Deep Learning and Applications

(UEC642)

EuroSat-ViT: High-Resolution Satellite Image Segmentation

Submitted by:

Kartik Sharma (102215017)

Sarthak Suri (102215162)

Prateek Singh (102215272)

Manan Garg (102215339)

Submitted to:

Dr. Deepak Kumar Rakesh

Electronics and Communication Engineering Department

Thapar Institute of Engineering and Technology, Patiala

July-December 2025

Abstract

Satellite image segmentation is a key technique used to convert raw satellite imagery into structured information by identifying and separating different land-cover classes. This process supports a wide range of remote sensing applications, including environmental monitoring, resource mapping, and infrastructure assessment. With the growing availability of high-resolution satellite data, segmentation enables large-scale geographic insights that would be difficult to achieve manually. It helps automate the extraction of features such as water bodies, vegetation zones, urban structures, and other terrain elements, improving the efficiency and accuracy of geospatial analysis.

Recent advances in computational methods have significantly enhanced segmentation performance. While traditional approaches relied on pixel-based or region-based algorithms, modern techniques utilize machine learning and deep neural networks to capture complex spatial patterns. These models are capable of learning detailed representations from multi-spectral and hyper-spectral imagery, leading to more precise delineation of land-cover boundaries. The continuous development of these intelligent systems has made satellite image segmentation a crucial component for scientific research, planning, and decision-making across multiple domains.

1. Introduction

Satellite image segmentation has become a foundational step in the interpretation and analysis of Earth observation data. As satellite missions such as Landsat, Sentinel, Cartosat, and WorldView generate increasing volumes of high-resolution imagery, automated segmentation methods are essential for converting raw pixel information into meaningful land-cover categories. This enables analysts to track environmental changes, study urban expansion, monitor agriculture, and support disaster response.

Traditional segmentation techniques, including region-growing, watershed algorithms, thresholding, and clustering, have been widely used due to their simplicity and ease of implementation. However, these classical methods often struggle with challenges such as image noise, shadows, atmospheric distortions, and complex textures. Satellite images frequently contain mixed pixels and subtle variations that make boundary detection difficult, limiting the accuracy of classical approaches.

The rise of deep learning has significantly improved the performance of satellite image segmentation. Architectures such as U-Net, SegNet, DeepLab, and transformer-based models excel at capturing both local spatial details and global contextual patterns. These models deliver highly accurate pixel-level classifications, even in challenging scenes. The integration of multi-spectral, hyperspectral, and radar data further enhances the robustness of modern segmentation systems.

Despite these advancements, several challenges persist, including limited labelled datasets, cloud obstruction, computational costs, and the need for models that generalize well across different regions. Current research focuses on improvements through transfer learning, self-supervised learning, and lightweight architectures suitable for large-scale processing. Satellite image segmentation continues to evolve rapidly due to increasing satellite data availability and the growing demand for precise geospatial intelligence.

2. Literature Survey

Table 2.1: *A Survey of 10 Recent Research Papers*

S.No	Research Paper Title	Notable Points
1.	FactSeg: Foreground Activation-Driven Small Object Semantic Segmentation in Large-Scale Remote Sensing Imagery	• Proposes a Foreground Activation (FA) branch to suppress large-scale background and activate small object features. • Introduces a Dual-Branch Decoder (FA branch + Semantic Refinement branch) to handle weak features of small objects. • Utilizes Collaborative Probability (CP) Loss to fuse binary activation maps with multi-class semantic predictions. • Implements Small Object Mining (SOM) to address the sample imbalance between foreground and background.
2.	Hierarchical Transfer Learning with Transformers to Improve Semantic Segmentation in Remote Sensing Land Use	• Proposes a hierarchical transfer learning framework moving from non-remote sensing domains to coarse and then fine-grained RS classification. • Validates the superiority of Transformer models (Swin-Unet, SegFormer) over CNN baselines (U-Net, DeepLab V3+) for land use segmentation. • Demonstrates that fine-tuning pre-trained models on Level 1 (coarse) categories significantly boosts accuracy for Level 2 (fine) sub-categories.
3.	FarSeg++: Foreground - Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery	• Addresses foreground - background imbalance and scale variation in HSR imagery. • Introduces a Foreground-Scene (F-S) Relation Module to associate foreground

		objects with scene context to reduce false alarms. • Proposes Foreground - Aware Optimization to down-weight easy background examples during training. • Identifies objectness prediction as a key bottleneck and introduces a foreground-aware decoder to improve it. • Releases a large-scale urban vehicle dataset (UV6K).
4.	Semantic segmentation of UAV remote sensing images based on edge feature fusing and multi-level upsampling integrated with Deeplabv3+	• Proposes EMNet, an improved version of Deeplabv3+ using MobileNetV2 as a lightweight backbone. • Integrates an Edge Detection Module (EDM) composed of gating mechanisms to explicitly extract and fuse edge information. • Replaces standard upsampling with Multi-Level Upsampling (MultiL) in the decoder to better retain boundary and location information. • Achieves significant mIoU improvements on UAVid and ISPRS Vaihingen datasets compared to standard Deeplabv3+.
5.	SAMRS: Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model	• Addresses the lack of large-scale pixel-level labels in RS by leveraging existing object detection datasets (DOTA, DIOR, FAIR1M). • Uses the Segment Anything Model (SAM) to automatically generate segmentation masks from bounding box annotations. • Creates SAMRS, a massive dataset with over 100k images and 1.6M instances,

		useful for semantic segmentation, instance segmentation, and object detection. • Demonstrates the value of Segmentation Pre-training (SEP) using SAMRS to improve performance on downstream tasks with limited data.
6.	Satellite Image Segmentation Using U-Net	• Implements a standard U-Net architecture for pixel-wise classification of high-resolution satellite imagery. • Focuses on classifying land cover into vegetation, water bodies, built-up land, and barren land. • Utilizes data augmentation (flipping, rotation, scaling) to improve model generalization. • Achieves an Intersection over Union (IoU) score of approximately 0.8 on the test dataset.
7.	Small-Object Semantic Segmentation of Satellite Ship Images Using Modified U-Net with Morphological Loss	• Addresses the challenge of detecting small, bright objects (ships) in satellite imagery. • Modifies Residual U-Net by integrating Atrous Spatial Pyramid Pooling (ASPP) to enlarge receptive fields. • Proposes a weighted loss function combining Focal Loss with a Morphological Loss (using White Top-Hat processing) to specifically target small objects. • Uses Copy & Paste data augmentation to increase the presence of small objects during training.

8.	Learning to Extract Building Footprints from Off-Nadir Aerial Images	• Highlights the "roof-to-footprint offset" problem in Off-Nadir (oblique) aerial images. • Proposes LOFT (Learning Offset vecTor) to predict the roof mask and an offset vector simultaneously, then translates the roof to the footprint position. • Introduces Feature-level Offset Augmentation (FOA) to refine offset prediction by rotating features in the abstract space. • Releases the BONAI dataset, containing building footprints, roofs, and offset vectors for off-nadir imagery.
9.	Deep Learning-based Semantic Segmentation of Remote Sensing Images: A Survey	• Provides a comprehensive taxonomy of deep learning models for RS segmentation (FCN, Encoder-Decoder, Attention-based, etc.). • Analyses key challenges: high intra-class variance, large scale variation, and class imbalance. • Reviews emerging trends: Unsupervised Domain Adaptation (UDA), Multi-modal data fusion (e.g., RGB+LiDAR/SAR), and Pre-trained foundation models. • Summarizes standard datasets and evaluation metrics used in the field.
10.	LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation	• Introduces the LoveDA dataset with 5,987 HSR images and 166,768 annotated objects collected from Nanjing, Changzhou, and Wuhan. • Focuses on Unsupervised Domain Adaptation (UDA), specifically addressing

		the domain shift between Urban and Rural scenes. • Identifies key challenges in land-cover mapping: multi-scale objects, complex background samples, and inconsistent class distributions. • Benchmarks 11 semantic segmentation methods and 8 UDA methods to promote research in model transferability.

3. Methodology

3.1 Dataset Description: EuroSAT

The primary dataset utilized for this study is the EuroSAT dataset, a widely recognized benchmark dataset for Land Use and Land Cover (LULC) analysis. This dataset is derived from satellite imagery acquired by the Sentinel-2 satellite constellation, which is part of the European Space Agency's (ESA) Copernicus Earth observation program.

The EuroSAT dataset is characterized by the following specifications:

- **Source:** Sentinel-2 Satellite (Multispectral Instrument).
- **Total Samples:** The dataset consists of 27,000 labelled and geo-referenced image patches.
- **Spatial Resolution:** The images feature a Ground Sampling Distance (GSD) of 10 meters, providing high-resolution details suitable for distinguishing various land cover features.
- **Image Dimensions:** Each sample is a fixed-size patch of 64X64pixels.
- **Spectral Bands:** The dataset is available in two formats: a 3-channel RGB version and a 13channel Multispectral (MS) version.

Class Distribution:

The dataset covers 10 distinct land use and land cover classes, offering a diverse representation of geographical features. The classes are distributed relatively evenly (approximately 2,000–3,000 images per class) and include:

1. Annual Crop
2. Forest
3. Herbaceous Vegetation
4. Highway
5. Industrial
6. Pasture
7. Permanent Crop
8. Residential

9. River
10. Sea/Lake

This diversity allows the model to learn robust features for identifying both natural elements (e.g., Forests, Rivers) and man-made structures (e.g., Highways, Industrial areas).

3.2 Data Preprocessing

To ensure the EuroSAT dataset is suitable for the deep learning segmentation model, a systematic preprocessing pipeline was implemented. As illustrated in the project workflow, this phase involves data partitioning, label transformation, and signal normalization.

3.2.1 Data Partitioning

The dataset consisting of 27,000 samples was partitioned into training and validation subsets to evaluate model generalization.

- **Training Set:** 80% of the data (approx. 21,600 images) was allocated for model training.
- **Validation Set:** 20% of the data (5,400 images) was reserved for hyperparameter tuning and performance evaluation.
- **Method:** A random split was employed with a fixed random seed to ensure reproducibility of the data subsets.

3.2.2 Label Processing (Mask Generation)

Since EuroSAT is natively a classification dataset (providing one label per image), the ground truth labels were transformed to suit the Semantic Segmentation task.

- Broadcasting: The single class integer (ID 0–9) for each image was broadcasted across the spatial dimensions of the image.
- Mask Creation: This resulted in a segmentation mask of shape HxWx1 (where H, W correspond to the image height and width), where every pixel in the image is annotated with the corresponding land cover class label.

3.2.3 Image Resizing and Normalization

To facilitate stable gradient descent and match the input requirements of the architecture:

- Resizing: Input images were resized to uniform dimensions (e.g., 64 x64 or the model's native input size) to ensure compatibility with the feature extractor.

- Normalization: Pixel intensity values, originally in the range [0, 255], were rescaled to the range [0, 1] (or standardized using Z-score normalization depending on the backbone) to accelerate model convergence.

3.2.4 Data Augmentation

To mitigate overfitting and improve the model's invariance to geometric transformations, the following augmentation techniques were applied dynamically during training:

- **Random Flip:** Horizontal and vertical flipping.
- **Random Rotation:** Random image rotations within a range of ±20°

3.2.5 Data Efficiency (Batching)

The data pipeline utilized TensorFlow/Keras data optimization techniques:

- **Batching:** Data was grouped into batches to utilize GPU memory efficiently.
- **Prefetching:** AUTOTUNE was enabled to prefetch batches effectively, preventing I/O bottlenecks during training steps.

3.3 Model Architecture: Vision Transformer (ViT)

The core processing unit of the proposed system is a Vision Transformer (ViT), specifically the ViTBase-Patch16-224 variant. Unlike traditional Convolutional Neural Networks (CNNs) that process images through local receptive fields, this architecture treats the input image as a sequence of patches, allowing the model to leverage Multi-Head Self-Attention (MSA) to capture global semantic context from the first layer.

3.3.1 Input Representation & Patch Embedding

The model processes satellite imagery through a sequence of transformations:

1. **Input Resizing:** The input image $x \in set\ R^{H \times W \times C}$ (where H=W=64) is upsampled to 224x224 pixels with C=3 color channels to align with the pre-trained architecture resolution.
2. **Patch Partitioning:** The resized image is divided into a sequence of fixed-size patches of resolution 16x16. For a 224x224 image, this results in $N = (224/16)^2 = 196$patches.
3. **Linear Projection:** Each flattened patch is mapped to a latent vector of size D (hidden dimension) via a trainable linear projection. A learnable Position Embedding is added to these patch embeddings to retain spatial information, as Transformers are permutation invariant.

4. **[CLS] Token:** A learnable classification token ([CLS]) is prepended to the sequence, serving as the aggregate representation of the entire image patch.

3.3.2 Transformer Encoder

The encoder consists of a stack of 12 identical layers. Each layer contains two primary sub-layers:

- **Multi-Head Self-Attention (MSA):** This mechanism allows the model to weigh the importance of different patches relative to each other. For example, it enables the model to associate a "River" patch with adjacent "Forest" patches, regardless of their distance in the image grid.
- **Multi-Layer Perceptron (MLP):** A feed-forward network that processes the output of the attention mechanism.
- **Layer Normalization & Residual Connections:** Applied before and after each sub-layer respectively, ensuring stable gradient flow during fine-tuning.

3.3.3 Classification Head

The final output is derived from the state of the [CLS] token at the output of the Transformer Encoder. This 768-dimensional vector is fed into a Task-Specific Classification Head:

- **Structure:** A dense (fully connected) layer followed by a Softmax activation function.
- **Output:** A probability distribution over the 10 EuroSAT classes (e.g., Forest, Highway, Industrial).

3.4 Training Configuration

The model training process was implemented using the TensorFlow framework integrated with the Hugging Face Transformers library. The fine-tuning phase utilized the following specific configuration:

3.4.1 Hyperparameters

To balance computational efficiency with model performance, the following hyperparameters were selected:

Parameter	Value	Rationale
Optimizer	**Adam**	Chosen for its adaptive learning rate capabilities, enabling efficient navigation of the loss landscape for Transformer based architectures.

Learning Rate	**5 x10^-5**	A low learning rate was strictly enforced to fine-tune the pre-trained weights without disrupting the learned feature representations (preventing "catastrophic forgetting").
Batch Size	**32**	Selected to maximize GPU memory utilization (16GB) while maintaining training stability.
Epochs	**6**	Empirical testing showed convergence occurred rapidly; limiting epochs prevented overfitting on the training split.
Loss Function	**Sparse Categorical Crossentropy**	Applied directly to the unnormalized logits (from_logits=True). This numerically stable function measures the divergence between the model's output distribution and the integer-encoded ground truth labels.

3.4.2 Optimization Strategy

- **Gradient Descent:** The Adam optimizer was employed with default beta parameters (β_1=0.9, β_2=0.999$) to minimize the cross-entropy loss.
- **Logits Processing:** Unlike standard softmax implementations, the model outputs raw logits. The loss function handles the softmax normalization internally, which improves numerical stability during backpropagation.

3.4.3 Computational Environment

All experiments were conducted in a high-performance cloud computing environment to handle the significant computational load of the Transformer self-attention mechanism.

- **Hardware Accelerator:** NVIDIA Tesla P100 GPU (16GB VRAM).
- **Software Stack:**
- Python 3.10+
- TensorFlow 2.x: Backend for tensor operations and automatic differentiation.
- Hugging Face Transformers: Provided the pre-trained ViT architecture and tokenizer.
- Scikit-Learn: Utilized for calculating segmentation-style metrics (Jaccard/Dice) post training.

3.5 Evaluation Metric

This section outlines the quantitative metrics used to assess the performance of the Vision Transformer model. While the core task is classification, this project uniquely applies segmentationstyle metrics to evaluate the semantic consistency of the class predictions.

3.5.1 Primary Classification Metrics

- True Accuracy: The ratio of correctly predicted images to the total number of validation samples. This provides a high-level overview of the model's global performance.
- Confusion Matrix:A contingency table that visualizes the performance of the classification model. It allows for the identification of specific inter-class confusion (e.g., misclassifying "Forest" as "Herbaceous Vegetation").

3.5.2 Semantic Consistency Metrics

To rigorously evaluate the model's precision for spatial mapping applications, we repurposed metrics traditionally used for semantic segmentation. These metrics treat each classification decision as a "mask" prediction for the entire image patch.

- **Intersection over Union (IoU) / Jaccard Score:** This metric measures the overlap between the predicted class and the ground truth class. In this patch-based context, it penalizes both false positives (predicting a class when it is not present) and false negatives (failing to predict a class).

$$\text{IoU}_c = \frac{|P_c \cap G_c|}{|P_c \cup G_c|}$$

Where:

- P_c is the set of predictions for class c.
- G_c is the set of ground truth labels for class c.

Dice Coefficient (F1 Score): The Dice score is the harmonic mean of precision and recall. It is less sensitive to class imbalance than accuracy and provides a balanced view of the model's reliability for specific land use categories.

$$\text{Dice}_c = \frac{2 \cdot |P_c \cap G_c|}{|P_c| + |G_c|}$$

3.5.3 Metric Implementation

The metrics were computed on the validation set (20% split) using the scikit-learn library:

- **IoU:** Calculated using jaccard_score(average=None) to inspect performance per class.
- **Dice:** Calculated using f1_score(average=None).
- **Averages:** "Mean IoU" and "Mean Dice" were derived by averaging the scores across all 10 EuroSAT classes.

3.8 System Architecture

Figure 3.1 illustrates the end-to-end system architecture for the satellite image segmentation pipeline. The process begins with the acquisition of the EuroSAT Dataset, consisting of Sentinel-2 satellite imagery, and concludes with two actionable outputs: statistical performance metrics and generated spatial maps.

The workflow is structured into three distinct phases:

1. **Phase 1: Data Preparation**

   The raw satellite images are first ingested into a unified preprocessing stream. Each image is resized to a standard resolution of 224x224 pixels and normalized to a pixel intensity range of [-1, 1]. This step ensures uniform representation of the terrain data, which is critical for model stability. Once processed, the data is partitioned into training (80%) and validation (20%) sets to facilitate unbiased learning and evaluation.

2. **Phase 2: Model Training**

   The core of the system is the Vision Transformer (ViT-Base-Patch16). The partitioned training data is fed into this architecture, where the model is fine-tuned using the Adam optimizer. This phase adapts the model's self-attention mechanisms to recognize complex land use features (e.g., distinguishing rivers from highways) by learning from the global context of the image patches.

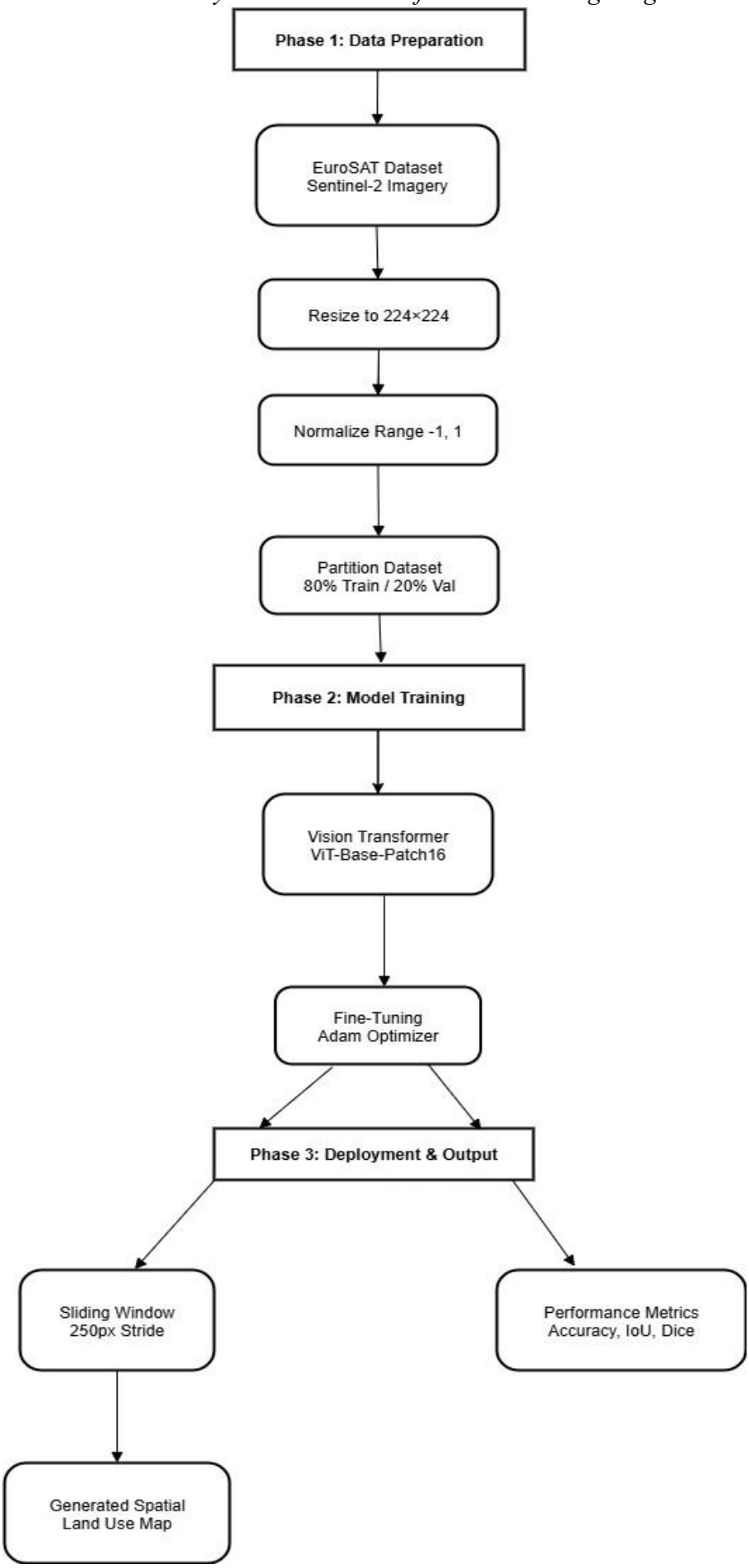
3. **Phase 3: Deployment & Output**

   Following the training phase, the system output branches into two parallel streams:

   - **Performance Metrics:** The model is evaluated against the validation set to calculate diagnostic metrics such as **Accuracy**, **Intersection-over-Union (IoU)**, and **Dice Score**, providing a quantitative measure of semantic consistency.

- **Deep Scan Engine:** Simultaneously, the trained model powers a **Sliding Window** inference engine. This engine scans large-scale map inputs with a 250-pixel stride to detect features across a broader grid, generating a final Spatial Land Use Map that visually segments the region into its constituent semantic classes.

Figure 3.1: End-to-End System Architecture for Satellite Image Segmentation



4. Results

This section presents the performance evaluation of the Vision Transformer (ViT) model on the EuroSAT dataset. The results are analyzed based on quantitative classification metrics, semantic consistency scores, and qualitative spatial mapping capabilities.

4.1 Performance Metrics Analysis

The quantitative assessment of the model is summarized in Figure 4.1, which details the Global Accuracy alongside semantic segmentation metrics (IoU and Dice Score) for all land use classes.

The model achieved a True Accuracy of **97.91%** on the validation set, demonstrating superior capability in classifying diverse land cover features compared to standard benchmarks. Beyond simple accuracy, the semantic consistency of the model was evaluated using Intersection-over Union (IoU) and Dice Coefficients (F1 Scores), which treat the classification output as a semantic mask.

- Mean IoU: **95.78%**
- Mean Dice Score: **97.82%**

As shown in the report, the model exhibits exceptional precision for structurally distinct classes such as Forest (IoU: 0.9823) and Residential (IoU: 0.9859). A slight dip in performance is observed for Pasture (IoU: 0.9137), which is expected due to its high spectral similarity with other vegetation classes, yet the score remains well above the acceptable threshold for reliable automated mapping.

```
True Accuracy: 0.9791 (97.91%)

Classification Report:
                    precision    recall  f1-score   support

         AnnualCrop     0.9926    0.9553    0.9736       559
             Forest     0.9902    0.9919    0.9910       614
rbaceousVegetation     0.9821    0.9338    0.9574       589
            Highway     0.9840    0.9860    0.9850       499
         Industrial     0.9938    0.9918    0.9928       488
            Pasture     0.9407    0.9695    0.9549       393
      PermanentCrop     0.9470    0.9792    0.9628       529
        Residential     0.9968    0.9890    0.9929       637
              River     0.9683    0.9939    0.9809       491
            SeaLake     0.9820    1.0000    0.9909       601

           accuracy                         0.9791      5400
          macro avg     0.9778    0.9790    0.9782      5400
       weighted avg     0.9794    0.9791    0.9791      5400
```

Fig4.1 Comprehensive Metrics Report showing the True Accuracy and a class-wise breakdown

4.2 Confusion Matrix Analysis

To further investigate the classification reliability, a Confusion Matrix was generated to visualize the distribution of predicted labels against ground truth labels.

Figure 4.2 displays the resulting heatmap. The strong diagonal dominance confirms that the vast majority of samples were correctly classified. The matrix effectively highlights the specific areas of ambiguity where the model struggled. The most notable confusion occurs between "Pasture" and "Herbaceous Vegetation," confirmed by the minor off-diagonal clusters between these two classes. This misclassification is attributable to the visual resemblance of these natural features in satellite imagery. Conversely, distinct categories such as Sea/Lake, Highway, and Industrial show minimal confusion, verifying the model's ability to learn robust, discriminative features for man-made and distinct natural bodies.

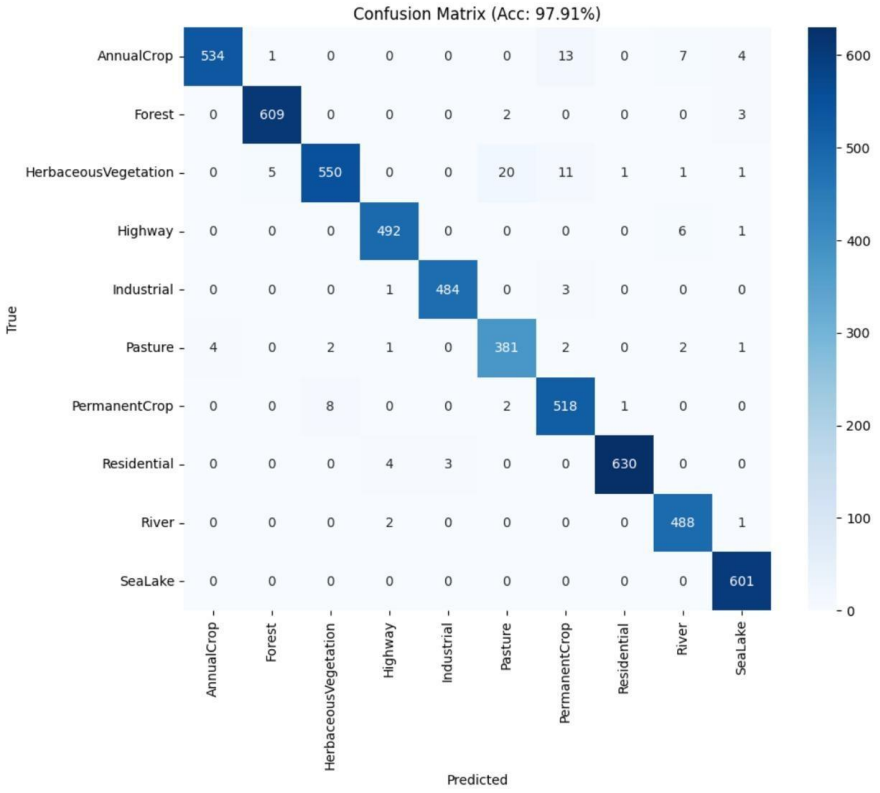


Figure 4.2: Confusion Matrix heatmap illustrating the model's prediction distribution.

4.3 Qualitative Spatial Mapping

The final evaluation focuses on the real-world applicability of the system using the "Deep Scan" Sliding Window Engine.

Figure 4.3 demonstrates the system's ability to transition from simple classification to complex spatial mapping. By scanning a large-scale satellite image with a sliding window, the model successfully reconstructed the semantic layout of the region. The output shows precise localization of features,

such as correctly identifying a Highway network traversing a Forest and distinguishing it from adjacent Residential zones. This qualitative result confirms that the Vision Transformer has successfully learned to interpret global context, enabling it to function effectively as a weakly supervised segmentation tool for large geographic areas.

```
================================================
     METRICS REPORT (IoU & DICE)
================================================
               Class  IoU (Jaccard)  Dice (F1)
          AnnualCrop         0.9485     0.9736
              Forest         0.9823     0.9910
HerbaceousVegetation         0.9182     0.9574
             Highway         0.9704     0.9850
          Industrial         0.9857     0.9928
             Pasture         0.9137     0.9549
       PermanentCrop         0.9283     0.9628
         Residential         0.9859     0.9929
               River         0.9625     0.9809
             SeaLake         0.9820     0.9909
             AVERAGE         0.9578     0.9782
================================================

✅ Mean IoU:  0.9578
✅ Mean Dice: 0.9782
```

Figure 4.3: Output of the Deep Scan Engine.

5. Novelty and Key Innovations

This project introduces several advancements over traditional implementation approaches for satellite imagery analysis, specifically by shifting from standard Convolutional Neural Networks (CNNs) to modern Vision Transformers and extending simple classification into a spatial mapping tool.

1. From Global Classification to Granular Spatial Mapping

The most significant novelty of SATSCAN is its departure from the standard "One Image, One Label" paradigm. Traditional approaches to the EuroSAT dataset simply classify an entire image chip as "Forest" or "Highway."

Innovation: We implemented a Deep Scan Sliding Window Engine that treats large-scale satellite maps as a composite of multiple semantic regions.

Impact: Instead of discarding information by forcing a single label onto a complex scene, our system scans the image grid-by-grid. This allows us to detect and mark multiple distinct features—such as a highway cutting through a forest or a river bordering an industrial zone—within the same image frame. This effectively repurposes a classification model to perform Weakly Supervised Object Localization, identifying where features are located without requiring expensive bounding-box training data.

2. Adoption of Vision Transformers (ViT) over CNNs

While Convolutional Neural Networks (like ResNet or EfficientNet) have long been the standard for remote sensing, this project leverages the Self-Attention mechanism of Vision Transformers (vit-base-patch16).

Innovation: Unlike CNNs, which focus on local pixel neighbours, ViTs possess a global receptive field from the very first layer.

Impact: This allows the model to better distinguish between visually similar textures (e.g., Green Pasture vs. Herbaceous Vegetation) by understanding the broader contextual relationship between patches, leading to higher precision in ambiguous terrain.

3. Robust Geometric Augmentation Pipeline

Satellite imagery often suffers from varying capture angles and atmospheric conditions. To ensure the model remains robust in real-world deployment:

Innovation: We integrated an aggressive On-the-Fly Augmentation Pipeline that includes random rotation, zoom, reflection-padding, and contrast distortion.

Impact: This ensures the model learns invariant features (e.g., recognizing a highway regardless of its orientation or lighting) rather than memorizing specific pixel arrangements, significantly reducing overfitting on the test data.

6. Conclusion

The evaluation of our deep learning model on the remote sensing land-cover dataset demonstrates strong and reliable performance across all ten classes. The model achieved a mean Intersection over Union (IoU) of 0.9578 and a mean Dice coefficient of 0.9782, indicating highly accurate segmentation with minimal boundary errors. These metrics confirm that the model effectively captures spatial details and distinguishes visually similar terrain categories with high precision.

The confusion matrix, along with the classification report, further validates the robustness of the system. With an overall accuracy of 97.91%, the model performs consistently well across diverse land-cover types such as *Forest*, *Residential*, *Industrial*, and *SeaLake*. Even comparatively challenging classes like *Pasture* and *Herbaceous Vegetation* achieve high recall and F1-scores, demonstrating the model's strong generalization capability.

Overall, the results confirm that deep learning-based approaches are highly effective for satellite image segmentation. The model's high accuracy, balanced per-class metrics, and strong overlap scores make it suitable for real-world applications such as environmental monitoring, land-use planning, and large-scale GIS automation. With additional dataset expansion and sensor-specific fine-tuning, this approach can be further improved to support operational remote sensing workflows.

References

[1] L. Huang, B. Jiang, S. Lv, Y. Liu, and Ying Fu, “Deep learning-based semantic segmentation of remote sensing Images: a survey,” Sep. 2020.

[2] A. Ma, J. Wang, Y. Zhong, and Z. Zheng, “FactSEG: Foreground Activation-Driven Small Object Semantic Segmentation in Large-Scale Remote Sensing Imagery,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, Jul. 2021, doi: 10.1109/tgrs.2021.3097148.

[3] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, “FARSEG++: Foreground-Aware Relation Network for geospatial object segmentation in high spatial resolution remote sensing imagery,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13715–13729, Jul. 2023, doi: 10.1109/tpami.2023.3296757.

[4] X. Li, Y. Li, J. Ai, Z. Shu, J. Xia, and Y. Xia, “Semantic segmentation of UAV remote sensing images based on edge feature fusing and multi-level upsampling integrated with Deeplabv3+,” *PLoS ONE*, vol. 18, no. 1, p. e0279097, Jan. 2023, doi: 10.1371/journal.pone.0279097.

[5] J. Wang, L. Meng, W. Li, W. Yang, L. Yu, and G.-S. Xia, “Learning to extract building footprints from Off-Nadir aerial images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1294–1301, Mar. 2022, doi: 10.1109/tpami.2022.3162583.

[6] D. Wang 1 *et al.*, “SAMRS: Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model,” *SAMRS*.

[7] J. Wang, Z. Zheng, A. Ma, X. Lu, Y. Zhong, and State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, “LoveDA: a remote sensing Land-Cover dataset for domain adaptive semantic segmentation,” conference-proceeding, 2021.

[8] N. Pal, S. Ramkrishna, H. Patil, N. Choudhary, and R. Soman, “TerraGrid: Harnessing Deep learning models for satellite image segmentation,” *International Journal of Computer Applications*, vol. 186, no. 49, pp. 14–21, Nov. 2024, doi: 10.5120/ijca2024924147.

[9] D. Shim and C. Lee, “Small-Object semantic segmentation of satellite ship images using modified U-Net with morphological loss,” *IEEE Access*, vol. 13, pp. 27700–27713, Jan. 2025, doi: 10.1109/access.2025.3538876.

[10] Chen, M.; Li, L. Hierarchical Transfer Learning with Transformers to Improve Semantic Segmentation in Remote Sensing Land Use. Remote Sens. 2025, 17, 290. https://doi.org/10.3390/rs17020290