On-board Satellite Image Classification for Earth Observation: A Comparative Study of Pre-trained Vision-Transformer Models

Thanh-Dung Le, Vu Nguyen Ha, Ti Ti Nguyen, Geoffrey Eappen, Prabhu Thiruvasagam, Luis M. Garces-Socarras, Hong-fu Chou, Jorge L. Gonzalez-Rios, Juan Carlos Merlano-Duncan, Symeon Chatzinotas

***Abstract*—Remote sensing image classification 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, offering enhanced performance and efficiency. This study focuses on identifying the most effective pre-trained model for land use classification for onboard-satellite (On-Air) processing, emphasizing achieving high accuracy, computational efficiency, and robustness against noisy data—conditions commonly encountered during On-Air inference. Through extensive experimentation, we compared traditional CNN-based models, ResNet-based models, and various pre-trained vision Transformer models. Our findings demonstrate that pre-trained Transformer models, particularly MobileViTV2 and EfficientViT-M2, outperform models trained from scratch in accuracy and efficiency. These models achieve high performance with reduced computational requirements and exhibit greater resilience during inference under noisy conditions. While MobileViTV2 excelled on clean validation data, EfficientViT-M2 proved more robust when handling noise, making it the most suitable model for EO tasks. In conclusion, EfficientViT-M2 is the optimal choice for reliable and efficient On-Air remote-sensing image classification, achieving 98.76% accuracy, precision, and recall. Specifically, EfficientViT-M2 delivered the highest performance across all metrics, excelled in training efficiency (1,000s) and inference time (10s), and demonstrated greater robustness (overall robustness score at 0.79).**

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

I. Introduction

Remote sensing image classification (RSIC) has long been a critical application in EO systems, with traditional methods primarily relying on CNNs and other DL techniques. These approaches have been well-documented and validated over the years, demonstrating significant capabilities in processing and classifying remote sensing (RS) data with reasonable accuracy and efficiency [1]–[3]. However, as the field has evolved, so too have the methods, with recent advances emphasizing the use of Transformer-based architectures, which have begun to surpass traditional CNNs in many image classification (IC) tasks, including RS [4], [5].

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 (NLP) with models like BERT [6] and GPT [7]. This paradigm shift is now extending into computer vision, where pre-trained vision Transformer (ViT) models are proving to be highly effective [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].

In RS, while pre-trained models, especially pre-trained ViT models, hold great potential for enhancing classification performance and reducing computational demands during inference, their application on board satellites is constrained due to the significant power consumption required. Implementing such large models onboard is particularly challenging, as seen with the pioneering use of neural networks in satellite-based EO systems. For instance, the Φ-Sat-1 mission employed the CloudScout model, a CNN-based neural network, for image segmentation, marking the first instance of On-Air DL [11]. Similarly, the Φ-Sat-2 mission utilized a convolutional autoencoder for image compression, further illustrating the practical challenges of deploying computationally intensive models in space environments [12]. One helpful method to enhance computational efficiency in such downstream tasks is transfer learning, where a pre-trained model is fine-tuned for a specific application, such as land use classification. This approach has shown consistent improvements in performance across various tasks, as it leverages the pre-learned features from large-scale datasets, requiring only minimal adaptation to the new task [13].

This study aims to identify the most effective pre-trained model for RSIC, explicitly focusing on land use classification applications with On-Air processing. 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 instrument imperfections. As shown in Fig. 1, 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 to On-Air inference. In this setup, the EO satellite records sensor data, which is then preprocessed by the onboard processor, eliminating duplicates and preparing the data for level 0/1 processing. The preprocessed data is subsequently transmitted

Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg, Luxembourg

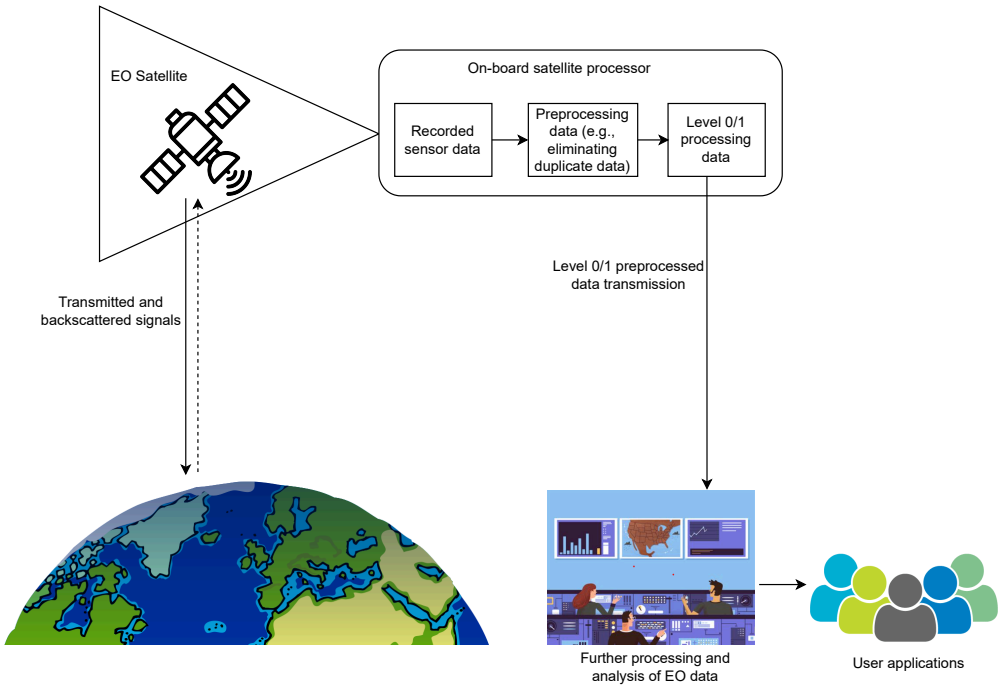

Fig. 1: EO system architecture in which the recorded EO data is preprocessed and inference at the satellite (Fig from Prabhu). **Redrwaing the Figure, removing the level of pre-processing**

back to Earth, where further processing and analysis are conducted, ultimately supporting various user applications. By leveraging the trained model, we aim to optimize the classification performance directly within the satellite, ensuring efficient and accurate land use classification in real-time.

To achieve this, we conducted 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 complexity, computational efficiency, and performance. A key aspect of our evaluation is the robustness of these models when faced with noisy data during inference, as this scenario closely mirrors the real-world conditions of satellite EO systems.

Our findings indicate that pre-trained vision Transformer models significantly outperform models trained from scratch, particularly in efficiency and accuracy. Among the models tested, MobileViTV2 and EfficientViT-M2 emerged as the top performers for On-Air EO processing. These models deliver high performance with lower computational demands and exhibit greater robustness in the presence of noisy test data. While MobileViTV2 achieved higher validation performance on clean data, EfficientViT-M2 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 On-Air EO processing tasks 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 RSIC in real-world satellite operations.

II. Related Works

Recent RS advancements have focused on enhancing CNN models through transfer learning, structural modifications, and integrating attention mechanisms. For instance, [14] proposed a self-attention fused CNN architecture optimized explicitly for land use and land cover classification, incorporating advanced data augmentation and custom-designed CNN models to improve classification performance. Similarly, [15] 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 RSI community. To address this, [16] proposed the MGC method, which employs an 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 [17], which utilizes fully pre-trained deep convolutional networks 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 IMP, 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, [18] demonstrated that MLPs could significantly enhance discriminative capabilities and alleviate distribution shifts, thereby improving classification performance in RS tasks. Integrating ViT models and advanced transfer learning techniques has become increasingly popular as the field progresses. For instance, [19] 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.

Some studies have focused on generating synthetic data or integrating multiple data sources to maximize data utility further. For example, [20] 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, [21] 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. [22] proposed 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, [23] introduced 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, recent research has emphasized the importance of interpretability in RS models. [24] developed 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 RSIC, incorporating Vision Transformers, advanced transfer learning methods, and synthetic data generation 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 On-Air EO processing, ensuring high performance and reliability in challenging operational environments.

III. Methodology

A. Dataset

Public Benchmark for Land Use and Land Cover Classification Dataset [25]. The EuroSAT dataset is a large-scale benchmark dataset designed explicitly for land use and land cover classification derived from Sentinel-2 satellite imagery. It comprises 27,000 geo-referenced labeled images, each measuring 64x64 pixels and spanning 13 spectral bands. The dataset is categorized into 10 classes. Each class contains 2000–3000 images, including industrial buildings, residential buildings, annual crops, permanent crops, rivers, seas and lakes, Herbaceous vegetation, highways, pastures, and forests across Europe. Due to its compact image size and diverse class representation, EuroSAT is particularly well-suited for developing and evaluating DL models intended for onboard-processing EO missions. This makes it a valuable resource for applications such as real-time environmental monitoring, disaster response, and precision agriculture, where in-situ processing capabilities are crucial for timely decision-making.

B. Machine Learning Models

In this study, we utilize a diverse set of machine learning (ML) models, including both models trained from scratch and pre-trained models, to evaluate their performance on IC tasks. The models range from traditional convolutional neural networks (CNNs) to advanced architectures like Vision Transformers (ViTs) and hybrid models that combine convolutional layers with transformer-based processing. This selection allows us to explore the effectiveness of different architectures, from lightweight and efficient models to more complex, high-capacity networks, in handling IC challenges.

1) Training from Scratch:

- **CNN:** A basic convolutional neural network 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:** 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) [26]:** 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 [27]:** A Vision Transformer 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 [28]:** An efficient vision transformer combining convolutional layers with local window attention mechanisms optimized for balancing performance and computational efficiency with approximately 4 million parameters.
- **MobileViTV2 [29]:** A hybrid model leveraging convolutional and transformer-based processing, utilizing depthwise separable convolutions and self-attention mechanisms for accurate and efficient IC.
- **xLSTM [30]:** Integrates convolutional patch embedding with advanced LSTM-based layers to capture spatial and sequential dependencies for comprehensive IC.
- **EfficientNet-B2 [31]:** A highly efficient convolutional neural network using depthwise separable convolutions and squeeze-and-excitation layers to minimize parameter usage while maximizing performance.
- **ResNet50-DINO [32]:** A ResNet-50 model trained using self-supervised learning with DINO, optimized for feature representation without labeled data, enabling robust classification performance.
- **EfficientViT-L2 [33]:** Combines convolutional and transformer-based architectures with techniques like fused MBConv and Lite Multi-Head Attention to balance efficiency and representational power effectively.
- **SwinTransformer [34]:** A hierarchical vision transformer that divides images into non-overlapping patches, utilizing shifted window-based self-attention to capture both local and global context.
- **Vision Transformer (ViT) [35]:** 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) [36]. These metrics are calculated for each class individually and then aggregated using macro-averaging as follows,

$$\text{Accuracy (Acc)} = \frac{\sum_{i=1}^{N} \text{TP}_i}{\sum_{i=1}^{N} (\text{TP}_i + \text{FP}_i + \text{FN}_i)}$$

$$\text{Precision} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}$$

$$\text{Recall} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}$$

By using macro-averaging, 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 neural network robustness to common corruptions and perturbations [37]. The robustness score measures the degradation in accuracy caused by a specific perturbation, ϕ. For a given model, m, let A_{clean} represent the model's accuracy on the original (clean) test dataset, and A_ϕ denote the accuracy of the model on the test set that has been modified with perturbation ϕ. These perturbations can stem from natural environmental conditions encountered during deployment or adversarial modifications of the original test samples. Following the methodology outlined in [38], we define the robustness score of model f with respect to perturbation ϕ as follows:

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

A robustness score close to one indicates that the model is highly resistant to perturbation, demonstrating its ability to maintain performance in challenging conditions.

IV. Experimental Setup

The dataset was split into training and testing sets with a 70/30 ratio to ensure a balanced and comprehensive evaluation of the model's performance.

For model optimization, different strategies were employed based on the model architecture. The CNN and ResNet models were optimized using the Stochastic Gradient Descent (SGD) optimizer [39], 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, were optimized using the AdamW optimizer [40]. 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 (γ) of 0.1.

The Compact Convolutional Transformer (CCT) [26] and SmallViT models [27] were trained using code adapted from the ViT-PyTorch repository, which provides reliable and optimized implementations for Vision Transformers. Other pre-trained models were 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

TABLE I: Comparison of Machine Learning Models

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

for a fair and direct comparison of their performance across the different experiments. [1]

All experiments were 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 were trained and tested using an Nvidia RTX A5000 GPU with 48 GB of GDDR6 memory, providing sufficient computational resources for most 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 were necessary for the two larger pre-trained Vision Transformer 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 was 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.

After identifying the best model with lower computational complexity and high performance, we further assess their robustness during inference onboard satellites. This analysis evaluates the models' performance under noisy conditions, which is common in real-world satellite operations. 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 the image. This evaluation helps us understand how well the models can perform under varying noise levels, which is critical for reliable satellite-based IC, as shown in Fig. 2. The noise and severity level was implemented

[1] 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

based on the code [2] from [37].

To further assess the robustness and generalization capabilities of the selected models, we applied a series of data augmentation techniques, as illustrated in Fig. 3. The augmentations include Standard Transformed, Rand Augment [41], Random Erasing [42], 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, which mimic real-world variations that the models may encounter during satellite IC tasks. 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

The heatmap shown in Table II Fig. 4 provides a detailed comparison of the computational complexity of various models, with a particular emphasis on total parameters, estimated total size, FLOPs, training time, and inference time, all displayed on a logarithmic scale. 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 Vision Transformer 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 are crucial. 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. 5 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 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, 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 vision transformer architecture. Unlike conventional Vision Transformers (ViT), which rely on a global attention mechanism across the entire image, the SwinTransformer 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. **Explain why SwinTransformer is better than ViT**

From the results presented in Table II, and III, Fig. 4, and 5, we confirm that MobileViT and EfficientViT are the two best models for IC. These models excel due to their lower computational complexity, shorter training and inference times, and superior performance across all evaluation metrics. 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 previous experiments. 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. 6 shows the confusion matrices for EfficientViT-M2 and MobileViTV2, highlighting their exceptional performance in IC. Both models exhibit vital accuracy across all classes, with most predictions aligning closely with the true labels. In the confusion matrix for EfficientViT-

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

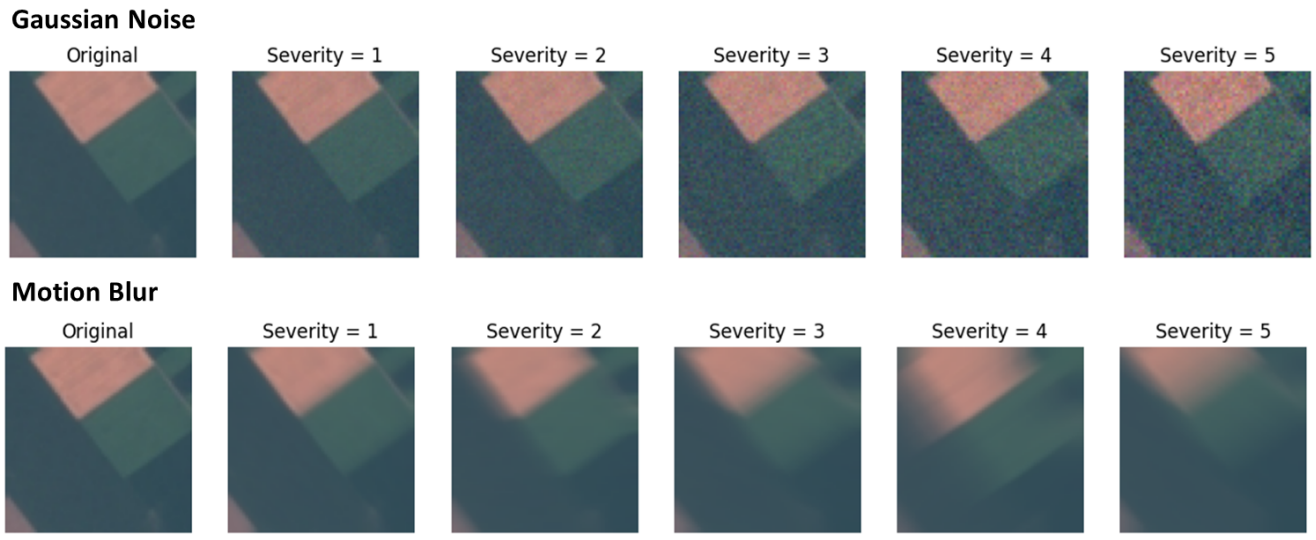

Fig. 2: Different noise levels.

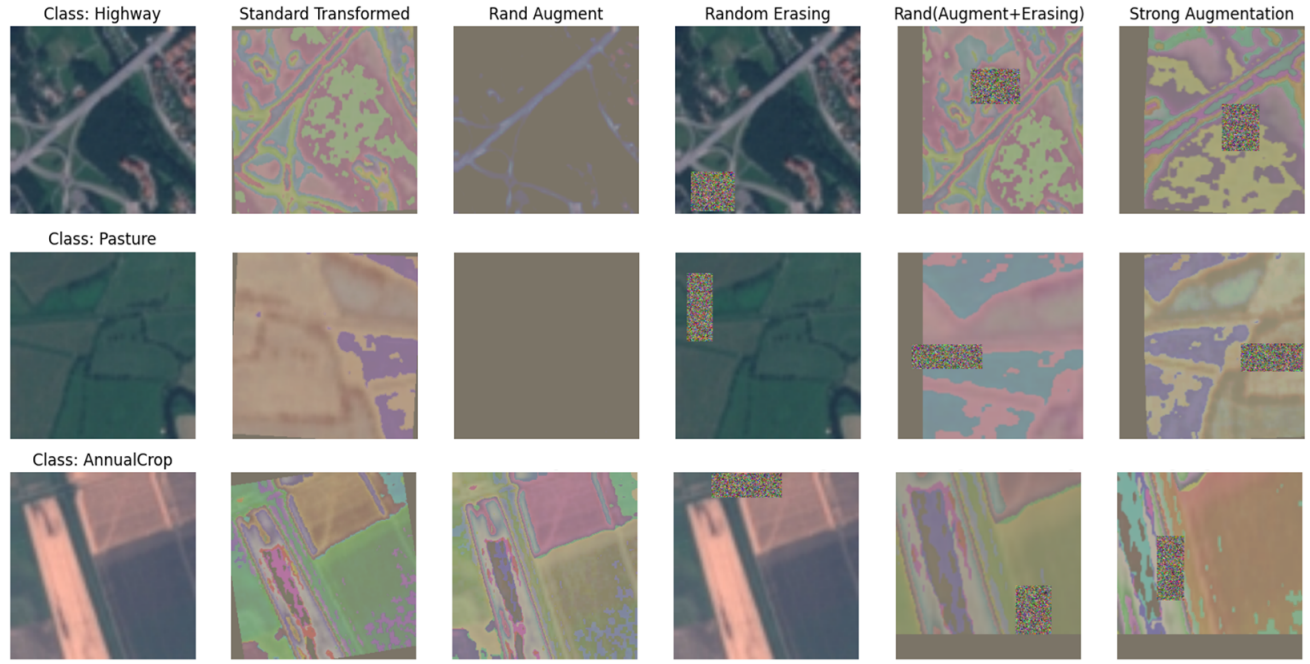

Fig. 3: Different augmentation techniques.

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. **Explain why EfficientViT and MobileViT are better than large ViT-based models**

The experiment results, as presented in Tables V and VI, reveal that applying data augmentation techniques during training and inference generally leads to a slight reduction in model performance when evaluated on clean test data. MobileViTV2 and EfficientViT-M2's baseline performance—achieved without augmentation—yields 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

TABLE II: Summarization of model computation complexity, estimated total size, training, and inference times

Models	Input Size	Total Parameters	Estimated Total 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		
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

Fig. 4: Model complexity comparison.

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),

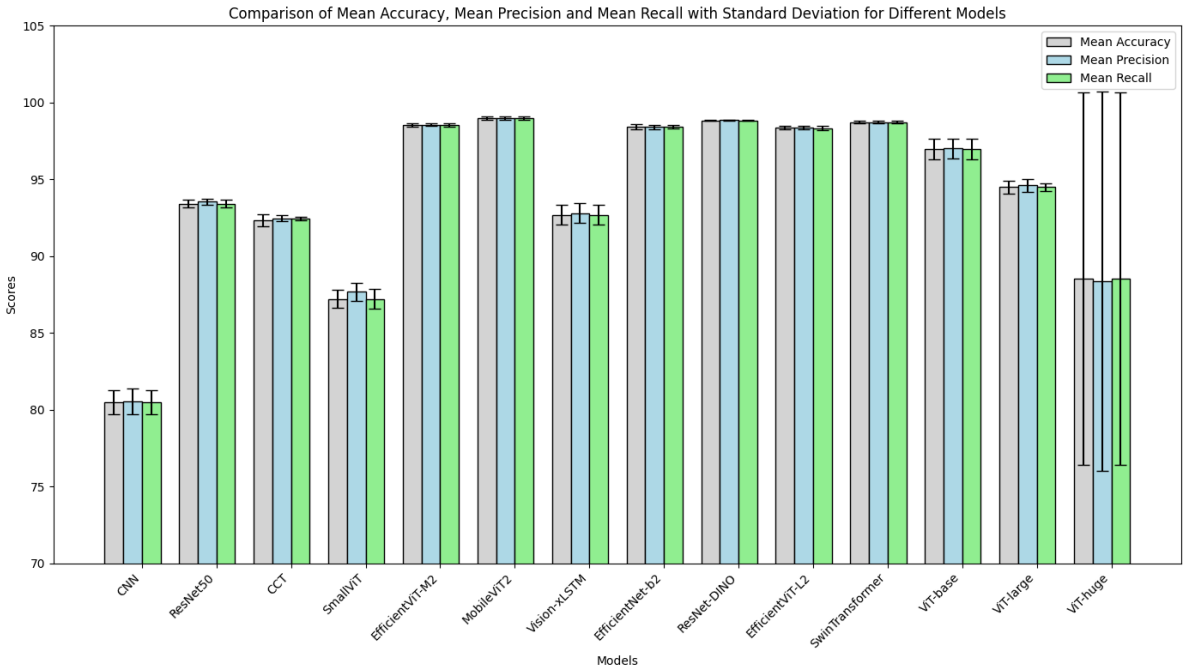

Fig. 5: Statistical comparison for model performance.

TABLE III: Experiment results for the best model's performance

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.

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 severity 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 perfor-

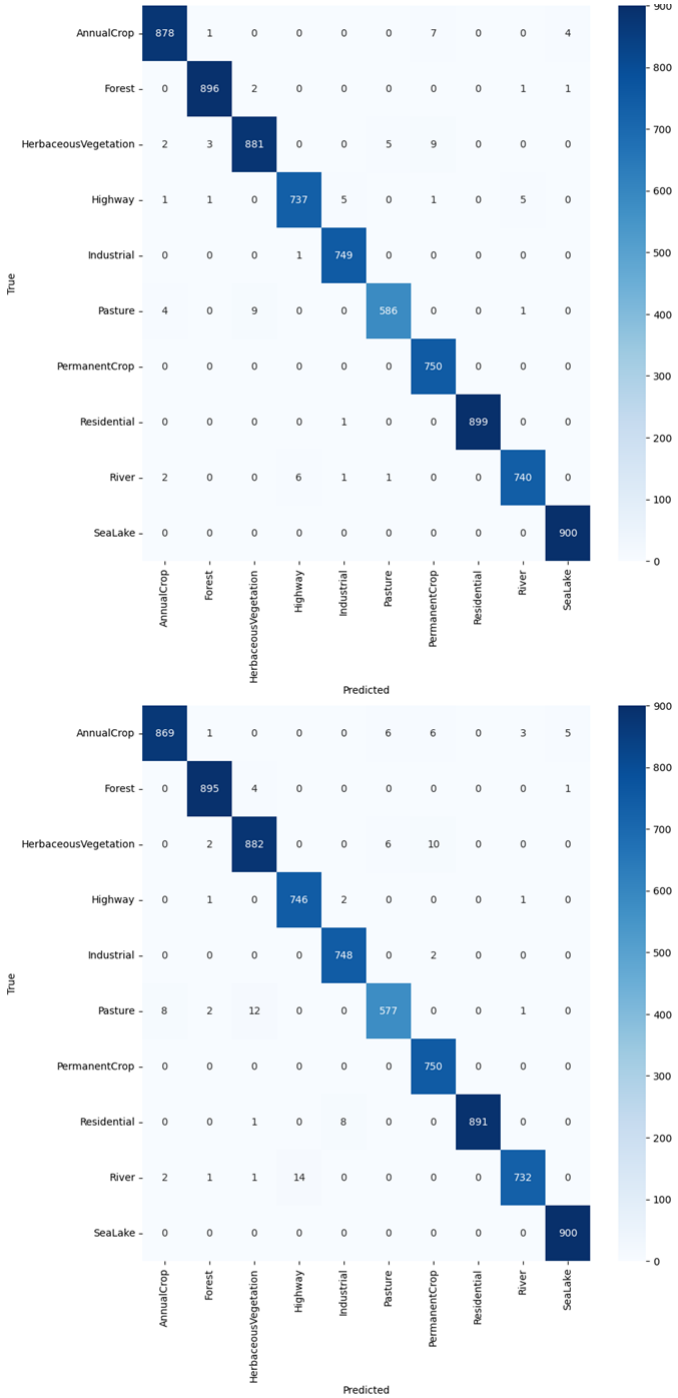

Fig. 6: Confusion matrix from MobileViTV2 (top) and EfficientViT-M2 (bottom) performance.

mance than MobileViTV2. While MobileViTV2 remains robust at lower severity levels, it experiences a more pronounced decline in performance as severity increases, particularly with 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-

TABLE IV: Performance comparison with the-state-of-the-art models

Models	Accuracy	Precision	Recall
EfficientViT-M2	*98.76*	*98.77*	*98.76*
MobileViTV2	**99.09**	**99.09**	**99.09**
Few-shot MLP [18]	79.62	x	x
Attention+CNN [14]	89.5	x	x
ACL [43]	95.46	x	x
MGC [16]	96.41	x	x
GeRSP [21]	97.87	x	x
Constrastive Learning [20]	96	x	x
GeoSystemNet [22]	95.32	x	x
Transformer+CNN [19]	95.48	x	x
MoE-ViT [23]	98.1	x	x
Self-Attention CNN [15]	90.3	89.05	88.59
SIBNet [44]	97.8	97	96.97
CNN-SHAP [24]	94.72	93.73	94.10

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

TABLE V: 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 VI: 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.

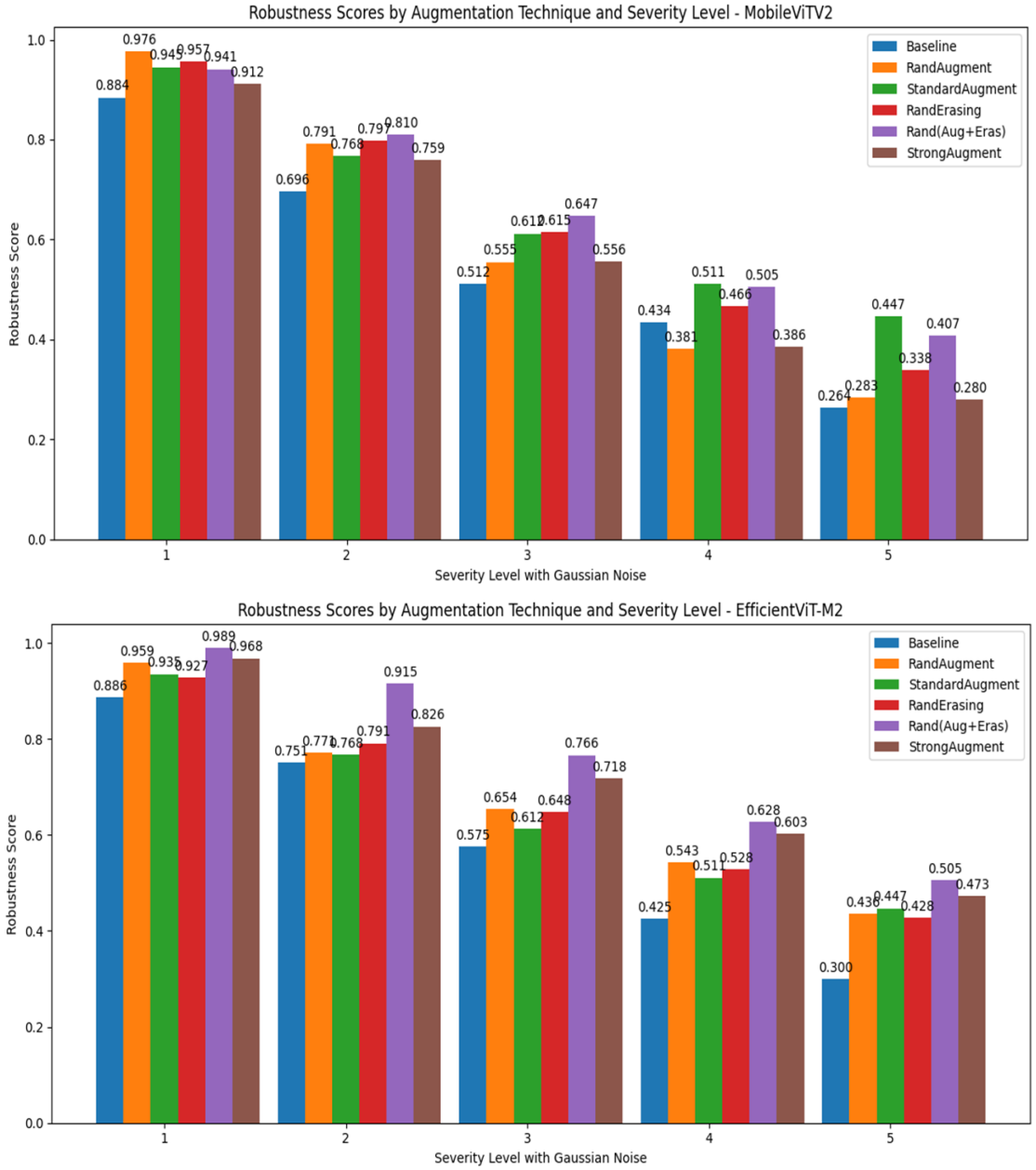

Fig. 7: MobileViT (Top) and EfficientViT (Bottom) robustness with Gaussian noisy inference data

M2 model in practical applications, mainly onboard satellite systems, 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 [11], 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 [12], 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. These factors make EfficientViT-M2 a powerful model for IC and a practical and feasible option for real-world deployment in resource-constrained environments like satellite missions.

VI. Conclusions

In conclusion, the EfficientViT-M2 model is optimal for implementing IC in On-Air EO processing missions. Its superior robustness to noise, demonstrated across varying

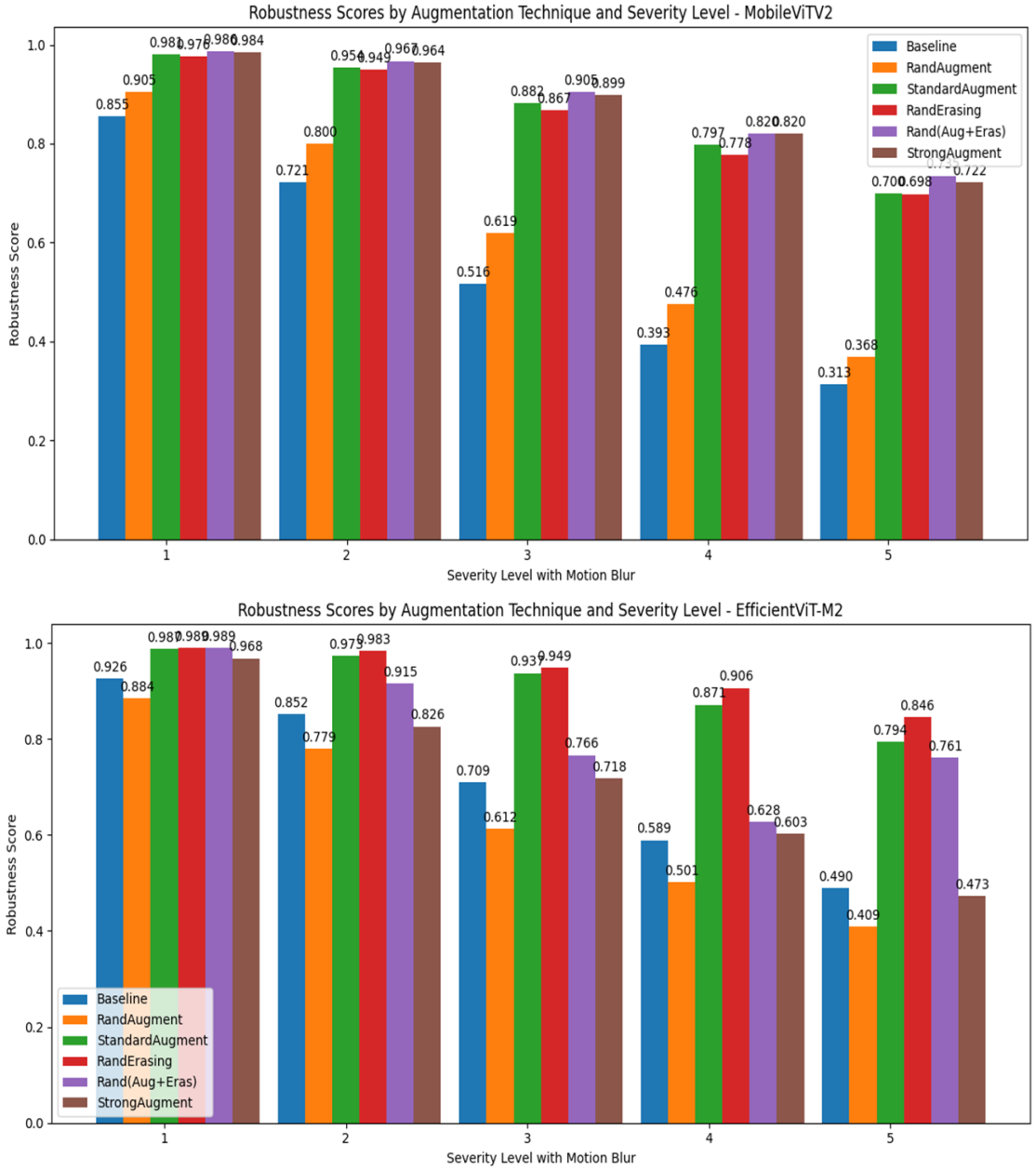

Fig. 8: MobileViT (Top) and EfficientViT (Bottom) robustness with motion blur noisy inference data

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.

Limitations and Future Works:

One limitation of this study is the challenge of adapting deep neural networks 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) [45]. 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 on-board satellite systems. Future research should investigate the application of EfficientViT in multi-tasking scenarios to balance performance with computational efficiency.

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 potential of transformer-based approaches within a multi-modal DL framework to utilize better and integrate diverse types of EO data, as demonstrated by [46], [47].

ACKNOWLEDGMENT

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

REFERENCES

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

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

[3] W. Li, Z. Wang, Y. Wang, J. Wu, J. Wang, Y. Jia, and G. Gui, "Classification of high-spatial-resolution remote sensing scenes method using transfer learning and deep convolutional neural network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1986–1995, 2020.

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

[5] X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, "Vision-language models in remote sensing: Current progress and future trends," *IEEE Geoscience and Remote Sensing Magazine*, 2024.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[8] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[9] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[10] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, *et al.*, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.

[11] G. Giuffrida, L. Fanucci, G. Meoni, M. Batič, L. Buckley, A. Dunne, C. Van Dijk, M. Esposito, J. Hefele, N. Vercruyssen, *et al.*, "The ϕ-sat-1 mission: The first on-board deep neural network demonstrator for satellite earth observation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[12] G. Guerrisi, F. Del Frate, and G. Schiavon, "Artificial intelligence based on-board image compression for the ϕ-sat-2 mission," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

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

[14] H. M. Albarakati, M. A. Khan, A. Hamza, F. Khan, N. Kraiem, L. Jamel, L. Almuqren, and R. Alroobaea, "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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[15] S. Rubab, M. A. Khan, A. Hamza, H. M. Albarakati, O. Saidani, A. Alshardan, A. Alasiry, M. Marzougui, and Y. Nam, "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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[16] Z. Li, B. Hou, W. Li, Z. Wu, B. Ren, and L. Jiao, "Mgc: Mlp-guided cnn pre-training using a small-scale dataset for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

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

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

[19] A. C. Depoian, C. P. Bailey, and P. Guturu, "Land use classification efficient vision transformer," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 2922–2925.

[20] L. Fan, K. Chen, D. Krishnan, D. Katabi, P. Isola, and Y. Tian, "Scaling laws of synthetic images for model training... for now," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7382–7392.

[21] Z. Huang, M. Zhang, Y. Gong, Q. Liu, and Y. Wang, "Generic knowledge boosted pre-training for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[22] 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.

[23] J. Yu, Y. Zhuge, L. Zhang, P. Hu, D. Wang, H. Lu, and Y. He, "Boosting continual learning of vision-language models via mixture-of-experts adapters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 219–23 230.

[24] 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 Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[25] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.

[26] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," *arXiv preprint arXiv:2104.05704*, 2021.

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

[28] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "Efficientvit: Memory efficient vision transformer with cascaded group attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 420–14 430.

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

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

[31] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[32] M. Goldblum, H. Souri, R. Ni, M. Shu, V. Prabhu, G. Somepalli, P. Chattopadhyay, M. Ibrahim, A. Bardes, J. Hoffman, *et al.*, "Battle of the backbones: A large-scale comparison of pretrained models across

computer vision tasks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[33] 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.

[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

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

[37] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=HJz6tiCqYm

[38] A. Laugros, A. Caplier, and M. Ospici, "Are adversarial robustness and common perturbation robustness independent attributes?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[39] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *International conference on machine learning*. PMLR, 2016, pp. 1225–1234.

[40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*.

[41] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html

[42] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[43] Y. Xu, H. Bi, H. Yu, W. Lu, P. Li, X. Li, and X. Sun, "Attention-based contrastive learning for few-shot remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[44] S. Rubab, M. A. Khan, A. Hamza, H. M. Albarakati, O. Saidani, A. Alshardan, A. Alasiry, M. Marzougui, and Y. Nam, "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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[45] D. Wang, J. Zhang, M. Xu, L. Liu, D. Wang, E. Gao, C. Han, H. Guo, B. Du, D. Tao, *et al.*, "Mtp: Advancing remote sensing foundation model via multi-task pretraining," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[46] 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 Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[47] 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 Transactions on Geoscience and Remote Sensing*, 2024.