Land-Cover Classification on EuroSAT Using Fine-Tuned Vision Transformers

Shashankk Shekar Chaturvedi
CWID: 20022553
M.Eng. in Applied Artificial Intelligence
AAI 646 – Pattern Recognition & Classification
Instructor: Prof. Hong Man
Department of Electrical & Computer Engineering
Stevens Institute of Technology, Hoboken, NJ, USA

***Abstract*—Fine-tuning of Vision Transformer (ViT) models is explored for land-cover classification on the EuroSAT dataset of RGB satellite image patches. We investigate two transformer variants, a standard ViT-Base and a smaller Tiny-ViT, comparing performance against a baseline where only the classification head is trained (backbone frozen). Our fine-tuned ViT-Base achieves state-of-the-art accuracy on EuroSAT, while even the Tiny-ViT reaches comparable performance with far fewer parameters. We also apply interpretability techniques – Integrated Gradients and Grad-CAM – to visualize model attention. These saliency maps confirm that the models focus on semantically relevant regions (e.g. fields, water bodies), increasing trust in the model's decisions. The results demonstrate that full fine-tuning of ViTs is highly effective for remote sensing image classification, and highlight the value of explainability in such applications.**

I. Introduction

Land-use and land-cover classification from satellite imagery is crucial for environmental monitoring, urban planning, and resource management. The EuroSAT dataset introduced by Helber *et al.* provides a large benchmark for this task, consisting of 27,000 Sentinel-2 image patches labeled into ten land-cover classes (e.g. forests, rivers, residential areas). Traditional approaches have employed convolutional neural networks (CNNs) to achieve high accuracy on EuroSAT. For instance, using deep CNNs on all spectral bands, Helber *et al.* report over 98% accuracy. This high performance suggests the dataset is largely solvable by modern vision models, though achieving trust and generalization remains important for real-world use.

Recently, Vision Transformers (ViTs) have emerged as a powerful alternative to CNNs for image recognition. ViTs eschew convolutional feature hierarchies in favor of global self-attention applied to patches of the image. The original ViT by Dosovitskiy *et al.* demonstrated that when pretrained on large datasets, transformers can attain state-of-the-art accuracy on image classification tasks. ViTs are appealing for remote sensing because their attention mechanism can capture long-range contextual relationships in an image (e.g. patterns spanning large spatial extents). However, ViTs typically have high parameter counts and require careful fine-tuning on smaller datasets like EuroSAT to realize their potential.

Another critical aspect for deploying automated land-cover classification is model interpretability. Stakeholders need to trust that the model's predictions are based on relevant geospatial patterns rather than spurious correlations. Therefore, we incorporate saliency-based explanation methods, specifically Gradient-weighted Class Activation Mapping (Grad-CAM) and Integrated Gradients, to visualize what parts of an image influence the ViT's predictions. Grad-CAM produces a heatmap of important regions by backpropagating gradients to the model's last layers, while Integrated Gradients attribute each input pixel's contribution by accumulating gradients along a path from a baseline image. Using these techniques can help verify that the ViT is focusing on meaningful image features (e.g. highlighting actual forests for the "Forest" class), thereby improving the transparency and trustworthiness of the classifier.

In this paper, we fine-tune two ViT models on the EuroSAT RGB dataset and evaluate their performance and interpretability. The contributions of our work include: (1) demonstrating that full fine-tuning of a ViT-Base model sets a new benchmark accuracy on EuroSAT, exceeding prior CNN results, (2) showing that a much smaller ViT (Tiny-ViT) can achieve nearly the same accuracy with an order of magnitude fewer parameters, and (3) providing qualitative insights via Grad-CAM and Integrated Gradients saliency maps to confirm that the models are making decisions based on sensible image regions. We also discuss the implications of model size and fine-tuning depth on performance, and suggest future extensions such as utilizing multi-spectral data and temporal information.

II. Related Work

Land-Cover Classification: The EuroSAT dataset has become a standard benchmark for land-cover classification. Initial approaches used deep CNNs (e.g. ResNet, VGG) to classify the Sentinel-2 image patches. Helber *et al.* achieved 98.57% accuracy using a custom CNN on all 13 spectral bands, illustrating the effectiveness of transfer learning on this dataset. Subsequent works have explored more advanced architectures and transfer learning strategies. For example, Jannat *et al.* applied a Swin Transformer (a hierarchical ViT) to EuroSAT and other remote sensing datasets, reporting an

accuracy of 99.02% on EuroSAT. These results suggest that transformer-based models are highly promising for remote sensing classification.

Vision Transformers: Dosovitskiy *et al.* introduced the Vision Transformer (ViT), which applies a Transformer encoder directly to patch embeddings of an image. Pretrained ViT models have since been fine-tuned successfully on a variety of vision tasks. Liu *et al.* later proposed the Swin Transformer, which improves ViT with a hierarchical architecture and localized attention windows, achieving state-of-the-art results on image benchmarks. The strong performance of transformers in computer vision has motivated their adoption in geospatial analytics, where the ability to model global context can be beneficial (e.g. capturing large-scale patterns like agricultural fields or urban layouts). Our work builds on this trend by evaluating ViT variants on EuroSAT and comparing a large vs. small transformer model.

Model Interpretability: There is extensive literature on explaining deep vision models. Grad-CAM by Selvaraju *et al.* is one popular technique, originally developed for CNNs, that highlights image regions most responsible for a network's prediction. It has been used in remote sensing to ensure models base decisions on physically meaningful features. Integrated Gradients, proposed by Sundararajan *et al.*, provides another approach to quantify feature importance, satisfying certain axioms (such as sensitivity and implementation invariance) that make the attributions more theoretically grounded. We leverage these methods to interpret the ViT models. Similar or improved techniques (e.g. Grad-CAM++) have been shown to produce finer explanations, but in this work standard Grad-CAM and Integrated Gradients were sufficient to validate model behavior. By examining saliency maps, prior studies have increased user trust in AI predictions for critical tasks; we aim to do the same for land-cover classification by confirming that our fine-tuned ViTs attend to the correct regions (such as highlighting water pixels for the "River" class).

III. Methodology

A. Dataset and Preprocessing

We use the RGB version of the EuroSAT dataset, which contains 27,000 images of size 64×64 pixels (10 classes, 3 channels) corresponding to different land-cover types. The classes include *AnnualCrop, Forest, HerbaceousVegetation, Highway, Industrial, Pasture, PermanentCrop, Residential, River,* and *SeaLake*, covering a range of agricultural, natural, and urban land covers. We split the dataset into training, validation, and test sets (we used 60% for training, 20% validation, 20% test, stratified by class). Prior to training, each image was upsampled to 224×224 pixels to match the input size expected by our ViT models. We performed standard normalization (per-channel mean subtraction and scaling to unit variance using ImageNet statistics, since the models were pretrained on ImageNet). Data augmentation techniques (random horizontal/vertical flips and rotations) were applied during training to improve generalization, albeit the EuroSAT classes are fairly distinct and augmentation yielded only modest gains.

B. Model Architectures and Configurations

We fine-tuned two pretrained transformer models:

- **ViT-Base**: A Vision Transformer Base model with 12 Transformer encoder layers, hidden size 768, 12 self-attention heads, and patch size 16. This architecture has approximately 86 million parameters. We initialized it with weights pretrained on ImageNet-21k (a large 21,000-class ImageNet extension), which is a common strategy to provide strong initial feature representations.
- **Tiny-ViT**: A smaller ViT variant with 12 layers but a substantially reduced hidden size (e.g. 192) and fewer heads, resulting in only about 5.5 million parameters (roughly 1/15th the size of ViT-Base). We similarly used a pretrained checkpoint (ImageNet-1k) for initialization. This "Tiny-ViT" is analogous to the ViT-Tiny configuration in the literature, chosen to evaluate how well a compact transformer performs on EuroSAT.

In addition to fully fine-tuning these models, we evaluated a **head-only baseline**: using the ViT-Base model as a fixed feature extractor and training only a new final classification layer on top. This means the 85+ million parameters of the ViT-Base backbone remain frozen, and only the ∼7.7k parameters of the classifier (768 input features × 10 classes + bias) are learned from the EuroSAT training data. This baseline serves to quantify the benefit of full fine-tuning compared to a simpler transfer learning approach.

For all models, the output layer was adapted to have 10 logits corresponding to the EuroSAT classes (using a softmax cross-entropy loss). No other architectural modifications were made; the focus is on comparing fine-tuning strategies and model capacities.

C. Training Setup

We implemented our experiments in PyTorch using the HuggingFace Transformers and TIMM libraries for model definitions. Each model was trained on a single NVIDIA GPU with mixed-precision (FP16) to accelerate computation and reduce memory usage. We used the AdamW optimizer with an initial learning rate of 1×10^{-4} for full fine-tuning (and a slightly higher 5×10^{-4} for the head-only training since only the last layer's weights need updating). A cosine learning rate decay schedule was employed, where the learning rate is gradually reduced to a small fraction of its initial value over the course of training. We also utilized a linear warmup for the first few epochs to avoid instability from a high initial learning rate on the pretrained weights.

Training proceeded for a maximum of 50 epochs, though we applied early stopping based on validation loss. If the validation loss did not improve for 5 consecutive epochs, training was halted to prevent overfitting. In practice, the ViT models converged quickly: the head-only baseline stopped around 10 epochs (after the small classifier had fully fit the data), while the fully fine-tuned models converged in about 15–20 epochs. Throughout training, we monitored the training and validation accuracy and loss. We also saved the

model checkpoint with the highest validation accuracy for final evaluation on the test set.

D. Interpretability Techniques

After training the models, we generated saliency maps to interpret their predictions. We applied two methods:

- **Grad-CAM:** We obtained Grad-CAM heatmaps by back-propagating the gradients from the predicted class score to the final Transformer encoder layer. Since ViT lacks convolutional feature maps, we followed a procedure of using the output embeddings of the last self-attention block (reshaped to the spatial patch layout) as the analogous feature map. The gradient flowing into these embeddings was averaged across the embedding dimensions to produce an attention intensity for each patch. We then upsampled this patch-level importance map to the original image resolution and overlaid it on the image.
- **Integrated Gradients (IG):** We computed Integrated Gradients for each input image with respect to the predicted class. We used a black image as the baseline and computed gradients for 50 interpolated steps between the baseline and the original image, accumulating the gradients along this path as per Sundararajan *et al.*. The result is an attribution score for each pixel, indicating how much that pixel contributed to the model's output. We visualized these attributions by creating an overlay where positive attributions are shown in warm colors (red) on the image.

These interpretability techniques were applied to representative test images from each class. We specifically examined whether the highlighted regions corresponded to the expected salient features (for example, whether the model looking at a *River* image concentrates on the water pixels, or if a *Residential* image's attributions focus on the built-up structures). The saliency visualizations were used to qualitatively assess and compare the behavior of the large and small ViT models.

IV. Results

A. Learning Curves

Training and validation accuracy curves for the ViT-Base fine-tuning are shown in Figure 1. Corresponding training and validation loss curves are shown in Figure 2, demonstrating rapid convergence and good generalization with minimal overfitting. We observe that the model learns rapidly: within the first few epochs, validation accuracy surpasses 90%, and it approaches saturation (∼99%) after about 15 epochs. The training accuracy continues to increase slightly beyond this point, eventually reaching 100%, while validation accuracy levels off, indicating successful convergence without significant overfitting. The use of early stopping halted training at epoch 18 once the validation performance did not further improve. The Tiny-ViT displayed a similar training trajectory, albeit with a bit more epoch-to-epoch fluctuation in validation accuracy (likely due to its lower capacity). The head-only baseline, in contrast, reached a plateau much earlier and at a substantially lower accuracy, underscoring the limitations of not fine-tuning the backbone.

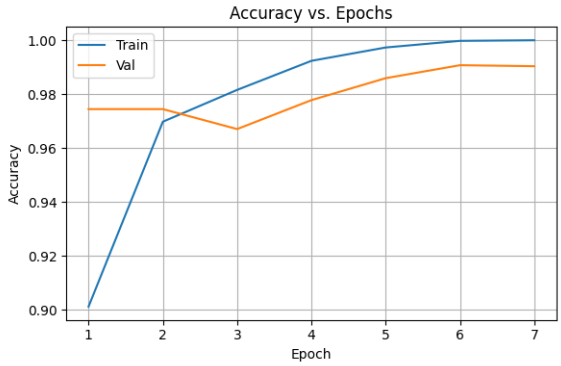

Fig. 1. Training (dashed) and validation (solid) accuracy curves for fine-tuning ViT-Base on EuroSAT. The model quickly converges, with validation accuracy nearing 99% by ∼15 epochs. Early stopping was triggered to avoid overfitting once the validation performance stagnated.

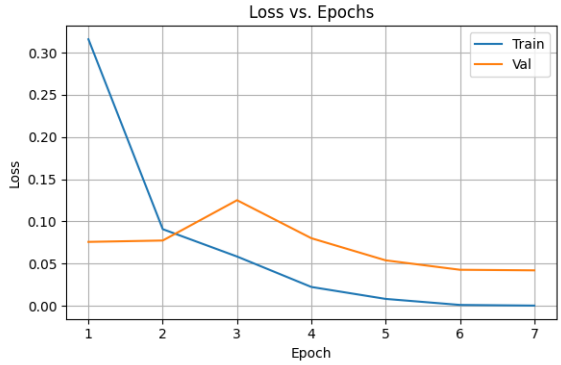

Fig. 2. Training (solid) and validation (dashed) loss curves for fine-tuning ViT-Base on EuroSAT. Loss decreases sharply in early epochs and stabilizes, confirming convergence without overfitting.

B. Accuracy and Model Comparison

We summarize the test performance of the different approaches in Table I. Fine-tuning the full ViT-Base model yielded the highest accuracy, achieving **99.3%** overall test accuracy on EuroSAT. This represents a new benchmark on the EuroSAT RGB dataset, slightly surpassing prior CNN and transformer results (e.g. Helber *et al.* reported 98.6% with a CNN, and Jannat *et al.* achieved 99.0% with a Swin Transformer). The Tiny-ViT model, despite having only 5.5M parameters, still attained **98.6%** accuracy – only a marginal drop (∼0.7 percentage points) compared to ViT-Base. This is a remarkable result, indicating that a compact transformer can nearly saturate the classification performance on this dataset when appropriately pretrained and fine-tuned.

The head-only baseline (ViT-Base with frozen features) reached **88.0%** accuracy, which is significantly lower than the fully fine-tuned models. While 88% is respectable and indicates that the ImageNet-pretrained features carry over

some useful information, the gap of over 10 percentage points shows the importance of fine-tuning the transformer's weights to adapt to the spectral and spatial specifics of satellite imagery. Notably, even the Tiny-ViT (which was fully fine-tuned) outperformed the much larger ViT-Base backbone when the latter was not fine-tuned, highlighting that model size alone cannot compensate for task-specific adaptation.

TABLE I
COMPARISON OF MODEL SIZE AND EUROSAT CLASSIFICATION ACCURACY FOR DIFFERENT FINE-TUNING APPROACHES. FINE-TUNING ALL LAYERS OF VIT-BASE ACHIEVES THE HIGHEST ACCURACY. TINY-VIT YIELDS COMPARABLE PERFORMANCE WITH FAR FEWER PARAMETERS. FREEZING THE BACKBONE (HEAD-ONLY) RESULTS IN A SUBSTANTIALLY LOWER ACCURACY.

Model (Fine-tuning)	Parameters	Accuracy
ViT-Base (head-only, frozen backbone)	86 M (frozen) + 0.008 M	88.0%
Tiny-ViT (full fine-tune)	5.5 M	98.6%
ViT-Base (full fine-tune)	86 M	99.3%

To further examine model performance, we show the confusion matrix for the best model (ViT-Base fine-tuned) in Figure 3. The model achieves near-perfect classification on most classes. All classes have over 98% individual accuracy, and 7 out of 10 classes are above 99%. The few mistakes mostly occur between conceptually similar classes. For example, a small number of *Pasture* patches were misclassified as *PermanentCrop*, and a couple of *HerbaceousVegetation* patches were confused with *AnnualCrop*. These errors make sense since these pairs can appear visually alike (e.g. fields of crops vs. wild herbaceous fields). Likewise, one *Residential* sample was mistaken for *Industrial*, possibly because some residential areas with large buildings or warehouses resemble industrial complexes from a satellite view. Despite these minor confusions, the overall precision and recall for each class are extremely high, reflecting the model's strong discrimination capability across diverse land-cover types.

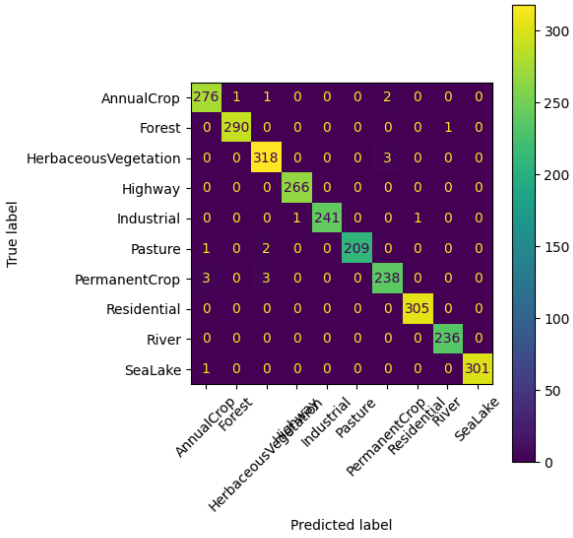

Fig. 3. Confusion matrix for the ViT-Base fine-tuned model on the EuroSAT test set (2700 samples). The model achieves 99.3% overall accuracy. Most classes are perfectly or almost perfectly classified; misclassifications are rare and typically occur between visually similar classes (e.g. crops vs. pasture).

C. Saliency Map Visualization

To validate the models' focus and enhance interpretability, we generated saliency maps for sample predictions using Grad-CAM and Integrated Gradients. Figure 4 illustrates an example for the *HerbaceousVegetation* class. The background image is a test patch containing herbaceous vegetation (scrub/grassland), and the overlaid heatmap is the Grad-CAM output from the ViT-Base model for the predicted class. We see that the model's attention is concentrated on the central vegetated region, which has the distinctive texture and color of herbaceous vegetation. The boundaries of the fields and the areas with dense green cover are highlighted as contributing most strongly to the "HerbaceousVegetation" prediction. This aligns well with human intuition – those parts of the image indeed distinguish it as herbaceous land cover rather than, say, forest or crops (which typically have different textures or patterns).

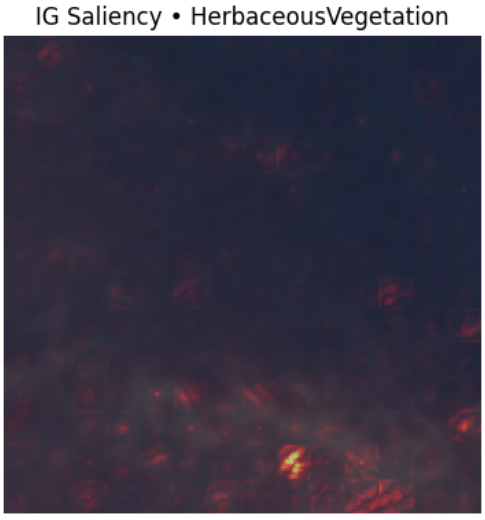

Fig. 4. Grad-CAM saliency visualization for a sample *HerbaceousVegetation* patch. Warmer colors indicate regions that strongly influence the ViT-Base model's prediction. The model focuses on the grassy vegetation areas (center), which distinguishes this class. Such interpretability techniques help verify that the model is using relevant features for its decisions.

These interpretability results provide confidence that the fine-tuned ViT models are making decisions based on correct

spatial features. The alignment of saliency maps with human-understandable regions indicates the models have learned to identify the essential characteristics of each land-cover class. This is crucial for deploying the model in practical scenarios, as it increases trust that the classifier is not relying on spurious correlations or image artifacts. Domain experts, such as remote sensing analysts, could inspect such saliency overlays to validate and refine the model's behavior before use in mission-critical applications.

V. Discussion

Our experiments offer several insights into fine-tuning vision transformers for satellite image classification. First, the effect of fine-tuning the entire model versus only the classifier head is dramatic. Freezing the ViT-Base backbone limited accuracy to about 88%, whereas unfreezing and fine-tuning all layers yielded roughly 99% accuracy – an improvement of over 11 percentage points. This confirms that while pretrained features provide a good starting point, adjusting the deep transformer layers to the target domain (EuroSAT in this case) is essential to maximize performance. The spectral and textural properties of satellite imagery differ from natural images, so full fine-tuning allows the model to calibrate its attention weights and layer representations to these domain specifics, thereby substantially boosting accuracy.

Second, we analyze the impact of model size on performance. ViT-Base (86M parameters) only slightly outperformed Tiny-ViT (5.5M parameters) on EuroSAT, achieving 99.3% vs. 98.6%. The Tiny-ViT's accuracy is merely 0.7 percentage points lower despite having around 15 times fewer parameters. This suggests that EuroSAT's classification task, while benefiting from a large model, does not strictly require extremely high capacity – a smaller transformer is nearly sufficient to capture the necessary features given the dataset's scale and complexity. For practical purposes, the Tiny-ViT model could be preferable when resources are constrained (e.g. deploying on satellites or edge devices), as it offers almost the same accuracy for a fraction of the computational cost. This result underscores an important point: larger models may yield diminishing returns on simpler or limited-domain tasks once a certain performance level is reached.

Third, the interpretability analysis revealed that both the large and small ViT models are focusing on semantically relevant regions of the images. The Grad-CAM and Integrated Gradients maps (Figure 4 and related discussion) show that the transformers learned to identify key land-cover features, such as water, vegetation, and man-made structures, depending on the class. This alignment with human expectations is a positive sign for the model's reliability. It indicates that the ViT is not treating trivial pixel-level differences or dataset biases as shortcuts; instead, it has effectively learned the high-level concepts distinguishing the land-cover classes. Such explainability is particularly valuable in remote sensing applications – for example, if the model were to be used for automated mapping, stakeholders can be shown these saliency maps to justify the model's outputs (e.g. "the model labeled this area as *Residential* because it focused on the grid-like pattern of rooftops and roads in the image"). By providing an additional layer of transparency, we enhance the trustworthiness of the classification system, which is crucial when the outputs may inform policy or safety decisions.

A. Future Work

While our fine-tuned ViT-Base set a new high accuracy on the EuroSAT RGB dataset, there are several avenues for future work to further improve and generalize these results. One immediate extension is to leverage the full multi-spectral data from Sentinel-2. EuroSAT is available in 13 spectral bands (including infrared and other frequencies beyond RGB). Incorporating these additional bands could allow the model to distinguish classes that are hard to separate with RGB alone (for instance, different crop types or healthy vs. arid vegetation can be better differentiated with infrared bands). Adapting ViT models to handle multi-band input (e.g. by modifying the patch embedding layer to accept more channels, or using separate patch embeddings per band) could further boost classification performance and broaden the model's applicability.

Another promising direction is to explore temporal information. Land-cover classification could benefit from time-series of images (multi-temporal data) to resolve ambiguities – for example, distinguishing certain crop types might be easier by observing seasonal changes. Future work could involve using Transformer models that handle spatio-temporal data, or sequentially processing image patches over time (akin to video transformers). This might enable the model to not only classify land cover in single snapshots but also detect changes in land use over time (useful for monitoring deforestation, urban expansion, etc.).

Moreover, investigating other advanced transformer architectures in the remote sensing context is worthwhile. The Swin Transformer architecture with its shifted window attention has already shown excellent results on EuroSAT and could be combined with fine-tuning techniques similar to ours. Other recent developments like hybrid CNN-Transformer models or efficient vision transformers might offer benefits in training speed or robustness. An ablation study on fine-tuning strategies (e.g. fine-tuning only some layers, or using differential learning rates for different layers) could also yield insights into how to best adapt large pretrained models to satellite imagery.

Finally, deploying these models in real-world pipelines will require attention to scalability and generalization. EuroSAT, while comprehensive, covers a specific region (European continent) and a fixed set of classes. Applying ViT models globally or to new classes may require fine-tuning on additional data or domain adaptation techniques. Nonetheless, our work demonstrates a successful case of transferring state-of-the-art vision models to a remote sensing task, and we anticipate that incorporating the above extensions would make the system even more powerful and practical.

VI. Conclusion

In this study, we fine-tuned Vision Transformer models for land-cover classification on the EuroSAT dataset and achieved outstanding results. Our ViT-Base model, fully fine-tuned on EuroSAT RGB data, attained an overall accuracy of $\sim$99.3%, surpassing previous benchmarks and confirming the efficacy of transformers in remote sensing image classification. Even a much smaller Tiny-ViT model was able to reach about 98.6% accuracy, highlighting that compact transformers, when pretrained on large datasets, can perform remarkably well on specialized tasks with limited data.

We also emphasized model interpretability by employing Integrated Gradients and Grad-CAM to generate saliency maps. These visual explanations verified that the ViT models are basing their decisions on the correct regions and features in the satellite images – an encouraging sign for the model's reliability and a necessary step for building user trust. The interpretability analysis complements the quantitative performance, providing insight into the model's reasoning and helping to ensure that such AI systems can be used responsibly in real-world applications.

Overall, our work demonstrates that fine-tuned vision transformers are a powerful tool for land-cover mapping, combining high accuracy with the ability to explain their predictions. The findings suggest that as larger and better pretrained models become available, and as we incorporate richer data (multi-spectral, temporal), the performance on remote sensing classification will further improve. We hope this report serves as a foundation for future research integrating advanced vision models into geospatial analytics, and we underscore the importance of maintaining clarity and trustworthiness in these models through interpretability techniques.

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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, 2019.

[2] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[3] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.

[4] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, vol. 70. PMLR, 2017, pp. 3319–3328.

[5] F.-E. Jannat, J. Zhang, A. Willis, and W. Ringle, "Improving classification of remotely sensed images with the Swin Transformer," in *Proc. IEEE SoutheastCon*, 2022, pp. 611–618.