtle structural and textural differences. Standard architectures show limited capability in modeling such fine-grained variations, particularly in complex landscapes where contextual reasoning is essential.

To address these challenges, we introduce the GeoLULC dataset and the GeoSwin model. A high-level overview of the proposed framework is illustrated in Fig. 1 and the main contributions of this work are as follows:

1. We introduce the GeoLULC dataset, a large-scale benchmark comprising 73 land-cover classes and 73,988 images. The dataset explicitly incorporates rare and visually complex categories to mitigate class imbalance and improve representativeness across diverse environments, enabling more reliable evaluation and generalization.
2. We present GeoSwin, a transformer-based model tailored for LULC classification. It leverages a Swin Transformer backbone to capture hierarchical local–global dependencies and generate robust feature representations suitable for large-scale and heterogeneous land-cover data.
3. We integrate a Geo-inspired Multi-Dimensional Feature Attention (MDFA) mechanism that jointly refines feature representations across complementary dimensions, enhancing fine-grained discrimination and improving the differentiation of visually similar land-cover categories in heterogeneous scenes.
4. Extensive experiments, ablation studies, and cross-dataset evaluations demonstrate the effectiveness, robustness, and generalization capability of the proposed dataset and model.

3. Related work

LULC has primarily relied on conventional ML methods such as support vector machines, random forests, and K-Nearest neighbors. These methods, as demonstrated by Saini et al. [36] and Alshari et al. [6], have been effective, with hybrid models showing improved accuracy. Ensemble-based frameworks [16] have also proven successful, particularly on datasets like UC-Merced. However, these ML methods face limitations, including the need for manual feature engineering, susceptibility to noise, and limited generalization across regions. Deep learning based models such Convolutional Neural Networks (CNNs) and Transformers have significantly advanced LULC classification by automating feature extraction and handling complex, high-dimensional data. CNN models, such as ResNet-50 and VGG-16, have been widely used with great success, as shown by studies like Fahmi et al. [15]. More recently, transformer-based architectures such as ViT and TinyViT [51] have outperformed traditional CNNs that require long-range dependencies and contextual understanding. Approaches such as the Optimal Guidance Whale Optimization Algorithm (OG-WOA) [50] and hybrid CNN-Transformer models [4, 18] have shown promise for enhancing model stability and generalization. Additionally, advanced transformer models such as ExViT [59] and ForestViT [28] have introduced novel approaches to improve performance on multi-label LULC tasks. However, issues such as dataset imbalance and computational inefficiency persist. Our proposed GeoLULC dataset and GeoSwin model aim to overcome these limitations, providing a comprehensive and efficient solution for large-scale LULC classification.

4. Methodology

Our proposed GeoSwin framework employs a Swin Transformer backbone tailored for large-scale land-cover classification. The backbone is augmented with a Geo-inspired Multi-Dimensional Feature Attention mechanism to enhance feature discriminability across heterogeneous scenes. The overall processing pipeline is illustrated in Fig. 2.

4.1. Problem Formulation

Given a set of input images $\mathcal{X} = \{x_i\}_{i=1}^N$ and corresponding labels $\mathcal{Y} = \{y_i\}_{i=1}^N$ with $y_i \in \{1, \ldots, C\}$ for C land cover classes, our goal is to learn a function $f_\theta : \mathcal{X} \to \mathbb{R}^C$ such that the predicted probability vector $\hat{\mathbf{y}}_i = f_\theta(x_i)$ accurately reflects the true class label y_i.

4.2. GeoSwin Transformer Backbone

We adopt a Swin Transformer backbone [29] specifically tailored for LULC classification. The backbone partitions input images into non-overlapping patches, embeds them into a high-dimensional space, and applies hierarchical self-attention within shifted windows.

Patch Embedding. Let an input image be $x_i \in \mathbb{R}^{H \times W \times 3}$. We divide it into M non-overlapping patches

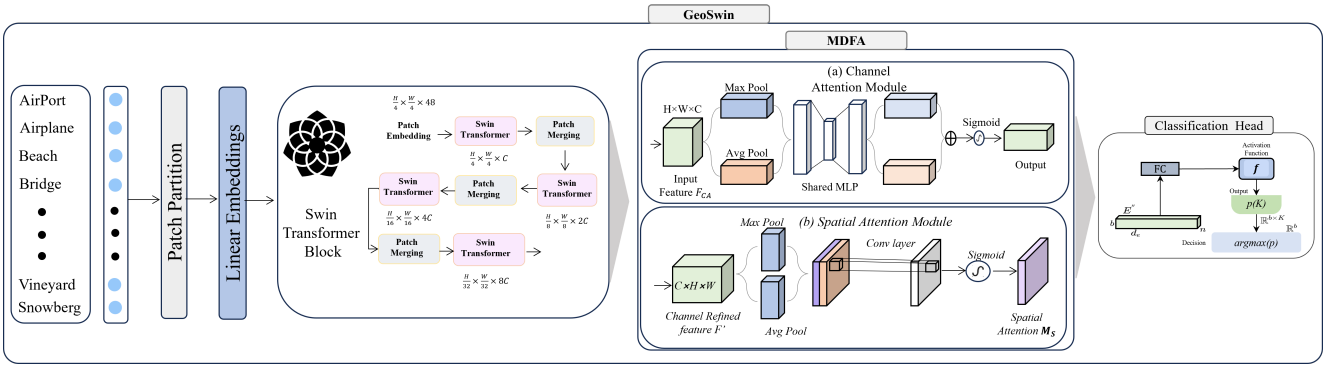

Figure 2. High level framework of the proposed GeoSwin architecture

of size $P \times P$:

$$\mathbf{x}_p^j \in \mathbb{R}^{P^2 \cdot 3}, \quad j = 1, \ldots, M, \quad M = \frac{HW}{P^2}.$$

Each patch is linearly projected to a d-dimensional embedding:

$$\mathbf{z}_0^j = \mathbf{x}_p^j \mathbf{E} + \mathbf{b}, \quad \mathbf{E} \in \mathbb{R}^{3P^2 \times d}, \mathbf{b} \in \mathbb{R}^d.$$

We also add learnable 2D positional encodings $\mathbf{p}_j \in \mathbb{R}^d$:

$$\mathbf{z}_0^j \leftarrow \mathbf{z}_0^j + \mathbf{p}_j.$$

Shifted Window Multi-Head Self-Attention. Let the feature map at layer l be $\mathbf{Z}^{(l)} \in \mathbb{R}^{M \times d}$. The features are partitioned into non-overlapping windows of size $W \times W$, forming K windows $\{\mathbf{Z}_k^{(l)}\}_{k=1}^K$:

$$\mathbf{Z}_k^{(l)} \in \mathbb{R}^{W^2 \times d}.$$

Within each window, multi-head self-attention is computed:

$$\mathrm{Attn}(\mathbf{Q}_k, \mathbf{K}_k, \mathbf{V}_k) = \mathrm{Softmax}\Big(\frac{\mathbf{Q}_k \mathbf{K}_k^\top}{\sqrt{d_k}} + \mathbf{B}_k\Big)\mathbf{V}_k,$$

where $\mathbf{Q}_k, \mathbf{K}_k, \mathbf{V}_k \in \mathbb{R}^{W^2 \times d_k}$, $d_k = d/h$, h is the number of attention heads, and $\mathbf{B}_k$ is the relative positional bias for window k.

To enhance cross-window interactions, windows are shifted by $\lfloor W/2 \rfloor$ pixels in alternating layers. Let the shifted feature be $\tilde{\mathbf{Z}}^{(l)}$:

$$\tilde{\mathbf{Z}}^{(l)} = \mathrm{Shift}(\mathbf{Z}^{(l)}, s = \lfloor W/2 \rfloor),$$

and attention is computed similarly in shifted windows.

Transformer Block. Each layer contains a residual self-attention block followed by a feed-forward network (FFN):

$$\mathbf{Z}^{(l)'} = \mathbf{Z}^{(l)} + \mathrm{Attn}(\mathrm{LN}(\mathbf{Z}^{(l)})), \tag{1}$$

$$\mathbf{Z}^{(l+1)} = \mathbf{Z}^{(l)'} + \mathrm{FFN}(\mathrm{LN}(\mathbf{Z}^{(l)'})), \tag{2}$$

where $\mathrm{LN}(\cdot)$ denotes Layer Normalization and FFN consists of two linear layers with GELU activation:

$$\mathrm{FFN}(\mathbf{x}) = \mathbf{W}_2\, \mathrm{GELU}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2.$$

Feature Aggregation. The final layer tokens $\mathbf{Z}_L \in \mathbb{R}^{M \times d}$ are aggregated via global average pooling to form the backbone feature vector:

$$\mathbf{f}_{\text{backbone}} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{Z}_L^j \in \mathbb{R}^d.$$

This hierarchical patch-wise attention mechanism captures both local fine-grained patterns and global contextual dependencies, providing a rich feature representation for downstream MDFA refinement and classification.

4.3. Multi-Dimensional Feature Attention

To improve discriminative representation learning, we introduce a Multi-Dimensional Feature Attention (MDFA) mechanism that refines features along different dimensions. Given an intermediate feature tensor $\mathbf{F} \in \mathbb{MDFAR}^{C \times H \times W}$ extracted from the backbone, MDFA sequentially applies channel attention, spatial attention, and transformer-level fusion.

4.3.1. Contextual Attention

Contextual attention models inter-channel dependencies by aggregating global contextual descriptors. We first compute channel-wise statistics using global average pooling and global max pooling:

$$\mathbf{f}_{\text{avg}}^c = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{F}_{c,i,j}, \tag{3}$$

$$\mathbf{f}_{\text{max}}^{c} = \max_{i,j} \mathbf{F}_{c,i,j}. \tag{4}$$

These descriptors are passed through a shared multi-layer perceptron (MLP):

$$\mathbf{M}_c = \sigma\left(\text{MLP}(\mathbf{f}_{\text{avg}}) + \text{MLP}(\mathbf{f}_{\text{max}})\right), \tag{5}$$

where $\mathbf{M}_c \in \mathbb{R}^C$ is the channel attention map and σ denotes the sigmoid function. The refined feature map is obtained via channel-wise reweighting:

$$\mathbf{F}_c = \mathbf{M}_c \odot \mathbf{F}. \tag{6}$$

4.3.2. Structural Attention

Structural attention emphasizes informative regions within the feature map. Channel-compressed spatial descriptors are computed as:

$$\mathbf{F}_{\text{avg}}^{s}(i,j) = \frac{1}{C}\sum_{c=1}^{C} \mathbf{F}_c(c,i,j), \tag{7}$$

$$\mathbf{F}_{\text{max}}^{s}(i,j) = \max_{c} \mathbf{F}_c(c,i,j). \tag{8}$$

The descriptors are concatenated and convolved to produce a spatial attention map:

$$\mathbf{M}_s = \sigma\left(\text{Conv}_{7\times7}\left([\mathbf{F}_{\text{avg}}^{s}; \mathbf{F}_{\text{max}}^{s}]\right)\right), \tag{9}$$

where $\mathbf{M}_s \in \mathbb{R}^{H\times W}$. Spatial refinement is then applied:

$$\mathbf{F}_{cs} = \mathbf{M}_s \odot \mathbf{F}_c. \tag{10}$$

4.3.3. Attention Fusion

Let $\mathbf{z}_L \in \mathbb{R}^{N\times d}$ denote the token representation from the final transformer layer, where N is the number of tokens and d is the embedding dimension. We further refine the representation through a residual multi-head self-attention (MHSA) fusion block:

$$\mathbf{z}_L^{(1)} = \mathbf{z}_L + \text{MHSA}(\text{LayerNorm}(\mathbf{z}_L)), \tag{11}$$

$$\mathbf{z}'_L = \mathbf{z}_L^{(1)} + \text{FFN}(\text{LayerNorm}(\mathbf{z}_L^{(1)})). \tag{12}$$

The multi-head self-attention is defined as:

$$\text{MHSA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)\mathbf{W}^O, \tag{13}$$

$$\text{head}_i = \text{Softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d_k}}\right)\mathbf{V}_i. \tag{14}$$

4.3.4. MDFA Formulation

The overall refinement process can be summarized as:

$$\mathbf{F}_{\text{MDFA}} = \mathcal{A}_{\text{Transformer}}\big(\mathcal{A}_{\text{Structural}}(\mathcal{A}_{\text{Contextual}}(\mathbf{F}))\big), \tag{15}$$

MDFA enables hierarchical refinement across complementary feature dimensions, improving discrimination of visually similar land-cover categories in heterogeneous scenes.

4.4. Classification Head

The final feature vector $\mathbf{f} \in \mathbb{R}^d$ from GeoSwin is passed through a two-layer multi-layer perceptron (MLP) with dropout and GELU activation:

$$\mathbf{z} = \text{Linear}_2\big(\text{GELU}(\text{Dropout}(\text{Linear}_1(\mathbf{f})))\big) \in \mathbb{R}^C, \tag{16}$$

producing logits over the C land-cover classes. The predicted class probabilities are obtained using the Softmax function:

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{z}), \quad \hat{y}_c = \frac{\exp(z_c)}{\sum_{k=1}^{C}\exp(z_k)}, \quad c = 1,\ldots,C. \tag{17}$$

We employ label-smoothed cross-entropy for regularization:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}^{\text{smooth}} \log \hat{y}_{i,c}, \quad y_{i,c}^{\text{smooth}} = (1-\epsilon)\mathbf{1}_{[c=y_i]} + \frac{\epsilon}{C}, \tag{18}$$

where ϵ is the smoothing factor.

Optimization is performed using AdamW with cosine annealing warm restarts:

$$\theta_{t+1} = \theta_t - \eta_t \frac{\partial \mathcal{L}}{\partial \theta_t}, \quad \eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\Big(1 + \cos\frac{t\pi}{T_0}\Big), \tag{19}$$

where η_t is the learning rate at step t, and T_0 is the restart period.

5. Dataset Decsription

Table 2 provides a summary comparison of different publicly available datasets, along with the proposed GeoLULC dataset. The comparison includes key attributes such as the number of classes, the number of samples, and the image resolution for each dataset. The FLAIR dataset contains 5 classes with a limited number of samples, though the resolution is unspecified. The Open Earth dataset includes 8 classes, with an unspecified number of samples and resolution. The DeepGlobe dataset has 7 classes and 1,146 samples, but its resolution is not provided. The RSSCN7 dataset contains 7 classes and 5,000 samples, with a resolution of 256×256 pixels. The SIRI WHU dataset contain

Table 1. Class Distribution in the GeoLULC-73 Dataset

Class	Samples	Class	Samples	Class	Samples
Agriculture	1000	Airplane	1018	Airport	1008
AnnualCrop	1000	BareLand	1000	Barn	1019
BaseballField	1020	BasketballCourt	1020	Beach	1003
Bridge	1018	Chaparral	1018	Church	1015
Cloudy	1009	CommercialArea	1001	DenseResidential	1018
Desert	1000	Farmland_Circular	1017	Farmland_Eroded	1014
Farmland_Rectangular	1019	Forest	1001	Freeway	1011
GolfCourse	1019	Grassland	1019	GreenhouseFarm	1013
Harbor	1000	HerbaceousVegetation	1008	Highway	1012
Industrial	1017	Intersection	1016	Island	1018
Lake	1013	Landmark	1020	Meadow	1017
MediumResidential	1016	MobilehomePark	1020	Mountain	1020
Overpass	1019	Palace	1016	Park	1014
ParkingLot	1014	Parkway	1016	Pasture	1018
PermanentCrop	1016	Pond	1012	Port	1008
PowerSubstation	1015	Railway	1017	RailwayStation	1016
Residential	1019	River	1014	Roundabout	1016
Runway	1011	Rural	1014	SeaIce	1010
Ship	1019	ShippingYard	1018	Snowberg	1009
SparseResidential	1013	Stadium	1019	StorageTanks	1020
SwimmingPool	1014	TennisCourt	1012	Terrace	1012
ThermalPowerStation	1013	TrackField	1012	TransmissionTower	1010
Urban	1020	UrbanLowAltitude	1019	Vineyard	1019
Water	1008	Wetland	1014	WindTurbine	1004
WindTurbineLowAltitude	1011				
Total Samples: 73,988					

10 classes, an unspecified number of samples, and a resolution of 256×256 pixels. EuroSAT features 10 classes and 27,000 samples, with a resolution of 64×64 pixels. Land CoverNet includes 12 classes and 200,000 samples, though the resolution is unspecified. The UCM (University of California, Merced) dataset has 21 classes and 2,100 samples, with a resolution of 256×256 pixels. The AID (Aerial Image Dataset) contains 30 classes, 10,000 samples, and a resolution of 600×600 pixels. BigEarthNet includes 43 classes and 590,000 samples, with an unspecified resolution. The NWPU dataset has 45 classes and 45,000 samples, with a resolution of 224×224 pixels.

Many datasets have been used previously in land use and land cover classification, but these datasets typically had a limited number of classes, which limiting the granularity and depth of the classification. Our proposed GeoLULC dataset significantly increases the number of classes, offering 73 distinct classes and 73,988 samples in total as given in Table 1. With a resolution of 224×224 pixels, this dataset provides a much more detailed and comprehensive framework for classification. By including a broader range of land-use and land-cover categories, GeoLULC enables more precise, nuanced analyses, making it a valuable comparatively. The increased number of classes enhances the ability to capture a diverse variety of land features, contributing to more accurate and reliable model evaluations in real-world applications. ***The GeoLULC dataset will be made publicly available, and the download link will be provided upon paper acceptance.***

6. Results

We evaluate the proposed GeoSwin framework on the newly introduced GeoLULC dataset, leveraging a workstation equipped with an NVIDIA RTX 3090 GPU, 64 GB of RAM, and an Intel i9 CPU. All models are implemented in PyTorch, and experiments are conducted with consistent training protocols, including data augmentation, AdamW optimization, and label-smoothed cross-entropy loss. To benchmark the performance of GeoSwin and demonstrate its generalization capability, we compare against several state-of-the-art methods on widely used publicly available LULC datasets. These datasets include RSSCN7 [67], EuroSat [24], SIRI-WHU [64], UCMerced Land Use [58], AID [53], and NWPU RESISC45 [11]. To evaluate performance, we used the following key metrics: Accuracy

Table 2. Summary of publicly available and specialized land use and land cover (LULC) datasets, including the proposed GeoLULC dataset, highlighting the number of classes, samples, and resolution for each dataset.

Year	Dataset	Classes	Samples	Resolution
2023	Open Earth [54]	8	...	....
2018	DeepGlobe [13]	7	1,146	....
2015	RSSCN7 [67]	7	5,000	256×256
2015	SIRI WHU [64]	10	...	256×256
2019	EuroSAT [24]	10	27,000	64×64
2010	WHU-RS19 [52]	19	1,005	....
2010	UCMercedLandUse [58]	21	2,100	256×256
2017	AID [53]	30	10,000	600×600
2018	PatternNet [65]	38	30,400	....
2019	BigEarthNet [42]	43	590k	....
2017	NWPU RESISC-45 [11]	45	31,500	224×224
2020	MLRSNet [35]	46	109,160	224×224
2026	**proposed GeoLULC**	**73**	**73,988**	**224×224**

(Acc), Precision (P), Recall (R), and F1-Score (F1) for fair comaprison with baselines.

6.1. Datasets

We utilized several remote sensing datasets to evaluate the performance of GeoSwin architecture with baselines. These datasets include RSSCN7 [67], EuroSat [24], Siri-WHU [64], UCMerced Land Use [58], AID [53], and NWPU RESISC45 [11]. The RSSCN dataset comprises 7 classes and 5000 images, representing specific geographical scenes, thereby enabling more comprehensive evaluation under environmental changes. EuroSat, a small land-use classification dataset, includes 10 classes and approximately 27,000 images. This dataset is frequently used for lidar-based land-use classification models because it contains a diverse set of objects across multiple classes. The Siri-WHU dataset contains 12 classes and 2400 images, providing greater variation in land-use classes and enriching the dataset while ensuring a wider range of land-cover scenarios for model evaluation.

The UCMerced Land Use dataset contains 21 land-cover classes, with 100 images per class, for a total of 2,100 images. This collection is carefully organized, meets academic standards, and offers an invaluable resource for land-cover classification. The AID dataset contains 10,000 images across 30 classes, spanning numerous land-cover categories. To address issues of small-class diversity and generalization in previous datasets, we also used a subset of the NWPU RESISC45 dataset. This dataset includes 45 classes (31,500 images) covering diverse, imbalanced land-cover categories. It is appropriate for evaluating model performance in complex, multi-class land-use situations.

Additionally, we used our proposed dataset, **GeoLULC**, which contains 73 classes and 73,988 images, further enhancing the model's evaluation with a more extensive and varied land-use classification framework. All datasets were split into 70% for training, 20% for validation, and 10% for testing to provide a robust model evaluation of the predictor.

6.2. Ablation Study on the GeoLULC Dataset

To systematically assess the performance, we perform an extensive ablation study on the GeoLULC dataset. Table 3 reports the performance of nineteen widely used convolutional architectures, such as lightweight networks (MobileNetV2, EfficientNet variants), deep residual networks (ResNet, DenseNet) and advanced convolutional designs (ConvNeXt). All baseline methods were re-implemented and evaluated on GeoLULC under identical experimental settings to ensure a fair comparison.

The results show that DenseNet architectures dominate the traditional baselines, with DenseNet121 achieving the highest accuracy at 86.2%, followed by DenseNet169 (85.4%) and DenseNet201 (85.2%). This suggests that dense connectivity, which facilitates feature reuse through direct layer-wise connections, is particularly effective for capturing the fine-grained textural patterns in remote sensing imagery. Conversely, EfficientNetV2 variants perform poorly (e.g., EfficientNetV2B3 at 5.7% accuracy), indicating that scaling strategies optimized for natural image datasets do not directly transfer to geospatial classification tasks.

In GeoSwin, we evaluate three model scales: Tiny (768 tokens), Base (1024 tokens), and Large (1536 tokens), both with and without MDFA attention. Incorporating attention consistently improves performance across all scales, with GeoSwinLarge w/ MDFA achieving the best results (accuracy 92.3%, F1 92.2%), demonstrating that the proposed attention mechanisms enhance feature discriminability for heterogeneous land-cover classes. Even the smaller GeoSwinTiny model benefits substantially from MDFA, improving accuracy from 87.8% to 91.3%, showing that the attention modules efficiently capture long-range dependencies without requiring excessive model capacity. These observations highlight that GeoSwin's hierarchical feature extraction and attention-driven refinement offer a robust solution for large-scale land-cover classification, outperforming all re-implemented baselines.

6.3. Comparative Evaluation on Public Remote Sensing Datasets

To further validate the effectiveness of GeoSwin, we conduct a comparative evaluation across six widely-used remote sensing datasets: UC-Merced, AID, EuroSAT, NWPU-RESISC45, SIRI-WHU, and RSSCN7. Tables 4 and 5 report precision, recall , F1-score, and overall accuracy for various state-of-the-art CNN, Transformer, and hybrid architectures alongside GeoSwin in Tiny, Base, and Large configurations. Across all datasets, GeoSwin consis-

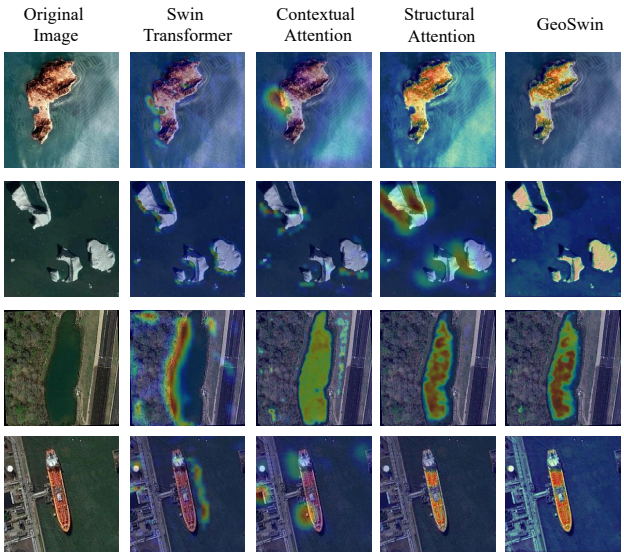

Figure 3. Comparison of attention map activations showing that GeoSwin provides more precise and localized results than Swin Transformer and its attention variants.

Table 3. Ablation study on the proposed GeoLULC dataset, comparing the performance of 19 baseline convolutional architectures and GeoSwin variants (Tiny, Base, Large) with and without the proposed MDFA module.

Model	Acc (%)	P (%)	R (%)	F1 (%)
EfficientNetV2B3	0.057	0.040	0.058	0.048
EfficientNetV2M [49]	0.121	0.092	0.121	0.105
ResNet101 [22]	0.299	0.282	0.300	0.291
ResNet50 [22]	0.329	0.317	0.326	0.322
EfficientNetV2S [49]	0.354	0.360	0.349	0.355
ConvNetXtSmall [30]	0.432	0.435	0.433	0.434
ConvNetXtBase [30]	0.540	0.580	0.547	0.563
ConvNetXtTiny [30]	0.455	0.468	0.450	0.459
ConvNetXtLarge [30]	0.558	0.595	0.561	0.577
VGG16 [40]	0.738	0.735	0.738	0.736
NasNetLarge [66]	0.775	0.779	0.777	0.778
InceptionV3 [47]	0.780	0.785	0.783	0.784
NasNetMobile [66]	0.784	0.785	0.785	0.785
Xception [12]	0.806	0.806	0.803	0.804
InceptionResNetV2 [48]	0.814	0.818	0.815	0.816
ResNet50V2 [23]	0.826	0.828	0.826	0.827
MobileNetV2 [37]	0.832	0.837	0.834	0.836
DenseNet201 [27]	0.852	0.853	0.850	0.852
DenseNet169 [27]	0.854	0.854	0.854	0.854
DenseNet121 [27]	0.862	0.860	0.860	0.860
GeoSwinTiny w/o Att	0.878	0.885	0.878	0.876
GeoSwinBase w/o Att	0.882	0.892	0.882	0.880
GeoSwinLarge w/o Att	0.906	0.914	0.906	0.904
GeoSwinTiny w/ Att	0.913	0.920	0.913	0.912
GeoSwinBase w/ Att	0.922	0.927	0.922	0.921
GeoSwinLarge w/ Att	0.923	0.934	0.923	0.922

tently achieves competitive or superior performance. For instance, on UC-Merced and AID, GeoSwinLarge attains 98.0% accuracy, outperforming most baselines. Similarly, on EuroSAT and NWPU-RESISC45, the Base and Large variants achieve 98.0% and 96.7% accuracy respectively. Even the smaller GeoSwinTiny demonstrates strong performance, indicating the robustness of hierarchical feature extraction combined with MDFA attention. These results demonstrate that GeoSwin generalizes effectively across diverse geospatial domains, consistently capturing the spectral and spatial patterns necessary for accurate land-cover classification while outperforming or matching the best-performing existing methods.

Table 4. Comparative analysis in terms of Precision (P), Recall(R), and F1- Score on different methods across multiple remote sensing datasets.

	UC-Merced			AID		
Method	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Inception V3 [5]	0.880	0.870	0.870	–	–	–
DenseNet201 [21]	0.950	0.940	0.940	0.900	0.890	0.890
DAS-RHDIS [43]	0.653	0.700	0.675	–	–	–
VGG19 [2]	0.740	0.740	0.730	–	–	–
DNNE-SWA [14]	0.973	0.973	0.973	–	–	–
ViT-CNN [1]	0.920	0.900	0.940	–	–	–
GAN [7]	0.950	0.938	0.954	0.975	0.976	0.974
Ensemble [16]	0.960	0.960	0.970	–	–	–
Dlan [18]	0.970	0.970	0.971	0.920	0.920	0.930
GeoSwinTiny	0.974	0.973	0.973	0.974	0.973	0.973
GeoSwinBase	0.978	0.973	0.973	0.978	0.977	0.977
GeoSwinLarge	0.980	0.980	0.980	0.980	0.980	0.980
	EuroSAT			NWPU-RESISC45		
Method	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Inception V3 [2]	0.750	0.750	0.750	0.710	0.700	0.700
DenseNet201 [21]	–	–	–	0.930	0.930	0.930
DNNE-SWA [14]	–	–	–	0.945	0.945	0.945
Dlan [18]	0.950	0.930	0.960	0.560	0.970	0.860
GeoSwinTiny	0.975	0.975	0.975	0.958	0.957	0.957
GeoSwinBase	0.981	0.980	0.980	0.966	0.966	0.966
GeoSwinLarge	0.979	0.979	0.979	0.967	0.967	0.967
	SIRI-WHU			RSSCN7		
Method	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
ResNet 50 [38]	0.940	0.942	0.940	0.927	0.928	0.927
ViT-CNN [1]	0.980	0.970	0.970	0.950	0.940	0.930
GeoSwinTiny	0.962	0.963	0.962	0.961	0.958	0.959
GeoSwinBase	0.983	0.983	0.983	0.962	0.960	0.960
GeoSwinLarge	0.979	0.979	0.979	0.964	0.964	0.964

6.4. Visual Results

Figure 3 presents a comparative visual analysis of attention map activations for the Swin Transformer and the proposed GeoSwin model. The Swin Transformer produces diffuse activations that struggle with precise localization, failing to clearly highlight key features, which results in poor focus on complex land cover structures. The addition of contextual attention improves the focus slightly, but the activations remain inconsistent and lack the necessary precision. Structural Attention offers some enhancement, sharpening

Table 5. Comparative analysis of different methods across multiple remote sensing datasets.

UC-Merced		AID	
Methods	**Accuracy**	**Methods**	**Accuracy**
inception V3 [5]	0.943	InceptionV3 [17]	0.860
ViT-CNN [1]	0.960	ViT [9]	0.942
VGG19 [34]	0.943	DNNE-SWA [14]	0.974
MJDCNN [56]	0.958	MJDCNN [56]	0.941
MCPSF [46]	0.975	MCPSF [46]	0.931
Ensemble model [16]	0.970	DFAGCN [55]	0.948
SAGGAN [20]	0.904	MSHNet [31]	0.969
Attention GAN [60]	0.977	Attention GAN [60]	0.845
DenseNet201 [21]	0.940	DenseNet201 [21]	0.890
GAN [7]	0.957	GAN [7]	0.977
GBN [44]	0.970	GBN [44]	0.954
DFAGCN [55]	0.974	DFAGCN [55]	0.948
XcelNet-KNN [3]	0.976	DLAN [18]	0.940
GeoSwinTiny	0.973	GeoSwinTiny	0.973
GeoSwinBase	0.977	GeoSwinBase	0.977
GeoSwinLarge	0.980	GeoSwinLarge	0.980
EuroSAT		**NWPU**	
Methods	**Accuracy**	**Methods**	**Accuracy**
Inception-V3 [17]	0.920	MJDCNN [56]	0.931
SAGGAN [20]	0.943	ResNet18 [62]	0.916
Ensemble Model [16]	0.980	CSCANet [25]	0.937
CNN [33]	0.960	DNNE-SWA [14]	0.945
Attention CNN [4]	0.895	Attention CNN [4]	0.917
CNN [1]	0.949	DenseNet201 [21]	0.930
PSO-optimized CNNs [41]	0.950	SceneNet [32]	0.895
MSMatchEfficientNet [19]	0.969	OGWOA–BiLSTM [50]	0.960
ResNet50 [17]	0.910	MSHNet [31]	0.939
DLAN [18]	0.940	DLAN [18]	0.960
GeoSwinTiny	0.975	GeoSwinTiny	0.957
GeoSwinBase	0.980	GeoSwinBase	0.966
GeoSwinLarge	0.979	GeoSwinLarge	0.967
SIRI-WHU		**RSSCN7**	
Methods	**Accuracy**	**Methods**	**Accuracy**
RFPNet [63]	0.915	AlexNet [39]	0.856
JoinG [26]	0.975	GLNet [45]	0.950
Transformer-based [1]	0.960	TEX-Net [8]	0.940
POVH [61]	0.866	POVH [61]	0.840
MRNet [26]	0.970	GoogleNet [39]	0.860
AlexNet [39]	0.902	DFAGCN [55]	0.941
ViT-CNN [1]	0.960	ViT-CNN [1]	0.950
GeoSwinTiny	0.962	GeoSwinTiny	0.958
GeoSwinBase	0.983	GeoSwinBase	0.960
GeoSwinLarge	0.979	GeoSwinLarge	0.964

the activations further, but it still struggles with capturing multi-scale details and finer structures. In contrast, the proposed GeoSwin model generates sharply localized, high-resolution, and semantically coherent activation maps. It excels in accurately highlighting boundaries, contours, and textures across different land cover types, showing significantly better results than the other models. GeoSwin outperforms the Swin Transformer, Contextual Attention, and Structural Attention, providing superior precision and accuracy in detecting and understanding land cover features.

7. Conclusion

We presented GeoLULC, a large and diverse benchmark dataset for LULC classification, comprising 73 classes and 73,988 high-resolution images, addressing the limitations of existing datasets in coverage and variability. To effectively utilize this dataset, we proposed GeoSwin, a Swin Transformer-based architecture enhanced with hierarchical attention modules that capture local and global contextual dependencies. Extensive experiments on GeoLULC and multiple public remote sensing benchmarks demonstrate that GeoSwin consistently outperforms state-of-the-art methods, highlighting its robustness, scalability, and ability to discriminate fine-grained land-cover classes. Ablation studies further confirm the effectiveness of the attention mechanisms and model scaling. Future work will includes exploring multi-modal and multi-temporal extensions to further improve land-cover discrimination and generalization.

References

[1] Adekanmi Adegun, Serestina Viriri, and Jules-Raymond Tapamo. Automated classification of remote sensing satellite images using deep learning based vision transformer. *Applied Intelligence*, 54(24):13018–13037, 2024. 7, 8

[2] Adekanmi Adeyinka Adegun, Serestina Viriri, and Jules-Raymond Tapamo. Review of deep learning methods for remote sensing satellite images classification: experimental survey and comparative analysis. *Journal of Big Data*, 10 (1):93, 2023. 7

[3] Bilal Ahmed, Tallha Akram, Syed Rameez Naqvi, Anas Alsuhaibani, Muhammad Attique Khan, and Naoufel Kraiem. Xcelnet14: a novel deep learning framework for aerial scene classification. *IEEE Access*, 12:196266–196281, 2024. 8

[4] Hussain Mobarak Albarakati, Muhammad Attique Khan, Ameer Hamza, Faheem Khan, Naoufel Kraiem, Leila Jamel, Latifah Almuqren, and Roobaea Alroobaea. A novel deep learning architecture for agriculture land cover and land use classification from remote sensing images based on network-level fusion of self-attention architecture. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:6338–6353, 2024. 2, 8

[5] Abebaw Alem and Shailender Kumar. Transfer learning models for land cover and land use classification in remote sensing image. *Applied Artificial Intelligence*, 36(1): 2014192, 2022. 7, 8

[6] Eman A Alshari, Mohammed B Abdulkareem, and Bharti W Gawali. Classification of land use/land cover using artificial intelligence (ann-rf). *Frontiers in Artificial Intelligence*, 5: 964279, 2023. 1, 2

[7] S Ansith and AA Bini. Land use classification of high resolution remote sensing images using an encoder based modified gan architecture. *Displays*, 74:102229, 2022. 7, 8

[8] Rao Muhammad Anwer, Fahad Shahbaz Khan, Joost Van De Weijer, Matthieu Molinier, and Jorma Laaksonen. Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS journal of photogrammetry and remote sensing*, 138:74–85, 2018. 8

[9] Yakoub Bazi, Laila Bashmal, Mohamad M Al Rahhal, Reham Al Dayil, and Naif Al Ajlan. Vision transformers for

remote sensing image classification. *Remote Sensing*, 13(3): 516, 2021. 8

[10] Rajeswari Harini Bikkasani, M Dhanya, and SV Veena. Analysis of long-term changes for land use and land cover using machine learning: A case study. In *2023 2nd international conference on applied artificial intelligence and computing (ICAAIC)*, pages 542–546. IEEE, 2023. 1

[11] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 5, 6

[12] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 7

[13] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 172–181, 2018. 6

[14] Burak Ekim and Elif Sertel. Deep neural network ensembles for remote sensing land cover and land use classification. *International Journal of Digital Earth*, 14(12):1868–1881, 2021. 7, 8

[15] Hisyam Fahmi and Wina Permana Sari. Analysis of deep learning architecture for patch-based land cover classification. In *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 1–5. IEEE, 2022. 1, 2

[16] Muhammad Fayaz, L Minh Dang, and Hyeonjoon Moon. Enhancing land cover classification via deep ensemble network. *Knowledge-Based Systems*, 305:112611, 2024. 1, 2, 7, 8

[17] Muhammad Fayaz, Junyoung Nam, L Minh Dang, Hyoung-Kyu Song, and Hyeonjoon Moon. Land-cover classification using deep learning with high-resolution remote-sensing imagery. *Applied Sciences*, 14(5):1844, 2024. 8

[18] Muhammad Fayaz, L Minh Dang, and Hyeonjoon Moon. Dlan: a dual attention network for effective land cover classification in remote sensing. *Knowledge-Based Systems*, page 113620, 2025. 2, 7, 8

[19] Pablo Gómez and Gabriele Meoni. Msmatch: Semisupervised multispectral scene classification with few labels. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:11643–11654, 2021. 8

[20] Dongen Guo, Ying Xia, and Xiaobo Luo. Gan-based semisupervised scene classification of remote sensing image. *IEEE Geoscience and Remote Sensing Letters*, 18(12):2067–2071, 2020. 8

[21] Nisha Gupta, Ajay Mittal, and Satvir Singh. Feature extraction for remote sensing image classification using variants of deep learning pre-trained models densenet-169, densenet-121 and densenet-201. *Library Progress International*, 43: 2048–2068, 2023. 7, 8

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 7

[24] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 5, 6

[25] Yan-E Hou, Kang Yang, Lanxue Dang, and Yang Liu. Contextual spatial-channel attention network for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. 8

[26] Wenyi Hu, Chunjie Lan, Tian Chen, Shan Liu, Lirong Yin, and Lei Wang. Scene classification of remote sensing image based on multi-path reconfigurable neural network. *Land*, 13 (10):1718, 2024. 8

[27] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 7

[28] Maria Kaselimi, Athanasios Voulodimos, Ioannis Daskalopoulos, Nikolaos Doulamis, and Anastasios Doulamis. A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring. *IEEE Transactions on Neural Networks and Learning Systems*, 34(7):3299–3307, 2022. 2

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[30] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 7

[31] Yuanyuan Lu, Yanhui Zhu, Hao Feng, and Yang Liu. Remote sensing scene classification using multi-domain sematic high-order network. *Image and Vision Computing*, 143:104948, 2024. 8

[32] Ailong Ma, Yuting Wan, Yanfei Zhong, Junjue Wang, and Liangpei Zhang. Scenenet: Remote sensing scene classification deep learning network using multi-objective neural evolution architecture search. *ISPRS Journal of Photogrammetry and Remote Sensing*, 172:171–188, 2021. 8

[33] Nzurumike L Obianuju, Nwojo Agwu, and Onyenwe Ikechukwu. Medium resolution satellite image classification system for land cover mapping in nigeria: a multi-phase deep learning approach. In *Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 2*, pages 1056–1072. Springer, 2021. 8

[34] Rafael Pires de Lima and Kurt Marfurt. Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. *Remote Sensing*, 12(1):86, 2019. 8

[35] Xiaoman Qi, Panpan Zhu, Yuebin Wang, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao, Ning Zang, and P Takis Mathiopoulos. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:337–350, 2020. 6

[36] Rashmi Saini and Suraj Singh. Land use land cover mapping and snow cover detection in himalayan region using machine learning and multispectral sentinel-2 satellite imagery. *International Journal of Information Technology*, 16(2):675–686, 2024. 1, 2

[37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 7

[38] Amsa Shabbir, Nouman Ali, Jameel Ahmed, Bushra Zafar, Aqsa Rasheed, Muhammad Sajid, Afzal Ahmed, and Saadat Hanif Dar. Satellite and scene image classification based on transfer learning and fine tuning of resnet50. *Mathematical Problems in Engineering*, 2021(1):5843816, 2021. 7

[39] Mayar A Shafaey, Mohammed A-M Salem, HM Ebeid, MN Al-Berry, and Mohamed F Tolba. Comparison of cnns for remote sensing scene classification. In *2018 13th International conference on computer engineering and systems (ICCES)*, pages 27–32. IEEE, 2018. 8

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[41] D Suganya and R Sugumar. Pso-optimized cnn for feature extraction and accurate classification of satellite images using machine learning. In *2024 International conference on computing and data science (ICCDS)*, pages 1–6. IEEE, 2024. 8

[42] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*, pages 5901–5904. IEEE, 2019. 6

[43] Gencer Sumbul, Mahdyar Ravanbakhsh, and Begüm Demir. Informative and representative triplet selection for multilabel remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021. 7

[44] Hao Sun, Siyuan Li, Xiangtao Zheng, and Xiaoqiang Lu. Remote sensing scene classification by gated bidirectional network. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):82–96, 2019. 8

[45] Huiming Sun, Yuewei Lin, Qin Zou, Shaoyue Song, Jianwu Fang, and Hongkai Yu. Convolutional neural networks based remote sensing scene classification under clear and cloudy environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 713–720, 2021. 8

[46] Xiongli Sun, Qiqi Zhu, and Qianqing Qin. A multi-level convolution pyramid semantic fusion framework for high-resolution remote sensing image scene classification and annotation. *IEEE Access*, 9:18195–18208, 2021. 8

[47] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 7

[48] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 7

[49] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 7

[50] VN Vinaykumar, J Ananda Babu, and Jaroslav Frnda. Optimal guidance whale optimization algorithm and hybrid deep learning networks for land use land cover classification. *EURASIP Journal on Advances in Signal Processing*, 2023 (1):13, 2023. 1, 2, 8

[51] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *European conference on computer vision*, pages 68–85. Springer, 2022. 1, 2

[52] Gui-Song Xia, Wen Yang, Julie Delon, Yann Gousseau, Hong Sun, and Henri Maître. Structural high-resolution satellite image indexing. In *ISPRS TC VII Symposium-100 Years ISPRS*, pages 298–303, 2010. 6

[53] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 5, 6

[54] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6254–6264, 2023. 6

[55] Kejie Xu, Hong Huang, Peifang Deng, and Yuan Li. Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10):5751–5765, 2021. 8

[56] Lu Xu, Yiyun Chen, Jiawei Pan, and Aji Gao. Multi-structure joint decision-making approach for land use classification of high-resolution remote sensing images based on cnns. *IEEE Access*, 8:42848–42863, 2020. 8

[57] Xingguang Yan, Jing Li, Andrew R Smith, Di Yang, Tianyue Ma, and Yiting Su. Rapid land cover classification using a 36-year time series of multi-source remote sensing data. *Land*, 12(12):2149, 2023. 1

[58] Yi Yang and Shawn Newsam. 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*, pages 270–279, 2010. 5, 6

[59] Jing Yao, Bing Zhang, Chenyu Li, Danfeng Hong, and Jocelyn Chanussot. Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep

learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 2

[60] Yunlong Yu, Xianzhi Li, and Fuxian Liu. Attention gans: Unsupervised deep feature learning for aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):519–531, 2019. 8

[61] Bushra Zafar, Rehan Ashraf, Nouman Ali, Muhammad Kashif Iqbal, Muhammad Sajid, Saadat Hanif Dar, and Naeem Iqbal Ratyal. A novel discriminating and relative global spatial image representation with applications in cbir. *Applied Sciences*, 8(11):2242, 2018. 8

[62] Jianming Zhang, Chaoquan Lu, Jin Wang, Xiao-Guang Yue, Se-Jung Lim, Zafer Al-Makhadmeh, and Amr Tolba. Training convolutional neural networks with multi-size images and triplet loss for remote sensing scene classification. *Sensors*, 20(4):1188, 2020. 8

[63] Xin Zhang, Yongcheng Wang, Ning Zhang, Dongdong Xu, and Bo Chen. Research on scene classification method of high-resolution remote sensing images based on rfpnet. *Applied Sciences*, 9(10):2028, 2019. 8

[64] Bei Zhao, Yanfei Zhong, Gui-Song Xia, and Liangpei Zhang. Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 54(4): 2108–2123, 2015. 5, 6

[65] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145:197–209, 2018. 6

[66] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018. 7

[67] Qin Zou, Lihao Ni, Tong Zhang, and Qian Wang. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and remote sensing letters*, 12(11):2321–2325, 2015. 5, 6