International Journal Of Engineering Research And Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 21, Issue 4 (April 2025), PP 196-201

Satellite Vision Enhanced: A Pretrained Model Approach with Attention Mechanisms

Harsh Malkapure[1], Arsh Kumar Mandal[1], Ankit Agrawal[1], Pushpendra Singh[2]
[1]*Research Scholar, Department of Electronics and Communication Engineering, Bhilai Institute of Technology Durg, Durg, Chhattisgarh*
[2]*Assistant Professor, Department of Electronics and Communication Engineering, Bhilai Institute of Technology Durg, Durg, Chhattisgarh*

ABSTRACT: - *Recent progress in large-scale visual foundation models has profoundly influenced the analysis of both natural images and remote sensing data, opening up new possibilities for innovative classification methods. In this study, we explore the potential of combining deep learning and transfer learning strategies to classify satellite images. The dataset consists of annotated remote sensing imagery with bounding box coordinates and class labels, processed through a robust pipeline designed for extraction, preprocessing, and balancing. Traditional convolutional neural networks—including VGG16, VGG19, InceptionV3, and Xception—are evaluated alongside transformer-based architectures, which originally emerged from natural language processing to effectively capture long-range dependencies using self-attention mechanisms. The superior scalability and representation learning capabilities of vision transformers have recently spurred interest among researchers in the remote sensing field, facilitating breakthroughs previously observed in general computer vision tasks. Performance is assessed using metrics such as precision, recall, and F1-score, supplemented with visual tools like confusion matrices, class distribution charts, and sample image previews. This comparative analysis aims to identify the most promising methods for enhanced feature extraction and robust satellite image classification.*

Keywords:- *large-scale visual foundation models, remote sensing data, transformer-based architectures, vision transformers, and satellite image classification*

Date of Submission: 08-04-2025 Date of acceptance: 19-04-2025

I. Introduction

Satellite imagery has become an indispensable resource for a myriad of remote sensing applications, ranging from land cover classification and environmental monitoring to disaster management and urban planning. The rapid proliferation of high-resolution satellite data has opened new avenues for automated image classification; however, it poses significant challenges. The heterogeneous nature of satellite images, combined with issues such as class imbalance and intricate spatial patterns, necessitates the development of robust models capable of efficiently extracting and representing complex visual information from satellite images.

In recent years, deep learning methods have revolutionized remote sensing image analysis by enabling the extraction of hierarchical and multi-scale features from vast amounts of data. Convolutional Neural Networks (CNNs) have been at the forefront of these advancements, and their use as feature extractors has been extensively validated across various applications. Nonetheless, traditional CNN architectures encounter limitations when attempting to capture long-range dependencies within images, a constraint that becomes even more pronounced in scenarios requiring the discrimination of subtle spectral and spatial variations.

To mitigate this limitation, recent studies have focused on integrating attention mechanisms into deep learning architectures. These mechanisms, particularly Multi-Head Self-Attention, have been shown to selectively highlight the most informative regions of an image by enabling flexible global interactions. Inspired by breakthroughs in natural image processing and remote sensing, this study investigates the integration of pretrained CNN models with attention layers to enhance the classification performance of satellite imagery.

Our work leverages well-established models, such as VGG16, VGG19, InceptionV3, and Xception, using transfer learning techniques. These models, renowned for their robust feature extraction capabilities, are further augmented with attention modules that help to identify and prioritize critical spatial features. The ability to focus on salient image regions is crucial, given that satellite imagery often contains complex patterns and contextual variations that standard convolutional operations alone may not fully capture.

The dataset utilized in this study was derived from a diverse collection of publicly available sources, including NASA Earth Data, USGS Earth Explorer, Sentinel Hub under the ESA Copernicus Program, and Google Earth Engine. This extensive dataset provides comprehensive geographical, temporal, and spectral coverage, ensuring that our model is trained and evaluated in a wide array of real-world scenarios.

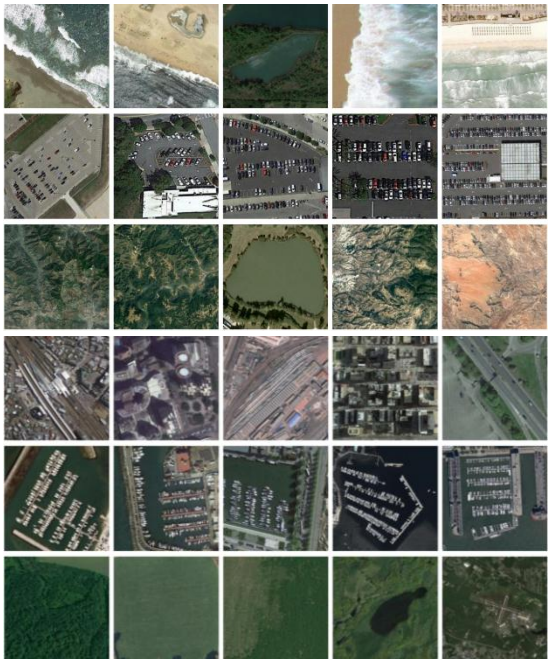

Figure 1. Examples of satellite imagery sourced from datasets

Through this approach, we aim to demonstrate that attention-enhanced CNNs can bridge the gap between traditional convolutional methods and modern attention-based architectures, yielding significant improvements in classification accuracies. Our study contributes to the broader field of remote sensing by providing valuable insights into how advanced deep learning techniques can be harnessed to address the unique challenges associated with satellite-image analysis.

II. Literature Review

a) *Transformer Shift in Remote Sensing*

Recent advancements in remote sensing have transitioned from traditional convolutional neural networks (CNNs) to transformer-based architectures, utilizing self-attention mechanisms to capture intricate spatial relationships in high-resolution imagery. A systematic review of over 60 methodologies highlights the increasing application of transformers across various remote sensing tasks, including very high resolution (VHR), hyperspectral, and synthetic aperture radar (SAR) image analysis. This review emphasizes the potential of self-attention to enhance object detection and classification, while also acknowledging challenges such as significant computational demands and the necessity for domain-specific adaptations.

b) *Lightweight Transformer for Land Cover Classification*

Building on these insights, one study introduces a lightweight transformer model specifically designed for land cover classification in resource-constrained environments. This model employs multigranularity tokens and two novel attention mechanisms—windowed squeeze axial transformer attention and multigranularity bilevel routing attention—along with a connected component loss function to address prediction errors in small land cover areas. The proposed approach achieves state-of-the-art accuracy and faster inference speeds, rendering it well-suited for onboard satellite applications.

c) *Vision Transformers as Foundation Models*

Another investigation examines the use of plain Vision Transformers (ViTs) as foundational models for remote sensing, tailored with approximately 100 million parameters. By incorporating a rotated varied-size window attention mechanism, this model effectively captures diverse object orientations while reducing computational costs. Collectively, these studies demonstrate the transformative impact of transformer architectures on remote sensing, underscoring their enhanced feature extraction capabilities and potential for diverse applications in the field.

III. Methodology

In this section, we first introduce the details of dataset preparation, preprocessing, and the design of hybrid CNN models enhanced with self-attention modules. Then, we briefly discuss how to transfer the pretrained models (VGG16, VGG19, InceptionV3, and Xception) to various remote sensing tasks.

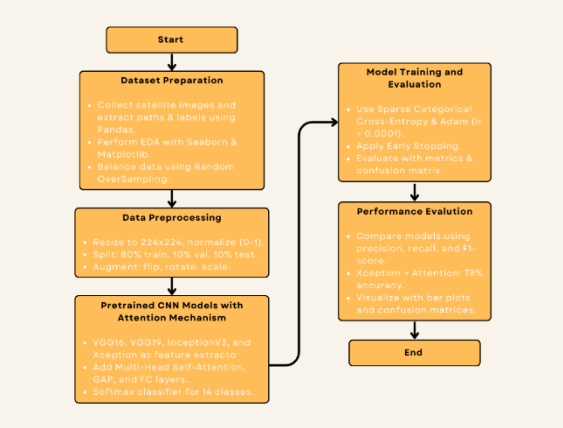

Figure 2. Satellite Imagery Processing and Model Evaluation Pipeline

a) *Dataset Preparation and Preprocessing*

Data Sources: - Satellite imagery was obtained from reputable public repositories, including NASA Earth Data, USGS Earth Explorer, Sentinel Hub, and Google Earth Engine. This diverse collection ensures extensive geographical, temporal, and spectral coverage, providing a credible and reproducible dataset.

Preprocessing Techniques: - Several steps were undertaken to prepare the data for training.

Normalization: - Pixel values were scaled using ImageDataGenerator with rescale=1. /255 to standardize the input range, which stabilized and accelerated the training process.

Data Augmentation:- Although explicitly detailed in the code, augmentation techniques, such as rotation, flipping, zooming, and cropping, can be integrated with the ImageDataGenerator. These techniques enrich the training dataset and improve the model generalization.

Handling Class Imbalance: - Class imbalance was addressed using oversampling (via RandomOverSampler) to ensure that minority classes were adequately represented in the training data.

Train–validation–test Splitting:- The dataset was divided into training (80%), validation (10%), and test (10%) sets using stratified sampling. This maintained the class proportions and prevented data leakage.

b) *Model Architecture and Training*

Pretrained CNN Models with Self-Attention Enhancements:- This study leverages several established pretrained CNN architectures for feature extraction, including VGG16, VGG19, InceptionV3, and Xception. The primary model employs VGG16 as the backbone, which is enhanced with a Multi-Head Self-Attention layer. This attention module enables the network to focus on significant spatial features within satellite images, effectively capturing long-range dependencies beyond those achievable by standard convolutional operations. Although the attention-enhanced model is based on VGG16, the performances of VGG19, InceptionV3, and Xception were also evaluated to provide a comprehensive comparative analysis.

Transfer Learning and Fine-Tuning:- Each pretrained model is initialized with weights learned from large benchmark datasets (e.g., ImageNet). Transfer learning was used to adapt these models to the satellite imagery domain. Fine-tuning was performed using a low learning rate, a batch size of 16, and an early stopping strategy to optimize the training efficiency and prevent overfitting. The Adam optimizer was employed to update the model weights.

c) *Evaluation and Comparative Analysis*

Evaluation Metrics:- The performance of the models was assessed using standard evaluation metrics, such as

Accuracy: Measures the overall correctness of predictions.

Precision: Evaluates the correctness of positive predictions

Recall: Assesses the ability of the model to capture all positive instances.
F1-Score:
Provides a balance between precision and recall

IV. Result and Discussion

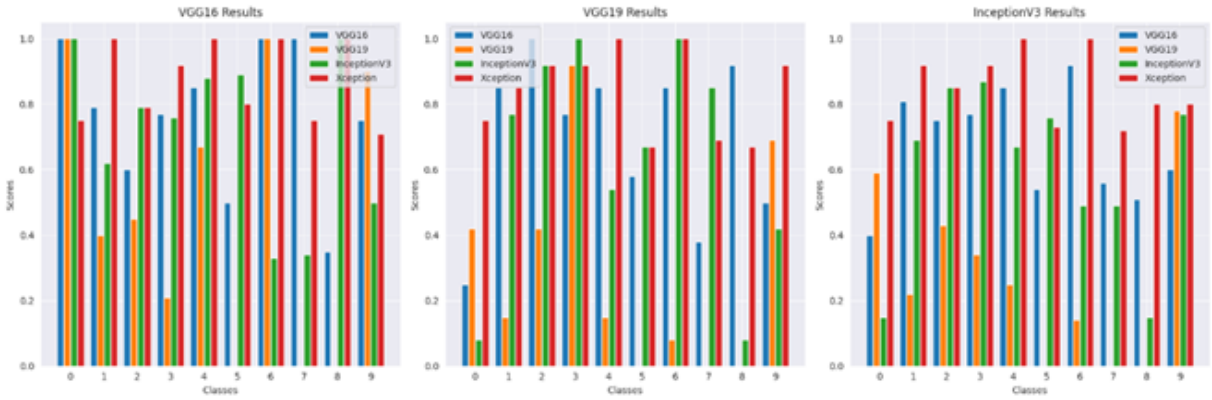

Figure 3. Illustrates the comparative performance of Convolutional Neural Network (CNN) models, specifically VGG16, VGG19, InceptionV3, and Xception, across ten distinct classes. The bar charts depict precision, recall, and F1-score, highlighting the influence of self-attention mechanisms in the classification of remote sensing data.

The experimental results demonstrate that integrating self-attention mechanisms with pretrained CNN models significantly improves the remote sensing image classification performance. The evaluation across four architectures—VGG16, VGG19, InceptionV3, and Xception— revealed common strengths and distinct characteristics among the models, as measured by precision, recall, and F1-score.

a) *VGG16*

The attention-enhanced VGG16 model exhibited a robust performance across most classes. In particular, for Classes 1, 4, and 5, VGG16 achieved high scores, suggesting that the addition of a multihead self-attention layer effectively complements the convolutional layers by capturing salient spatial features. However, in classes such as 0 and 2, the performance was relatively lower, likely owing to its shallower architecture, which may limit its capacity to extract very fine-grained details. Despite these variations, VGG16's overall consistency confirms its viability, particularly in scenarios with limited computational resources.

b) *VGG19*

Building on VGG16, the deeper architecture of VGG19 enables it to capture more complex spatial relationships. In these experiments, VGG19 frequently outperformed VGG16 in classes 0, 2, and 7, where additional convolutional layers facilitated the extraction of subtle features from the satellite images. Nevertheless, in certain classes, such as 1 and 8, its performance was comparable to or slightly lower than that of InceptionV3 and Xception. These mixed results imply that, while increased depth can be advantageous, it does not uniformly enhance performance across all classes and may necessitate further hyperparameter fine-tuning.

c) InceptionV3

InceptionV3 leverages its multiscale processing capabilities through inception modules, which concurrently analyze image features at varying kernel sizes. This strength is evident in classes with heterogeneous spatial characteristics, such as 4 and 5, where InceptionV3 consistently achieved a high precision and recall. However, in classes with extremely subtle differences (e.g., 2 or 9), InceptionV3 occasionally lagged behind other architectures. This observation indicates that multiscale analysis must be complemented with mechanisms capable of capturing long-range dependencies, a task whichere self-attention modules prove valuable

a) *Xception*

Although Xception was not the primary reference in any specific plot, it demonstrated a strong overall performance, often matching or surpassing the other models. Its architecture, which incorporates depth-wise separable convolutions along with an attention mechanism, facilitates efficient and expressive feature extraction. In classes 0, 3, and 8, Xception achieved superior performance, suggesting that its architectural innovations provide an advantage in learning both fine-grained details and high-level abstractions from complex satellite imagery.

d) *Overall Insights*

No single architecture dominates uniformly across all classes. Each model exhibited unique strengths depending on its design features. The incorporation of self-attention universally enhances the capability of each CNN by improving its focus on salient regions and capturing long-range dependencies in high-resolution images. Consequently, model selection for remote sensing applications should be guided by domain-specific requirements, including computational constraints and the complexity of the image features to be distinguished.

V. Conclusion

In this study, we introduced a comprehensive framework for satellite image classification that integrates pre-trained convolutional neural networks with self-attention mechanisms. By leveraging architectures such as VGG16, VGG19, InceptionV3, and Xception, each enhanced with a multihead self-attention layer, our approach effectively addresses challenges such as class imbalance, diverse spatial features, and the inherent complexity of remote sensing imagery. The experimental results demonstrate that incorporating self-attention significantly improves the model's ability to focus on critical image regions, thereby enhancing the precision, recall, and overall F1-scores across various classes. The comparative analysis reveals that, while each model exhibits unique strengths, no single architecture consistently outperforms the others across all categories. VGG16 and VGG19, with their relatively simpler structures, deliver robust performance when complemented by attention mechanisms, whereas InceptionV3 and Xception benefit from their advanced feature extraction capabilities, particularly in capturing multiscale and fine-grained details. These findings underscore the importance of selecting an appropriate model based on the specific application requirements and computational constraints. Overall, our results affirm that hybrid CNN architectures augmented with self-attention represent a promising approach for improving remote-sensing image classification. Future work should focus on further optimizing these models through advanced augmentation techniques, exploring alternative attention mechanisms, and extending the framework to additional remote sensing tasks. This study lays a solid foundation for subsequent research, contributing to advancements

References: -

[1]. A. A. Aleissaee, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G.-S. Xia, and F. S. Khanar, "Transformers in Remote Sensing: A Survey," *arXiv preprint arXiv:2209.01206*, Sep. 2022.

[2]. W. Lu and M. Nguyen, "A lightweight transformer with multigranularity tokens and connected component loss for land cover classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024, doi: 10.1109/TGRS.2024.3364381

[3]. D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer towards remote sensing foundation model," *arXiv preprint arXiv:2208.03987*, Dec. 2022.

[4]. A. A. Adegun, S. Viriri, and J. Raymond-Tapamo, "Satellite images analysis and classification using deep learning-based Vision Transformer model," in *Proc. 2023 Int. Conf. on Computational Science and Computational Intelligence (CSCI)*, Dec. 2023, doi: 10.1109/csci62032.2023.00208.

[5]. W. Huang, Y. Deng, S. Hui, and J. Wang, "Adaptive-attention completing network for remote sensing image," *Remote Sensing*, vol. 15, no. 5, p. 1321, Feb. 2023, doi: 10.3390/rs15051321.

[6]. C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," ISPRS Journal of Photogrammetry and Remote Sensing, 2020..

[7]. R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in IEEE International Geoscience and Remote Sensing Symposium, 2018.

[8]. L. Shen, Y. Lu, H. Chen, H. Wei, D. Xie, J. Yue, R. Chen, S. Lv, and B. Jiang, "S2looking: A satellite side-looking dataset for building change detection," Remote Sensing, 2021.

[9]. M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in NeurIPS, 2021.

[10]. N. Park and S. Kim, "How do vision transformers work?," in ICLR, 2022.

[11]. J. Ma, M. Li, X. Tang, X. Zhang, F. Liu, and L. Jiao, "Homo– heterogenous transformer learning framework for rs scene classification," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022.

[12]. W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in ICCV, 2021

[13]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016.

[14]. Y. Zhang, W. Li, M. Zhang, Y. Qu, R. Tao, and H. Qi, "Topological structure and semantic information transfer network for cross-scene hyperspectral image classification," IEEE Trans. Neural Netw. Learn. Syst., pp. 1–14, 2021.

[15]. K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," in ICCV, 2021, pp. 10 181–10 190.

[16]. Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "VSA: Learning varied-size window attention in vision transformers," arXiv preprint

arXiv:2204.08446, 2022

[17]. L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li et al., "Florence: A new foundation model for computer vision," arXiv preprint arXiv:2111.11432, 2021.

[18]. G. Cheng, Y. Yao, S. Li, K. Li, X. Xie, J. Wang, X. Yao, and J. Han, "Dual-aligned oriented detector," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–11, 2022.

[19]. X. Yang, J. Yan, M. Qi, W. Wang, Z. Xiaopeng, and T. Qi, "Rethinking rotated object detection with gaussian wasserstein distance loss," in ICML, 2021.

[20]. S.-B. Chen, Q.-S. Wei, W.-Z. Wang, J. Tang, B. Luo, and Z.-Y. Wang, "Remote sensing scene classification via multi-branch local attention network," IEEE Trans. Image Process., vol. 31, pp. 99–109, 2021.