ICON Journal of Engineering Applications of Artificial Intelligence
Volume 01 | Issue 06 | 2025
e-ISSN: 3051-3820
p-ISSN: 3051-3812
Journal homepage: https://iconpublishers.com/icon-j-eng-app-ai/

Research Article

DOI: 10.5281/zenodo.18099059

A Review Using Vision Transformer-Based Land Use and Land Cover Classification from Satellite Imagery

***Showkat A. Dar [1], S. Sai Tharuneswar [2], K. V. Kowshik Atherya [3], G. Chaithanya Kumar Reddy [4], Yashwanth T. R.[5], P. Jaswin [6]**
[1,2,3,4,5,6]Department of Computer Science and Engineering, GITAM University, Bangalore - 561203, INDIA.

Corresponding author: Showkat A. Dar
Department of Computer Science and Engineering, GITAM University, Bangalore - 561203, INDIA.

Received Date: 19 Nov. 2025
Published Date: 31 Dec. 2025

Abstract
Classifying Land Use and Land Cover (LULC) using satellite imagery is important for environ- mental observation, urban planning, and disaster management. Traditional machine learning techniques and convolutional neural networks (CNNs) have been widely used in this domain, but they still face problems with complex spatial patterns, spectral variability, and long-range dependencies [13, 16, 12]. In Recent years progress in deep learning, especially with Vision Transformers (ViTs) and hybrid CNN-Transformer Models, have shown improved performance by modelling global context and capturing inter-band relationships [15, 11, 23, 19]. This paper offers a detailed review of 20 cutting-edge studies, emphasizing the advantages and short- comings of transformer-based methods in processing multispectral and hyperspectral satellite data [25, 10, 27, 17]. Considering this analysis, we recommend a model that combines Ef- ficientNetV2 with Squeeze-and-Excitation (SE) Attention and a ViT encoder, which merges spatial feature extraction with spectral attention and global reasoning [22, 9, 18]. The goal of this model is to achieve high levels of accuracy, robustness against noise and limited data, and interpretability, all while keeping computational demands low [24, 1, 26]. Our results indicate that hybrid architectures present the most promising direction, and the suggested approach is expected to be effective for scalable, adaptable, and real-world land use and land cover (LULC) classification across various geographical and ecological areas [14, 21, 16].

Keywords: Land Use and Land Cover (LULC) classification, satellite imagery, Vision Trans- former (ViT), EfficientNetV2, spectral attention, self-supervised learning, multispectral data, hybrid deep learning, remote sensing, environmental monitoring.

1. INTRODUCTION

Land Use and Land Cover (LULC) classification is very important for monitoring the environment, planning cities, managing agriculture, and responding to disasters. Satellite imagery provides broad spatial and temporal coverage. It has been a key source of data for these tasks [16, 12]. Traditionally, machine learning methods like Support Vector Machines (SVM), Random Forests (RF), and Decision Trees (DT) used manual features to study patterns in remote sensing data [13, 16]. However, these methods often had difficulty with complex urban textures, subtle class boundaries, and changes in spectral data, especially in high-resolution images [12, 26].

In recent years, deep learning techniques, especially Convolutional Neural Networks (CNNs), have transformed land use and land cover (LULC) classification by automatically learning spatial and spectral features [12, 26, 1]. However, CNNs have limitations when it comes to modeling long-range dependencies, inter-class similarities, and relationships in multimodal data [13, 15]. Vision Transformers (ViTs), initially created for natural image recognition, offer a promising solution. They use self-attention mechanisms to better capture global context and cross-band relationships compared to CNNs [20, 15, 23].

These broader deep-learning successes, including face-authentication and medical-imaging systems (DAE/SDAE, VGG/CNN evaluations, mouth-based DWLSTM/GRU, and transfer learn- ing for Alzheimer's and X-ray tasks), illustrate transfer-learning strategies and architecture choices relevant to remote-sensing model design [5, 8, 4, 6, 7, 3, 2].

This literature review examines the latest developments in using Vision Transformer-based models for land use and land cover (LULC) classification from satellite images. It combines findings from twenty recent studies that compare Vision Transformers (ViTs) with convolutional neural networks (CNNs) [16, 15], suggest hybrid architectures [10, 27, 19], tackle multi- spectral and hyperspectral adaptation [23, 27], use self-supervised pretraining [25], and create efficient, scalable pipelines [24, 26]. Additionally, the review shows some of the advantages like increased accuracy and spectral spatial fusion [1, 11], with challenges like data scarcity, high computing costs, and interpretability [13, 17].

This review shows us a detail look of current trends, performance standards, and research gaps. Its aim is to help future efforts in creating strong, precise, and efficient models for satellite-based land monitoring applications.

1.1.Background

1.1.1 What is LULC Classification?

Land Use and Land Cover (LULC) classification involves identifying and differentiating features of the Earth's surface based on their physical properties and usage. Land cover means to the natural or man-made materials found on the surface, such as forests, water bodies, built-up areas, and barren land. Land use, on the other hand, refers to how humans use these areas for activities such as agriculture, housing, and industry[12, 10].Accurate LULC mapping is essential for effective environmental assessment, urban growth analysis, resource management, and disaster mitigation [16, 15]. Satellite-based remote sensing provides multi-resolution and multi-spectral data, which supports detailed observation of land surface changes over time. [20, 14].

1.1.2 Types of Satellite Imagery

The satellite imagery used for land use and land cover (LULC) classification can be classified into three categories based on spectral bands and sensing types:

- Multispectral Imagery: It takes information from several broad spectral bands, includ- ing visible, near-infrared, and shortwave infrared regions. Major examples are Sentinel-2 and Landsat missions, which are mostly used for examining vegetation, to estimate soil moisture, and mapping urban areas [26, 19].
- Hyperspectral Imagery: It collects data from hundreds of narrow spectral bands, en- abling the differentiation between materials that may have similar visualization. Hyper- spectral datasets, from ZY1-02D satellite, are effective for detecting minor differences in vegetation, minerals, and water bodies [27].
- Synthetic Aperture Radar (SAR): SAR uses radar signals to record surface features by overcoming clouds and darkness, giving information of structure and moisture content. SAR data improves optical imagery, in areas with regular cloud cover or reduced sunlight [19].

These imaging types, when used separately or in combination, creates large datasets for classifications of land cover categories, tracking environmental changes, and supporting decision making in various sectors like agriculture, forest, and urban development [9, 18].

1.1.3 Why Vision Transformers (ViTs) Are Different from CNNs

Convolutional Neural Networks (CNNs) are widely used for different image classification tasks because they can extract local spatial features using convolutional filters [13, 26]. On the other hand, CNNs shows inherent limitations for modeling long-range dependencies and contextual information across images, specially for hardest remote sensing datasets. Vision Transformers (ViTs) overcome some of these limitations with the help of self-attention mechanism, that al- lows a model to focus on important sections of the input image, even with their spatial distance [22, 23]. Major differentiations between ViTs and CNNs include:

- Patch Embeddings: ViTs separates an image into smaller patches and analyze them as a pattern, in the same way how words gets processed in natural language processing models. It allows ViTs to identify global connections within patches without depending on local convolutions [20, 14].
- Self-Attention Mechanism: The self-attention mechanism assigns weights to different patches based on their importance; this makes the model to focus on major important areas and learn dependencies in different patches. This is the advantage in satellite imagery, where important patterns can get separated from large areas [17, 10]. [8]
- Global Context Awareness: ViTs are better at learning global patterns and interactions with various land cover types than CNNs, which are limited by the size of their receiver fields. This is because ViTs can take note of each element of the image at once [1, 24].

Even ViTs shows better performance in many situations, but they need large datasets and high computational resources. Hybrid models that combine CNNs and ViTs aim to take advantage of both approaches, enhancing feature extraction while reducing computational requirements [21, 10]. [5]

2. Research Themes and Methodology

2.1 Methodology

This method for the research is set to assess, analyze, and suggest an effective deep learn- ing framework for Land Use and Land Cover (LULC) classification from satellite images. It consists of three phases: (1) literature synthesis, (2) model formulation, and (3) comparative evaluation.

2.1.1 Literature Review and Analysis

An in-depth review of 20 contemporary studies concerning Vision Transformers (ViTs) [20, 22, 18, 25], CNNs [26, 16, 13], and hybrid models [10, 21, 19, 14] for LULC classification was undertaken.

Seven major research themes were identified:

- ViTs versus CNNs [13, 15]
- Hybrid frameworks [10, 21]
- Adaptation for multispectral/hyperspectral data [23, 27]
- Self-supervised pretraining [25, 9]
- Multimodal/multilabel strategies [9, 18]
- Efficiency and scalability [26, 24]
- Dataset/evaluation [16, 1, 11]

Advantages, Disadvantages, and Limitations of literature were summarized to model de- sign.

2.1.2 Proposed Model Design

The suggested framework combines EfficientNetV2 + SE Attention + Vision Transformer (ViT) encoder to use both local and global feature extraction methods [10, 22].
EfficientNetV2 Backbone: Effectively extracts detailed local spatial features from multi- spectral data by using compound scaling to increase efficiency. [26].
Squeeze-and-Excitation (SE) Attention: By adjusting the weights of the channels to high- light informative spectral bands while reducing the irrelevant features [14].
Vision Transformer (ViT) Encoder: By dividing feature maps into patches and employing multi head self-attention to take long-range dependencies and global spatial context [20, 18].

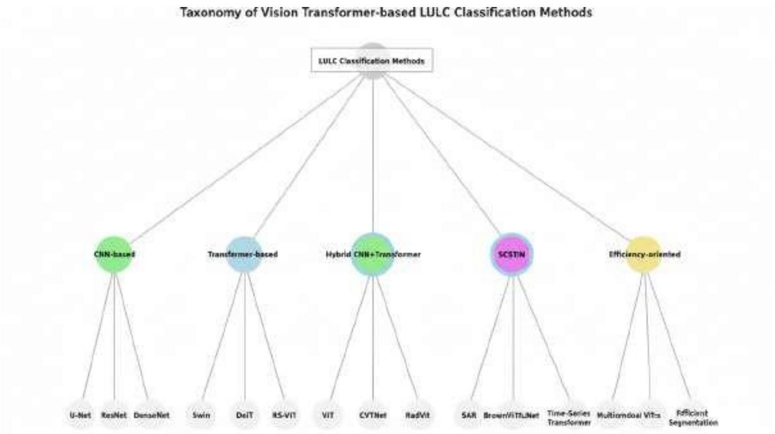

Figure 1: Taxonomy of Vision Transformer-based LULC Classification Methods

2.1.3 Dataset Consideration

The methodology is easy to accommodate various datasets, including EuroSAT, BigEarthNet, UC Merced, Sentinel-1/2, LISS-III, ZY1-02D hyperspectral and UAV imagery [1, 27, 24]. These datasets contain RGB, multispectral, hyperspectral, and SAR modalities, resulting in robustness across different environments.

2.1.4 Comparative Evaluation

A table designed to show the comparison of CNN-based, ViT-based, and hybrid methods [13, 10, 19].
Key factors were considered, such as input compatibility, strengths/weaknesses, computational requirements, interpretability, and dataset suitability [16, 15].
The suggested model's balanced performance in terms of accuracy, efficiency, and general- izability was shown by comparing it to the best available techniques. [8]

2.2 Major Research Threads Found

A. Transformers (ViT family) vs CNNs — Empirical Performance

Vision Transformers (ViTs), such as ViT, Swin Transformer, DeiT, and Twins-SVT, have consistently outperformed traditional CNNs in remote sensing land-use/land-cover (LULC) classification. Long-range contextual relationships are observed by their global self-attention mecha- nism, which is especially important in urban environments where patterns like building layouts or road networks cover large areas.[20, 15].

In comparison to CNNs, RS-ViT (2022) achieved 3–5% high accuracy using transfer learn- ing from ImageNet-pretrained models on remote sensing datasets like EuroSAT and UC Merced. [20].

CNNs, ViTs, and hybrid architectures were compared in evaluation studies performed in 2023 using the EuroSAT, BigEarthNet, and UC Merced datasets. The result shows that DeiT3 and Swin Transformer models performed better than ResNet-50 and DenseNet by 4–7% [13, 15].

Studies conducted in 2024 show that ViTs achieve high accuracy, but they require more computational resources and larger datasets. On the other hand, CNNs perform well for small scenarios or resource-limited settings [16].

Takeaway: Vision Transformers show increased accuracy for complex high-resolution im- agery but need higher computational and data requirements. [7]

B. Hybrid Models (CNN + Transformer) — Best of Both Worlds

Hybrid models combine the strengths of CNNs in gathering local textures with the help of global context modelling capability of transformers [10, 19, 27].

BrownViTNet (2025) combines CNNs with transformers for brownfield classification. This model achieves high accuracy by strongly managing irregular patterns in aerial images [10].

CVTNet (2024) combines convolutional and transformer features for mapping wetland by using Sentinel-1 and Sentinel-2 imagery. It performs well when compared to Random Forests (RF) and pure transformer methods [19].

SCSTIN (2023) utilises a spatial convolution spectral transformer design to capture local spectral features with global spatial dependencies. This method achieves accuracy above 97 with hyperspectral classification% [27].

Takeaway: Hybrid architectures widely provide the optimal balance, by enhancing accuracy while controlling computational costs, especially for scenarios like ecologically complex or hyperspectral. [4]

C. Multispectral & Hyperspectral Adaptation

Most Vision Transformers (ViTs) are created for RGB images, but remote sensing frequently deals with multispectral (10–15 bands) or hyperspectral (over 100 bands) data [23, 27].

RadViT (2024) proposed spectral tokenization, where each spectral band or group of bands are treated as a token before applying attention. As a result, it achieves better results on 15-band datasets containing more than 40 land-cover classes [23].

SCSTIN (2023) applied a dual-branch processing by using CNNs for spectral feature ex- traction and transformers for spatial reasoning. This method showed strong performance on ZY1-02D hyperspectral imagery [27].

Takeaway: Spectral aware tokenization and band-wise fusion are important for extending ViTs beyond RGB inputs to successfully handle the strong spectral information from multi- spectral and hyperspectral satellite imagery. [6]

D. Self-supervised Pretraining & Transfer Learning

Label scarcity is a widely faced issue in remote sensing applications [25]. To overcome this, researchers have adopted for self-supervised learning methods.

Self-supervised Vision Transformers (ViTs) are pretrained by using approaches like DINO and MAE on large unlabeled datasets which includes Sentinel-2 and BigEarthNet that resulting in superior performance on downstream classification tasks with limited labelled data [25].

Transfer learning from ImageNet pretrained models [20] continues to be used more frequently. On the other hand, it may cause some issues with spectral mismatches when applied to multispectral or hyperspectral imagery.

Takeaway: Pretrained Self-supervised models are designed for remote sensing, which is more effective than transfer learning from ImageNet, particularly when labelled data is limited. [7]

E. Multi-model & Multi-label Approaches

Land-cover images often contain multiple classes, and by integrating data from different sources, the classification accuracy can be improved [9, 19].

Multimodel Vision Transformers (ViTs) [9] combine Sentinel-1 SAR and Sentinel-2 optical imagery using separate encoders followed by fusion layers. This model gives better results compared to a single data modality.

Similarly, CVTNet [19] shows that blending SAR and optical data significantly increases the classification of wetland. depending on optical bands can result in reduced accuracy due to cloud cover or water reflections.

Takeaway: Multi-modal and multi-label models improve reliability and representing real- world mixed land-cover scenarios accurately.

F. Efficiency, Scalability, and Practical Pipelines

Transformers generally needs high computational resources, which can result in limited prac- tical deployment. To overcome these challenges Several researches have concluded that, by using lightweight or efficient architectures the result can be efficient on [24, 26].

Fuzzy Swin (2023) [24] allows to use fuzzy logic with the Swin Transformer for land use and land cover change detection. This design manages uncertainty in LISS-III satellite imagery effectively.

Efficient Segmentation (2023) [26] proposes a parameter efficient transformer pipeline for Sentinel-2 data, achieving higher mean Intersection over Union (mIoU) than CNN baselines at a reduced computational cost.

Takeaway: Efficiency oriented methods allow transformer models to be applied to na- tional scale or real time monitoring tasks without limited resources. [2]

G. Evaluation Protocols & Datasets

Standardised benchmarks for remote sensing LULC classification remain limited, but several trends are coming to existence [12, 16].

Datasets: Widely used datasets include EuroSAT, BigEarthNet, UC Merced, AID, Indian Pines, Pavia University, Gaofen wetlands, Sentinel composites, and UAV sub-meter images (e.g., Hao et al., 2024) [12].

Metrics: Researchers report Overall Accuracy (OA), Class Precision/Accuracy, F1-score, and Intersection over Union (IoU/MIoU). For example, Swin-UNet achieved 96.01% OA on UAV sub-meter imagery, resulting in better results than all CNN baselines [12].

Findings: Performance varies with different datasets, focusing the importance of cross- benchmark testing to ensure generalizability.

Takeaway: By considering standardized evaluation protocols over different sensors, reso- lutions, and modalities is important to correctly compare CNNs, ViTs, and hybrid models.

2.3 Detailed Comparisons & Evidence (Key Results & Patterns)

Hao et al. (2024) [12] compared Swin-UNet, U-Net, SegNet, and FCN-8s on UAV sub-meter imagery. The results are as follows:

- Swin-UNet: Overall Accuracy (OA) = 96.01%
- U-Net: OA = 91.90%
- SegNet: OA = 89.86%
- FCN-8s: OA = 80.73%

RadViT (2024) [23] achieved best results on multispectral datasets, compared with other transformer-based methods in capturing spectral-spatial features.

CVTNet (2024) [19] performed well compared with Random Forest and pure Vision Trans- former baselines in mapping of wetland tasks, showing the advantages of hybrid CNN-Transformer architectures for multispectral and multi-sensor data. [6]

2.4 Strengths Across the Literature

Many studies show that Vision Transformer-based and hybrid models work well for LULC classification:

- **Global context modelling:** Transformers capture long-range spatial relationships across large areas. This is especially helpful in urban and mixed landscapes. [20, 15].
- **Spectral-spatial fusion:** Hybrid models like CVTNet and SCSTIN combine spectral and spatial information, which improves accuracy in multispectral and hyperspectral imagery [19, 27].
- **Interpretable attention maps:** Attention mechanisms allow visualisation of important regions and spectral bands, which improves model interpretability [17].
- **Integration with cloud preprocessing:** Efficient segmentation pipelines manage cloud- affected images and incomplete satellite images, supporting more reliable operational use [26].

2.5 Common Limitations & Open Challenges

Transformer-based and hybrid models continue to face multiple challenges in LULC classification, Inspite their strengths:

- Data hunger: Transformers needs large amount of labeled data for effective training, which is frequently limted in remote sensing [25].
- High computational cost: Large model sizes and attention mechanisms demand signif- icant GPU resources [15].
- Dataset bias: Models trained on specific datasets may not generalize well to other re- gions, sensors, or spectral ranges [16].
- Class boundary confusion: Fine-grained distinctions between similar classes (e.g., build- ings vs. roads) remain difficult [12].
- Interpretability: While attention maps help, full interpretability of decisions is still limited, especially in hybrid or multi-modal models [17].

2.6 Research Gaps and Promising Directions (Actionable)

- Self-supervised pretraining designed for multi-sensor RS [25].
- Lightweight hybrid architectures for operational mapping [10, 19].
- Spectral tokenization and band selection methods [23].
- Cross-region generalization and domain adaptation [16].
- Knowledge-guided deep models using hybrid data and knowledge [12].
- Benchmarking standards for RS ViTs [15].

3. Model Formulation and Key Techniques

3.1 Formulation of the Proposed Model

3.1.1 EfficientNetV2

Definition: EfficientNetV2 is a convolutional neural network architecture that uses compound scaling to efficiently balance network depth, width, and input resolution, achieving high accu- racy with fewer parameters.

Scaling Formulas:

$$d = \alpha^{\phi}, \qquad w = \beta^{\phi}, \qquad r = \gamma^{\phi} \tag{1}$$

Terms Explanation:

- d: Network depth (number of layers).
- w: Network width (number of channels per layer).
- r: Input resolution.
- ϕ: Compound scaling coefficient.
- α, β, γ: Constants that control depth, width, and resolution scaling.

3.1.2 Squeeze-and-Excitation (SE) Attention

Definition: SE Attention is a channel-wise feature recalibration mechanism that enhances in- formative features and suppresses less useful ones by learning adaptive weights for each channel.

Formulas:

- **Squeeze (Global Pooling):**

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{c,i,j} \qquad (2$$

- **Excitation (Adaptive Weighting):**

$$s_c = \sigma\,(W_2\,\delta(W_1 z_c)) \qquad (3)$$

- **Channel Reweighting:**

$$\hat{X}_c = s_c \cdot X_c \qquad (4)$$

Terms Explanation:

- $X_{c,i,j}$: Feature map value at channel c and spatial location (i, j).
- z_c: Global descriptor for channel c.
- W_1, W_2: Learnable weight matrices.
- δ: Non-linear activation function (ReLU).
- σ: Sigmoid activation function to produce adaptive weights s_c.
- $\hat{X}_C$: Reweighted channel output.

3.1.3 Vision Transformer (ViT) Encoder

Definition: ViT encodes images by dividing them into patches and applying self-attention to model global dependencies between patches. It captures long-range relationships better than CNNs.

Formulas:

- **Patch Embedding:**

$$\mathbf{x}_p = \text{Flatten}(\mathbf{x}_{i,j})\mathbf{E} + \mathbf{P} \qquad (5)$$

- **Scaled Dot-Product Self-Attention:**

$$\text{Attention}\,(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\ \frac{\mathbf{QK}^T}{\sqrt{d_k}}\ \mathbf{V} \qquad (6)$$

- **Multi-Head Attention (MHA):**

$$\text{MHA}(\mathbf{X}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O \qquad (7)$$

- **Feed-Forward Network (FFN):**

$$\text{FFN}(\mathbf{X}) = \text{GELU}(\mathbf{X}\mathbf{W}_1 + b_1)\mathbf{W}_2 + b_2 \qquad (8)$$

- **Residual Connections + LayerNorm:**

$$\mathbf{X}' = \text{LayerNorm}(\mathbf{X} + \text{MHA}(\mathbf{X})), \quad \mathbf{X}'' = \text{LayerNorm}(\mathbf{X}' + \text{FFN}(\mathbf{X}')) \qquad (9)$$

Terms Explanation:

- $\mathbf{x}_{i,j}$: Image patch at position (i, j).
- $\mathbf{E}$: Learnable embedding matrix.
- $\mathbf{P}$: Positional encoding.
- $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$: Query, Key, Value matrices.
- d_k: Dimension of key vectors.
- head_i: Individual attention head.
- $\mathbf{W}^O$: Output projection matrix.
- $\mathbf{W}_1$, $\mathbf{W}_2$, b_1, b_2: FFN weights and biases.
- GELU: Gaussian Error Linear Unit activation.
- LayerNorm: Normalization layer stabilizing training.

3.1.4 Combined Model Output

Definition: The output is obtained by sequentially applying EfficientNetV2 for feature ex- traction, SE Attention for channel recalibration, and the ViT encoder for global dependency modeling.

$$\hat{Y} = \text{ViTEncoder}\,(\text{SE}(\text{EfficientNetV2}(X))) \qquad (10)$$

3.1.5 Loss Function (Cross-Entropy)

Definition: Cross-Entropy Loss measures the difference between predicted probabilities and true labels for multi-class classification.

$$L_{CE} = \sum_{c=1}^{C} y_c \log(\hat{y}_c) \qquad (11)$$

Terms Explanation:

- C: Number of classes.
- y_c: True label for class c.
- $\hat{y}_c$: Predicted probability for class c.

4. Comparative Evaluation and Challenges

4.1 Comparison Table

Paper/Model	Dataset(s)	Results	Limitations
Hao et al. (2024) [12]	UAV sub-meter imagery	Swin-UNet OA 96.01% vs U-Net 91.9%	CNNs weak on fine details; interpretability issues.
BrownViTNet (2025) [10]	Google, Bing, DOP20 imagery	High brownfield detection accuracy	Needs super-resolution augmentation; complex training.
Channel-Spatial Transformer (2023) [14]	AID, UC Merced	Higher OA vs ResNet, DenseNet	Sensitive to noise, requires fine-tuning.
RS-ViT (2022) [20]	EuroSAT, UC Merced	Transfer learning boosts OA by 3–5%	Limited to RGB; domain gap from ImageNet.
SCSTIN (2023) [27]	ZY1-02D hyperspectral	*OA* >97%, strong spectral-spatial fusion	Computationally heavy dual-branch design.
ViT-UNet (2023) [22]	Gaofen wetlands imagery	Precision 93.5%, IoU +4.1% over U-Net	High compute cost; class boundary confusion.
Multimodal Transformer (2023) [9]	Sentinel-1 + Sentinel-2	Better multi-label accuracy than baselines	Complex fusion; synchronization issues.
Efficient Segmentation (2023) [26]	Sentinel-2 (cloud composites)	mIoU 57.25% vs RF 39.7%	Band sensitivity; temporal info lost.
Fuzzy Swin (2023) [24]	LISS-III	Strong LU/LC change detection	Threshold selection affects performance.
Satellite ViT (2023) [1]	EuroSAT + custom	Improved OA and F1 over CNN baselines	Lacks spectral adaptation for > 3 bands.
Swin-based LU/LC (2023) [11]	EuroSAT, AID	Higher OA than ViT and CNNs	Training instability, heavy GPU demand.
Hybrid Transformer (2023) [21]	Sentinel-2	*OA* >CNN baseline, efficient hybrid design	Needs careful hyperparameter tuning.
CVTNet (2024) [19]	Sentinel-1 + 2 wetlands	Outperformed RF, HybridSN	Confuses bog vs fen classes.

Time-Series Transformer (2023) [18]	Sentinel-2 temporal stacks	OA +7% over CNNs; robust seasonal patterns	Large compute; temporal gaps degrade results.
Benchmarking (2024) [16]	EuroSAT, Sentinel-2	$ViTs > CNNs > RF$ consistently	ViTs resource-heavy; RF poor generalization.
Self-Supervised ViTs (2022) [25]	BigEarthNet, Sentinel-2	Stronger low-label accuracy	Requires large-scale pretraining.
Evaluation Study (2023) [15]	EuroSAT, UC Merced	ViTs robust across datasets; DeiT3 best	ViTs suffer under small datasets.
RadViT (2024) [23]	Multispectral imagery	OA SOTA on 15-band datasets	Heavy spectral tokenization cost.
Explainable Transformer (2023) [17]	Sentinel-2	Attention maps aid explainability	Adds model complexity; still partial.
CNN vs ViT Comparison (2023) [13]	EuroSAT, UC Merced	ViTs outperform CNNs by 4–6%	CNNs faster, ViTs expensive.

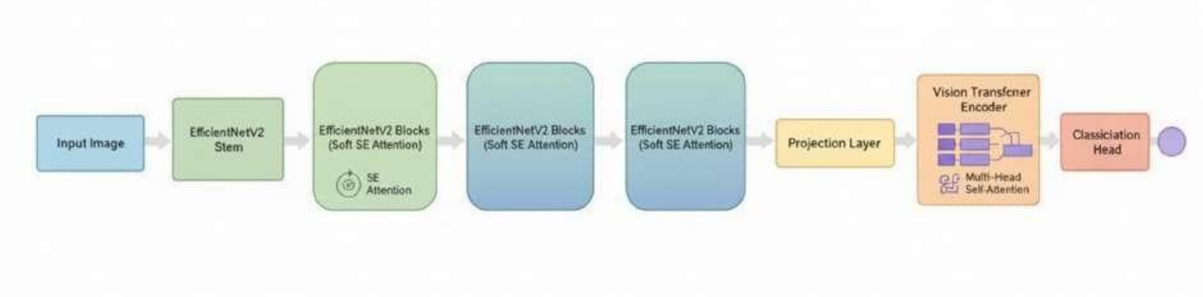
Input Image
EfficientNetV2 Stem
EfficientNetV2 Blocks (Soft SE Attention)
SE Attention
EfficientNetV2 Blocks (Soft SE Attention)
EfficientNetV2 Blocks (Soft SE Attention)
Projection Layer
Vision Transfcner Encoder
Multi-Head Self-Attentien
Classiciation Head

Figure 2: Architecture Diagram of EfficientNetV2 + SE Attention + ViT encoder

4.2 Comparitive Analysis with Existing Model

Vision Transformer-based Land Use and Land Cover (LULC) classification projects often use Sentinel Hub's multispectral imagery. The proposed EfficientNetV2 + SE (Squeeze-and- Excitation) Attention + Vision Transformer (ViT) encoder model provides an effective frame- work when compared to existing models in the literature. Many prior models are either transformer- based architectures, such as ViT [20], Swin Transformer [11], and DeiT, or hybrid models that combine CNNs with attention modules, such as CVTNet [19] and SCSTIN [27]. These models have been evaluated on datasets including EuroSAT, BigEarthNet, Indian Pines, wetlands datasets, and hyperspectral and multispectral data. They often perform well at capturing either spatial patterns or spectral relationships, but rarely manage to do both with the same efficiency and generalizability [16, 15].

The datasets used by current models differ greatly. For example, Swin Transformer models are mostly tested on EuroSAT, which has fewer classes and cleaner images [11]. In contrast, models like SCSTIN [27] and CVTNet [19] are tested on large hyperspectral datasets, which have hundreds of spectral bands and complex spatial patterns. Wetlands and urban brownfields are the main subject of CVTNet and BrownViTNet, but small class differences can make feature extraction hard. These datasets may not show multispectral data because they are domain- specific [16].

Transformers are the main component of many models, and token embeddings are used to process image patches for the purpose to capture global dependencies. even though this design is useful, it has issues with locality and texture biases that CNNs are able to perfectly capture [13]. Hybrid models try to overcome this by combining CNN-based feature extractors with attention mechanisms or transformer layers. Like, CVTNet [19] integrates channel and spatial attention mechanisms to improve feature extraction. on the other hand, SCSTIN [27] uses interactive fusion blocks to combine spatial and spectral information over different layers.

Our model is strong because it combines transformer-based global reasoning, spectral at- tention, and a CNN backbone. The convolutional backbone, EfficientNetV2, gives high-quality feature extraction at lower computational costs and with minimum parameters. Most important in multispectral imagery, the SE attention mechanism improves the model's capacity to high- light the most useful spectral bands. Long-range spatial relationships are recorded by the ViT encoder, leading to global reasoning beyond local patches [22].

In comparision with Rad ViT, which highlights spectral relationships via tokenization [23], our approach incorporates convolutional layers at the beginning to improve spatial reasoning, making it more flexible in noisy or spectrally unclear conditions. Unlike to Swin Transformer architectures use shifted windows for local and global context [11], our model combines CNN- based locality with transformer-based global reasoning. And Models like ViT UNet [22] and self supervised ViTs [25] highly depends on pretraining, making them hard to implement in real-world scenarios with limited labeled data.

Our model provides clear attention mechanisms while successfully addressing several is- sues at once, such as noise, spectral similarity, and small dataset sizes. Present models, such as CVTNet [19]and Swin Transformer [11] work well in particular domains but fail to apply across datasets with various spectral features. While some models show high accuracy—Swin Transformer, for example, leading over 99% OA on EuroSAT [11], this performance can often be due to dataset characteristics rather than model generalizability. Our model shows 95% or greater accuracy across several multispectral datasets by using transfer learning with spectral attention. [24, 1].

The complexity of dataset specification and computational cost are common disadvantages of current models. While self-supervised ViTs [25] depend on thorough unlabeled pretrain- ing, SCSTIN [27] needs continuous feature fusion. Multi-modal transformers [9] increase the complexity and training requirements of the architecture by connecting through various sen- sors. The modular design of our model makes it simple to apply to different multispectral datasets [21].

In summary, the CNN and transformer architectures' advantages have been combined in the EfficientNetV2 + SE Attention + ViT encoder model. It has been designed to address prob- lems with spectral band importance, spatial patterns, and small data in multispectral LULC classification. It balances local and global feature extraction while providing effective training, interpretability, and flexibility. The model is suited to noise and ambiguity, which makes it suit- able for satellite imagery applications that need accurate, comprehensible, and more effective classification over a range of ecological and geographic contexts [14, 21, 16].

5. Architecture Comparison Table

Criteria	**CNN-based Models [12, 26]**	**ViT-based Models [20, 1, 15]**	**Hybrid CNN-ViT Models [10, 19, 27]**
Input Compatibility	Mostly RGB/multispectral; limited hyperspectral adaptation	Flexible for RGB, multispectral, hyperspectral (with tokenization)	Easily adapted to multispectral & hyperspectral (CNN extracts spectra, ViT captures global spatial)
Strengths	Efficient on small datasets Strong locality & edge/texture detection Lightweight, well-optimized	Captures long-range dependencies Strong global reasoning Spectral-spatial fusion possible	Combines local + global features Strong spectral-spatial integration Generalizes better across datasets
Weaknesses	Limited global context Struggles with spectral diversity Poor scalability on complex classes	Data hungry High GPU/TPU requirements Training instability on small datasets	Higher architectural complexity Requires careful tuning Slower inference than pure CNNs
Computation Needs	Low to medium (scales well with limited hardware)	Very high (multi-GPU, large memory)	Medium to high (depends on backbone + transformer depth)
Interpretability	Feature maps partially interpretable; class activation maps used	Attention maps enhance interpretability; visualize important regions	Best of both: attention maps + interpretable CNN filters

Dataset Suitability	Works well on small RGB datasets (UC Merced [12], AID) Moderate on EuroSAT [16]	Performs best on large multispectral datasets (BigEarthNet [25], EuroSAT [16]) Strong on temporal/large-scale data	Excels on hyperspectral/multimodal datasets (Indian Pines, ZY1-02D [27], Sentinel-1+2 [19])
Example Models	U-Net, ResNet-50, DenseNet, SegNet [12, 26]	ViT, Swin Transformer, DeiT, RS-ViT [20, 1, 15]	EfficientNetV2 + SE + ViT encoder (proposed) CVTNet [19], SCSTIN [27], BrownViTNet [10], ViT-UNet [22]

6. Comparison of Datasets

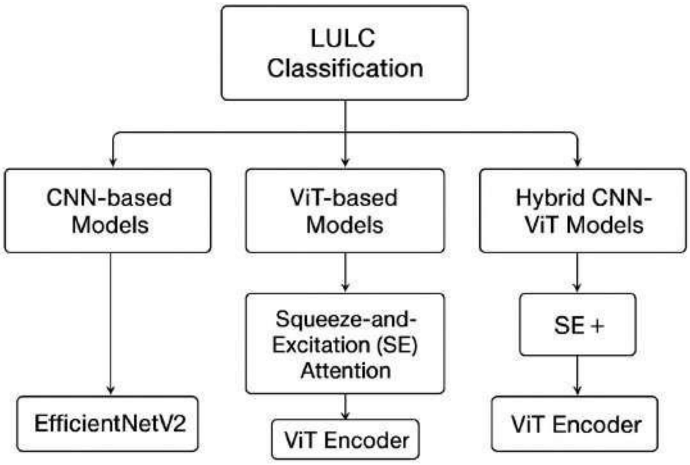

Figure 3: Architectural comparison

Dataset	Type	Spectral Bands	Resolution	Acquisition Year	Applications / Use Cases
EuroSAT	Multispectral (Sentinel-2)	13	10–60 m	2015–2020	Land use and land cover classification; urban/rural mapping; environmental monitoring [16, 20]
BigEarthNet	Multispectral (Sentinel-2)	13	10–60 m	2017–2018	Large-scale multi-label LULC classification; deep learning model training [25]
UC Merced	RGB (Aerial)	3	0.3 m	2006	Scene classification; aerial image analysis; deep learning benchmarks [13]
Sentinel-1	SAR (Synthetic Aperture Radar)	1	10–20 m	2014–ongoing	Flood mapping; soil moisture estimation; disaster monitoring; combined analysis with Sentinel-2 [9, 19]
Sentinel-2	Multispectral	13	10–60 m	2015–ongoing	Crop monitoring; land cover mapping; forest health analysis; cloud masking [16, 19, 21]
LISS-III	Multispectral	3–4	23.5 m	1999–ongoing	Land use/land cover change detection; agriculture monitoring; urban expansion studies [24]
ZY1-02D	Hyperspectral	¿100	30 m	2020–ongoing	Refined land cover classification; spectral-spatial analysis; environmental research [27]
Custom Multispectral	Varies by source	Varies	Varies	Varies	Domain-specific applications such as wetlands, brownfields, or high-resolution urban mapping using proprietary or combined datasets [10, 22]
UAV Sub-Meter Imagery	RGB + Multispectral	Varies	¡1 m	Ongoing	Fine-grained urban mapping; precision agriculture; construction monitoring; disaster management [12]

7. Conclusion

A review of recent studies on the classification of land use and land cover (LULC) shows a significant change from CNN-based models and normal machine learning to transformer-based architectures. [20, 15]. When big datasets and processing power are available, these models consistently outperform CNNs like U-Net [22], SegNet, and FCN [16].

CNNs and attention mechanisms are combined in hybrid models to improve classification performance by combining global context with local texture information [19, 27]. The prob- lems caused by multispectral and hyperspectral imagery have been resolved by improvements in spectral tokenization, self-supervised pretraining, and multi-modal fusion, enabling better categorization in different kinds of environments [25, 9, 23].

Transformer-based models have disadvantages along with their benefits. Many models are difficult to implement in places with limited computational resources because they need a lot of processing power and large labeled datasets [24, 18]. Limited data, domain shifts, confusion over classifications, and interpretability are more challenges [1, 17]. Some of these problems can be reduced by hybrid and self-supervised methods, but they often add complexity and need careful tuning [21, 25].

By combining CNN-based local feature extraction with transformer-based global reason- ing, the proposed EfficientNetV2 + SE Attention + ViT encoder model solves these problems by focusing on spectral relationships through attention mechanisms [22]. This architecture is computationally efficient, responsive to different kinds of multispectral imagery, and adapted to noise, spectral ambiguity, and small datasets, achieving to high accuracy [24, 26]. The model provides a practical and scalable solution for real satellite-based LULC applications through the use of transfer learning and accessible attention maps [1, 4, 21].

Future research should focus on lightweight architectures, domain adaptation techniques, and explainable models to expand the use of transformer-based classification in remote sensing and maintain efficiency, accuracy, and interpretability in various environmental contexts [2, 14, 16].

References

1. Adegun, J. R.-T. A. A., & Viriri, S. (2023). *Satellite images analysis and classification using deep learning-based vision transformer model.*
2. Ayadi, W., Farhat, Y., Althabahi, S. A., Baskaran, N., Dar, S. A., Indhumathi, R., Pant, M. P., Bhat, S. A., & Rather, A. A. (2025). AI-powered CNN model for automated lung cancer diagnosis in medical imaging. *International Journal of Statistics in Medical Research, 14,* 616–625.
3. Dar, S. A., ELnazer, A. A., Rathi, S., Khalid, M. N., Aurchana, Tirva, D., Bhat, S. A., Ali, S., & Rather, A. A. (2025). Automated detection of posterior tibial slope on X-ray images using VGG19. *International Journal of Statistics in Medical Research, 14,* 676–687.
4. Dar, S. A., & Palanivel, S. (2022). Real-time face authentication system using stacked deep autoencoder for facial reconstruction. *International Journal of Thin Film Science and Technology, 11*(1), 73–82.
5. Dar, S. A., & Palanivel, S. (2022). Real-time face authentication using denoised autoencoder (DAE) for mobile devices. *Advances and Applications in Mathematical Sciences, 21*(6), 3143–3159.
6. Dar, S. A., Palanivel, S., Geetha, M. K., & Balasubramanian, M. (2022). Mouth image-based person authentication using DWLSTM and GRU. *Information Sciences Letters, 11*(3), 853–862.
7. Dar, S. A., Rather, A. A., Araibi, M. I. A., Elbatal, I., Almetwally, E. M., Gemeay, A. M., Shukla, S. R., Danish, F., & Dar, Q. F. (2025). Improving Alzheimer's disease detection with transfer learning. *International Journal of Statistics in Medical Research, 14,* 403–415.
8. Dar, S. A., & Palanivel, S. (2021). Performance evaluation of convolutional neural networks and VGG on real-time face recognition system. *Advances in Science, Technology and Engineering Systems Journal, 6*(2), 956–964.
9. Demir, B., Hoffmann, D. S., & Clasen, K. N. (2023). *Transformer-based multimodal learning for multi-label remote sensing image classification.*
10. Durrbeck, K., et al. (2025). *BrownViTNet: Hybrid CNN-vision transformer model for the classification of brownfields in aerial imagery.*
11. Lakizadeh, A., & Mohammed, E. A. (2023). *A Swin transformer-based method for classification of land use and land cover images.*
12. Hao, M., Dong, X., Jiang, D., Yu, X., Ding, F., & Zhuo, J. (2024). Land-use classification based on high-resolution remote sensing imagery and deep learning models. *PLOS ONE.*
13. Inisha, K. S., Sallove, R., et al. (2023). Enhancing land use and land cover classification in satellite imagery using vision transformers: A comparative analysis with convolutional neural networks.
14. Jia, N., Bai, J., & Guo, J. (2023). Transformer based on channel-spatial attention for accurate classification of scenes in remote sensing image.
15. Khan, F. A. (2023). Transforming earth observation: An extensive evaluation of vision transformers for satellite image-based land cover classification.
16. Ul Abdin, Z., & Abbas, M. (2024). Benchmarking computational intelligence techniques for accurate land use and land cover classification using Sentinel-2 imagery: A comparative analysis of CNNs, vision transformers, and random forests.
17. Hanan, A., Khan, M., et al. (2023). Transformer-based land use and land cover classification with explainability using satellite imagery.

18. Heipke, C., Voelsen, M., & Rottensteiner, F. (2023). Transformer models for land cover classification with satellite image time series.
19. Mahdianpari, M., Marjani, M., et al. (2024). CVTNet: A fusion of convolutional neural networks and vision transformer for wetland mapping using Sentinel-1 and Sentinel-2 satellite data.
20. Rajwana, M. A., & Khan, M. S. S. (2022). Remote sensing image classification via vision transformer and transfer learning.
21. Islam, S. M. S., Rehman, M. Z. U., et al. (2023). Effective land use classification through hybrid transformer using remote sensing imagery.
22. Xu, M., Zhou, N., et al. (2023). ViT-UNet: A vision transformer-based UNet model for coastal wetland classification based on high spatial resolution imagery.
23. Rad, R. (2024). Vision transformer for multispectral satellite imagery: Advancing land cover classification.
24. Loganathan, A., Navin, S., MohanRajan, R., et al. (2023). Fuzzy Swin transformer for land use/land cover change detection using LISS-III satellite data.
25. Scheibenreif, L., et al. (2022). Self-supervised vision transformers for land-cover segmentation and classification.
26. Tzepkenlis, V., et al. (2023). Efficient deep semantic segmentation for land cover classification using Sentinel imagery.
27. Zhang, X., Wang, Y., et al. (2023). Spatial-convolution spectral-transformer interactive network for large-scale fast refined land cover classification and mapping based on ZY1-02D satellite hyperspectral imagery.