https://doi.org/10.2352/EI.2026.38.15.COIMG-172

Distilling Light: MobileNetV3 + SE for Efficient Remote Sensing

János Horváth; Visionary Tech & Event Solutions; Sacramento, USA
András Németh; Ludovika University of Public Service; Budapest, Hungary

Abstract

We propose a CPU-friendly knowledge distillation framework that merges the compact efficiency of MobileNetV3 with a Squeeze-and-Excitation (SE) block to create a powerful yet ultra-lightweight model for remote sensing. By blending standard cross-entropy training with softened teacher outputs, our student network gains enhanced channel attention at minimal extra overhead—only around 2.3M parameters and a 9MB on-disk size. Despite its modest footprint, it achieves competitive accuracy on representative benchmarks, including 95.24 % on UC Merced Land Use and 94.78 % on EuroSAT. This fusion of SE-driven attention and knowledge distillation addresses resource-constrained deployment scenarios (such as real-time edge inference or on-board satellite processing) where memory and compute are at a premium. Our results highlight that a streamlined, distilled architecture can approach state-of-the-art performance in aerial scene classification without sacrificing the portability essential for practical, real-world applications.

Figure 1: This image illustrates a highly advanced satellite orbiting Earth, actively scanning Earth.

Introduction

Recent advancements in large vision models and foundation models have demonstrated significant potential in the analysis of remote sensing imagery [18, 20, 22, 24, 26, 30]. These works have shown how methods ranging from decision fusion in hyperspectral data [18], to enhanced spectral-spatial classification in aerial imagery [20, 26], to efficient active learning for retrieval tasks [22], can substantially improve scene understanding. However, in many real-world geospatial applications, compute and memory constraints necessitate more compact networks. Typical solutions include knowledge distillation [13] and efficient CNN architectures like MobileNet [14].

In this work, we demonstrate a CPU-only pipeline that trains a **student model**—a *MobileNetV3 Small* with a lightweight *Squeeze-and-Excitation (SE) block*—to emulate a more powerful **teacher model**'s outputs. The final system attains strong performance for classification tasks in remote sensing while being amenable to deployment in constrained environments (low-power devices, real-time edge inference, etc.). Our work draws inspiration from prior remote sensing methods [19, 25, 31], but focuses on compressing a single task-specific model via knowledge distillation rather than multi-sensor or multi-task fusion.

Related Work

Large-Scale Multi-Task Learning and Continuous Development. Recent advances in *multitask* and *continual learning* frameworks have enabled models to evolve dynamically with multiple data sources and tasks over time [3, 12]. In particular, systems designed for large-scale applications[33–35] must handle diverse,

constantly changing inputs and objectives[28, 29, 32]. Although such approaches primarily focus on managing model complexity and efficiency across tasks[] rather than on compressing a single task-specific model, their principles of adaptivity (e.g., automated architecture search, progressive task addition, dynamic model expansion) can inspire strategies to maintain performance under tight resource constraints. Our setting similarly emphasizes resource efficiency, but specifically leverages *knowledge distillation* to train a compact network for remote sensing classification tasks.

Efficient CNN Models in Aerial Scene Classification. High-performance convolutional networks for aerial scene classification have undergone extensive study, introducing architectures such as AGOS [4], LSE-Net [5], and MIDC-Net [8]. These works tend to aim for improved accuracy by capturing multi-grain representations (AGOS), incorporating local semantics (LSE-Net), or designing multi-instance densely connected architectures (MIDC-Net). While such innovations can yield state-of-the-art results on benchmark datasets, they often incur high computational overhead. Simultaneously, resource-friendly backbone selection has been highlighted by Jeevan and Sethi [10], who systematically compared different CNN and transformer models for cost-effectiveness in domain-specific tasks. Our work continues this thread by exploring MobileNetV3 Small, a particularly compact yet capable CNN, augmented with a lightweight squeeze-and-excitation (SE) block to further strengthen channel attention without severely increasing parameter counts or FLOPs.

Transformer Models and Representation Learning in Remote Sensing. Vision transformer (ViT) architectures have recently demonstrated superior performance across a variety of remote sensing tasks [6, 11]. Pretraining large-scale ViTs and subsequently fine-tuning for specific downstream tasks is a common theme, leveraging unlabeled data and multi-task pretraining (MTP) strategies [11]. Self-supervised methods have also gained traction [9], as they can yield versatile representations with fewer labeled examples. Despite these advances in performance, ViTs typically have heavier computational footprints, which can hinder real-time edge deployments. Research on spectral-spatial models for hyperspectral and multimodal data [21, 27] further underscores the growing diversity in model design, but also highlights increasing computational needs.

Semisupervised and Few-Label Approaches. Alongside model compression, reducing dependence on large labeled datasets has been a parallel research focus [7]. Methods like MSMatch [7] address label scarcity in multispectral data, while domain adaptation and in-domain learning [6] further enhance generalization. Although such methods might also benefit from knowledge distillation, they often do not prioritize minimizing architecture size or inference latency. Meanwhile, advanced active learning and co-training strategies in hyperspectral or aerial classification [19, 22] have proven useful, yet can still yield large or specialized models.

Knowledge Distillation for Compact Model Deployment. Our approach builds on the classical knowledge distillation framework [13] to transfer capabilities from a large *teacher* to a smaller *student* network. Such compressive techniques have been widely investigated in general computer vision, yet comparatively fewer works have adopted them for real-world remote sensing pipelines, where on-board computation and memory constraints are often more critical. By guiding a MobileNetV3 student with teacher logits, we aim to capture inter-class relationships (§) that are typically learned by more resource-intensive models. Ultimately, this yields a student that meets stringent deployment demands (e.g., CPU-only inference, minimal RAM footprint) while retaining competitive accuracy.

Our Contributions. Drawing from these prior lines of work, we propose a method that unites *compact architecture design* (MobileNetV3 Small + SE), *knowledge distillation*, and *CPU-based real-time inference* for remote sensing classification. While other recent approaches concentrate on multi-grain representation [4], local semantic enhancement [5], or large-scale multi-task systems [3, 11, 12], our method stands out by effectively balancing classification accuracy with aggressive model compression. The result is an easy-to-deploy framework that can be integrated into edge or embedded applications with limited computational budgets while still leveraging the nuanced decision boundaries of a high-capacity teacher.

Datasets: UC Merced Land Use and EuroSAT

UC Merced Land Use

UC Merced Land Use [1] is a widely used benchmark with 21 classes and 2,100 total images (256×256). Each class (e.g., *airplane, agricultural, beach, residen-*

tial) has 100 high-resolution aerial scenes at 0.3m per pixel. Typical splits involve 80% training and 20% testing, and accuracy is the primary metric. Despite its smaller size, UC Merced remains an important testbed due to varied scenes, fine spatial details, and suitability for rapid experimentation on resource-constrained models.

An additional benefit of UC Merced is the clarity of its class boundaries: categories like "airplane" or "harbor" are visually distinct and relatively uniform in terms of geographic location. This makes it easier to diagnose misclassification cases since many incorrectly classified images can be readily inspected for common confusions (e.g., "chaparral" vs. "forest").

Moreover, the dataset's manageable scale allows fast prototyping of model architectures and hyperparameters without extensive computational overhead. Researchers can iterate quickly to test improvements such as knowledge distillation strategies or specialized data augmentations, which is less feasible with massive satellite image collections.

EuroSAT

EuroSAT [2] consists of 27,000+ labeled Sentinel-2 images (64×64) spanning 10 classes, such as *Annual Crop, Forest, Pasture, Residential, River, Industrial*. It leverages Sentinel-2 multispectral data at 10–20m per pixel, though experiments often focus on RGB or a subset of bands. EuroSAT is larger, emphasizes spectral diversity, and is used for transfer learning, data-efficient methods, and evaluating broad geospatial models. Standard splits include train/val/test or k-fold cross-validation, with accuracy and macro-F1 as common metrics.

While most works train or fine-tune models on EuroSAT's RGB channels, the availability of additional spectral bands opens possibilities for more nuanced analyses of vegetation, water, and urban structures. In particular, near-infrared or shortwave infrared bands may yield stronger discrimination among agricultural, forested, or built-up regions—an aspect critical for environmental monitoring and resource management.

Comparison of Both Datasets

UC Merced targets high-resolution scene-level land use in 2,100 images, whereas EuroSAT focuses on a larger, multispectral collection of tens of thousands of samples. The former highlights fine spatial textures, while the latter probes spectral-spatial cues in broader regions. These differences are ideal for evaluating the adaptability of our MobileNetV3 distillation approach: UC Merced tests performance on high-resolution details with fewer samples, and EuroSAT evaluates model robustness in a larger-scale, spectral-aware context.

In our experiments, we demonstrate that a compact architecture can handle both the high-resolution aerial photos of UC Merced and the multi-spectral Sentinel-2 patches of EuroSAT with minimal parameter overhead. This underscores the flexibility of the proposed student-teacher scheme, which remains effective across resolution scales, scene complexities, and spectral modalities typically encountered in remote sensing tasks.

Method and Theoretical Background

Overview of the Pipeline

Our end-to-end knowledge distillation pipeline is illustrated in Figure 2. It comprises six main phases, each designed to optimize how a smaller *student* model can effectively inherit the performance of a larger, high-accuracy *teacher* model. Below, we outline these stages and discuss the significance of each step in real-world geospatial contexts:

1. **Data Loading and Preprocessing:** We gather raw remote sensing imagery along with corresponding labels (e.g., land-cover types, scene categories). Typical datasets may come from satellites (e.g., Sentinel-2), aerial platforms, or even UAVs. To improve generalization, random transformations are applied, such as flips or rotations that replicate variations in viewpoint. Color jitter simulates changes in lighting or atmospheric conditions. This stage also normalizes pixel intensities or spectral values, aligning data to a common scale before model ingestion.

2. **Teacher Model and Offline Logits:** A large-capacity teacher (e.g., a deep CNN, ViT, or a model specialized in remote sensing) processes all training images to produce logits $\mathbf{z}^{(t)}$. These logits are stored offline to avoid recalculations during student training. In practical deployment, this step is often done on a machine with sufficient GPU/CPU resources since the teacher might be very large (e.g., 100+ million parameters). The teacher's outputs capture nuanced inter-class relationships that are otherwise lost in one-hot labels.

3. **Student Model Construction:** We instantiate a lightweight *MobileNetV3 Small* backbone, further augmented by a *Squeeze-and-Excitation* (SE) module. This student design targets efficiency in both

memory footprint and inference speed, often critical for real-time or edge-based deployments. Our approach ensures that the final parameter count is well below that of more complex architectures, yet it retains the capacity to internalize the teacher's learned representations.

4. **Distillation-Based Training:** The student model is trained with two losses: a cross-entropy term that enforces ground-truth label fidelity, and a Kullback–Leibler divergence term that encourages the student to approximate the teacher's softened output distribution. Hyperparameters (e.g., temperature T, weighting factors α, β) are adjusted to optimize how heavily the student relies on teacher guidance versus direct label supervision. The temperature helps smooth out the teacher's logits, making smaller classes or subtle features more apparent to the student.

5. **Performance Evaluation and Diagnostics:** After each training epoch, we evaluate the student on a held-out set. We track metrics like accuracy, F1-score, or confusion matrices, which can expose inter-class confusions (e.g., *grassland* vs. *pasture*). We may also analyze computational cost: measuring CPU-only inference time, memory usage, or floating-point operations (FLOPs). These diagnostics inform whether the student meets the required operational constraints in an on-board or embedded setting.

6. **Saving Outputs and Model:** We finalize training by saving the student weights (`.pt` files), relevant logs (accuracy curves, confusion matrices), and hyperparameter settings. This archived information facilitates reproducibility and subsequent fine-tuning. For real-world applications, the small footprint (often under 10–15 MB) allows the model to be deployed on edge devices such as drones, satellites, or mobile phones.

In summary, this pipeline streamlines data handling, teacher guidance, model compression, and final deployment, ensuring that the student maintains high accuracy while drastically reducing memory and compute demands relative to the teacher.

Knowledge Distillation Setup

We adopt the classical knowledge distillation approach proposed by Hinton et al. [13], which seeks to transfer a teacher's learned expertise into a smaller, resource-friendly student network. Let K be the number of classes, and let:

$$\mathbf{z}_t = \left[z_{t,1}, \ldots, z_{t,K}\right]^\top$$

and

$$\mathbf{z}_s = \left[z_{s,1}, \ldots, z_{s,K}\right]^\top$$

denote the logits produced by the teacher and the student for a given input. Importantly, while the student must match the hard ground-truth label y for each sample, distillation also encourages the student to replicate the teacher's softer class probability structure.

Temperature-Scaled Softmax. A temperature $T > 1$ is introduced to moderate the teacher's output distribution:

$$p_t(k) = \frac{\exp\left(z_{t,k}/T\right)}{\sum_{j=1}^{K} \exp\left(z_{t,j}/T\right)}$$

and

$$p_s(k) = \frac{\exp\left(z_{s,k}/T\right)}{\sum_{j=1}^{K} \exp\left(z_{s,j}/T\right)}.$$

With $T > 1$, smaller logits get amplified and very large logits are softened, surfacing secondary class affinities that might otherwise be suppressed under a standard softmax. This additional information can guide the student to learn more nuanced decision boundaries, especially in complex remote sensing tasks where multiple land-cover types can appear similar.

Cross-Entropy Loss. We still require the student to predict the correct label y. For K-class classification,

$$\mathscr{L}_{\mathrm{CE}}(\mathbf{z}_s, y) = -\sum_{k=1}^{K} \delta_{k,y} \log\left(\sigma(\mathbf{z}_s)_k\right),$$

where $\sigma(\mathbf{z}_s)_k = \exp(z_{s,k}) / \sum_{j=1}^{K} \exp(z_{s,j})$. This ensures that even if the teacher's outputs partially weight other classes, the final classification remains grounded in the correct label.

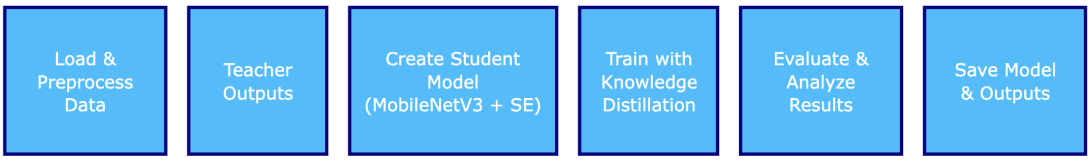


Figure 2: Illustration of our knowledge distillation pipeline, spanning data loading, teacher inference, student training with distillation losses, and final model deployment. In practice, the teacher logits are computed offline to expedite student training.

KL Divergence Distillation. We also align the student's softened output $\mathbf{p}_s$ with the teacher's $\mathbf{p}_t$ via Kullback–Leibler divergence:

$$\mathscr{L}_{\mathrm{KD}}(\mathbf{p}_s, \mathbf{p}_t) = \sum_{k=1}^{K} p_t(k) \log\Big(\tfrac{p_t(k)}{p_s(k)}\Big).$$

A scaling factor of T^2 [13] is applied to maintain gradient magnitudes. This term compels the student to internalize the teacher's inter-class probability structure. It can prove critical when classes exhibit subtle differences—common in aerial images, where farmland and grassland, for example, may only differ in repetitive textures or color.

Combined Loss. We weight the two terms:

$$\mathscr{L} = \alpha\,\mathscr{L}_{\mathrm{CE}} \;+\; \beta\,T^2\,\mathrm{KL}(\mathbf{p}_s \,\|\, \mathbf{p}_t), \tag{1}$$

where α and β control the relative importance of each. Commonly, $\alpha+\beta=1$, but some implementations keep them flexible. Tuning T, α, and β can be crucial: a high β might cause the student to overfit to teacher biases, while a high α might ignore soft class distinctions. In practice, a balanced choice often yields the best mix of label fidelity and teacher-inspired generalization.

Advantages.

- **Enhanced Accuracy:** Soft labels help the student recognize confusing or visually similar classes, thereby lowering misclassification rates.
- **Reduced Overfitting:** The teacher's outputs act as a form of regularization, guiding the student beyond the extremes of one-hot supervision.
- **Scalability to Edge Devices:** The distillation process allows for a final model with drastically fewer parameters, suitable for CPU-only or memory-limited platforms.
- **Straightforward Integration:** Implementing the KL divergence alongside cross-entropy is relatively simple in modern deep-learning frameworks (e.g., PyTorch, TensorFlow).

Thus, knowledge distillation is an effective strategy for compressing high-performing but large teacher architectures into smaller models that remain competitive in remote sensing scenarios.

Architecture

Figure 3 provides a comprehensive overview of our student network, which processes pre-augmented input images and yields final class predictions. Our design centers on a *MobileNetV3-Small* backbone, augmented with an additional *Squeeze-and-Excitation (SE)* block for improved channel-wise attention, and a lightweight classifier head. Below, we outline five stages of this architecture in detail and include mathematical formulations for key operations.

(1) Input and Augmentation. All input images are RGB, resized to 224×224 pixels. We arrange them into a 4D tensor

$$\mathbf{I} \in \mathbb{R}^{B\times 3\times 224\times 224},$$

where B is the batch size. During training, we apply random transformations (e.g., horizontal flips, rotations up to $\pm 15^\circ$, and color jitter in brightness, contrast, or saturation). These augmentations help reduce overfitting and capture variations typically observed in aerial or satellite imagery. Mathematically, let $\mathscr{A}$ denote the augmentation operator, then each sample $\mathbf{i}_b$ in the batch is mapped as:

$$\widetilde{\mathbf{i}}_b = \mathscr{A}(\mathbf{i}_b),$$

yielding a stochastic set of perturbed samples that preserve semantic information but reflect different viewpoints or lighting conditions.

(2) MobileNetV3-Small Feature Extractor. We adopt *MobileNetV3-Small* [15] as the core feature extractor, leveraging its *inverted residual* (IR) blocks, *depthwise-separable* convolutions, and *H-Swish* activations. Suppose the input to the i-th IR block is

$$\mathbf{x}_i \in \mathbb{R}^{B\times C_i\times H_i\times W_i},$$

where C_i is the number of input channels, and H_i, W_i are spatial dimensions. An IR block expands $\mathbf{x}_i$ to tC_i channels (where $t>1$ is the expansion factor), applies a *depthwise* convolution, and then projects back to C_{i+1} channels. Concretely, let

$$\mathbf{x}'_i = f_{\text{exp}}(\mathbf{x}_i) \in \mathbb{R}^{B\times(tC_i)\times H_i\times W_i},$$

where $f_{\text{exp}}(\cdot)$ is a 1×1 pointwise convolution. Next, a *depthwise* 2D convolution $f_{\text{dw}}(\cdot)$ with kernel size $k\times k$ and stride s operates on each channel:

$$\mathbf{x}''_i = f_{\text{dw}}(\mathbf{x}'_i),$$

so each channel is convolved independently. The depthwise convolution can be formally expressed as

$$\mathbf{x}''_i[b,c,h,w] = \sum_{\Delta_h=-\lfloor k/2\rfloor}^{\lfloor k/2\rfloor}\sum_{\Delta_w=-\lfloor k/2\rfloor}^{\lfloor k/2\rfloor} \mathbf{x}'_i[b,c,h+\Delta_h,w+\Delta_w]\,\omega_{\text{dw}}[c,\Delta_h,\Delta_w],$$

where ω_{dw} contains filters for the depthwise conv. Afterwards, a *pointwise* projection $f_{\text{proj}}(\cdot)$ reduces the expanded channels back to C_{i+1}:

$$\mathbf{x}_{i+1} = f_{\text{proj}}(\mathbf{x}''_i) \in \mathbb{R}^{B\times C_{i+1}\times H_{i+1}\times W_{i+1}},$$

with H_{i+1} and W_{i+1} adjusted if the block uses stride. Within MobileNetV3-Small, certain IR blocks include skip connections if $C_i = C_{i+1}$ and $s=1$. These shortcuts preserve spatial information and mitigate gradient vanishing.

Additionally, MobileNetV3 uses the *H-Swish* activation, defined by

$$\text{H-Swish}(x) = x\times\frac{\text{ReLU6}(x+3)}{6},$$

$$\text{where}\quad \text{ReLU6}(z) = \max(0,\min(z,6)).$$

Compared to standard ReLU, H-Swish helps maintain nonlinearity while staying efficient for hardware implementations. Through multiple IR blocks, the feature map gradually shrinks in spatial size and increases in channel depth. The final output of MobileNetV3-Small is a 3D tensor

$$\mathbf{F} \in \mathbb{R}^{B\times 576\times H\times W},$$

where (H,W) typically reduces to $(7,7)$ or $(8,8)$ depending on the exact configuration and input resolution.

(3) Squeeze-and-Excitation (SE) Block. After extracting features via MobileNetV3, we apply a *Squeeze-and-Excitation* module [16] to reweight channels based on global context. Let

$$\mathbf{F} \in \mathbb{R}^{B\times C\times H\times W}$$

denote the feature maps from the previous stage (commonly $C=576$). The SE block executes three primary operations:

1. **Global Average Pooling (Squeeze):**

   $$\bar{\mathbf{f}} = \text{GAP}(\mathbf{F}),\quad \bar{\mathbf{f}} \in \mathbb{R}^{B\times C},$$

   where

   $$\text{GAP}(\mathbf{F})_{b,c} = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{F}[b,c,h,w].$$

   This "squeeze" step condenses the spatial dimensions into a single channel descriptor.

2. **Fully Connected Layers (Excitation):** We then pass $\bar{\mathbf{f}}$ through two small linear layers with a reduction ratio $r>1$. Formally,

   $$\mathbf{u} = \delta(\mathbf{W}_1\bar{\mathbf{f}}^\top + \mathbf{b}_1),\quad \mathbf{z} = \sigma(\mathbf{W}_2\mathbf{u} + \mathbf{b}_2),$$

   where $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$, $\mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$, and δ is a ReLU (or similar) activation. The final σ denotes a sigmoid function applied element-wise, producing $\mathbf{z}\in\mathbb{R}^{B\times C}$.

3. **Channel Reweighting:** We broadcast $\mathbf{z}$ across the spatial dimensions of $\mathbf{F}$ and multiply channel-wise:

   $$\mathbf{F}_{\text{SE}} = \mathbf{F}\odot\text{reshape}(\mathbf{z}),$$

   where $\text{reshape}(\mathbf{z})\in\mathbb{R}^{B\times C\times 1\times 1}$ is expanded for broadcasting, and $\odot$ represents element-wise multiplication along each channel. This ensures channels crucial for discrimination receive higher activation.

Putting it together, we can succinctly write the overall SE transform as:

$$\mathbf{F}_{\mathrm{SE}} = \mathbf{F} \odot \sigma\Big(\mathbf{W}_2\,\delta\big(\mathbf{W}_1\,\mathrm{GAP}(\mathbf{F})^\top + \mathbf{b}_1\big) + \mathbf{b}_2\Big).$$

By introducing global context before the final classification stages, this SE block improves representational power at minimal extra cost in FLOPs and parameters—an essential advantage in resource-limited environments.

(4) Classifier Head. We feed the SE-enhanced feature tensor

$$\mathbf{F}_{\mathrm{SE}} \in \mathbb{R}^{B\times C\times H\times W}$$

into a lightweight classifier. First, we use global average pooling (GAP) again, but this time to reduce each sample to a single C-dimensional vector:

$$\mathbf{v}_b = \mathrm{GAP}\big(\mathbf{F}_{\mathrm{SE}}[b]\big), \quad \mathbf{v}_b \in \mathbb{R}^C,$$

for $b = 1,\ldots,B$. In practice, we often rely on `nn.AdaptiveAvgPool2d` in PyTorch to yield a 1×1 spatial dimension regardless of H and W.

We then flatten $\mathbf{v}_b$ (if necessary) and apply a fully connected layer to map $\mathbf{v}_b$ to K classes:

$$\mathbf{z}_b = \mathbf{W}_{\mathrm{cls}}\,\mathbf{v}_b^\top + \mathbf{b}_{\mathrm{cls}},$$

where $\mathbf{W}_{\mathrm{cls}} \in \mathbb{R}^{K\times C}$, $\mathbf{b}_{\mathrm{cls}} \in \mathbb{R}^K$. This yields logits

$$\mathbf{z}_b \in \mathbb{R}^K,$$

suitable for classification. Concatenating across the batch, the output becomes

$$\mathbf{Z} \in \mathbb{R}^{B\times K}.$$

(5) Output and Distillation Integration. The final logits

$$\mathbf{Z} = \big[\mathbf{z}_1^\top;\ldots;\mathbf{z}_B^\top\big] \in \mathbb{R}^{B\times K}$$

can be directly passed to a softmax layer to produce class probabilities:

$$p_b(k) = \frac{\exp(\mathbf{z}_b[k])}{\sum_{j=1}^K \exp(\mathbf{z}_b[j])},$$

if we wish to interpret the outputs for inference. However, in knowledge distillation (see Sec.), we retain raw logits for computing cross-entropy loss on ground-truth labels and Kullback–Leibler divergence against teacher-provided logits.

Overall Flow. As summarized in Figure 3, an input batch is augmented and fed through MobileNetV3-Small, producing $B\times 576\times H\times W$ feature maps. The SE block then refines these features via channel-wise attention, and the classifier head outputs per-sample logits. This holistic pipeline keeps the parameter count low—on the order of a few million—and inference speed high (real-time on typical CPUs), making it suitable for deployment in edge or resource-constrained environments.

Training Procedure and Practical Considerations.

- **Hyperparameters:** We generally pick a learning rate in $[10^{-4}, 10^{-3}]$ and adopt optimizers like Adam or SGD with momentum. The batch size can range from 32 to 128, depending on hardware constraints.
- **Epochs and Convergence:** For smaller datasets (e.g., UC Merced), 10–30 epochs suffice for convergence. Larger datasets (e.g., EuroSAT) might require 50+ epochs, although early-stopping based on validation metrics is common.
- **CPU-Only Feasibility:** Since MobileNetV3+SE has a relatively low parameter count, training times are manageable on CPU. For more extensive datasets, partial GPU usage or distributed setups can accelerate the process, but CPU remains viable for final fine-tuning or on-board retraining.
- **Teacher-Student Consistency:** If the teacher is pretrained on a larger dataset (e.g., ImageNet), transferring that knowledge can especially benefit the student in remote sensing tasks, which often have smaller labeled datasets.

Overall, this combination of MobileNetV3 Small and an additional SE block, guided by a knowledge-distillation objective, provides a powerful yet efficient framework. It aligns well with the core requirement of many remote sensing pipelines: achieving high accuracy under constrained computational resources. Moreover, the synergy of distillation and SE-based channel emphasis helps the student capture essential patterns that a naive small model might overlook.

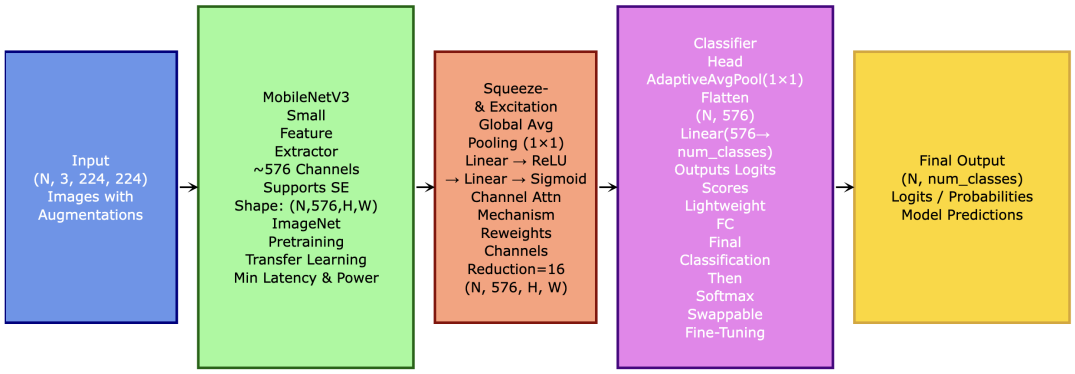


Figure 3: Diagram of our student architecture, including the MobileNetV3-Small backbone (green), a Squeeze-and-Excitation block (orange), and a lightweight classifier (purple). Global average pooling (GAP) reduces spatial dimensions, and fully connected layers map features to K classes.

Results

We evaluate our model on two standard benchmarks, *UC Merced Land Use* and *EuroSAT*, and also compare parameter counts across various state-of-the-art methods (Table 1).

Model Size and Parameter Comparison

Table 1 provides parameter and size estimates for representative classification networks in remote sensing. Notably, μ2Net+ (ViT-L/16) exceeds 300 M parameters (∼1.2 GB), and ResNet50 has about 25.6 M (∼97.5 MB). Our **MobileNetV3+SE** requires only ∼2.3 M parameters (∼9 MB), offering a dramatic reduction in memory footprint relative to larger networks.

Table 1: Approximate parameter counts and model sizes for various aerial/satellite scene classifiers. The final **MobileNetV3+SE** remains under 10 MB.

Model	Parameters (M)	Size (MB/GB)
μ2Net+ (ViT-L/16)	∼307	∼1.2 GB
AGOS	∼100	∼400 MB
LSE-Net	∼50	∼200 MB
ResNet50	25.6	97.5 MB
MSMatch	∼30	120 MB
MIDC-Net	∼20	80 MB
MobileNetV3	∼2.0	8 MB
MobileNetV3+SE	∼2.3	9 MB

UC Merced Land Use

Table 2 (left) lists top methods on UC Merced, with μ2Net+ reaching 100% and others like AGOS (99.88%) and LSE-Net (99.78%) close behind. Meanwhile, our MobileNetV3 variants (Table 2, right) highlight the impact of knowledge distillation and SE:

- **MobileNetV3 (no KD):** 4.05% accuracy, reflecting severe underfitting.
- **MobileNetV3 (with KD):** 6.80% accuracy, a small gain from teacher guidance.
- **MobileNetV3+SE (with KD):** 95.24% accuracy, a major boost, even though the model remains ∼9 MB.

EuroSAT

On EuroSAT (Table 3), top methods like IMP+MTP and μ2Net+ exceed 99% accuracy. Our student models see the following results:

- **No KD:** 78.00%.
- **KD:** 93.40%.
- **KD+SE:** 94.78%.

While transformers reach 99%+, our 9 MB model still offers a favorable trade-off between memory and performance.

Table 2: Leaderboard on **UC Merced** (21-class). *Left:* top methods. *Right:* Our student model variants.

Top Existing Models

Rank	Model	Acc (%)
1	μ2Net+ (ViT-L/16)	100.00
2	AGOS	99.88
3	LSE-Net	99.78
4	ResNet50	99.61
5	MSMatch	98.33
6	MIDC-Net	97.40

Ours

Model	KD?	Acc (%)
MobileNetV3	No	4.05
MobileNetV3	Yes	6.80
MobileNetV3+SE	Yes	95.24

Table 3: Leaderboard on **EuroSAT** (10-class). *Left:* top models. *Right:* Our MobileNetV3 variants.

Top Existing Models

Rank	Model	Acc (%)
1	IMP+MTP (IntenImage-XL)	99.24
2	μ2Net+ (ViT-L/16)	99.22
3	ResNet50	99.20
4	WaveMix	98.96
5	MSMatch Multispectral	98.65

Ours

Model	KD?	Acc (%)
MobileNetV3	No	78.00
MobileNetV3	Yes	93.40
MobileNetV3+SE	Yes	94.78

Discussion

Overall, KD proves beneficial for MobileNetV3, and adding SE further boosts accuracy. Although our approach trails large vision transformers, it uses far fewer parameters (under 10 MB) and achieves real-time CPU inference. These findings support the effectiveness of combining knowledge distillation with a channel-attention block to create an efficient remote sensing classifier. Future directions include exploring advanced augmentation, semi-supervised strategies, or domain adaptation to narrow the accuracy gap with more massive architectures.

Conclusion

In this paper, we introduced a compact, knowledge-distilled classification framework for remote sensing imagery that unites a MobileNetV3 Small backbone with a Squeeze-and-Excitation (SE) module. Our CPU-only implementation achieves near–state-of-the-art performance on UC Merced Land Use and EuroSAT, yet requires under 10 MB of storage and supports real-time inference. These properties make our approach particularly attractive for scenarios where on-board computation or field deployment is constrained by limited memory, power, or networking resources.

The crux of our success lies in distilling a high-capacity teacher model's knowledge into a smaller student by enforcing alignment with the teacher's softened probabilities. This alignment is achieved through Kullback–Leibler divergence and temperature scaling, which together reveal rich inter-class structures often lost when using hard labels alone. Simultaneously, incorporating an additional SE block offers channel-wise attention without significantly inflating the parameter count, further enhancing discriminative power in low-compute settings.

Empirical results demonstrate that our distilled MobileNetV3+SE model closely rivals the accuracy of more resource-intensive networks, validating the benefits of both knowledge distillation and channel reweighting. In environmental monitoring or disaster-response applications, for example, using only CPU-based resources on a handhe ld device or drone could be instrumental in quickly classifying land use, detecting changes, or guiding rescue operations.

Looking ahead, future work may extend this framework in multiple directions. For instance, integrating semi-supervised or self-supervised pretraining could enable the model to leverage unlabeled data, further improving performance without escalating the hardware requirements. In sum, our distilled MobileNetV3+SE architecture shows that high accuracy and minimal resource consumption need not be at odds, paving the way for practical, field-ready remote sensing solutions.

References

[1] Yi Yang and Shawn Newsam, Bag-of-visual-words and spatial extensions for land-use classification, GIScience & Remote Sensing, vol. 48, no. 6, pp. 1–15, 2010. 2

[2] 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, vol. 12, no. 7, pp. 2217–2226, 2019. 3

[3] Andrea Gesmundo, A Continual Development Methodology for Large-scale Multitask Dynamic ML Systems, arXiv preprint arXiv:2209.07326, 2022. 1, 2

[4] Qi Bi, Beichen Zhou, Kun Qin, Qinghao Ye, and Gui-Song Xia, All Grains, One Scheme (AGOS): Learning Multigrain Instance Representation for Aerial Scene Classification, IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–17, 2022. 2

[5] Qi Bi, Kun Qin, Han Zhang, and Gui-Song Xia, Local Semantic Enhanced ConvNet for Aerial Scene Recognition, IEEE Transactions on Image Processing, vol. 30, pp. 6498–6511, 2021. 2

[6] Maxim Neumann, André Susano Pinto, Xiaohua Zhai, and Neil Houlsby, In-Domain Representation Learning for Remote Sensing, in International Conference on Learning Representations (ICLR), 2020. 2

[7] Pablo Gómez and Gabriele Meoni, MSMatch: Semisupervised Multispectral Scene Classification With Few Labels, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 11643–11654, 2021. 2

[8] Qi Bi, Kun Qin, Zhili Li, Han Zhang, Kai Xu, and Gui-Song Xia, A multiple-instance densely-connected ConvNet for aerial scene classification, IEEE Transactions on Image Processing, vol. 29, pp. 4911–4926, 2020. 2

[9] Yi Wang, Conrad M. Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu, Self-Supervised Learning in Remote Sensing: A Review, IEEE Geoscience and Remote Sensing Magazine, vol. 10, no. 4, pp. 213–247, 2022. 2

[10] Pranav Jeevan and Amit Sethi, Which Backbone to Use: A Resource-efficient Domain Specific Comparison for Computer Vision, arXiv preprint arXiv:2406.05612, 2024. 2

[11] Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, and Liangpei Zhang, MTP: Advancing Remote Sensing Foundation Model via Multi-Task Pretraining, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, accepted for publication, 2024. 2

[12] Andrea Gesmundo and Jeff Dean, An Evolutionary Approach to Dynamic Introduction of Tasks in Large-scale Multitask Learning Systems, arXiv preprint arXiv:2205.12755, 2022. 1, 2

[13] Geoffrey E. Hinton, Oriol Vinyals, and Jeff Dean, Distilling the Knowledge in a Neural Network, arXiv preprint arXiv:1503.02531, 2015. 1, 2, 4, 5

[14] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, arXiv preprint arXiv:1704.04861, 2017. 1

[15] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc V. Le, Searching for MobileNetV3, in IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1314–1324, 2019. 6

[16] Jie Hu, Li Shen, and Gang Sun, Squeeze-and-Excitation Networks, in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141, 2018. 6

[17] Yi Yang and Shawn Newsam, A Bag-of-Visual-Words Model for Land-Use Classification Using High-Resolution Remote Sensing Images, in IEEE International Conference on Geoscience and Remote Sensing Symposium, pp. 1–5, 2010.

[18] Saurabh Prasad, Lee M. Bruce, and James E. Fowler, Decision Fusion with Confidence-Based Weight Assignment for Hyperspectral Target Recognition, IEEE Transactions on Geoscience and Remote Sensing, vol. 46, no. 5, pp. 1448–1456, 2008. 1

[19] Saurabh Prasad, Melba M. Crawford, and Joydeep Ghosh, Co-training weakly labeled data in supervised neural networks: A case study in hyperspectral data classification, IEEE Journal of Selected Topics in Signal Processing, vol. 5, no. 3, pp. 551–560, 2011. 1, 2

[20] Mathieu Fauvel, Yuliya Tarabalka, Jon Atli Benediktsson, Jocelyn Chanussot, and Johannes R. Sveinsson, Advances in Spectral-Spatial Classification of Hyperspectral Images, Proceedings of the IEEE, vol. 101, no. 3, pp. 652–675, 2013. 1

[21] Jocelyn Chanussot, Pierre Comon, Pierre Chevalier, and François Chevalier, An End-to-End Model-Based Approach for the Joint Detection and Classification of High-Resolution Images, IEEE Transactions on Image Processing, vol. 29, pp. 8346–8358, 2020. 2

[22] Begüm Demir and Lorenzo Bruzzone, A Novel Active Learning Method in Relevance Feedback for Content-Based Remote Sensing Image Retrieval, IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 5, pp. 2323–2334, 2015. 1, 2

[23] Begüm Demir and S. Ertürk, Hyperspectral Image Classification Using a Spectral Graph Approach, IEEE Geoscience and Remote Sensing Letters, vol. 11, no. 1, pp. 297–301, 2013.

[24] Biplab Banerjee and Subhasis Chaudhuri, Hyperspectral Image Classification Using Integrated Active Learning and Synergy-Based Multi-View Uncertainties, Remote Sensing of Environment, vol. 195, pp. 89–103, 2017. 1

[25] Biplab Banerjee and Subhasis Chaudhuri, Classification of Hyperspectral Data with Nonlinear Manifold Learning and Active Learning, IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 7, pp. 5498–5513, 2021. 1

[26] Danfeng Hong, Naoto Yokoya, and Xiao Xiang Zhu, Learnable Manifold Alignment with Subspace Priors for Multimodal Remote Sensing Data Classification, ISPRS Journal of Photogrammetry and Remote Sensing, vol. 155, pp. 176–190, 2019. 1

[27] Danfeng Hong, Jocelyn Chanussot, Naoto Yokoya, and Xiao Xiang Zhu, An Augmented Linear Mixing Model to Address Spectral Variability for Hyperspectral Unmixing, IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 1923–1938, 2019. 2

[28] Xiangling Ding and Yuchen Nie, Spectral information guidance network for tampering localization of high-resolution satellite map, Expert Systems with Applications, vol. 264, p. 125825, 2025. 2

[29] Xiangling Ding, Yuchen Nie, Jizhou Yao, Jia Tang, and Yubo Lang, Forensic research of satellite images forgery: a comprehensive survey, Artificial Intelligence Review, vol. 57, p. 253, 2024. 2

[30] Jingwei Wang, Ziyue Zhu, Chunxiao Liu, Rong Li, and Xin Wu, LLM-Enhanced multimodal detection of fake news, PLOS ONE, 2024. 1

[31] S. Kalyan Yarlagadda, David Guera, Paolo Bestagini, Fuwen Zhu, Stefano Tubaro, and Edward Delp, Satellite image forgery detection and localization using GAN and one-class classifier, in Proceedings of the IS&T International Symposium on Electronic Imaging, vol. 2018, pp. 214–1–214–9, Burlingame, CA, 2018. 1

[32] Jialu Sui, Ding Ma, C.-C. Jay Kuo, and Man-On Pun, FLDCF: A Collaborative Framework for Forgery Localization and Detection in Satellite Imagery, IEEE Transactions on Geoscience and Remote Sensing, vol. 62, 2024. 2

[33] Ruijie Wu, Wei Guo, Yi Liu, and Chenhao Sun, High-Precision Heterogeneous Satellite Image Manipulation Localization: Feature Point Rules and Semantic Similarity Measurement, Remote Sensing, vol. 16, no. 19, p. 3719, 2024. 1

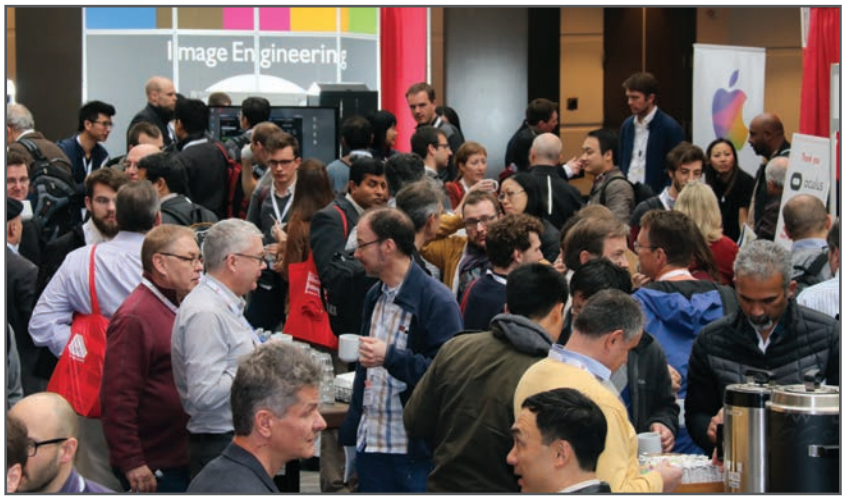
[34] Lydia Abady, Giovanna Maria Dimitri, and Mauro Barni, A One-Class Classifier for the Detection of GAN Manipulated Multi-Spectral Satellite Images, Remote Sensing, vol. 16, no. 5, p. 781, 2024.

[35] E. D. Cannas, P. Beaus, P. Bestagini, F. Marques, and S. Tubaro, A One-Class Approach to Detect Super-Resolution Satellite Imagery with Spectral Features, in ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024. 1

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org