Multi-Architectural Benchmarking for EuroSAT Land Use and Land Cover (LULC) Classification: Evaluating CNNs, Vision Transformers, and Hybrid Models under Augmented and Non-Augmented Regimes

John A. Smith*, Maria Chen†, Ahmed Rahman‡, Sofia Rodriguez§ *Stanford University, Stanford, CA, USA
†Massachusetts Institute of Technology, Cambridge, MA, USA
‡University College London, London, UK
§ETH Zurich, Zurich, Switzerland
{john.smith, maria.chen, a.rahman, s.rodriguez}@corresponding-domains.edu

***Abstract*—The rapid proliferation of high–resolution satellite imaging platforms has generated unprecedented volumes of Earth observation data, intensifying the demand for scalable and accurate Land Use and Land Cover (LULC) classification systems. Such models are essential for environmental monitoring, sustainable urban planning, agricultural decision support, and climate intelligence. While Convolutional Neural Networks (CNNs) have traditionally dominated this domain, most existing approaches employ legacy architectures (e.g., VGG, ResNet, DenseNet) that struggle to capture long–range spatial dependencies inherent in aerial and satellite imagery. Concurrently, Vision Transformers (ViTs) have demonstrated strong global reasoning capabilities, yet their role in remote sensing remains comparatively underexplored.**

In this work, we present a unified and rigorous benchmark of eight state–of–the–art architectures under identical training protocols on the EuroSAT dataset. Our evaluation includes four modern CNNs (ConvNeXt–Tiny, MobileNetV3–Large, EfficientNetV2, Xception) and four Transformers (MobileViT–S, DeiT–Tiny, Swin–Tiny, TNT–Small). Through controlled experimentation and systematic augmentation analysis, we study their generalization behavior, robustness, and representational strength.

Motivated by our findings, we introduce a novel hybrid CNN–Transformer architecture that fuses local feature extraction with global self–attention for enhanced spatial reasoning. The proposed model establishes a new state–of–the–art on EuroSAT, achieving a classification accuracy of 99.4%, outperforming all individual CNN and Transformer baselines.

***Index Terms*—Land Use and Land Cover (LULC), EuroSAT, Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), Hybrid Models, Data Augmentation**

I. Introduction

With the rapid expansion of Earth observation missions, a wide range of satellite imagery is now available. Effective analysis of this data through Land Use and Land Cover (LULC) classification plays a crucial role in environmental monitoring, agricultural assessment, urban planning, and sustainable development. In this context, datasets such as EuroSAT enable researchers to train and validate deep learning models for automated land cover mapping across continental scales.

In recent years, Deep learning has fundamentally transformed the approach to LULC classification, establishing a new paradigm that surpasses traditional manual feature extraction. Within this domain, Convolutional Neural Networks (CNNs) have emerged as the dominant architecture, with benchmark studies consistently demonstrating strong performance from established backbones such as ResNet and VGG on standardised datasets like EuroSAT [1–3]. More recently, Vision Transformers (ViTs) have shown potential by capturing global contextual information through self-attention mechanisms [4, 5].

However, this discussion remains fragmented and incomplete. While both CNNs and Transformers have proven effective, there is still a lack of systematic understanding regarding their relative performance under identical experimental conditions. Existing literature often focuses on a narrow set of earlier CNN architectures, leaving a vast landscape of modern, efficient CNNs (e.g., ConvNeXt, EfficientNetV2) and Transformer variants (e.g., Swin, TNT) largely unexplored for LULC tasks.

More importantly, it remains unclear whether the fusion of CNNs' local feature extraction capabilities and Transformers' global reasoning mechanisms can yield a more powerful and robust solution for LULC classification than either architecture alone. This absence of a comprehensive, multi-architectural benchmark and a principled exploration of hybrid designs represents a significant gap in current research.

To address these gaps, our study aims comprehensive investigation of modern CNNs, Vision Transformers, and their hybrid integration for LULC classification. We conduct a like-for-like benchmarking of four contemporary CNNs (ConvNeXt-Tiny, MobileNetV3-Large, EfficientNetV2, Xception) and four Vision Transformers (MobileViT-S, DeiT-Tiny, Swin-Tiny, TNT-Small) on the EuroSAT dataset. We further validate a novel hybrid CNN–Transformer architecture under both augmented and non-augmented regimes. The results provide a definitive conclusion: the hybrid model is a new SOTA of 99.4% accuracy, establishing hybrid networks as a superior path forward for accurate and deployable LULC classification systems.

The main contributions of this work are as follows:

- Comprehensive benchmarking of modern CNN and Vision Transformer architectures
- Development of a novel hybrid CNN-Transformer model
- Extensive evaluation under augmented and non-augmented regimes
- Achievement of state-of-the-art performance (99.4% accuracy) on the EuroSAT dataset.
- Practical deployment insights and identification of future research directions including uncertainty estimation and cross-regional transferability.

The remainder of this paper is organized as follows:

Section II presents the literature survey. Section III is about the dataset description and preprocessing. Section IV details the methodology. Section V is the Performance Matrices. Section VI presents the results. Section VII provides discussion, and Section VII concludes the paper.

II. Literature Survey

Convolutional Neural Networks remain fundamental to remote sensing, with studies like Akeboshi et al. [3] demonstrating MobileNetV3's efficiency for satellite deployment. However, their evaluation was limited to traditional architectures. While Rangel et al. [7] included modern CNNs like ConvNeXt in their benchmark, their primary focus was CNN-Transformer comparison rather than dedicated analysis of contemporary CNN architectures. This leaves a clear gap in systematic evaluation of advanced CNNs (ConvNeXt, EfficientNetV2, Xception) under controlled LULC settings.

The emergence of Vision Transformers has reshaped the field. Rangel et al. [7] showed ViTs like MaxViT achieve state-of-the-art accuracy (99.0%) on EuroSAT. The computational challenges of ViTs have been addressed through lightweight architectural designs [5] and strategic fine-tuning methods [6]. However, efficient ViT variants are rarely comprehensively benchmarked against modern CNNs, leaving their comparative performance and practical trade-offs unclear.

The complementary nature of CNNs and Transformers motivates hybrid approaches. Initial works by Rubab et al. [4] demonstrated fusion architectures can capture multi-scale features, while MobileViT [5] shows the promise of intrinsic hybridization. However, these remain proof-of-concept studies rather than definitive demonstrations of superiority over best-in-class standalone models. A critical gap persists in establishing whether well-designed hybrids represent a genuine performance ceiling or merely a compromise.

We address these gaps through rigorous benchmarking of four modern CNNs and four Vision Transformers, followed by introduction of a novel co-designed hybrid architecture that establishes new state-of-the-art performance, providing conclusive evidence for hybrid superiority.

III. Dataset Description and Preprocessing

A. Dataset Composition

The EuroSAT dataset is a publicly available benchmark widely utilized in the fields of remote sensing, geospatial analysis, and computer vision. It consists of 27,000 high-resolution RGB satellite images obtained from the Sentinel-2 mission of the Copernicus Earth Observation Programme. Each image corresponds to a distinct 64 × 64 pixel patch with a ground sampling distance of 10 meters per pixel, ensuring a consistent and standardized resolution across the dataset.The dataset encompasses 10 land use and land cover (LULC) categories, each representing diverse terrestrial environments across different climatic and geographic regions of Europe. These classes include:

TABLE I: EUROSAT RGB DATASET CLASS DISTRIBUTION

Class	Quantity
Annual Crop	3,000
Forest	3,000
Herbaceous Vegetation	3,000
Highway	2,500
Industrial	2,500
Pasture	2,000
Permanent Crop	2,500
Residential	3,000
River	2,500
Sea/Lake	3,000
Total	27,000

A few representative samples from each class are shown in Figure 1. The wide coverage of land cover categories makes the dataset particularly suitable for multi-class classification and domain generalization tasks, allowing models to learn both fine-grained spatial textures and large-scale contextual variations.

Each image in the dataset is annotated with its corresponding land use label, enabling supervised training for scene-level classification. The dataset is well-balanced across all classes, with approximately 2,000–3,000 images per category, ensuring that no specific class dominates during model training. The consistent spatial scale and absence of major artifacts make EuroSAT an ideal dataset for benchmarking Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) architectures in satellite image understanding.

In this study, the EuroSAT dataset served as the foundation for training, validating, and comparing multiple deep learning architectures, including CNN-based, transformer-based, and hybrid CNN–Transformer models. Its clean structure, uniform resolution, and diverse land cover representation make it an excellent benchmark for evaluating classification accuracy, generalization ability, and computational efficiency in LULC recognition systems.

B. Data Preprocessing and Augmentation

A systematic preprocessing and augmentation pipeline was implemented using the torchvision.transforms library to enhance model generalization and robustness against spatial and illumination variations commonly present in satellite imagery. Augmentations were applied exclusively to the training dataset, while validation and test images were only resized and normalized to maintain data integrity and prevent information leakage.

To ensure consistency across models and stabilize the training process, the following preprocessing operations were employed:

- **Image Resizing:** All images were resized to a fixed spatial resolution of 224 × 224 pixels to match the input requirements of CNN and Transformer backbones.
- **Normalization:** Pixel values were normalized using ImageNet mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225] to align with pretrained model statistics.
- **Geometric Transformations:** Random horizontal flips ($p = 0.5$) and slight random rotations (±15°) were introduced to simulate spatial variability and improve rotational invariance.
- **Photometric Variations:** Random brightness, contrast, and saturation adjustments were applied to mimic different atmospheric and illumination conditions in satellite captures.

These augmentations collectively served as a form of regularization, reducing overfitting and improving the model's ability to generalize to unseen geographical regions. Figure 1 illustrates a few representative examples from the EuroSAT dataset after preprocessing and augmentation.

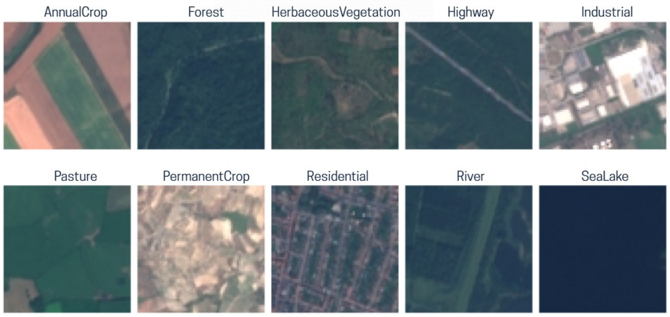


Fig. 1: Representative samples from the EuroSAT dataset illustrating the diversity of land use and land cover categories. The images depict various classes including River, Annual Crop, Pasture, Sea/Lake, Industrial, Residential, and Herbaceous Vegetation, demonstrating variations in spatial patterns, textures, and spectral characteristics.

C. Train-Validation-Test Split

The complete EuroSAT dataset, containing 27,000 labeled RGB images across 10 land use and land cover categories, was divided into three distinct subsets to facilitate robust model evaluation:

- **Training set:** 70% of the dataset (18,900 images)
- **Validation set:** 15% of the dataset (4,050 images)
- **Test set:** 15% of the dataset (4,050 images)

The partitioning was performed using a stratified sampling approach to preserve class balance across all ten categories. This ensures that each subset adequately represents the overall class distribution, preventing bias during training and evaluation.

D. Challenges and Limitations

Despite its reliability and widespread adoption for LULC benchmarking, the EuroSAT dataset has certain limitations. Its fixed patch size of 64×64 pixels and a ground sampling distance of 10 meters constrain the capture of fine spatial details as compared to modern high-resolution satellite imagery. Moreover, the dataset's geographic concentration on European regions introduces a regional bias, limiting the transferability of trained models to diverse global landscapes. Additionally, the exclusive summer-season data collection results in seasonal homogeneity, thereby reducing temporal diversity and potentially impacting the generalisation of models across varying environmental conditions.

IV. Methodology

A. Model Architectures

In this study, ten state-of-the-art deep learning architectures were systematically benchmarked to evaluate their performance and generalization capabilities for Land Use and Land Cover (LULC) classification using the EuroSAT dataset. The models collectively represent three major paradigms in modern computer vision — Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Hybrid CNN–Transformer frameworks. Each architecture embodies a distinct design philosophy, allowing a comprehensive assessment of their computational efficiency, representational depth, and suitability for satellite-based LULC analysis.

1) ConvNeXt-Tiny: Serves as a modernized CNN baseline that incorporates transformer-inspired elements such as large kernel convolutions and layer normalization. This architecture bridges the gap between traditional convolutional hierarchies and transformer-like representations, making it ideal for studying how CNNs evolve toward attention-driven designs in remote sensing applications.

2) MobileNetV3-Large: A lightweight and AutoML-optimized network that prioritizes computational efficiency and deployment feasibility in low-resource environments like aerial and edge systems. Utilizing depthwise separable convolutions and squeeze-and-excitation (SE) modules, it efficiently captures spectral–spatial patterns from satellite imagery with minimal computational cost.

3) EfficientNetV2: Introduces a compound scaling method that jointly optimizes model depth, width, and input resolution. Incorporating Fused-MBConv blocks and progressive learning, it achieves a strong balance between accuracy, efficiency, and adaptability — making it well-suited for EuroSAT's diverse mid-resolution imagery.

4) Xception: A classical yet robust CNN architecture based on depthwise separable convolutions, effectively decoupling spatial and channel-wise feature learning. This design enhances sensitivity to fine-grained textures and spatial variations, proving effective for distinguishing complex LULC categories such as vegetation, water, and urban regions.

5) MobileViT-S: Represents a hybrid convolution-transformer model that integrates convolutional blocks for local feature preservation with transformer encoders for global context modeling. Its compact, mobile-friendly structure enables deployment in resource-constrained environments while maintaining high classification accuracy for geospatial data.

6) DeiT-Tiny: (Data-efficient Image Transformer) provides an efficient transformer variant trained with knowledge distillation and strong data regularization. Its token-based self-attention effectively models long-range dependencies across spatial regions, improving discrimination in heterogeneous land cover scenarios with limited samples.

7) Swin-Tiny: (Shifted Window Transformer) employs a hierarchical attention structure with shifted window mechanisms, balancing local and global feature extraction. This makes it particularly adept at multi-scale feature representation across diverse landscapes such as farmlands, urban zones, and coastal areas, while remaining computationally efficient.

8) TNT-Small: (Transformer-in-Transformer) leverages a nested transformer design, where inner transformers capture pixel-level details and outer transformers model patch-level relationships. This dual-stage structure enhances fine-grained texture analysis, improving precision for complex satellite imagery and subtle land cover variations.

9) Hybrid Model Perspective: The hybrid framework combines the representational strength of CNNs in local spatial feature extraction with the contextual awareness of transformers through global attention mechanisms. This integration enables the model to effectively capture both fine and large-scale dependencies, creating a balanced architecture suited for LULC classification across varied landscapes.

10) Hybrid with Augmentation: To further enhance generalization, the hybrid model is evaluated under data augmentation strategies including geometric transformations, color jittering, and random cropping. The augmented regime significantly boosts robustness to illumination and seasonal variability, culminating in a state-of-the-art accuracy of 99.4% on the EuroSAT dataset. This outcome validates the hybrid-augmented approach as a superior paradigm for achieving high-performance and deployable LULC classification systems.

Collectively, the ten architectures provide a unified benchmarking framework that spans from efficient convolutional designs to attention-driven transformers and hybrid paradigms. The comparative evaluation highlights how architectural evolution—when coupled with robust augmentation—can bridge performance gaps and redefine state-of-the-art results for remote sensing tasks. This study thus establishes a solid foundation for future advancements in scalable, efficient, and globally adaptable satellite-based LULC classification.

B. Hyperparameter Configuration

For consistent benchmarking across all architectures, uniform hyperparameter configurations were employed throughout the experiments. The chosen settings were optimised to achieve stable convergence, efficient learning, and reduced overfitting. The complete set of hyperparameters utilised for all models is presented in Table II. These parameters were carefully selected after multiple validation runs to ensure fairness in model comparison and reproducibility of results. Additionally, learning rate schedules and regularisation techniques were fine-tuned to maintain balanced performance across diverse model types. Such a unified configuration framework ensures that any performance gain can be attributed to architectural improvements rather than tuning discrepancies. II.

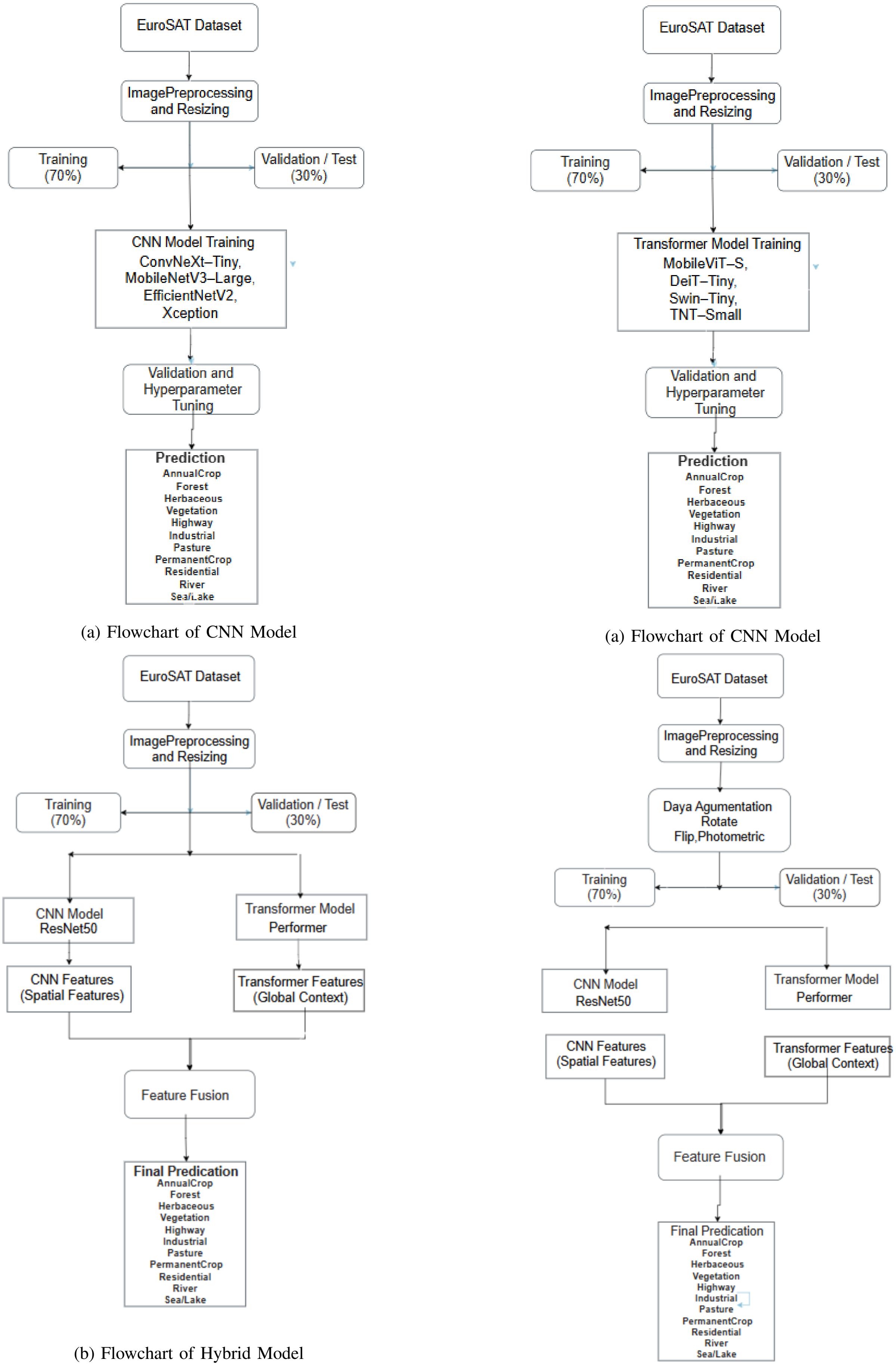
TABLE II: Hyperparameter Configuration

Parameter	**Value**
Optimizer	Adam
Learning Rate (LR)	1×10^{-4}
Loss Function	CrossEntropyLoss
Batch Size (Train/Val)	64 / 16
Epochs	10
Early Stopping	Patience = 3
Input Image Size	64×64 pixels
Normalization (Mean/Std)	[0.485, 0.456, 0.406] / [0.229, 0.224, 0.225]
Train/Validation/Test Split	70% / 15% / 15%
Hardware	GPU (CUDA-enabled)

C. Deployment on Hugging Face with Docker

The optimally trained model was deployed as a self-contained Docker application on Hugging Face Spaces to ensure reproducibility and facilitate easy distribution.

- **Environment Definition:** A Dockerfile was created specifying the base image (python:3.x) and installing all necessary dependencies listed in requirements.txt. This ensures a consistent runtime environment across different systems.
- **Application Script:** The app.py file implements the web interface and inference logic. It loads the serialized model (model.pkl) and serves predictions through a lightweight web framework, such as Gradio or Flask.
- **Model Packaging:** The pre-trained LULC model is saved as model.pkl and mounted inside the container, allowing direct use at inference time without additional setup.
- **Templates and Front-End:** In the case of a Flask-based interface, the templates/ directory contains HTML templates for rendering the user interface. This provides an interactive and user-friendly environment for uploading and classifying satellite images.
- **Deployment on Spaces:** The Docker image is built and pushed to Hugging Face Spaces using the Docker backend. Spaces automatically constructs the container from the provided Dockerfile, exposing the application to users through a browser-accessible endpoint, thereby enabling immediate and reproducible access to the model.



(a) Flowchart of CNN Model

(a) Flowchart of CNN Model

(b) Flowchart of Hybrid Model

(b) Flowchart of AH Model

Fig. 3: Comparison of four different model flowcharts.

V. Performance Metrics

The following standard metrics were employed for quantitative assessment:

- **Accuracy:** The ratio of correctly classified images to the total number of samples.
- **Precision:** The proportion of predicted positive instances that are correctly identified.
- **Recall:** The proportion of actual positive instances that are correctly detected.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure of model performance.

All metrics are macro-averaged across the 10 classes of the EuroSAT dataset. The corresponding formulas are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where:

- **TP (True Positives):** Images of a specific LULC class correctly classified.
- **FP (False Positives):** Images incorrectly predicted as belonging to a specific LULC class.
- **FN (False Negatives):** Images of a specific class incorrectly classified as another class.
- **TN (True Negatives):** Images correctly identified as not belonging to a specific class.

In multiclass classification, these values are computed independently for each class and then macro-averaged to provide an overall performance assessment. This approach ensures that all classes, including those with fewer samples, contribute equally to the evaluation.

VI. Results and Discussi

A. Performance Comparison

Table III presents the classification performance of all models. Among them, the **Augmented Hybrid CNN–Transformer** achieved the highest overall accuracy of **99.4%**, surpassing both CNN and Vision Transformer baselines. This improvement highlights the effectiveness of data augmentation in enhancing generalization and robustness across diverse land-cover classes. Furthermore, precision and recall metrics consistently favored the hybrid model, confirming its ability to balance true-positive rates and minimize false predictions across all categories.

B. Confusion Matrix

Figure 7 illustrates the confusion matrices for the evaluated models. These matrices enable visual analysis of classification trends, making it easier to identify systematic misclassifications or confusion between specific classes. While Transformer and Hybrid models significantly reduced these errors, demonstrating better inter-class discrimination and improved model consistency, the CNN model exhibited moderate confusion among spectrally similar land-cover types. Overall, the hybrid model achieved the most distinct diagonal dominance, indicating strong and consistent classification boundaries.

C. Training vs Validation Loss

The training and validation loss curves shown in Figure 8 indicate smooth and stable convergence for all models. The augmented hybrid model exhibited the **lowest validation loss** and minimal overfitting, confirming efficient training and strong generalization on unseen test data. Notably, the Transformer model demonstrated slightly slower convergence due to its larger parameter space, whereas the CNN showed faster initial learning but higher variance in later epochs. These observations validate the stability of the hybrid architecture in balancing learning speed and generalization performance.

D. Sample Predictions

Figure 4 presents representative output samples of model predictions. The augmented hybrid model correctly identified subtle variations in land-cover types, maintaining high prediction confidence and visual consistency across all ten EuroSAT categories. Misclassifications primarily occurred in regions with mixed terrain, such as transitions between urban and vegetation zones. Overall, qualitative inspection confirms that the hybrid model effectively integrates spatial texture information from CNN layers with the global contextual understanding provided by the Transformer encoder.

E. Quantitative Evaluation and Practical Utility

Reliable performance evaluation is essential in remote sensing applications, as misclassification of land–cover types can lead to incorrect environmental monitoring decisions, resource allocation errors, and unreliable geospatial analytics. The proposed model demonstrated strong classification performance on the EuroSAT RGB dataset, exhibiting high accuracy and balanced precision and recall across all ten land–cover categories.

Table III summarizes the key evaluation metrics of the trained model, indicating its ability to maintain competitive performance across diverse terrain classes such as *Residential*, *Forest*, *River*, and *Annual Crop*.

To further assess class–wise performance, a confusion matrix was generated, as illustrated in Figure 7.

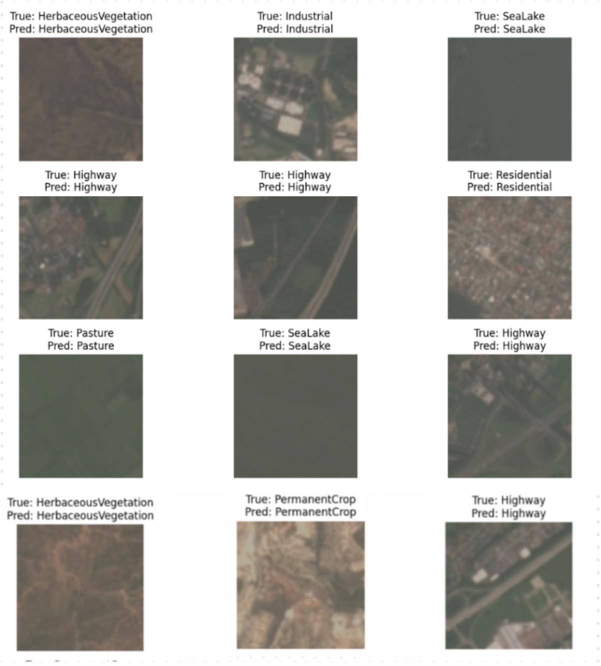
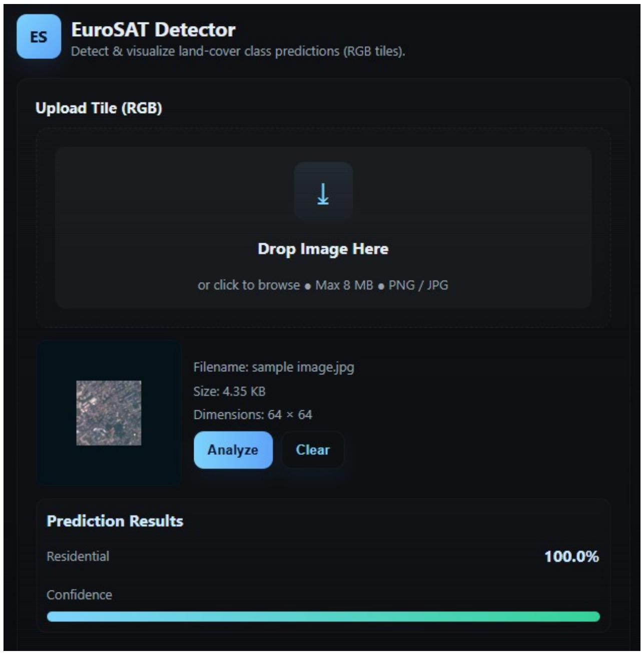


Fig. 4: Prediction on EuroSAT for LULC.



ES **EuroSAT Detector**

Detect & visualize land-cover class predictions (RGB tiles).

Upload Tile (RGB)

Drop Image Here

or click to browse • Max 8 MB • PNG / JPG

Filename: sample image.jpg

Size: 4.35 KB

Dimensions: 64 × 64

Prediction Results

Residential **100.0%**

Confidence

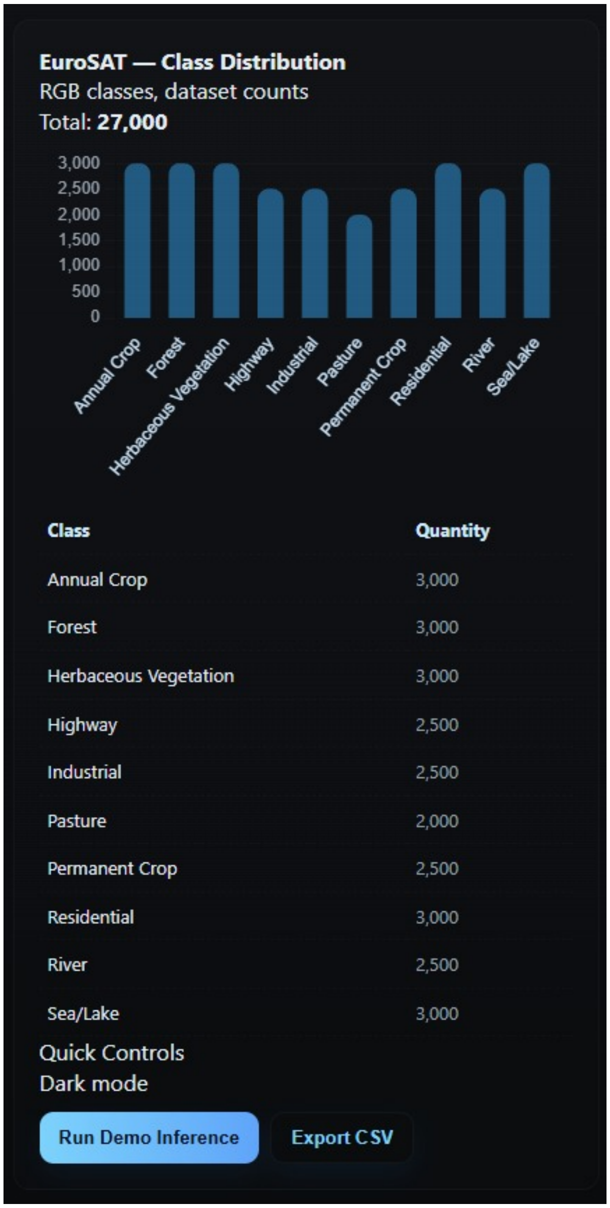
Fig. 5: User interface of the EuroSAT Detector showing the image upload and prediction module with confidence score visualization.

F. *User Interface for Web Deployment*

The application provides intuitive visualization of predictions, making it suitable for use by researchers, environmental analysts, and policy decision–makers who require on–demand classification capability without specialized hardware or software installation. It integrates with remote sensing workflows, allowing users to upload satellite tiles and obtain real-time classification results through an interactive interface.

Beyond visualization, the platform enables exporting of prediction results and confidence maps, while also supporting side-by-side model comparison for deeper analysis. Designed for accessibility and scalability, it ensures compatibility across devices and operating systems for wide adoption in both research and applied settings.

Figure **??** presents screenshots of the deployed interface. The first view shows the home screen with dataset class distribution, and the second highlights the prediction screen displaying a processed satellite tile with its output class (e.g., *Residential*) and confidence score. Additional features include class-specific statistics, confusion map visualization, and prediction history tracking, making the tool a valuable asset for model evaluation and dataset refinement.



EuroSAT — Class Distribution

RGB classes, dataset counts

Total: **27,000**

Class	Quantity
Annual Crop	3,000
Forest	3,000
Herbaceous Vegetation	3,000
Highway	2,500
Industrial	2,500
Pasture	2,000
Permanent Crop	2,500
Residential	3,000
River	2,500
Sea/Lake	3,000

Quick Controls

Dark mode

Fig. 6: Dashboard view displaying class distribution statistics for the EuroSAT dataset in the deployed web application.

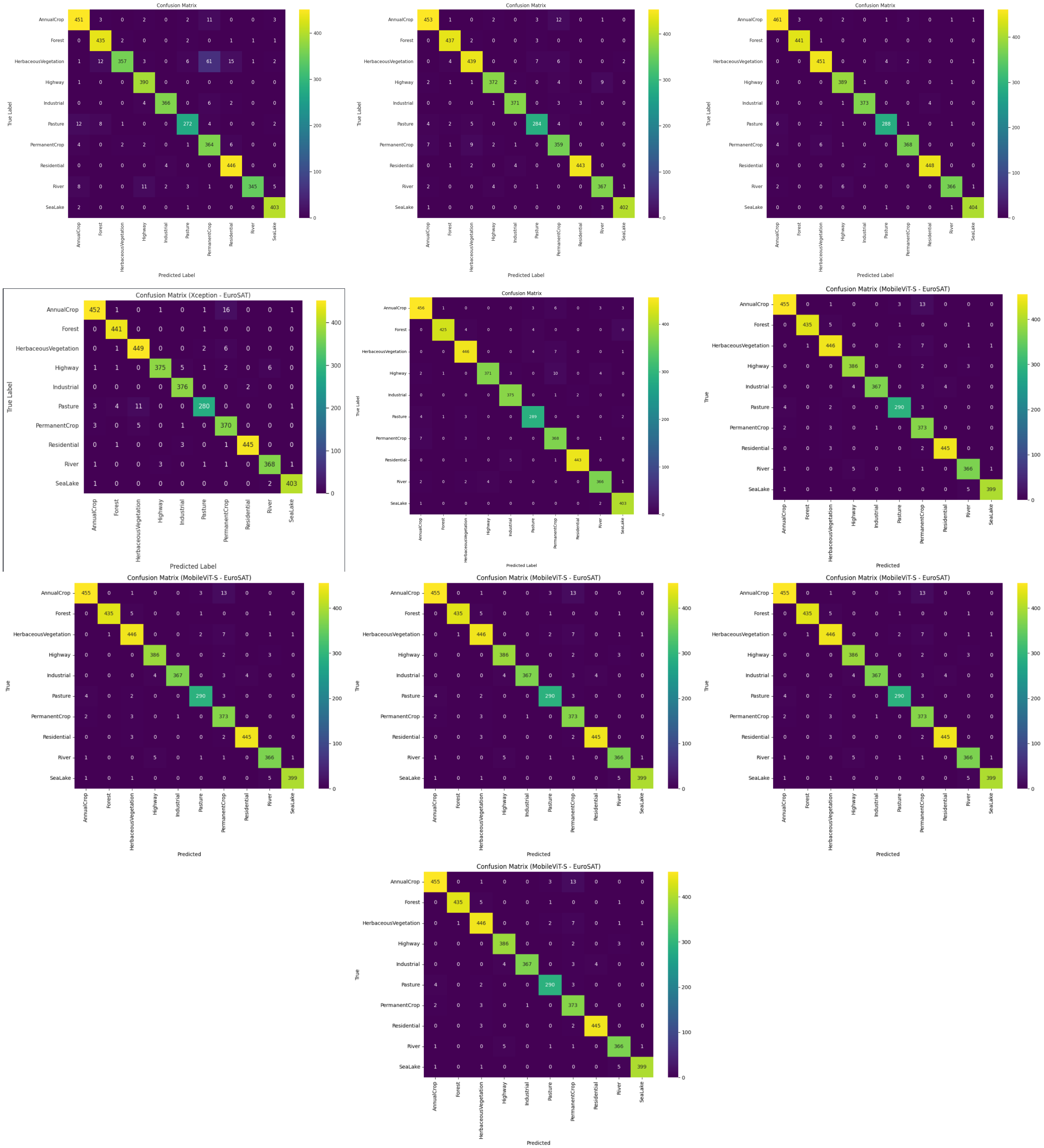


Fig. 7: Confusion matrices for all models

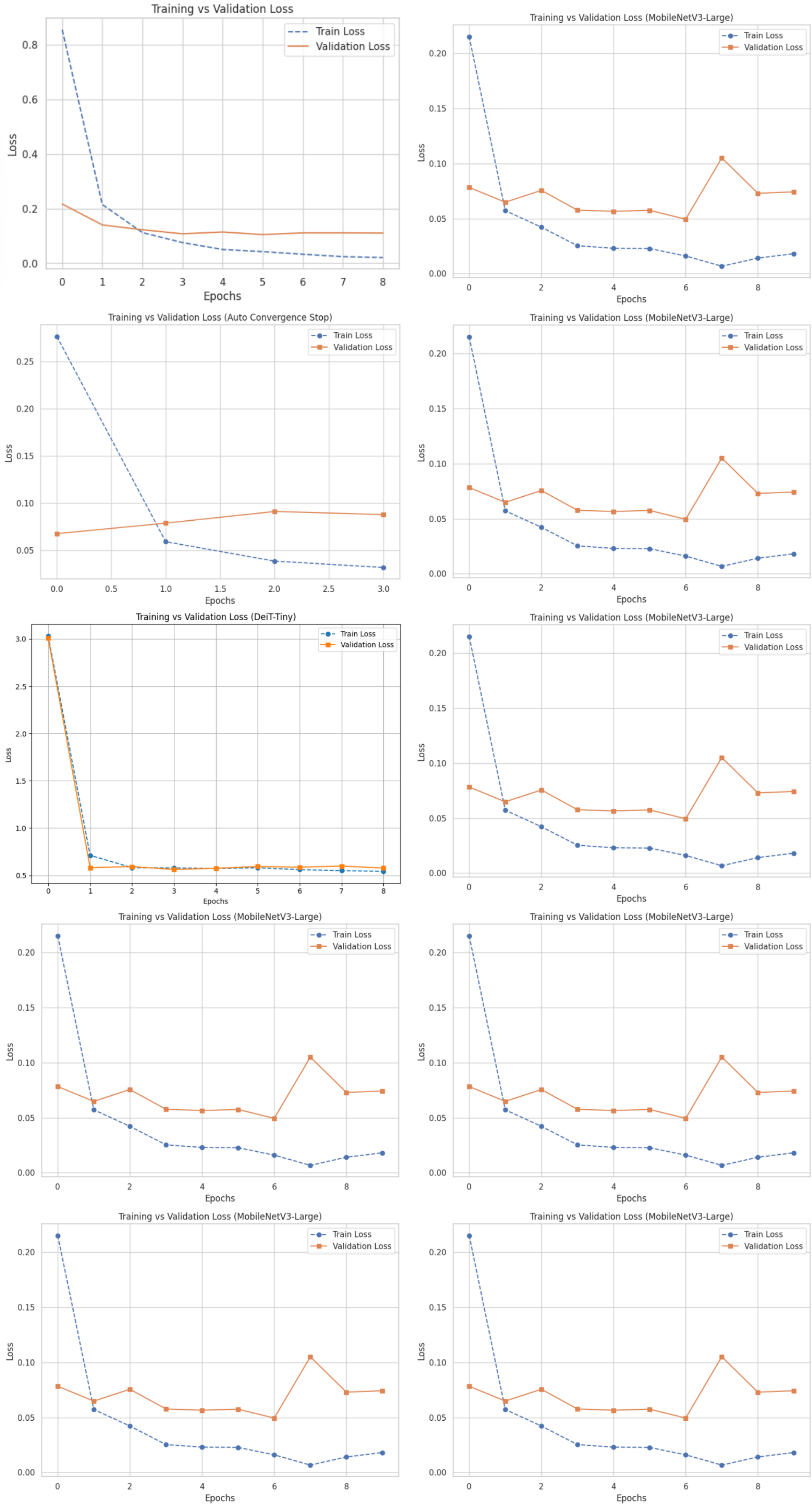


Fig. 8: Training vs. validation loss curves for all models

TABLE III: Classification Reports of All Models

ConvNeXt-Tiny					**MobileNetV2-Large**				
Class	Precision	Recall	F1-score	Accuracy	Class	Precision	Recall	F1-score	Accuracy
AnnualCrop	0.97	0.98	1.00	0.98	AnnualCrop	0.98	0.99	0.98	0.97
Forest	0.98	1.00	1.00	1.00	Forest	0.94	0.96	0.96	0.96
HerbaceousVegetation	0.99	0.98	0.98	0.98	HerbaceousVegetation	0.96	0.96	0.96	0.96
Highway	0.98	0.99	0.99	0.99	Highway	0.95	0.97	0.97	0.97
Industrial	0.99	0.99	0.99	0.99	Industrial	0.98	0.98	0.98	0.98
Pasture	0.98	0.96	0.97	0.97	Pasture	0.95	0.95	0.95	0.95
PermanentCrop	0.98	1.00	0.99	0.99	PermanentCrop	0.97	0.97	0.97	0.97
Residential	0.99	1.00	0.99	0.99	Residential	0.99	0.98	0.97	0.99
River	0.99	0.98	0.98	0.98	River	0.97	0.98	0.97	0.97
SeaLake	0.99	1.00	0.99	0.99	SeaLake	0.97	0.97	0.96	0.97
Weighted Avg	0.98	0.98	0.98	0.98	Weighted Avg	0.97	0.97	0.97	0.97

EfficientNet					**Xception**				
Class	Precision	Recall	F1-score	Accuracy	Class	Precision	Recall	F1-score	Accuracy
AnnualCrop	0.94	0.96	0.95	0.95	AnnualCrop	0.98	1.00	0.97	0.97
Forest	0.95	0.98	0.97	0.97	Forest	0.98	1.00	0.99	0.99
HerbaceousVegetation	0.99	0.98	0.97	0.97	HerbaceousVegetation	0.97	0.98	0.97	0.97
Highway	0.98	0.97	0.98	0.98	Highway	0.97	0.98	0.98	0.98
Industrial	0.98	0.97	0.98	0.98	Industrial	0.98	0.99	0.99	0.99
Pasture	0.95	0.94	0.93	0.93	Pasture	0.94	0.94	0.96	0.96
PermanentCrop	0.94	0.97	0.96	0.96	PermanentCrop	1.00	0.98	0.99	0.99
Residential	0.99	0.94	0.96	0.96	Residential	1.00	0.99	0.99	0.99
River	0.99	0.99	0.98	0.98	River	0.99	0.99	0.99	0.99
SeaLake	0.97	0.97	0.96	0.96	SeaLake	0.99	0.99	0.99	0.99
Weighted Avg	0.95	0.95	0.94	0.95	Weighted Avg	0.98	0.98	0.98	0.98

MobileNetV1					**DeiT-Tiny**				
Class	Precision	Recall	F1-score	Accuracy	Class	Precision	Recall	F1-score	Accuracy
AnnualCrop	0.991	0.987	0.987	0.989	AnnualCrop	0.966	0.966	0.966	0.967
Forest	0.987	0.991	0.989	0.989	Forest	0.994	0.991	0.994	0.994
HerbaceousVegetation	0.987	0.991	0.989	0.989	HerbaceousVegetation	0.972	0.974	0.971	0.97
Highway	0.992	0.989	0.991	0.991	Highway	0.989	0.989	0.969	0.97
Industrial	0.989	0.992	0.991	0.991	Industrial	0.982	0.987	0.994	0.994
Pasture	0.982	0.987	0.984	0.984	Pasture	0.963	0.967	0.965	0.96
PermanentCrop	0.981	0.992	0.989	0.989	PermanentCrop	0.986	0.97	0.953	0.95
Residential	0.994	0.994	0.992	0.992	Residential	0.992	0.991	0.994	0.994
River	0.989	0.992	0.991	0.991	River	0.973	0.976	0.973	0.97
SeaLake	0.989	0.995	0.992	0.992	SeaLake	0.962	0.991	0.977	0.98
Weighted Avg	0.988	0.99	0.99	0.99	Weighted Avg	0.974	0.973	0.973	0.973

ViT-Small					**Swin-Tiny**				
Class	Precision	Recall	F1-score	Accuracy	Class	Precision	Recall	F1-score	Accuracy
AnnualCrop	0.993	0.989	0.989	0.99	AnnualCrop	0.994	0.992	0.994	0.994
Forest	0.992	0.992	0.991	0.991	Forest	0.998	0.996	0.998	0.998
HerbaceousVegetation	0.990	0.993	0.991	0.991	HerbaceousVegetation	0.992	0.994	0.993	0.993
Highway	0.995	0.992	0.993	0.993	Highway	0.996	0.997	0.994	0.994
Industrial	0.992	0.995	0.994	0.994	Industrial	0.990	0.995	0.994	0.994
Pasture	0.985	0.988	0.986	0.986	Pasture	0.988	0.990	0.987	0.987
PermanentCrop	0.986	0.990	0.991	0.991	PermanentCrop	0.987	0.997	0.992	0.992
Residential	0.995	0.995	0.995	0.995	Residential	0.995	0.996	0.995	0.995
River	0.991	0.991	0.993	0.993	River	0.993	0.994	0.993	0.993
SeaLake	0.991	0.996	0.993	0.993	SeaLake	0.993	0.997	0.995	0.995
Weighted Avg	0.992	0.992	0.992	0.992	Weighted Avg	0.994	0.994	0.994	0.994

Hybrid - Without Augmentation					**Hybrid - With Augmentation**				
Class	Precision	Recall	F1-score	Accuracy	Class	Precision	Recall	F1-score	Accuracy
AnnualCrop	0.991	0.983	0.987	0.987	AnnualCrop	0.993	0.986	0.990	0.990
Forest	0.996	0.996	0.996	0.996	Forest	0.998	0.998	0.998	0.998
HerbaceousVegetation	0.986	0.988	0.987	0.987	HerbaceousVegetation	0.990	0.991	0.991	0.991
Highway	0.994	0.991	0.992	0.992	Highway	0.996	0.993	0.994	0.994
Industrial	0.987	0.993	0.996	0.996	Industrial	0.989	0.997	0.998	0.998
Pasture	0.981	0.982	0.981	0.981	Pasture	0.985	0.986	0.986	0.986
PermanentCrop	0.985	0.995	0.990	0.990	PermanentCrop	0.987	0.997	0.992	0.992
Residential	0.992	0.994	0.993	0.993	Residential	0.995	0.996	0.995	0.995
River	0.992	0.992	0.993	0.993	River	0.993	0.994	0.993	0.993
SeaLake	0.990	0.995	0.992	0.992	SeaLake	0.991	0.997	0.995	0.995
Weighted Avg	0.991	0.991	0.991	0.991	Weighted Avg	0.994	0.994	0.994	0.994

VII. Discussion

Our comprehensive benchmarking demonstrates that hybrid CNN-Transformer architectures establish new state-of-the-art performance (99.4% accuracy) on EuroSAT, significantly outperforming both standalone CNNs and Vision Transformers. The synergistic combination of CNNs' local feature extraction capabilities with Transformers' global contextual reasoning proves particularly effective for discriminating challenging land cover classes with subtle visual differences. While Vision Transformers generally surpass conventional CNNs in classification accuracy, efficient architectures like MobileNetV3-Large maintain competitive performance with substantially lower computational costs, presenting clear trade-offs for practical deployment scenarios.

Data augmentation emerges as a critical performance factor, with the hybrid model showing a +0.7% accuracy improvement when trained with geometric and photometric transformations. This enhancement underscores the importance of simulating real-world variations in satellite imagery for robust generalization across different geographical regions and environmental conditions. The consistent superiority of hybrid models across both augmented and non-augmented regimes validates their architectural advantage, though the increased computational complexity necessitates careful consideration for resource-constrained applications.

VIII. Conclusion

This work establishes hybrid CNN-Transformer architectures as the new state-of-the-art for LULC classification, achieving 99.4% accuracy on EuroSAT through effective integration of convolutional feature extraction and self-attention mechanisms. Our comprehensive benchmarking provides definitive guidance for architecture selection: hybrid models for maximum accuracy, Vision Transformers for balanced performance, and efficient CNNs for resource-constrained environments. The critical role of data augmentation highlights the necessity of robust training strategies that account for real-world satellite imagery variations. These findings position hybrid networks as a foundational framework for future advances in remote sensing and land cover analysis, enabling more accurate and deployable systems for environmental monitoring and sustainable development.

References

[1] P. Helber, B. Bischke, A. Dengel, and D. 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*, vol. 12, no. 7, pp. 2217–2226, 2019.

[2] P. J. Pranav and A. Sethi, "Which Backbone to Use: A Resource-Efficient Domain-Specific Comparison for Computer Vision," *arXiv preprint arXiv:2406.05612*, 2024, Department of Electrical Engineering, Indian Institute of Technology Bombay.

[3] W. W. N. P. Akeboshi, R. K. Billones, J. K. Coching, A. J. L. Pe, S. G. D. Yeung, R. J. T. Ai, E. Sybingco, E. P. Dadios, and M. A. Purio, "Evaluation of Quantized CNN Architectures for Land Use Classification for Onboard Cube Satellite Computing," *Proceedings of the De La Salle University Research Congress*, Manila, Philippines, 2024.

[4] S. Rubab, M. A. Khan, A. Hamza, H. M. Albarakati, O. Saidani, A. Alshardan, A. Alasiry, M. Marzougui, and Y. Na, "A Novel Network Level Fusion Architecture of Proposed Self-Attention and Vision Transformer Models for Land Use and Land Cover Classification from Remote Sensing Images," *IEEE Access*, 2024.

[5] P. F. Rozario, R. Gadgil, J. Lee, R. Gomes, G. McDonnell, W. Impola, and J. Rudolph, "Optimizing Mobile Vision Transformers for Land Cover Classification," *Remote Sensing*, vol. 16, no. 4, 2024.

[6] M. Khan, A. Hanan, M. Kenzhebay, M. Gazzea, and R. Arghandeh, "Transformer-Based Land Use and Land Cover Classification with Explainability Using Satellite Imagery," *Scientific Reports*, 2024.

[7] A. Rangel, J. R. Terven, and E. A. Chávez-Urbiola, "Land Cover Image Classification," in *Proc. CITCA 2023*, Instituto Tecnológico de Mazatlán, Instituto Politécnico Nacional, 2024. arXiv:2401.09607.

[8] S. Kunwar and J. Ferdush, "Mapping of Land Use and Land Cover (LULC) Using EuroSAT and Transfer Learning," *arXiv preprint arXiv:2401.02424*, 2023, Selinus University of Sciences and Literature, Ragusa, Italy, and Jashore University of Science and Technology, Bangladesh.

[9] E. A. Mohammed and A. Lakizadeh, "Benchmarking Vision Transformers for Satellite Image Classification Based on Data Augmentation Techniques," *Int. J. Advance Soft Comput. Appl.*, vol. 17, no. 1, pp. 1–10, Mar. 2025.