# Onboard Satellite Image Classification for Earth Observation: A Comparative Study of ViT Models

Thanh-Dung Le, *Senior Member, IEEE*, Vu Nguyen Ha, *Senior Member, IEEE*, Ti Ti Nguyen, *Member, IEEE*, Duc-Dung Tran, *Member, IEEE*, Hung Nguyen-Kha, Luis M. Garces-Socarras, *Member, IEEE*, Juan Carlos Merlano-Duncan, *Senior Member, IEEE*, Symeon Chatzinotas, *Fellow, IEEE*

***Abstract*—Remote sensing (RS) image classification (IC) is a critical component of Earth observation (EO) systems, traditionally dominated by convolutional neural networks (CNNs) and other deep learning (DL) techniques. However, the advent of Transformer-based architectures and large-scale pre-trained models has significantly shifted the trend by offering enhanced performance and efficiency. Hence, this study focuses on identifying the most effective pre-trained model for land use classification in onboard satellite processing, emphasizing achieving high accuracy, computational efficiency, and robustness against noisy data—conditions commonly encountered during satellite-based inference. Through extensive experimentation, we compare the performance of traditional CNN-based, ResNet-based, and various pre-trained vision Transformer models. Our findings demonstrate that pre-trained Vision Transformer (ViT) models, particularly MobileViTV2 and EfficientViT-M2, outperform models trained from scratch in terms of accuracy and efficiency. These models achieve high performance with reduced computational requirements and exhibit greater resilience during inference under noisy conditions. While MobileViTV2 has excelled on clean validation data, EfficientViT-M2 has proved to be more robust when handling noise, making it the most suitable model for onboard satellite EO tasks. Our experimental results demonstrate that EfficientViT-M2 is the optimal choice for reliable and efficient RS-IC in satellite operations, achieving 98.76% of accuracy, precision, and recall. Precisely, EfficientViT-M2 delivers the highest performance across all metrics, excels in training efficiency (1,000s) and inference time (10s), and demonstrates greater robustness (overall robustness score of 0.79). Consequently, EfficientViT-M2 consumes 63.93% less power than MobileViTV2 (79.23 W) and 73.26% less power than SwinTransformer (108.90 W). This highlights its significant advantage in energy efficiency. Reproducible codes for data augmentation, noisy data, training, and inference are available from our shared Github repository [1].**

***Index Terms*—Earth observation, remote sensing, Transformers, on-board processing, pre-trained ViT, model robustness**

## I. Introduction

THE increasing demand for satellites to support EO and RS missions has driven a rapid expansion in the number of Low Earth Orbit (LEO) satellites [1]. These missions are critical for various applications, including environmental monitoring, disaster response, precision agriculture, and scientific research, where high-frequency, high-resolution data is essential. A key component of EO systems is RS-IC, which plays a fundamental role in analyzing satellite-acquired data. Traditionally, IC in RS has relied on CNNs and other DL techniques, which have been widely validated and proven effective for processing and classifying RS data with high accuracy and efficiency [2]–[4].

However, recent advances have shifted towards Transformer-based architectures, surpassing conventional CNNs in several IC tasks, including those in RS applications [5], [6]. The rise of large-scale pre-trained models has marked a new era in artificial intelligence (AI), achieving breakthroughs across various domains, particularly in natural language processing with models like BERT [7]. This paradigm shift extends into computer vision, where pre-trained ViT models demonstrate superior performance across a wide range of tasks [8], [9]. These models benefit from extensive pre-training on large datasets, allowing them to learn rich contextual representations that can be fine-tuned for specific tasks with minimal additional training. The success of these models is primarily attributed to advancements in computational power, the availability of vast amounts of data, and innovations in model architecture and efficiency [10]. While pre-trained ViT models hold great potential for enhancing classification performance and reducing computational demands during inference, their application in onboard satellite systems (OSS) is constrained due to such large-scale models' high power consumption and computing resource requirements. Implementing these models for OSS real-time decision-making remains challenging. Furthermore, a major operational bottleneck for LEO satellites is their dependence on ground stations for data transmission, leading to limited communication opportunities. The short visibility window results in prolonged connectivity loss, delaying access to critical data and hindering timely responses [11].

Satellite Internet Providers (SIPs) emergence, e.g. SpaceX and OneWeb, offers the potential to overcome these limitations by enabling continuous (24/7) connectivity, allowing on-demand data access [12]. However, seamless connectivity alone is insufficient to meet the modern EO and RS requirements, which demand low-latency, real-time decision-making capabilities. This highlights the need for advanced onboard processing solutions that can autonomously analyze data, prioritize critical information, and take immediate action, such as refocusing during the subsequent revisit, without relying on delayed ground-based interventions [13].

Therefore, most onboard neural networks (NNs) are designed to be lightweight, prioritizing computational efficiency

This work was funded by the Luxembourg National Research Fund (FNR), with the granted SENTRY project corresponding to grant reference C23/IS/18073708/SENTRY.

Thanh-Dung Le, Vu Nguyen Ha, Ti Ti Nguyen, Duc-Dung Tran, Hung Nguyen-Kha, Luis M. Garces-Socarras, Juan Carlos Merlano-Duncan, Symeon Chatzinotas are with the Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg, Luxembourg (Corresponding author. Email: thanh-dung.le@tamucc.edu).

[1] https://github.com/ltdung/SnT-SENTRY

over model complexity. For example, the Φ-Sat-1 mission employed the CloudScout model, a CNN-based NN, for onboard image segmentation on the Intel Movidius Myriad 2 vision processing unit (VPU), making it the first instance of DL deployed on a satellite [14]. Similarly, the Φ-Sat-2 mission implemented a convolutional autoencoder for image compression to minimize data transmission needs [15] and tested on three different hardware, including graphic processing unit (GPU) NVIDIA GeForce GTX 1650, VPU Myriad 2, and central processing unit (CPU) Intel Core i7-6700. These examples highlight the practical limitations in deploying more complex models, such as Transformer-based architectures and onboard satellites, due to restricted power budgets and limited processing resources. Consequently, while CNN-based models can achieve reasonable performance, they cannot often capture the intricate patterns and contextual dependencies required for more advanced tasks, limiting their effectiveness in RS.

The computational demands of ViTs - especially in model complexity and resource consumption - are one of the primary challenges in their onboard deployment. Unlike CNNs, which are relatively lightweight and optimized for image processing, ViTs require substantial computational resources, particularly for large models with numerous parameters and attention mechanisms. A recent study proposed a solution to this challenge by introducing RepSViT, a lightweight ViT-based model designed for onboard image processing. With only 3.77 million parameters and a computational cost of just 0.6 GFLOPs, RepSViT demonstrates that it is possible to deploy a compact version of ViT for satellite missions, while still capturing some of the advantages of Transformer architectures. However, such models, though promising, still face challenges when it comes to processing power and energy efficiency, making them less suited for complex tasks unless further model compression and optimization techniques are applied [16]. One promising approach to onboard processing is using neuromorphic processors, which rely on converting traditional NNs into Spiking NNs (SNNs). Currently, this conversion is best supported for CNN-based models, as demonstrated in recent studies on energy-efficient neuromorphic computing for space applications [16]–[18]. While these lightweight CNN-based models are advantageous due to their low power consumption and reduced computational requirements, they often fall short in prediction accuracy and robustness compared to more complex Transformer-based models, better suited for capturing patterns and contextual dependencies in RS data.

One promising approach to enhance computational efficiency for such downstream tasks is transfer learning, where a pre-trained model is fine-tuned for a specific application. This method leverages the rich, hierarchical features learned from large-scale datasets, minimizing the need for extensive retraining and enabling rapid adaptation to new tasks with significantly fewer labeled samples [19]. Transfer learning has consistently demonstrated improvements in performance across various tasks, making it particularly beneficial for resource-constrained environments like OSS, where both data and computational power are limited.

Building on this strategy, this study aims to identify the most effective pre-trained model for RS-IC, explicitly focusing on land-use classification applications in OSS. The goal is to find a model that provides high accuracy, operates efficiently with limited computational resources, and demonstrates robustness under noisy conditions, which is common in satellite-based inference due to environmental factors or hardware imperfections. The process begins with the experimental training of the model on the ground, where various pre-trained models are rigorously evaluated. Once the most effective model is identified, it is deployed in the onboard satellite processor for real-time inference. In this setup, the EO satellite records sensor data, which the onboard processor then preprocesses. By leveraging the trained model on the ground, we aim to optimize the inference's classification performance and power consumption efficiency directly within the satellite processor, ensuring efficient and accurate real-time RS-IC.

To achieve this, we conduct extensive experiments comparing traditional CNN-based, ResNet-based, and Transformer-based models for IC. The study then delves into various pre-trained ViT models to identify the optimal balance between model computational efficiency and performance. A key aspect of our evaluation is the robustness of these models when faced with noisy data during on-board inference.

Our findings indicate that pre-trained ViT models significantly outperform models trained from scratch, particularly in classification accuracy, precision, and recall. Among the models tested, MobileViTV2 [20] and EfficientViT-M2 [21] have emerged as the top performers for onboard satellite EO. These models deliver high performance with lower computational demands and exhibit greater robustness in the presence of noisy test data. While MobileViTV2 has achieved higher validation performance on clean data, EfficientViT-M2 has proved more resilient during inference with noisy data, making it the most suitable model for our specific application.

In conclusion, EfficientViT-M2 is identified as the optimal model for EO-OSSs due to its balance of training efficiency, high accuracy, and robustness under challenging inference conditions. This model offers a practical solution for enhancing the reliability and effectiveness of RS-IC in real-world satellite operations. We summarize our contributions in terms of translational AI for onboard satellite processing as follows:

- **Energy-aware pretrained ViT selection:** A systematic exploration of pre-trained ViTs highlights EfficientViT-M2 as the best accuracy-per-power consumption option for onboard EO, satisfying the constrained power budgets of onboard processing.
- **Comprehensive robustness validation:** Inference reliability is proven under unseen perturbations, across multiple benchmark EO datasets, and within a SatCom-in-the-loop framework that injects DVB-S2(X) channel impairments - demonstrating stable performance from sensor capture to downlink.

The remainder of this paper is organized as follows. Section II provides a brief survey of related works, serving as the basis for highlighting this paper's contributions. Section III presents the methodologies, including the dataset, machine learning (ML) models, and evaluation metrics employed in this comprehensive study. The experimental setup is explained in

Section IV, followed by the numerical results and discussion in Section V. Finally, Section VI concludes the paper.

## II. Related Works

Recent advancements in RS have focused on enhancing CNN models through transfer learning, structural modifications, and integrating attention mechanisms. For instance, [22] proposed a self-attention fused CNN architecture optimized explicitly for hyperspectral land use and land cover classification, incorporating advanced data augmentation and custom-designed CNN models to improve classification performance. Similarly, [23] introduced a CNN architecture that combines dense blocks with inverted bottleneck residuals, further enhancing deep feature extraction for RS data.

One of the critical benefits of transfer learning is its ability to mitigate the domain gap between natural and RS images. This gap has been a significant challenge due to the lack of large-scale, widely recognized benchmarks like ImageNet within the RS community. To address this, [24] proposes the MGC method, which employs a multilayer perceptron (MLP) to guide the pre-training of a CNN using small-scale RS datasets. By leveraging attention guidance, MGC effectively directs CNN branches to focus on foreground regions, thereby learning more discriminative representations. Another notable example is the transfer learning approach presented by [25], which utilizes fully pre-trained deep CNNs for land-use classification in high spatial resolution imagery. This method addresses the common issue of separation between the feature descriptor and classifier parts in transferred CNNs by pre-training both components, resulting in faster convergence and improved accuracy without sacrificing performance.

Despite the success of CNNs, the inherent limitations of these models, particularly when initialized with ImageNet pretraining, highlight the need for more specialized approaches in RS. Domain gaps between RS and natural images often lead to suboptimal performance when using ImageNet pre-training, prompting researchers to explore alternative transfer learning strategies. In this context, MLP models have shown promise in addressing challenges such as cross-domain few-shot classification. For example, [26] demonstrated that MLPs could significantly enhance discriminative capabilities and alleviate distribution shifts, thereby improving classification performance in RS tasks. Furthermore, integrating ViT models and advanced transfer learning techniques has become increasingly popular as the field progresses. For instance, [27] trained a model from scratch using a Transformer-based architecture combined with a CNN classifier layer, demonstrating the effectiveness of such hybrid approaches in processing EuroSAT images. Additionally, some studies have focused on generating synthetic data or integrating multiple data sources to maximize data utility further. For example, [28] explored the scaling laws of synthetic images generated by state-of-the-art text-to-image models, highlighting the potential of synthetic data in scenarios with limited authentic images. Similarly, [29] introduced the GeRSP framework, which combines self-supervised and supervised pretraining branches to learn robust representations from both RS and natural images, enhancing model initialization for various downstream tasks.

Moreover, addressing the challenges of data scarcity and computational efficiency has led to the development of innovative models. [30] proposes the GeoSystemNet model, which applies DL in the context of geosystem analysis to overcome labeled data scarcity in high-resolution RS data classification. In parallel, [31] introduces a parameter-efficient continual learning framework that dynamically expands pre-trained models, such as CLIP, with Mixture-of-Experts adapters to handle new tasks while preserving zero-shot recognition capabilities. In addition to improving model performance with computational efficiency, recent research has emphasized the importance of interpretability in RS models. [32] develops an interpretable DL framework for land use classification using SHAP (Shapley Additive Explanations), offering insights into how different spectral bands affect model predictions. While CNN-based approaches continue to dominate RS-IC, incorporating ViTs, advanced transfer learning methods, and synthetic data represents a significant evolution in the field. These developments enhance model robustness and generalizability and pave the way for more efficient and reliable RS applications.

Given the rapid advancements in ViT models and their increasing application in RS, there is a critical need for comprehensive studies that extensively compare existing ViT models, particularly in onboard processing. Current trends emphasize model complexity reduction through downscaling, making these models more suitable for resource-constrained environments such as satellites. However, despite these efforts, there remains a significant gap in understanding how these optimized models perform under real-world conditions, where noisy data is a common challenge. Therefore, it is essential to evaluate these models based on their computational efficiency and rigorously test their robustness during inference with noisy data. Such an analysis will provide valuable insights into the practical deployment of ViT models for onboard satellite EO, ensuring high performance and reliability in challenging operational environments.

## III. Methodology

This section details the methodological framework used in evaluating pre-trained models for land use classification in onboard satellite processing. We begin by outlining the dataset and preprocessing steps, followed by a list of the ML models that can be utilized for RS-IC. Additionally, we present the evaluation metrics employed to assess model performance in terms of accuracy, computational efficiency, and robustness, particularly under noisy conditions commonly encountered in satellite-based inference.

### A. Dataset

The EuroSAT dataset [33], a public benchmark for land use and land cover classification, will be used in this study. Specifically designed for land use and land cover classification, the EuroSAT dataset is derived from Sentinel-2 satellite imagery and comprises 27,000 geo-referenced labeled images, each measuring 64x64 pixels and spanning 13 spectral bands. It is categorized into 10 classes, including industrial buildings, residential buildings, annual crops, permanent crops, rivers, seas and lakes, herbaceous vegetation, highways, pastures, and

forests across Europe. With its compact image size and diverse class representation, EuroSAT is particularly well-suited for developing and evaluating deep learning models intended for onboard satellite processing in EO missions. In addition, the models will also be tested on the PatternNet dataset [34], which focuses on image retrieval using 30,400 high-resolution, 256x256 pixel images spanning 38 categories. The dataset's broad and diverse coverage enables the development of retrieval-focused approaches for addressing more complex, fine-grained remote sensing challenges. It is particularly valuable for applications such as real-time environmental monitoring and precision agriculture, where onboard processing capabilities are essential for timely decision-making.

### *B. Machine Learning Models*

CNNs have long been the benchmark for visual data processing. However, recent studies demonstrate that ViTs can achieve comparable or even superior performance in image classification tasks [35]. In this study, we evaluate a diverse set of ML models, including both models trained from scratch and pre-trained models, across various IC tasks. The selection spans traditional CNNs to advanced architectures like ViTs and hybrid models combining convolutional layers with Transformer-based processing. This allows us to assess the effectiveness of lightweight, efficient models alongside more complex, high-capacity networks in addressing IC challenges.

*1) Training from Scratch:*

- **CNN:** A basic CNN designed for IC, consisting of two convolutional layers followed by batch normalization, ReLU activation, and max pooling, with fully connected layers at the end for classification.
- **ResNet-14:** A smaller version of the ResNet framework, termed ResNet-14, with 14 layers, including two residual blocks in each of the three hidden layers, optimizing feature extraction and gradient flow through the network.
- **Compact Transformer (CCT) [36]:** A hybrid model that combines convolutional layers with transformer-based processing, utilizing convolutional tokenization followed by transformer encoder layers for robust feature extraction and classification.
- **Small ViT [37]:** A small ViT model designed for efficiency, using a small patch tokenization process and a lightweight transformer with self-attention mechanisms to classify images effectively.

*2) Pretrained ViT Models:*

- **EfficientViT-M2 [21]:** An efficient ViT combining convolutional layers with local window attention mechanisms optimized for balancing performance and computational efficiency with approximately 4 million parameters.
- **MobileViTV2 [20]:** A hybrid model leveraging convolutional and transformer-based processing, utilizing depthwise separable convolutions and self-attention mechanisms for accurate and efficient IC.
- **xLSTM [38]:** Integrates convolutional patch embedding with advanced LSTM-based layers to capture spatial and sequential dependencies for comprehensive IC.
- **EfficientNet-B2 [39]:** A highly efficient CNN using depthwise separable convolutions and squeeze-and-excitation layers to minimize parameter usage while maximizing performance.
- **ResNet50-DINO [40]:** A ResNet-50 model trained using self-supervised learning with DINO, optimized for feature representation without labeled data, enabling robust classification performance.
- **EfficientViT-L2 [41]:** Combines convolutional and transformer-based architectures with techniques like fused MBConv and Lite Multi-Head Attention to balance efficiency and representational power effectively.
- **SwinTransformer [42]:** A hierarchical ViT that divides images into non-overlapping patches, utilizing shifted window-based self-attention to capture both local and global context.
- **ViT [43]:** A transformer model that divides images into patches, processing them through multiple transformer layers with self-attention to achieve high-capacity IC.

Table I summarizes various ML models by comparing their advantages and disadvantages. CNNs are ideal for simple, low-complexity tasks due to their ease of implementation and low computational cost, whereas ResNet-14 balances complexity and performance for mid-scale tasks. Compact Transformers and SmallViTs, while offering flexible and efficient architectures, may struggle with large datasets or higher complexity tasks. EfficientViT models, including M2 and L2, are well-suited for resource-constrained environments and high-performance tasks, respectively, but they require careful consideration of their computational limits. MobileViTV2 and Vision-xLSTM are powerful for specific tasks but demand significant computational resources. SwinTransformer and ViT models excel in capturing detailed and complex features, making them suitable for high-capacity IC tasks, though they come with high computational costs. Each model has its niche, making it ideal for different types of applications based on specific needs and resource availability.

### *C. Evaluation Metrics*

To comprehensively evaluate the performance of our multi-class classification model across 10 classes, we employ three key metrics: accuracy, precision, and recall (sensitivity) [44]. These metrics are calculated for each class individually and then aggregated using macro-averaging to assess the model's performance as follows,

$$\text{Accuracy} = \sum_{k=1}^{K} \frac{\text{TP}_k}{N},$$

$$\text{Precision} = \frac{1}{N} \sum_{i=1}^{K} N_k \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k},$$

$$\text{Recall} = \frac{1}{N} \sum_{i=1}^{K} N_k \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k},$$

where $N$ is the total number of data points across all classes. $K$ is the total number of classes. $N_k$ is the number of data points in class $k$. $TP_k$ is True Positives, $FP_k$ is False Positives, $FN_k$ is False Negatives for class $k$, respectively.

TABLE I: COMPARISON OF MACHINE LEARNING MODELS

| Models | Advantages | Disadvantages |
|---|---|---|
| **CNN** | - Simple and easy to implement<br>- Low computational cost<br>- Good for small-scale tasks | - Limited to basic features<br>- Less effective for complex data<br>- Overfitting with small datasets |
| **ResNet-14** | - Efficient with residual blocks<br>- Good for deep networks<br>- Avoids vanishing gradient problem | - Higher computational cost than CNN<br>- More complex to train<br>- May still require large datasets |
| **CCT** | - Hybrid model benefits from convolution and transformer<br>- Captures both local and global features<br>- Flexible for various tasks | - More complex architecture<br>- Higher memory consumption<br>- Slower inference time |
| **SmallViT** | - Efficient transformer architecture<br>- Good balance of accuracy and speed<br>- Suitable for medium-sized datasets | - Limited capacity for very large data<br>- Higher training time compared to simpler models |
| **EfficientViT-M2** | - Lightweight and efficient<br>- Combines benefits of convolutions and self-attention<br>- Suitable for devices with limited resources | - Less powerful than larger transformers<br>- May struggle with highly complex datasets |
| **MobileViTV2** | - Strong performance on visual tasks<br>- Efficient for both training and inference<br>- Good balance of complexity and accuracy | - Higher memory footprint<br>- Requires more computational resources |
| **Vision-xLSTM** | - Captures both spatial and sequential dependencies<br>- Good for tasks needing sequence modeling<br>- Robust feature extraction | - Very high computational cost<br>- Long training times<br>- Complex architecture |
| **EfficientNet-B2** | - Highly efficient network<br>- Optimizes performance with low parameters<br>- Scalable to different levels of complexity | - May not capture the most complex features<br>- Requires larger datasets to perform optimally |
| **ResNet50-DINO** | - Self-supervised learning<br>- Strong feature representation<br>- Adaptable to various tasks | - Very high computational demand<br>- Large memory requirements<br>- Slow inference time |
| **EfficientViT-L2** | - Combines convolutional and transformer architectures<br>- Efficient computation with robust performance<br>- Scales well with large datasets | - Large model size<br>- High memory and computational requirements |
| **SwinTransformer** | - Hierarchical structure for better feature extraction<br>- Strong for both local and global contexts<br>- Scales efficiently with input size | - Very high memory and computational demand<br>- Complex to implement and optimize |
| **ViT** | - High representational power<br>- Strong at capturing complex patterns<br>- Scalable for larger datasets | - Requires very large datasets for optimal performance<br>- High computational and memory demands |

Using weighted precision and recall, we ensure that each class is given equal importance, thereby providing a balanced evaluation of the model's classification capabilities across the entire dataset. These macro-averaged evaluation metrics will be used to select the best models in the final analysis.

To assess the model's robustness, we adopt a benchmarking approach for evaluating NN robustness to common corruptions and perturbations [45]. The robustness score measures the degradation in accuracy caused by a specific perturbation, $\phi$. For a given model, $m$, let $A_{\text{clean}}$ represent the model's accuracy on the original (clean) test dataset, and $A_\phi$ denote the accuracy of the model on the test set that has been modified with perturbation $\phi$. These perturbations can stem from natural environmental conditions encountered during deployment or adversarial modifications of the original test samples. Following the methodology outlined in [46], we define the robustness score of model $f$ with respect to perturbation $\phi$ as:

$$R_f^\phi = \frac{A_\phi}{A_{\text{clean}}}. \tag{1}$$

A robustness score of model $f$ for perturbation $\phi$, namely $R_f^\phi$, close to one indicates that the model is highly resistant to perturbation, demonstrating its ability to maintain performance in challenging conditions.

## IV. EXPERIMENTAL SETUP

This section presents the experimental setup used to evaluate the performance of the models. We describe the hardware and software configurations, along with the specific parameters used for training and inference. Additionally, we outline the data augmentation techniques and noise simulations applied to mimic real-world conditions in satellite-based EO tasks. The experiments focus on comparing traditional CNNs, ResNet-based models, and pre-trained ViT models, including MobileViTV2 and EfficientViT-M2, to assess their effectiveness in achieving high accuracy, computational efficiency, and robustness under noisy conditions.

### A. ML Model Configuration

First, to ensure a balanced and comprehensive evaluation of the model's performance, the dataset was split into training and testing sets with a 70/30 ratio. For model optimization, different strategies have been employed based on the model architecture. The CNN and ResNet models were optimized using the Stochastic Gradient Descent (SGD) optimizer [47], configured with a learning rate of 1e-3, momentum of 0.9, and a weight decay of 5e-4. In contrast, the ViT-based models, including the pre-trained variants, have been optimized using the AdamW optimizer [48]. This setup also utilized a learning rate 1e-3 and a weight decay 5e-4, ensuring consistency across

different model types. A step learning rate scheduler was applied to further enhance training efficiency with a step size of 7 epochs and a decay factor ($\gamma$) of 0.1.

The CCT [36] and SmallViT models [37] have been trained using code adapted from the ViT-PyTorch repository, which provides reliable and optimized implementations for ViT. Other pre-trained models have been sourced from Huggingface's model hub or the official repositories in the respective papers. This approach ensured that the models were implemented consistently, with standardized architectures and preprocessing steps, allowing for a fair and direct comparison of their performance across the different experiments.[2]

### B. Hardware setup

All experiments have been conducted on a workstation equipped with an Intel(R) Xeon(R) W-11855M CPU, operating at 3.20GHz with 12 cores and supported by 64 GB of RAM. The models have been trained and tested using an Nvidia RTX A5000 GPU with 48 GB of GDDR6 memory, providing sufficient computational resources for most of the tasks. This setup offered a robust and efficient environment for handling most models' training and evaluation processes.

However, due to their increased complexity and size, more extensive computational resources are necessary for the two larger pre-trained ViT models (ViT-Large and ViT-Huge). These models were parallelly trained on a high-performance computing workstation with the following specifications: an Intel Xeon w9-3475X CPU featuring 36 cores and 72 threads, a maximum turbo frequency of 4.8 GHz, and an 82.5 MB cache. The system is also equipped with 512 GB of DDR5 RAM, clocked at 4800 MHz, and dual NVIDIA RTX™ 6000 Ada Generation GPUs, each with 48 GB of GDDR6 memory. This advanced hardware configuration ensured efficient training of the larger models, leveraging parallel processing and extensive memory resources.

### C. Noisy Data for Inference

After selecting the best model with lower computational complexity and high performance, we assess its robustness during onboard satellite inference. This evaluation focuses on the model's ability to maintain performance under noisy conditions commonly encountered in real-world satellite operations. Although the model has been trained on a clean dataset, the test data used during inference may be corrupted by noise introduced by natural events or instrument movement. This robustness analysis is critical to ensure reliable model performance in dynamic and unpredictable environments. Specifically, we introduce two types of noise into the inference test data: Gaussian noise and instrument noise represented by motion blur. For every kind of noise, we define five levels of severity, as illustrated in Fig. 1. This evaluation helps us understand how well the models can perform under varying noise levels, which is critical for reliable satellite-based IC. The noise and severity level was implemented based on the code[3] from [45].

To further assess the robustness and generalization capabilities of the selected models, we applied a series of data augmentation techniques, as illustrated in Fig. 2. The augmentations include Standard Transformed, Rand Augment [49], Random Erasing [50], Rand (Augment+Erasing), and Strong Augmentation, which combine all four augmentations designed to simulate challenging variations in the input data. These transformations introduce different levels of distortion and occlusion, such as color shifts, random erasures, and severe augmentations. By evaluating the models on these augmented datasets, we aim to determine their ability to maintain performance when faced with altered or corrupted inputs, thereby ensuring their reliability in diverse and unpredictable conditions.

## V. Results and Discussions

Table II provides a detailed comparison of the computational complexity of various models, with a particular emphasis on total parameters, estimated total size, floating point operations per second (FLOPs), training time, and inference time. Notably, the ViT-large and ViT-huge models stand out for their substantial complexity. They are characterized by high parameter counts, large memory footprints, and extensive computational operations (FLOPs), resulting in significantly longer training and inference times. In contrast, smaller models like EfficientViT-M2 and MobileViT2, despite being part of the ViT family, demonstrate much lower complexity across all metrics. These models are designed to be more lightweight, with fewer parameters and reduced FLOPs, making them more efficient in terms of both computational and memory resources. Compared to the pre-trained ViT models, EfficientViT-M2 and MobileViT2 balance performance and efficiency, providing a viable alternative for limited computational resources or faster processing times. This comparison underscores the diversity within Transformer-based models and highlights the importance of selecting models that align with the specific needs of a given application.

Based on five runs for each model, the experimental results from Fig. 3 clearly demonstrate that fine-tuning pre-trained models consistently outperforms training models from scratch. Pre-trained models deliver superior performance and exhibit excellent stability across metrics, including mean accuracy, precision, and recall. Among the pre-trained models, MobileViT-V2, SwinTransformer, and EfficientViT models stand out as the top performers, with MobileViT-V2 leading with the highest mean accuracy (98.97%), mean precision (98.97%), and mean recall (98.97%). However, it is essential to note that increasing the model size does not necessarily lead to better performance. For instance, while the ViT-base model performs well, its larger counterparts, ViT-large

[2]Available codes and pre-trained models:
ViT-Pytorch: https://github.com/lucidrains/vit-pytorch/tree/main
EfficientViT-M2: https://github.com/microsoft/Cream/tree/main/EfficientViT
MobileViTV2: https://huggingface.co/docs/transformers/en/model_doc/mobilevitv2
Vision-xLSTM: https://github.com/NX-AI/vision-lstm
EfficientNet-b2: https://huggingface.co/google/efficientnet-b2
ResNet-DINO: https://github.com/facebookresearch/dino
EfficientViT-L2: https://github.com/mit-han-lab/efficientvit
SwinTransformer: https://huggingface.co/docs/transformers/en/model_doc/swin
ViT-base: https://huggingface.co/google/vit-base-patch16-224
ViT-large: https://huggingface.co/google/vit-large-patch16-224
ViT-huge: https://huggingface.co/google/vit-huge-patch14-224-in21k

[3]https://github.com/hendrycks/robustness

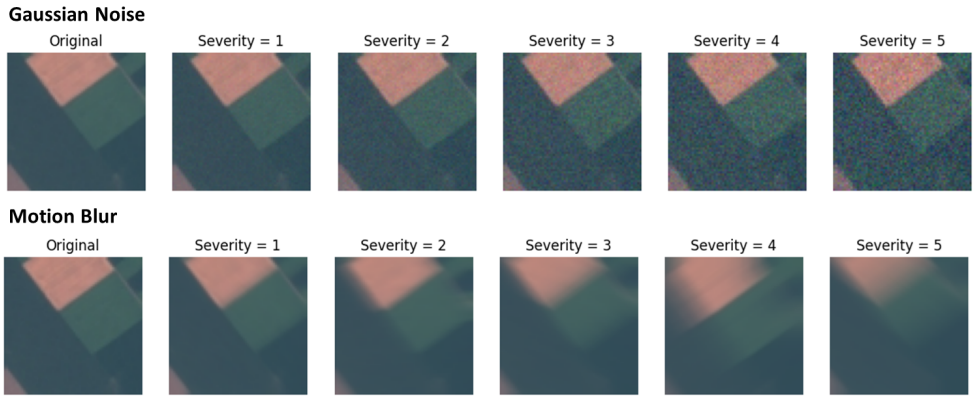


Fig. 1: Different noise levels with Gaussian and motion blur.

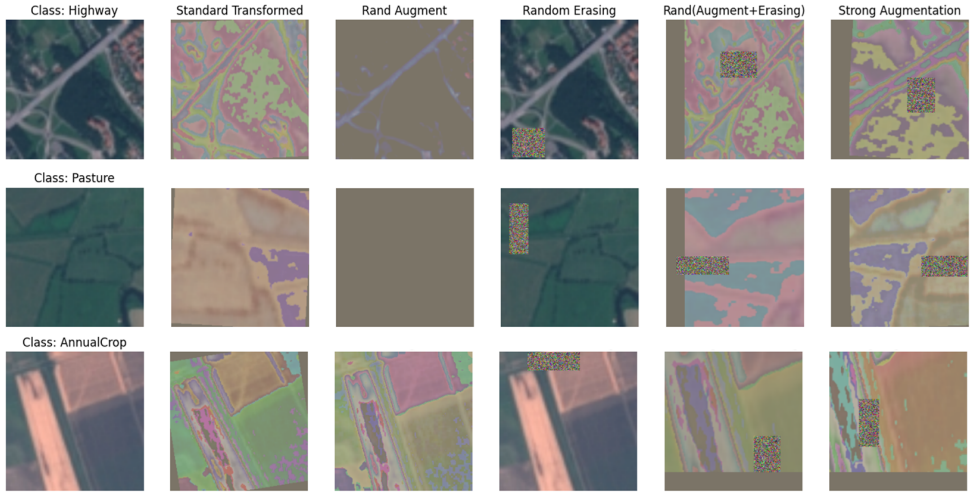


Fig. 2: Different augmentation techniques.

TABLE II: SUMMARY OF MODEL COMPUTATION COMPLEXITY, ESTIMATED TOTAL SIZE, TRAINING, AND INFERENCE TIMES

| Models | Input Size | Total Parameters | Size (MB) | FLOPs | Training (s) | Inference (s) |
|---|---|---|---|---|---|---|
| CNN | 64x64 | 66,330 | 0.71 | 0.93 MFLOPs | 233 | 7 |
| ResNet | 64x64 | 195,738 | 10.01 | 117.88 MFLOPs | 487 | 6.7 |
| CCT (Compact CNN-Transformer) | 64x64 | 1,507,211 | 7.50 | 62.86 MFLOPs | 241 | 23 |
| SmallViT | 64x64 | 2,764,562 | 21.22 | 213.84 MFLOPs | 821 | 21 |
| EfficientViT-M2 (Microsoft) | 224x224 | 3,964,804 | 38.19 | 203.53 MFLOPs | 1000 | 10 |
| MobileViTV2 (Apple) | 256x256 | 4,393,971 | 259.30 | 1.84 GFLOPs | 2,096 | 16 |
| Vision-xLSTM | 224x224 | 6,090,098 | 141.66 | 1.86 GFLOPs | 431,941 | 33 |
| EfficientNet-b2 (Google) | 260x260 | 7,715,084 | 252.86 | 32.92 MFLOPs | 3,100 | 18 |
| ResNet-DINO (Facebook) | 224x224 | 23,528,522 | 272.54 | 4.14 GFLOPs | 2,294 | 15 |
| EfficientViT-L2 (MiT) | 224x224 | 60,538,026 | 457.56 | 6.99 GFLOPs | 6000 | 47 |
| SwinTransformer (Microsoft) | 224x224 | 86,753,474 | 636.38 | 15.47 GFLOPs | 21,573 | 538 |
| ViT-base (Google) | 224x224 | 85,806,346 | 505.40 | 16.87 GFLOPs | 8,582 | 46 |
| ViT-large (Google)* | 224x224 | 303,311,882 | 1642.31 | 59.70 GFLOPs | 9,271 | 60 |
| ViT-huge (Google)* | 224x224 | 630,777,610 | 3454 | 162.00 GFLOPs | 13,533 | 75 |

* Parallelly trained on a higher computing workstation with the following specifications:
**CPU:** Intel Xeon w9-3475X (82.5 MB Cache, 36 cores, 72 threads, 4.8 GHz, 300 W), 512 GB DDR5 RAM, 4800 MHz
**GPU:** Dual NVIDIA RTX™ 6000 Ada Generation, 48 GB GDDR6

and ViT-huge, do not offer improved performance. ViT-huge shows a lower mean accuracy (88.54%) and more significant variability, indicating unstable performance. This suggests that extensive models can become less reliable due to overfitting or the challenges in optimizing such complex architectures. Therefore, the results emphasize balancing model size with performance rather than assuming that larger models will inherently perform better.

Based on the best run for each model in Table III, the summarized experimental results identify MobileViTV2, EfficientViT-M2, and SwinTransformer as the top-performing models. MobileViTV2 (Apple) achieves the highest overall performance, with a perfect training accuracy of 100%, leading to test accuracy, precision, and recall, each at 99.09%. SwinTransformer follows closely, boasting a training accuracy of 99.8% and a test accuracy, precision, and recall of 98.83%. EfficientViT-M2 ranks third, delivering impressive results with a training accuracy of 99.99% and test accuracy, precision,

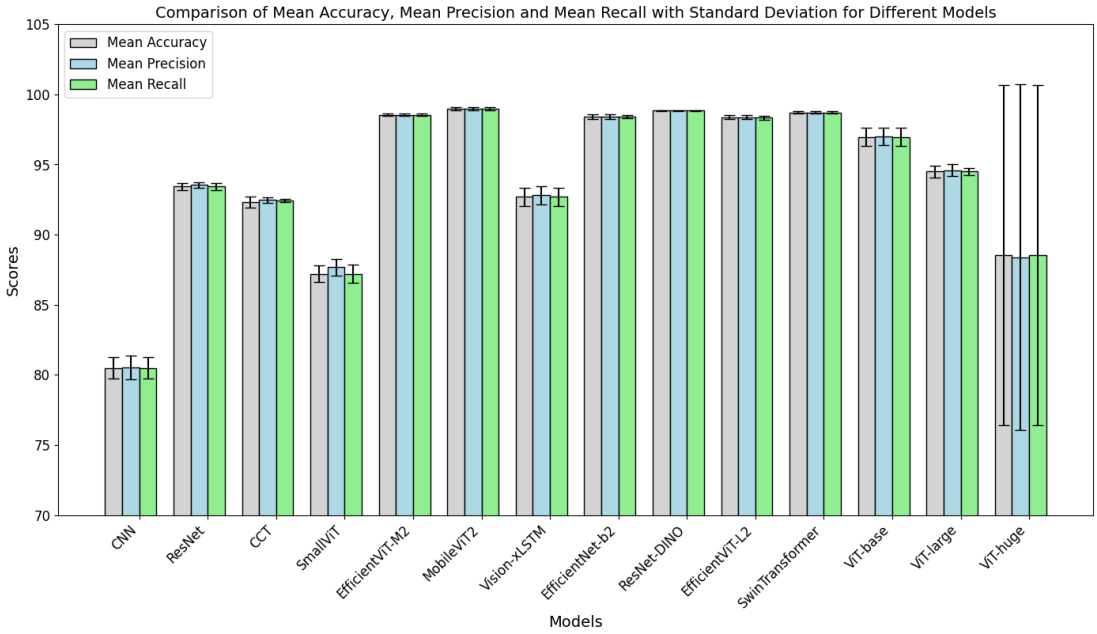


Fig. 3: Statistical comparison for model performance.

TABLE III: EXPERIMENTAL RESULTS FOR THE BEST MODEL'S PERFORMANCE ON EUROSAT

| Models | Train Loss | Test Loss | Train Accuracy | Accuracy | Precision | Recall |
|---|---|---|---|---|---|---|
| CNN | 0.469 | 0.534 | 83.73 | 81.82 | 81.07 | 81.19 |
| ResNet | 0.217 | 0.206 | 93.5 | 93.88 | 93.89 | 93.88 |
| CCT (Compact CNN-Transformer) | 0.038 | 0.26 | 99.02 | 92.61 | 92.7 | 92.61 |
| SmallViT | 0.25 | 0.42 | 90.98 | 86.49 | 86.9 | 86.49 |
| EfficientViT-M2 (Microsoft) | 0.002 | 0.039 | *99.99* | **98.76** | **98.77** | **98.76** |
| MobileViTV2 (Apple) | 0.00036 | 0.034 | **100** | **99.09** | **99.09** | **99.09** |
| Vision-xLSTM | 0.34 | 0.41 | 86.9 | 85.8 | 85.9 | 85.8 |
| EfficientNet-b2 (Google) | 0.007 | 0.07 | 99.9 | 98.47 | 98.49 | 98.47 |
| ResNet-DINO (Facebook) | 0.28 | 0.18 | 94.2 | 94.9 | 94.5 | 94.9 |
| EfficientViT-L2 (MiT) | 0.0006 | 0.081 | *99.99* | 98.14 | 98.16 | 98.15 |
| SwinTransformer (Microsoft) | 0.009 | 0.034 | 99.8 | *98.83* | *98.83* | *98.83* |
| ViT-base (Google) | 0.007 | 0.05 | 99.96 | 98.45 | 98.47 | 98.45 |
| ViT-large (Google) | 0.002 | 0.04 | 100 | 98.55 | 98.55 | 98.55 |
| ViT-huge (Google) | 0.105 | 0.127 | 97.5 | 96.39 | 96.39 | 96.39 |

**Bold** denotes the best values.
*Italic and underline* denote the second best values.
**Bold and underline** denote the third best values.

and recall, all at 98.76%. When comparing the top four large pre-trained models, the SwinTransformer notably outperforms its counterparts despite having a smaller parameter count of 86 million. It surpasses the ViT-base, ViT-large, and ViT-huge models, which do not achieve better results despite their larger parameter sizes. This highlights the effectiveness of the Swin Transformer's design, which utilizes a hierarchical ViT architecture. Unlike conventional ViT, which relies on a global attention mechanism across the entire image, the Swin-Transformer divides images into non-overlapping patches and hierarchically processes them. This method allows it to handle images at various scales and resolutions efficiently, resulting in better generalization and improved task performance. The superior performance of the SwinTransformer, even with fewer parameters, underscores the advantages of its architecture in effectively capturing both local and global image features, making it a more efficient and effective model compared to the larger, conventional attention-based ViT models.

Fig. 4 compares power consumption in W during the inference on GPU. Among the pretrained ViT-based models, EfficientViT-M2 stands out for its significantly lower power consumption, registering at 29.04 $\pm$ 0.96 W. This makes it the most energy-efficient choice compared to other ViT-based models, such as MobileViTV2, which consumes 79.23 $\pm$ 1.45 W, and even more power-intensive models like ViT-huge, which uses 552.96 $\pm$ 6.19 W. The EfficientViT-M2 model also compares favorably with models trained from scratch, such as CNNs and ResNets, which exhibit power consumption of 15.42 $\pm$ 0.52 W and 24.63 $\pm$ 1.63 W, respectively. This demonstrates that EfficientViT-M2 achieves a comparable level of power efficiency during inference despite its more complex architecture and superior performance. This balance of computational power and efficiency makes it an optimal choice for onboard satellite processing, where energy consumption is a critical constraint.

From the results presented in Tables II, and III, and Fig. 3, and 4 we confirm that MobileViT and EfficientViT are the two best models for IC on EuroSat. These models excel due to their lower computational complexity, shorter training and inference times, and superior performance across all evaluation metrics.

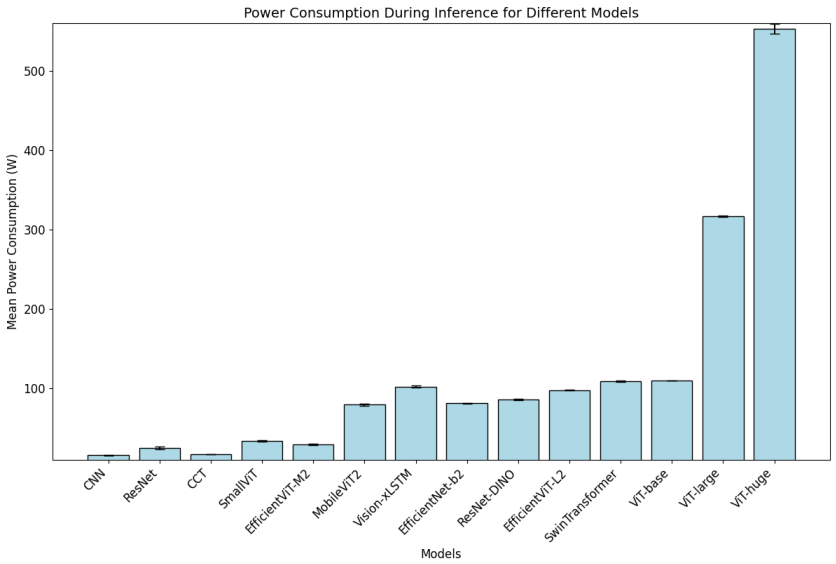


Fig. 4: Statistical comparison for power consumption during inference.

Notably, compared to state-of-the-art models, as shown in Table IV, the exceptional performance of EfficientViT-M2 and MobileViTV2 becomes evident. In the EuroSAT IC task, MobileViTV2 achieves the highest scores across all evaluation metrics, outperforming all other models in the comparison. EfficientViT-M2 also performs exceptionally well, securing the second-highest results. These findings highlight the effectiveness of both EfficientViT-M2 and MobileViTV2 in achieving top-tier performance with relatively lower computational complexity and faster training and inference times, as observed in our experiment results. This confirms that these two models are currently the best choices for high-performance IC tasks, providing a significant advantage over existing methods.

Additionally, Fig. 5 shows the confusion matrices for EfficientViT-M2 and MobileViTV2, highlighting their exceptional performance in IC on EuroSat. Both models exhibit vital accuracy across all classes, with most predictions aligning closely with the true labels. In the confusion matrix for EfficientViT-M2, we observe that classes like *AnnualCrop*, *Forest*, *PermanentCrop*, and *SeaLake* are classified with nearly perfect accuracy, showing minimal misclassification. Similarly, the confusion matrix for MobileViTV2 shows even fewer errors, with several classes like *Forest*, *PermanentCrop*, and *SeaLake* being classified almost flawlessly. A few minor misclassifications are observed in both models, particularly in categories like *Pasture* and *River*, but these are relatively small compared to the overall high performance. These matrices underscore the effectiveness of both models in handling complex IC tasks, with MobileViTV2 slightly outperforming EfficientViT-M2 in terms of fewer misclassifications, further affirming its position as the top-performing model.

Furthermore, Table V and the accompanying confusion matrix, Fig. 6, clearly demonstrate that EfficientViT-M2 achieves near-perfect scene recognition on PatternNet. With 99.52% Top-1 accuracy, precision, and recall, the model sits within 0.14 percentage points of the best-reported score while outperforming all remaining baselines, including Swin-Transformer (99.23%) and ResNet-50 (98.23%) [54]. The confusion matrix perfectly classifies 38 classes, confirming exceptionally consistent predictions across every land-use category. This combination of accuracy, uniform class-level reliability, and the architecture's lightweight design, again highlights EfficientViT-M2 as an excellent choice for onboard EO applications where both performance and computational efficiency are critical.

The experiment results, as presented in Tables VI and VII, reveal that applying data augmentation techniques during training and inference generally leads to a slight re-

TABLE IV: PERFORMANCE COMPARISON WITH THE-STATE-OF-THE-ART MODELS (TOP-1) ON EUROSAT.

| Models | Accuracy | Precision | Recall |
|---|---|---|---|
| EfficientViT-M2 | *98.76* | *98.77* | *98.76* |
| MobileViTV2 | **99.09** | **99.09** | **99.09** |
| Few-shot MLP [26] | 79.62 | x | x |
| Attention+CNN [22] | 89.5 | x | x |
| ACL [51] | 95.46 | x | x |
| MGC [24] | 96.41 | x | x |
| GeRSP [29] | 97.87 | x | x |
| Constrastive Learning [28] | 96 | x | x |
| GeoSystemNet [30] | 95.32 | x | x |
| Transformer+CNN [27] | 95.48 | x | x |
| MoE-ViT [31] | 98.1 | x | x |
| Self-Attention CNN [23] | 90.3 | 89.05 | 88.59 |
| SIBNet [52] | 97.8 | 97 | 96.97 |
| CNN-SHAP [32] | 94.72 | 93.73 | 94.10 |
| CNN-SVM [53] | 97.91 | x | x |

**Bold** denotes the best values.
*Italic and underline* denote the second best values.

TABLE V: PERFORMANCE COMPARISON WITH THE-STATE-OF-THE-ART MODELS (TOP-1) ON PATTERNNET.

| Models | Accuracy | Precision | Recall |
|---|---|---|---|
| EfficientViT-M2 | ***99.52*** | ***99.52*** | ***99.52*** |
| MobileViTV2 | **99.66** | **99.66** | **99.66** |
| EfficientViT-b2 | *99.54* | *99.54* | *99.53* |
| ResNet-DINO | 96.01 | 96.01 | 96.00 |
| SwinTransformer | 99.23 | 99.23 | 99.23 |
| ResNet-50 [54] | 98.23 | 97.95 | x |

**Bold** denotes the best values.
*Italic and underline* denote the second best values.
**Bold and underline** denote the third best values.

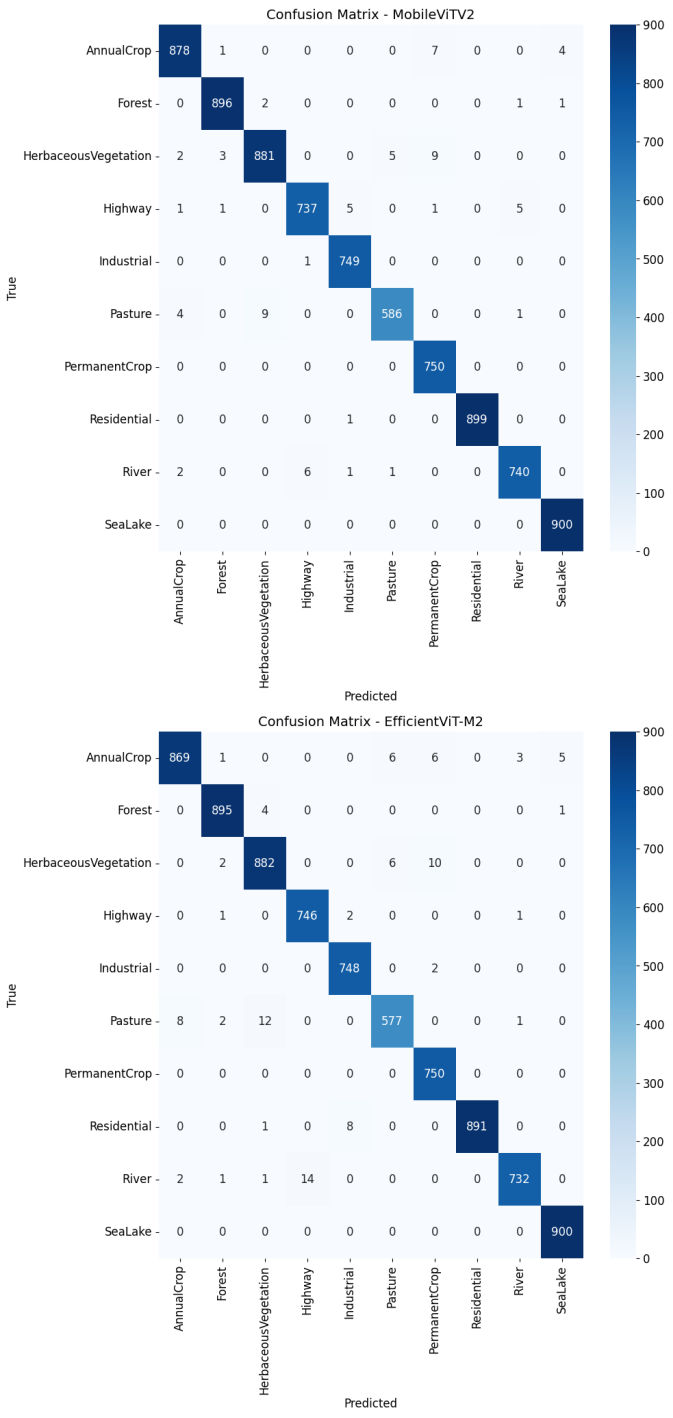


Fig. 5: Confusion matrix from MobileViTV2 (top) and EfficientViT-M2 (bottom) performance on EuroSat.

TABLE VI: MOBILEVITV2 PERFORMANCE WITH DIFFERENT AUGMENTATION TECHNIQUES

| MobileViTV2 | Accuracy | Precision | Recall |
|---|---|---|---|
| Baseline | **99.09** | **99.09** | **99.09** |
| RandAugment | 98.68 | 98.68 | 98.68 |
| StandardAugment | 98.23 | 98.25 | 98.23 |
| RandErasing | 98.73 | 98.73 | 98.73 |
| Rand(Aug+Eras) | 98.39 | 98.42 | 98.39 |
| StrongAugment | 98.24 | 98.25 | 98.24 |

**Bold** denotes the best values.

TABLE VII: EFFICIENTVIT PERFORMANCE WITH DIFFERENT AUGMENTATION TECHNIQUES

| EfficientViT-M2 | Accuracy | Precision | Recall |
|---|---|---|---|
| Baseline | **98.76** | **98.77** | **98.76** |
| RandAugment | 98.59 | 98.6 | 98.59 |
| StandardAugment | 97.39 | 97.44 | 97.39 |
| RandErasing | 98.22 | 98.24 | 98.22 |
| Rand(Aug+Eras) | 97.96 | 97.99 | 97.96 |
| StrongAugment | 97.09 | 97.14 | 97.1 |

**Bold** denotes the best values.

duction in model performance when evaluated on clean test data. MobileViTV2's and EfficientViT-M2's baseline performance—achieved without augmentation—yield the highest accuracy. However, among the various augmentation techniques tested, RandErasing for both models demonstrates the most effective results, providing performance metrics (98.73% for MobileViTV2 and 98.22% for EfficientViT-M2) that are closest to the baseline. This indicates that RandErasing introduces sufficient variability to enhance robustness without significantly compromising the models' ability to perform well on clean data. Although there is a slight decrease compared to the baseline, using RandErasing offers a good balance, maintaining high accuracy while preparing the models to handle real-world scenarios with noisy inputs.

When comparing the robustness of MobileViTV2 and EfficientViT-M2 under Gaussian noise conditions, as shown in Fig. 7, both models demonstrate varying levels of resilience depending on the severity of the noise and the specific augmentation technique employed during training. At lower severity levels (1 and 2), both models maintain high robustness scores across all augmentation techniques. However, EfficientViT-M2 slightly outperforms MobileViTV2, particularly at severity level 1, where it achieves a robustness score of 0.989 with the Rand(Aug+Eras) technique, compared to MobileViTV2's 0.976. As the severity of the Gaussian noise increases, a general decline in robustness is observed for both models; nevertheless, EfficientViT-M2 consistently maintains a higher robustness score, especially at severity levels 3 to 5, where the Rand(Aug+Eras) and StrongAugment techniques prove most effective. At severity level 5, EfficientViT-M2 achieves a robustness score of 0.505 using Rand(Aug+Eras), significantly higher than MobileViTV2's score of 0.447 with StandardAugment. This suggests that EfficientViT-M2 exhibits greater resilience to high levels of Gaussian noise, especially when trained with augmentation techniques that introduce substantial variability. Conversely, MobileViTV2 performs comparably well at lower noise levels and demonstrates strong performance with StandardAugment at lower severity, but it faces more significant challenges as noise severity increases. Nonetheless, RandErasing provides MobileViTV2 with a balanced performance, helping it maintain reasonable robustness even under more challenging conditions. Overall, while both models show robustness against Gaussian noise, EfficientViT-M2 consistently maintains higher robustness scores at elevated noise levels, indicating its potential advantage in more demanding noisy environments.

Under motion blur noise, as depicted in Fig. 8, both MobileViTV2 and EfficientViT-M2 display resilience at lower sever-

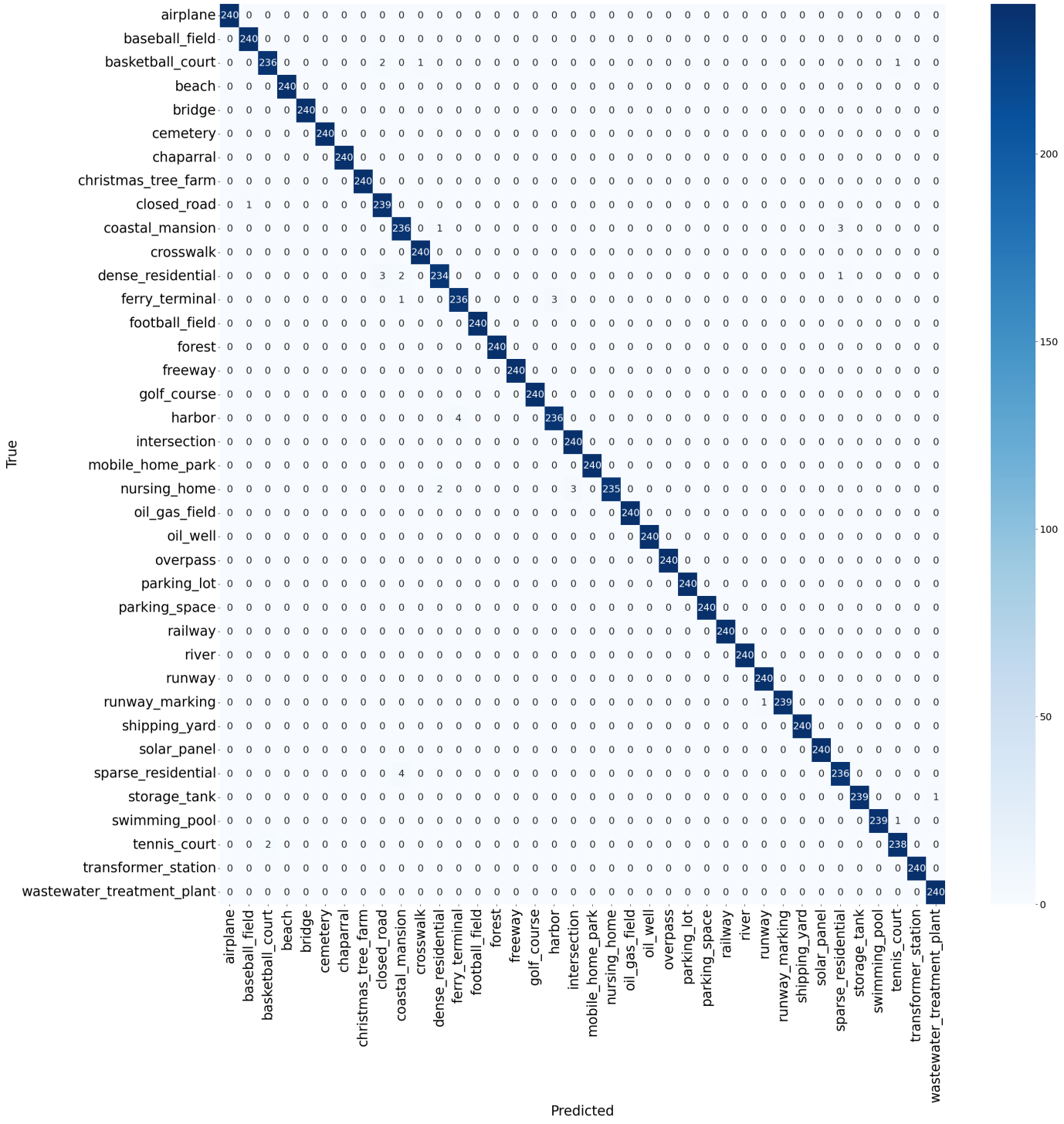


Fig. 6: Confusion matrix of EfficientViT-M2 on PatternNet.

ity levels, though a noticeable decline in robustness is observed as the severity increases. At severity level 1, both models achieve high robustness scores, with EfficientViT-M2 slightly outperforming MobileViTV2, reaching near-perfect scores with StandardAugment (0.989) and RandErasing (0.980). In contrast, MobileViTV2 peaks at a robustness score of 0.984 with Rand(Aug+Eras). As the severity of motion blur intensifies, EfficientViT-M2 consistently demonstrates superior robustness, particularly with RandErasing, maintaining higher robustness scores across all severity levels. At severity level 5, EfficientViT-M2 achieves a robustness score of 0.846 with StandardAugment, significantly outperforming MobileViTV2's top score of 0.722 with Rand(Aug+Eras). These results indicate that EfficientViT-M2 is better equipped to handle severe motion blur noise, maintaining a more stable performance than MobileViTV2. While MobileViTV2 remains robust at lower severity levels, it experiences a more pronounced decline in performance as severity increases, particularly with

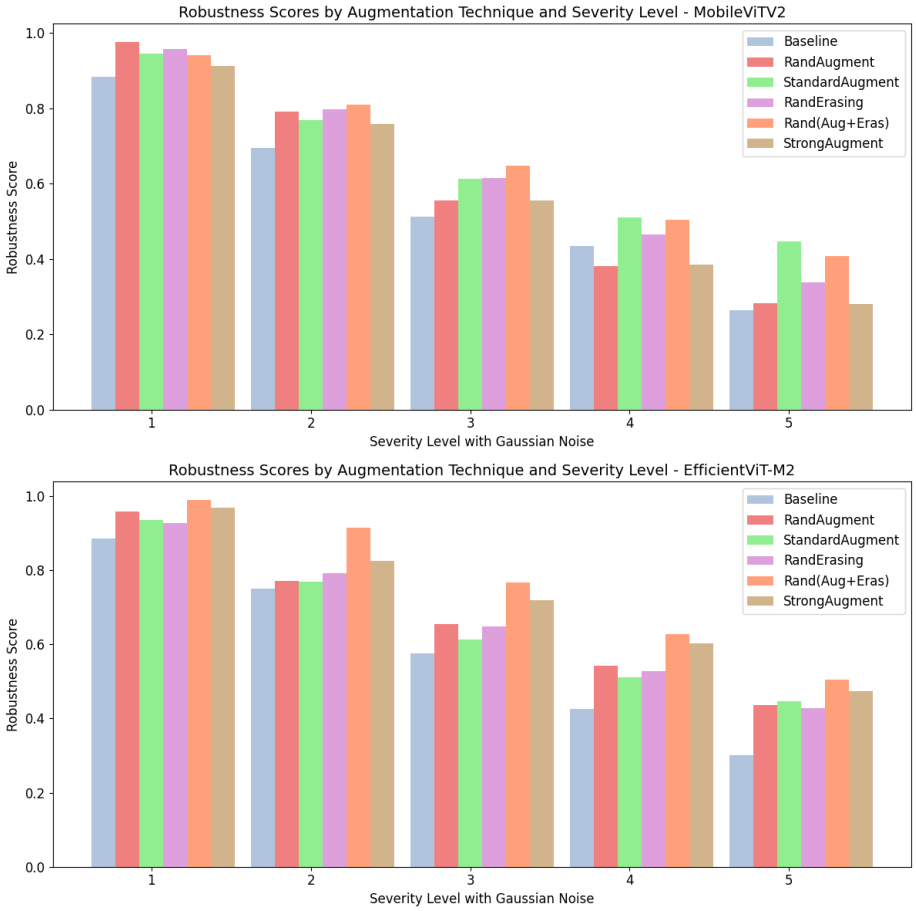


Fig. 7: MobileViT (Top) and EfficientViT (Bottom) robustness with Gaussian noisy inference data.

RandAugment and Baseline, where robustness scores drop more sharply. Nevertheless, when trained with StrongAugment, MobileViTV2 shows improved performance at higher severity levels, though it still falls short of EfficientViT-M2. Overall, EfficientViT-M2 proves to be more resilient to motion blur noise, particularly at higher severity levels, making it a more reliable choice in scenarios where such noise is prevalent. MobileViTV2, while strong at lower noise levels, may benefit from more robust augmentation techniques to enhance its performance under severe motion blur conditions.

The results from the experiments on Gaussian noise and motion blur conclusively demonstrate the superior robustness of the EfficientViT-M2 model compared to MobileViTV2. Across varying levels of noise severity, EfficientViT-M2 consistently outperformed MobileViTV2, particularly at higher noise levels where maintaining model performance becomes increasingly challenging. These findings underscore the capability of EfficientViT-M2 to reliably handle noisy environments, making it a robust and dependable choice for real-world applications where various noise factors may compromise image quality. This enhanced robustness positions EfficientViT-M2 as a leading model for deployment in challenging operational settings, such as on-board satellite IC, where consistent and accurate performance is critical.

Furthermore, the feasibility of deploying the EfficientViT-M2 model in practical applications, mainly OSS, is underscored by its relatively low computational complexity and compact size. With an estimated total size of 38.19 MB and a computational complexity of 203.53 MFLOPs, EfficientViT-M2 offers a significant advantage over earlier models used in pioneering satellite missions. For instance, the Φ-Sat-1 mission [14], which utilized the CloudScout model for image segmentation, required approximately 4.67 GFLOPs and had an estimated total size of 123.98 MB. In contrast, EfficientViT-M2 requires substantially less computational power and storage, making it a more efficient choice for similar tasks. Additionally, compared to the Φ-Sat-2 mission [15], which employed a convolutional autoencoder with 81.2 MFLOPs for vision processing on an Intel Movidius Myriad 2 VPU, EfficientViT-M2's slightly higher complexity is justified by its enhanced robustness and performance capabilities. While the Φ-Sat-2 model took nearly 11 seconds to compress an image with a compression ratio of 8 on a VPU, EfficientViT-M2's processing requirements are more aligned with modern GPU capabilities, completing complex tasks much faster and with greater accuracy. Furthermore, the RepSViT model [16] utilizes 3.77 million parameters and consumes 600 MFLOPs of computational costs for onboard image processing, but EfficientViT-M2 offers a ≈3x reduction in complexity compared to RepSViT. This lower computational cost, coupled with its efficient architecture, positions EfficientViT-M2 as a promising model for satellite EO, offering the necessary balance between performance and resource efficiency for real-time processing in space environments. These factors make

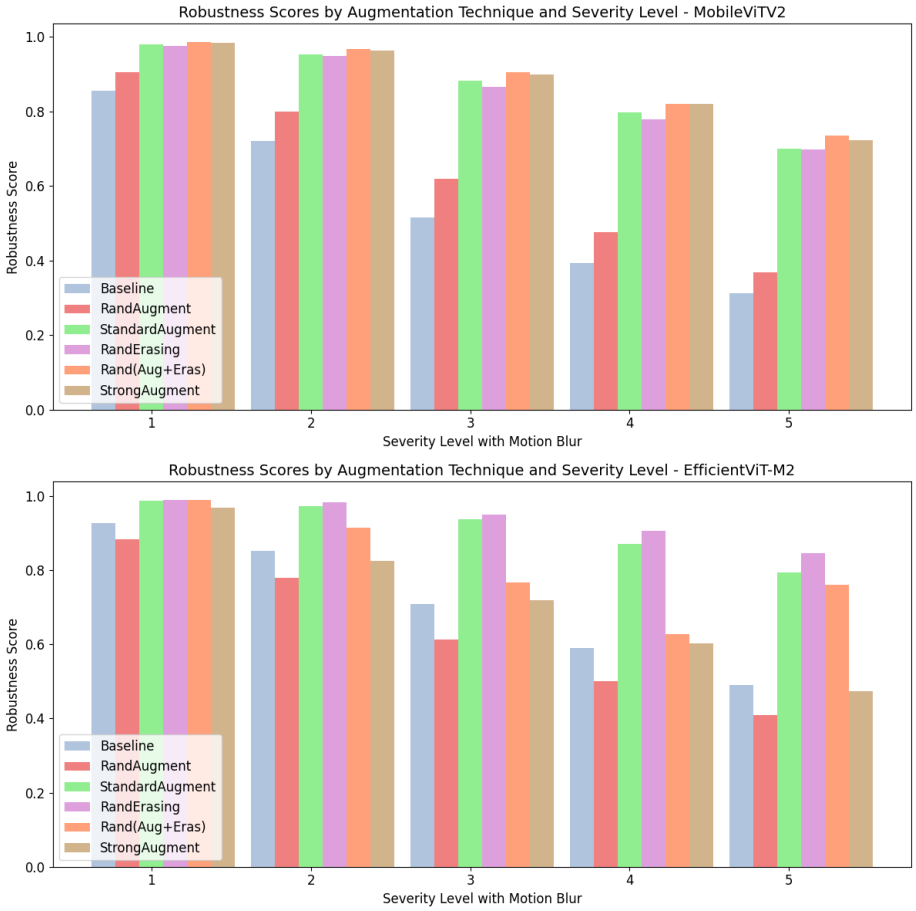


Fig. 8: MobileViT (Top) and EfficientViT (Bottom) robustness with motion blur noisy inference data

EfficientViT-M2 a powerful model for IC and a practical and feasible option for real-world deployment in resource-constrained environments like satellite missions.

Lastly, the robustness analysis was extended beyond synthetic perturbations by subjecting EfficientViT-M2 to an end-to-end satellite communications (SatCom) workflow that emulates the DVB-S2(X) downlink with realistic channel coding, RF impairments, and compression losses. EfficientViT demonstrated highly efficient and robust performance for EO-IC under realistic SatCom conditions, explicitly using the DVB-S2(X) transmission standard, incorporating typical channel impairments. By systematically evaluating EfficientViT across varying image compression levels and practical channel SNR-to-Shannon-capacity ratios, the model consistently achieved high classification accuracy, remaining above 97% under optimal transmission scenarios and degrading gradually as source and channel loss increased. The consistently high accuracy from empirical results across the full SatCom operating, modeled by an exponential curve fitting framework, reveals that EfficientViT-M2 maintains strong task-oriented effectiveness and reliability for EO applications in bandwidth and noise-constrained SatCom environments, validating its practicality for real-world deployments [55].

## VI. Conclusions

In this paper, we study and conduct extensive experiments comparing traditional CNN-based, ResNet-based, and Transformer-based models for IC. Our analysis delves into various pre-trained ViT architectures to identify the optimal balance between model complexity, computational efficiency, and performance. A key focus of our evaluation is assessing the robustness of these models under noisy data conditions during inference, which closely mirrors real-world scenarios encountered in satellite EO systems.

In conclusion, the EfficientViT-M2 model is optimal for implementing IC in onboard satellite EO missions. Its superior robustness to noise, demonstrated across varying levels of Gaussian noise and motion blur, ensures reliable performance in the challenging conditions often encountered in space. Additionally, EfficientViT-M2's low computational complexity (203.53 MFLOPs) and compact size (38.19 MB) make it highly suitable for resource-constrained environments, offering significant efficiency gains over pioneering models used in previous satellite missions like Φ-Sat-1 and Φ-Sat-2. The model's ability to maintain high accuracy and classification performance with minimal computational overhead further cements its practicality for real-world deployment, where consistent and accurate image processing is critical. As satellite EO demands increasingly sophisticated yet efficient algorithms, EfficientViT-M2 stands out as a leading solution, combining advanced ML capabilities with the feasibility of successful on-board implementation.

One limitation of this study is the challenge of adapting deep NNs to the dynamic nature of wireless communication

systems. Frequent changes in the environment and data distribution can necessitate on-device retraining, adding complexity. Future work should address this by focusing on managing lossy and dynamic transmission errors through practical communication system modeling.

Another limitation is the study's exclusive focus on IC, without exploring the potential of multitask learning, as exemplified by the Multi-Tasks Pre-trained Model (MTP) [56]. Although the EfficientViT achieves slightly lower accuracy compared to the MTP model's 99.30%, the MTP model's incorporation of both CNNs and Transformers results in over 300 million parameters and significantly higher computational complexity, limiting its practicality in resource-constrained environments like OSS. Future research should investigate the EfficientViT application in multitasking scenarios for computational efficiency balancing.

Finally, the study is limited by its focus on single-modality data, potentially restricting its ability to leverage the expanding wealth of multimodal EO data. Future research should explore the Transformer-based approaches within a multimodal DL framework to utilize better and integrate diverse types of EO data [57], [58]. Diffusion models also warrant investigation, as their iterative denoising process can generate high-fidelity, cross-modal representations that boost downstream performance [59]. Moreover, lightweight state-space sequence models offer a lower-complexity alternative to standard Transformers while still capturing long-range dependencies [60], making them well-suited to resource-constrained EO deployments.

## REFERENCES


[1] S. Sadek, "New satellite market forecast anticipates 1,700 satellites to be launched on average per year by 2030 as new entrants and incumbents increase their investment in space," [Online]. Available: https://shorturl.at/6PVma, 2021.

[2] Y. Boualleg, M. Farah, and I. R. Farah, "Remote sensing scene classification using convolutional features and deep forest classifier," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1944–1948, 2019.

[3] C. Xu, G. Zhu, and J. Shu, "A lightweight and robust lie group-convolutional neural networks joint representation for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.

[4] W. Li, *et al.*, "Classification of high-spatial-resolution remote sensing scenes method using transfer learning and deep convolutional neural network," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 1986–1995, 2020.

[5] K. Xu, P. Deng, and H. Huang, "Vision transformer: An excellent teacher for guiding small networks in remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[6] X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, "Vision-language models in remote sensing: Current progress and future trends," *IEEE Geosci. Remote Sens. Mag.*, 2024.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *North American Associ. Comput. Linguistics: Human Language Technol. (NAACL-HLT)*. ACL, 2019, pp. 4171–4186.

[8] S. Khan, *et al.*, "Transformers in vision: A survey," *ACM Comput. Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[9] Y. Liu, *et al.*, "A survey of visual transformers," *IEEE Trans. Neural Netw. Learning Systems*, 2023.

[10] X. Han, *et al.*, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.

[11] H. Al-Hraishawi, H. Chougrani, S. Kisseleff, E. Lagunas, and S. Chatzinotas, "A survey on nongeostationary satellite systems: The communication perspective," *IEEE Commun. Surveys & Tuts.*, vol. 25, no. 1, pp. 101–132, 2022.

[12] H. Chougrani, *et al.*, "Connecting space missions through ngso constellations: feasibility study," *Frontiers in Commun. Netw.*, vol. 5, p. 1356484, 2024.

[13] G. Fontanesi, *et al.*, "Artificial intelligence for satellite communication and non-terrestrial networks: A survey," *arXiv preprint arXiv:2304.13008*, 2023.

[14] G. Giuffrida, *et al.*, "The $\phi$-sat-1 mission: The first on-board deep neural network demonstrator for satellite earth observation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[15] G. Guerrisi, F. Del Frate, and G. Schiavon, "Artificial intelligence based on-board image compression for the $\phi$-sat-2 mission," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2023.

[16] Y. Pang, *et al.*, "Repsvit: An efficient vision transformer based on spiking neural networks for object recognition in satellite on-orbit remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[17] E. Lagunas, *et al.*, "Performance evaluation of neuromorphic hardware for onboard satellite communication applications," *arXiv preprint arXiv:2401.06911*, 2024.

[18] F. Ortiz *et al.*, "Energy-efficient on-board radio resource management for satellite communications via neuromorphic computing," *IEEE Trans. Machine Learning Commun. Network.*, 2024.

[19] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Trans. Pattern Analysis Machine Intell.*, 2024.

[20] S. Mehta and M. Rastegari, "Separable self-attention for mobile vision transformers," *arXiv preprint arXiv:2206.02680*, 2022.

[21] X. Liu, *et al.*, "Efficientvit: Memory efficient vision transformer with cascaded group attention," in *IEEE Conf. Comput. Vision Pattern Recognition*, 2023, pp. 14 420–14 430.

[22] H. M. Albarakati, *et al.*, "A novel deep learning architecture for agriculture land cover and land use classification from remote sensing images based on network-level fusion of self-attention architecture," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2024.

[23] S. Rubab, *et al.*, "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 J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2024.

[24] Z. Li, *et al.*, "Mgc: Mlp-guided cnn pre-training using a small-scale dataset for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, 2024.

[25] B. Zhao, B. Huang, and Y. Zhong, "Transfer learning with fully pretrained deep convolution networks for land-use classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1436–1440, 2017.

[26] S. Bai, W. Zhou, Z. Luan, D. Wang, and B. Chen, "Improving cross-domain few-shot classification with multilayer perceptron," in *IEEE Int. Conf. Acoustics, Speech Signal Processing (ICASSP)*. IEEE, 2024, pp. 5250–5254.

[27] A. C. Depoian, C. P. Bailey, and P. Guturu, "Land use classification efficient vision transformer," in *IEEE Int. Geosci. Remote Sens. Symposium (IGARSS)*. IEEE, 2023, pp. 2922–2925.

[28] L. Fan, *et al.*, "Scaling laws of synthetic images for model training... for now," in *IEEE/CVF Conf. Comput. Vision Pattern Recognition (CVPR)*, 2024, pp. 7382–7392.

[29] Z. Huang, M. Zhang, Y. Gong, Q. Liu, and Y. Wang, "Generic knowledge boosted pre-training for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, 2024.

[30] S. A. Yamashkin, A. A. Yamashkin, V. V. Zanozin, M. M. Radovanovic, and A. N. Barmin, "Improving the efficiency of deep learning methods in remote sensing data analysis: geosystem approach," *IEEE Access*, vol. 8, pp. 179 516–179 529, 2020.

[31] J. Yu, *et al.*, "Boosting continual learning of vision-language models via mixture-of-experts adapters," in *IEEE/CVF Conf. Comput. Vision Pattern Recognition (CVPR)*, 2024, pp. 23 219–23 230.

[32] A. Temenos, N. Temenos, M. Kaselimi, A. Doulamis, and N. Doulamis, "Interpretable deep learning framework for land use and land cover classification in remote sensing using shap," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[33] 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. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, 2019.

[34] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 197–209, 2018.

[35] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Informat. Process. Systems*, vol. 34, pp. 12 116–12 128, 2021.

[36] A. Hassani, *et al.*, "Escaping the big data paradigm with compact transformers," *arXiv preprint arXiv:2104.05704*, 2021.

[37] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," *arXiv preprint arXiv:2112.13492*, 2021.

[38] B. Alkin, M. Beck, K. Pöppel, S. Hochreiter, and J. Brandstetter, "Vision-lstm: xlstm as generic vision backbone," *arXiv preprint arXiv:2406.04303*, 2024.

[39] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Int. Conf. Machine Learning (ICML)*. PMLR, 2019, pp. 6105–6114.

[40] M. Goldblum, *et al.*, "Battle of the backbones: A large-scale comparison of pretrained models across comput. vision tasks," *Advances in Neural Informat. Process. Systems*, vol. 36, 2024.

[41] H. Cai, J. Li, M. Hu, C. Gan, and S. Han, "Efficientvit: Multi-scale linear attention for high-resolution dense prediction," *arXiv preprint arXiv:2205.14756*, 2022.

[42] Z. Liu, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, 2021, pp. 10 012–10 022.

[43] A. Dosovitskiy, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learning Represent. (ICLR)*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[44] C. Goutte and et. al., "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *European Conf. Informat. Retrieval*. Springer, 2005, pp. 345–359.

[45] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Int. Conf. Learning Represent., (ICLR)*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=HJz6tiCqYm

[46] A. Laugros, A. Caplier, and M. Ospici, "Are adversarial robustness and common perturbation robustness independent attributes?" in *IEEE/CVF Int. Conf. Comput. Vision Workshops (ICCV)*, 2019, pp. 0–0.

[47] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *Int. Conf. Machine Learning*. PMLR, 2016, pp. 1225–1234.

[48] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Int. Conf. Learning Represent.*

[49] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Advances in Neural Informat. Process. Systems (NeurIPS)*, 2020.

[50] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *AAAI Conf. Artificial Intell.*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[51] Y. Xu, *et al.*, "Attention-based contrastive learning for few-shot remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, 2024.

[52] S. Rubab, *et al.*, "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 J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2024.

[53] P. P. Tumpa and M. S. Islam, "Lightweight parallel convolutional neural networkwith svm classifier for satellite imagery classification," *IEEE Trans. Artificial Intell.*, 2024.

[54] C. Peng, Y. Li, L. Jiao, and R. Shang, "Efficient convolutional neural architecture search for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 6092–6105, 2020.

[55] T. T. Nguyen, *et al.*, "A semantic-loss function modeling framework with task-oriented machine learning perspectives," *arXiv preprint arXiv:2503.09903*, 2025.

[56] D. Wang, *et al.*, "Mtp: Advancing remote sensing foundation model via multi-task pretraining," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2024.

[57] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[58] S. K. Roy, A. Sukul, A. Jamali, J. M. Haut, and P. Ghamisi, "Cross hyperspectral and lidar attention transformer: An extended self-attention for land use and land cover classification," *IEEE Trans. Geosci. Remote Sens.*, 2024.

[59] D. Tang, *et al.*, "Aerogen: enhancing remote sensing object detection with diffusion-driven data generation," *arXiv preprint arXiv:2411.15497*, 2024.

[60] J. Yao, D. Hong, C. Li, and J. Chanussot, "Spectralmamba: Efficient mamba for hyperspectral image classification," *arXiv preprint arXiv:2404.08489*, 2024.