Towards Lightweight Vision Models for Ecological Earth Observation

Patrick Vincent Ndowo and Tushar Shinde
IIT Madras Zanzibar
shinde@iitmz.ac.in

Abstract

Efficient deployment of deep learning models for real-time Earth observation is critical for ecological monitoring, biodiversity conservation, and sustainable land-use management. However, resource constraints on satellites, UAVs, and edge devices pose significant challenges due to limited compute, memory, and bandwidth. This work systematically investigates model compression strategies, including knowledge distillation (KD), pruning, and quantization-aware training (QAT), for satellite image classification tasks central to ecological applications such as habitat mapping and land-cover change detection. We conduct experiments on the EuroSAT and UC Merced datasets, analyzing trade-offs between model size, accuracy, and inference efficiency. Our results show that compact CNNs, such as MobileNetV2 trained via distillation, achieve competitive accuracy relative to larger ResNet-18 models while substantially reducing parameter count and memory footprint. While aggressive pruning and post-training quantization degrade performance, combining KD with QAT preserves accuracy even under low-bit precision, yielding lightweight and robust models suitable for on-board or field deployment. These findings provide practical guidance for developing compact, ecologically focused vision models, enabling scalable, accessible, and efficient AI for global environmental monitoring and conservation.

1. Introduction and Related Work

Earth observation (EO) using satellite and aerial imagery plays a central role in monitoring land use, forest health, agriculture, biodiversity, and climate change impacts [20]. Deep vision models, including CNNs (e.g., ResNet [7]) and Transformers (e.g., ViT [4]), have achieved state-of-the-art performance on such tasks [8, 19]. However, their substantial computational and memory requirements limit deployment in low-resource regions and on edge devices such as UAVs or local monitoring stations, where cloud connectivity and power budgets are restricted.

Model compression has therefore emerged as a critical enabler of efficient EO inference. We consider three principal strategies: **Knowledge Distillation (KD):** transfers knowledge from a large teacher network to a compact student model, preserving accuracy while reducing computation [6, 12]. In EO tasks, KD enables lightweight models to approximate high-resolution feature extraction needed for fine-grained land-cover and habitat classification. **Pruning:** removes redundant parameters, either unstructured (individual weights) or structured (filters/channels), thereby reducing both model size and inference latency [1]. Structured pruning is especially beneficial for deployment on UAVs and IoT sensors tasked with continuous environmental and ecological monitoring. **Quantization:** reduces numerical precision (e.g., 32-bit $\rightarrow$ 8-bit or 4-bit), compressing model size and lowering inference energy [3, 5, 14]. Quantization-Aware Training (QAT) substantially mitigates accuracy degradation, which is crucial for detecting subtle vegetation, habitat, or urban changes in EO pipelines.

Despite progress, most prior work evaluates these techniques in isolation. Few studies systematically examine their *combined* effects on satellite vision models under real-world deployment constraints [15]. Recent efforts focus on lightweight EO architectures and edge inference in agriculture and forestry, but a rigorous, cross-architecture benchmarking for ecological monitoring remains absent.

Our Contributions.

- We provide the first joint benchmarking of KD, pruning, and quantization across CNNs (ResNet, MobileNetV2) and Transformers (ViT, CLIP) on two diverse remote sensing datasets (EuroSAT, UC Merced).
- We establish that MobileNetV2 distilled from ResNet-50 achieves accuracy nearly on par with ResNet-18 on EuroSAT, offering a practical lightweight model for EO tasks such as biodiversity and land-use monitoring.
- We characterize quantization limits: QAT sustains accuracy down to 4-bit precision, while 1-bit models consistently fail to converge, thereby defining compression boundaries for low-power devices in ecological monitoring pipelines.

The remainder of this paper is structured as follows: Section 2 presents our compression pipeline. Section 3 intro-

duces datasets and training protocols. Section 4 reports results and ablation studies. Section 5 concludes with broader implications and future research directions.

2. Method

2.1. Model Architectures

We consider both convolutional neural networks (CNNs) and Transformer-based architectures to ensure broad applicability across Earth observation (EO) tasks. High-capacity teacher models include ResNet-50 [7] and EfficientNet-V2-S [18], pretrained on ImageNet, which provide rich representations for satellite and UAV imagery. Student and baseline models include ResNet-18, a lightweight custom ResNet-9, and MobileNetV2 [11], all well-suited for resource-limited EO deployments. We further evaluate Transformer-based models (ViT-B/16 and CLIP ViT-B/32 [10]) to assess the generality of compression across architectures and modalities. These models are applied to datasets spanning land use, agriculture, forest and habitat monitoring, and biodiversity assessment.

2.2. Model Compression Techniques

Knowledge Distillation. We adopt standard knowledge distillation (KD) to transfer teacher knowledge into compact student models, improving performance under resource constraints. The loss function combines cross-entropy with softened teacher–student alignment:

$$\mathcal{L}_{\text{KD}} = (1-\alpha)\,\mathcal{L}_{\text{CE}}(s,y) + \alpha\,T^2\,\text{KL}\Big(\sigma\big(\tfrac{s}{T}\big), \sigma\big(\tfrac{t}{T}\big)\Big), \quad (1)$$

where s and t are student and teacher logits, y is the ground-truth label, T controls distribution softness, and α balances the two terms. In EO tasks, KD enables smaller models to retain fine-grained distinctions critical for land cover and ecological habitat mapping.

Network Pruning. Pruning removes low-magnitude weights based on global L1-norm, with different sparsity levels. We also evaluate structured pruning, which eliminates entire channels or attention rows in ViT-B/16, guided by a composite layer-importance score to mitigate Transformer sensitivity. Less informative layers are pruned first, maximizing compression with minimal accuracy degradation [1]. Structured pruning is particularly valuable for UAVs and IoT sensors conducting continuous ecological surveillance, such as deforestation tracking or wetland monitoring.

Quantization. We employ two schemes: Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT) [5]. PTQ statically reduces precision of weights and dynamically quantizes activations to 8-bit integers. QAT fine-tunes KD-initialized student models under simulated low-bit constraints, exploring $\{8, 4, 3, 2, 1\}$-bit settings [16]. QAT is particularly critical in EO, where even small spectral differences can signal vegetation stress, habitat degradation, or urban expansion.

Algorithm 1 Budget-Aware Iterative Compression (BAIC)

Input: Full-precision model M_{FP}, budget B, sparsity levels S, bit-widths Q, validation set D_{val}
Initialize: Valid configurations $C_{valid} \leftarrow \emptyset$
for each sparsity $s_i \in S$ **do**
  $M_{pruned} \leftarrow \text{Prune}(M_{FP}, s_i)$
  $M_{pruned} \leftarrow \text{FineTune}(M_{pruned}, D_{val}, \text{epochs=3})$
  for each bit-width $b_j \in Q$ **do**
    $M_{quant} \leftarrow \text{QAT}(M_{pruned}, b_j, D_{val})$
    $B_{final} \leftarrow \text{GetSize}(M_{quant})$
    if $B_{final} \leq B$ **then**
      $Acc_{final} \leftarrow \text{Evaluate}(M_{quant}, D_{val})$
      $C_{valid} \leftarrow C_{valid} \cup \{(s_i, b_j, Acc_{final})\}$
    end if
  end for
end for
return $\arg\max_{(s,b,Acc)\in C_{valid}}(Acc)$

2.3. Adaptive Compression Pipeline

We introduce the **Budget-Aware Iterative Compression (BAIC)** framework, which automatically searches for compression configurations balancing accuracy, size, and latency under a resource budget B [9, 13, 17]. Starting from a full-precision model M_{FP}, BAIC jointly explores pruning sparsity S and quantization bit-widths Q, selecting configurations that maximize validation accuracy while respecting B:

$$\begin{aligned} &\max_{s_i\in S, b_j\in Q} \text{Accuracy}(M_{\text{quant}}(M_{\text{pruned}}(M_{FP}, s_i), b_j)), \\ &\text{s.t.} \quad \text{Size}(M_{\text{quant}}) \leq B. \end{aligned} \quad (2)$$

Algorithm 1 summarizes the process. Each candidate model is pruned, fine-tuned, and quantized; only feasible configurations are retained and ranked. BAIC provides a scalable pipeline for real-world EO deployments, enabling UAV-based monitoring, satellite edge inference, and low-bandwidth agricultural and ecological sensing [2].

3. Experimental Setup and Evaluation

We systematically evaluate compression strategies on modern vision architectures, quantifying trade-offs between efficiency (model size, sparsity, bit-width) and task-specific performance in Earth observation (EO). Our evaluation considers not only classification accuracy but also resource utilization and robustness. All experiments are implemented in PyTorch and executed on NVIDIA GPUs with at least 24 GB memory.

3.1. Datasets and Preprocessing

We use two widely adopted EO benchmarks, chosen for their ecological relevance and complementary characteristics in spatial resolution, scene complexity, and class diversity.

EuroSAT [8] consists of 27,000 Sentinel-2 RGB images across 10 land-cover classes, including residential areas, forests, rivers, industrial zones, and farmland. Images (64×64 pixels) are resized to 224×224 depending on model requirements. This dataset was selected for its diverse geographic coverage and seasonal variations, making it ideal for testing robustness in ecological monitoring tasks like vegetation health and deforestation tracking. Preprocessing includes ImageNet normalization and augmentations (random crops, flips, color jitter) to improve generalization.

UC Merced Land-Use [19] contains 2,100 aerial images over 21 balanced categories such as farmland, commercial areas, and forests. Its fine-grained spatial detail and varied U.S. landscapes make it suitable for evaluating compression on complex scenes relevant to habitat and land-use change studies. Similar preprocessing and augmentations are applied.

For both datasets, we split as $D = D_{\text{train}} \cup D_{\text{val}} \cup D_{\text{test}}$ with $|D_{\text{train}}| = 0.7|D|$, $|D_{\text{val}}| = |D_{\text{test}}| = 0.15|D|$. Stratified sampling preserves class distribution across splits.

3.2. Training and Evaluation Protocol

Let θ denote model parameters optimized over loss $\mathcal{L}$. We train models with AdamW for Transformers and KD-based students, and SGD with momentum 0.9 for baseline CNNs and QAT fine-tuning. Learning rates are searched logarithmically over $[5 \cdot 10^{-6}, 1 \cdot 10^{-2}]$ with weight decay 10^{-4}. Training lasts 3–25 epochs with early stopping when validation accuracy plateaus. We assess both performance and efficiency. On test samples $(x_i, y_i) \in D_{\text{test}}$, Top-1 Accuracy is:

$$\text{Accuracy} = \frac{1}{|D_{\text{test}}|} \sum_{(x_i,y_i)} \mathbf{1}[\hat{y}_i = y_i], \tag{3}$$

where $\hat{y}_i$ is the predicted label. For pruning, sparsity is defined as:

$$\text{sparsity} = \frac{\#\text{pruned weights}}{\#\text{total weights}}. \tag{4}$$

This evaluation framework jointly measures accuracy, sparsity, and resource savings, capturing the accuracy–efficiency trade-offs essential for practical EO deployment, including UAV-based monitoring, biodiversity and habitat assessment, forest health monitoring, and precision agriculture.

4. Results and Discussion

We evaluate compression strategies in terms of accuracy, memory footprint, and robustness on Earth observation (EO) datasets. Results are presented across baseline training, knowledge distillation (KD), pruning, quantization, and their combinations.

Table 1. Baseline ResNet-18 performance on EuroSAT. Fine-tuning significantly improves accuracy.

Configuration	Accuracy (%)
Pretrained (No Fine-tuning)	15.52
From Scratch	89.89
Fine-tuned Pretrained	**95.81**

Table 2. Knowledge distillation on EuroSAT. MobileNetV2 via KD matches ResNet-18 while reducing complexity.

Configuration	Accuracy (%)
ResNet-18 (Fine-tuned)	**95.81**
ResNet-18 (Scratch)	89.89
ResNet-18 (via KD)	93.96
MobileNetV2 (via KD)	95.74
ResNet-9 (via KD)	55.56

4.1. Baseline Performance and Transfer Learning

Table 1 shows ResNet-18 results on EuroSAT. A pretrained model without fine-tuning performs poorly (15.52%), reflecting the domain gap between ImageNet and EO imagery. Training from scratch achieves 89.89%. Fine-tuning a pretrained model yields the best performance (95.81%), confirming that transfer learning provides efficient convergence and high accuracy. These baselines establish reference points for subsequent compression experiments.

4.2. Knowledge Distillation

KD transfers knowledge from a teacher f_{θ_T} to a compact student f_{θ_S} using a softened distribution loss $\mathcal{L}_{KD}$. Results (Table 2) show that MobileNetV2 distilled from ResNet-50 achieves 95.74%, nearly matching ResNet-18. Intra-family distillation also improves accuracy over scratch training. Extremely small students (ResNet-9) fail to retain performance, highlighting capacity limits. KD therefore enables efficient small models while preserving accuracy and reducing inference cost, particularly useful for edge EO applications.

4.3. Pruning

We define sparsity s as the fraction of pruned weights. Table 3 compares ResNet-18 and CLIP ViT-B/32. ResNet-18 maintains 93.36% even at 80% sparsity, while CLIP accuracy drops drastically to 70.00% at 40% sparsity. CNNs tolerate pruning due to filter redundancy, whereas ViTs are sensitive because pruning disrupts global attention dependencies. Thus, CNNs are preferable for aggressive pruning in edge EO tasks.

Table 3. Impact of pruning on ResNet-18 and CLIP ViT-B/32. CNNs show higher robustness.

Model	Sparsity	Accuracy (%)
ResNet-18	0%	95.33
ResNet-18	40%	**95.36**
ResNet-18	80%	93.36
CLIP ViT-B/32	0%	92.77
CLIP ViT-B/32	40%	70.00

Table 4. QAT results for ResNet-18 (EuroSAT). Accuracy holds at 4-bit but drops sharply at lower precision.

Bit-Width	Accuracy (%)
8-bit	**87.56**
4-bit	87.20
3-bit	81.70
2-bit	71.48
1-bit	11.11

Table 5. ResNet-18 (distilled) on UC Merced. Accuracy remains stable under extreme compression.

Setting	Accuracy (%)
Baseline (Distilled)	**100.00**
Pruned (50%)	**100.00**
Pruned (87.5%)	**100.00**
QAT (4-bit)	98.81
QAT (2-bit)	**100.00**

4.4. Quantization

Table 4 shows quantization-aware training (QAT) results for ResNet-18 on EuroSAT. Performance remains strong down to 4 bits (87.20%), but degrades at 2 bits (71.48%) and collapses at 1 bit (11.11%). Post-training quantization (PTQ) at 8 bits yields only 62.57%, underscoring the importance of QAT. On UC Merced (Table 5), compression has negligible effect: even extreme pruning (87.5%) or 2-bit QAT maintains >99% accuracy, reflecting higher dataset redundancy and lower variability. Thus, dataset complexity strongly influences compression tolerance.

4.5. Combined Pruning and Quantization

Table 6 shows combined pruning and QAT on EuroSAT. Moderate compression (50% sparsity, 4-bit) yields 76.72%, but aggressive settings (75%, 2-bit) reduce accuracy to 61.37%. Non-linear degradation arises from compounded pruning and quantization errors. Overall, CNNs remain more resilient than Transformers, KD is vital for small models, QAT is preferred over PTQ, and dataset complexity governs compression tolerance.

Guidelines for Edge Deployment

For efficient EO deployment: 1) Use KD to compress models while maintaining accuracy. 2) Combine pruning with QAT for CNNs when retraining is possible. 3) Avoid aggressive unstructured pruning or ultra-low-bit quantization for Transformers. 4) Prefer QAT over PTQ to ensure stable low-bit performance. 5) Tailor compression to dataset characteristics, as redundancy and variability strongly affect robustness.

Table 6. Combined pruning and QAT on ResNet-18 (EuroSAT). Accuracy degrades under extreme compression.

Sparsity	Bit-Width	Accuracy (%)
25%	4-bit	67.04
25%	2-bit	63.74
50%	4-bit	**76.72**
50%	2-bit	62.06
75%	4-bit	76.59
75%	2-bit	61.37

In summary, integrating KD, pruning, and QAT offers a principled path to deploying accurate, lightweight models for ecological monitoring and sustainable EO applications.

5. Conclusion and Future Directions

In this work, we systematically evaluated knowledge distillation (KD), pruning, and quantization on CNN and Transformer architectures across two satellite imagery benchmarks (EuroSAT, UC Merced). Our results highlight clear trade-offs between accuracy, sparsity, and bit-width. KD enables compact student models to closely match larger teachers (e.g., MobileNetV2 distilled from ResNet-50 reaches 95.74% vs. 95.81% for ResNet-18). Pruning combined with quantization-aware training (QAT) consistently outperforms post-training quantization, particularly under low-bit regimes. We also find that compression robustness strongly depends on dataset complexity, underscoring the need for context-aware deployment strategies for ecological monitoring, habitat mapping, and sustainable environmental assessment.

Future directions include: (i) iterative pipelines integrating pruning, fine-tuning, and QAT for maximal compression; (ii) hardware-aware methods tailored to edge accelerators and GPUs; (iii) scaling evaluations to larger, high-resolution datasets such as BigEarthNet and SEN12MS to enable comprehensive monitoring of forests, urban expansion, and biodiversity hotspots; (iv) exploring lightweight Transformers (e.g., MobileViT, TinyViT, EfficientFormer) with higher inherent efficiency suitable for UAV- and satellite-based ecological surveys, and investigating alternative pruning strategies specifically designed for Transformer architectures to mitigate their observed sensitivity to unstructured pruning; and (v) developing theoretical bounds on accuracy loss as a function of sparsity and bit-width to provide rigorous guidelines for deploying compressed EO models in conservation and sustainable land management applications.

References

[1] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? *Proceedings of machine learning and systems*, 2: 129–146, 2020. 1, 2

[2] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*, 2019. 2

[3] Gabriel Dax, Srilakshmi Nagarajan, Hao Li, and Martin Werner. Compression supports spatial deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:702–713, 2022. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[5] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-power computer vision*, pages 291–326. Chapman and Hall/CRC, 2022. 1, 2

[6] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[8] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian 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*, 12(7):2217–2226, 2019. 1, 3

[9] Pallavi Jain, Diego Marcos, Dino Ienco, Roberto Interdonato, Aayush Dhakal, Nathan Jacobs, and Tristan Berchoux. Aligning geo-tagged clip representations and satellite imagery for few-shot land use classification. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 319–323. IEEE, 2024. 2

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[11] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 2

[12] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 1

[13] Tushar Shinde. Adaptive quantization and pruning of deep neural networks via layer importance estimation. In *Workshop on Machine Learning and Compression, NeurIPS 2024*, . 2

[14] Tushar Shinde. High-performance lightweight vision models for land cover classification with coresets and compression. In *TerraBytes-ICML 2025 workshop*, . 1

[15] Tushar Shinde. Model compression meets resolution scaling for efficient remote sensing classification. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1200–1209, 2025. 1

[16] Tushar Shinde. Towards optimal layer ordering for efficient model compression via pruning and quantization. In *2025 25th International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2025. 2

[17] Tushar Shinde and Sukanya Tukaram Naik. Adaptive quantization of deep neural networks via layer importance estimation. In *International Conference on Computer Vision and Image Processing*, pages 220–233. Springer, 2024. 2

[18] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 2

[19] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010. 1, 3

[20] Xuming Zhang, Yuanchao Su, Lianru Gao, Lorenzo Bruzzone, Xingfa Gu, and Qingjiu Tian. A lightweight transformer network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023. 1