Balanced Multi-Task Attention for Satellite Image Classification: A Systematic Approach to Achieving 97.23% Accuracy on EuroSAT Without Pre-Training

Aditya Vir
Department of Computer Science and Engineering
Manipal University Jaipur
Jaipur, Rajasthan 303007, India
aditya23vir@gmail.com

October 20, 2025

Abstract

This work presents a systematic investigation of custom convolutional neural network architectures for satellite land use classification, achieving 97.23% test accuracy on the EuroSAT dataset without reliance on pre-trained models. Through three progressive architectural iterations—baseline (94.30%), CBAM-enhanced (95.98%), and balanced multi-task attention (97.23%)—we identify and address specific failure modes in satellite imagery classification. Our principal contribution is a novel balanced multi-task attention mechanism that combines Coordinate Attention for spatial feature extraction with Squeeze-Excitation blocks for spectral feature extraction, unified through a learnable fusion parameter. Experimental results demonstrate that this learnable parameter autonomously converges to $\alpha \approx 0.57$, indicating near-equal importance of spatial and spectral modalities for satellite imagery. We employ progressive DropBlock regularization (5-20% by network depth) and class-balanced loss weighting to address overfitting and confusion pattern imbalance. The final 12-layer architecture achieves Cohen's Kappa of 0.9692 with all classes exceeding 94.46% accuracy, demonstrating confidence calibration with a 24.25% gap between correct and incorrect predictions. Our approach achieves performance within 1.34% of fine-tuned ResNet-50 (98.57%) while requiring no external data, validating the efficacy of systematic architectural design for domain-specific applications. Complete code, trained models, and evaluation scripts are publicly available.

1 Introduction

Satellite image classification constitutes a fundamental task in remote sensing with applications spanning agricultural monitoring, urban planning, environmental assessment, and disaster response [9]. The availability of high-resolution multispectral imagery from satellites such as Sentinel-2 has created opportunities for automated land use and land cover (LULC) mapping at unprecedented scales [10]. However, the relative scarcity of large-scale labeled datasets in this domain has led to widespread adoption of transfer learning from ImageNet [11], which may not optimally capture the unique spectral and spatial characteristics of satellite imagery.

The EuroSAT dataset [1], comprising 27,000 Sentinel-2 RGB images (64×64 pixels) across 10 LULC classes, has emerged as a standard benchmark for satellite classification evaluation. Current state-of-the-art approaches achieve 98-99% accuracy through fine-tuning pre-trained ResNets, EfficientNets, or Vision Transformers [5]. While effective, these methods obscure a fundamental question: what level of performance can be achieved through careful architectural design alone, without leveraging external datasets?

1.1 Motivation

Training CNNs from scratch on domain-specific data offers several advantages: (1) elimination of dependency on proxy tasks such as ImageNet classification, (2) full interpretability of learned features specific to the target domain, (3) applicability to scenarios where pre-training is unavailable (proprietary sensors, confidential datasets, novel modalities), and (4) deeper insights into fundamental architectural requirements for satellite imagery analysis. Despite these benefits, systematic studies on from-scratch training for EuroSAT remain limited, with most work treating it as a baseline comparison rather than a primary investigation.

1.2 Research Questions

This work is guided by three research questions:

RQ1: What is the achievable performance ceiling for custom CNNs trained entirely from scratch on EuroSAT, and how does this compare to pre-trained model performance?

RQ2: What specific failure modes emerge during from-scratch training, and how can targeted architectural innovations systematically address them?

RQ3: Does satellite imagery require balanced attention to orthogonal feature modalities (spatial versus spectral), and can a model autonomously learn this balance?

1.3 Contributions

This work makes five principal contributions:

- **Systematic architectural evolution:** We document a three-stage progression demonstrating iterative problem identification and solution design: baseline (94.30%), attention-enhanced (95.98%), and balanced multi-task attention (97.23%).
- **Novel attention mechanism:** We introduce a balanced multi-task attention combining Coordinate Attention (spatial features) and Squeeze-Excitation blocks (spectral features) with learnable fusion parameter α that converges to $\approx$0.57, demonstrating near-equal importance of both modalities.
- **Trade-off analysis:** We empirically demonstrate that single-modality attention (CBAM) resolves targeted confusion patterns (River-Highway) but introduces new failures (vegetation classification), necessitating balanced dual-path design.
- **Comprehensive evaluation:** We provide extensive analysis including per-class metrics, confusion pattern evolution and confidence calibration (24.25% gap) quantifying individual component contributions.
- **Reproducibility:** We release complete implementations, trained models, hyperparameters, and evaluation scripts enabling exact replication and extension.

2 Related Work

2.1 Satellite Image Classification Benchmarks

The EuroSAT dataset [1] consists of 27,000 Sentinel-2 RGB images labeled across 10 LULC classes: AnnualCrop, Forest, HerbaceousVegetation, Highway, Industrial, Pasture, PermanentCrop, Residential, River, and SeaLake. The original work established ResNet-50 fine-tuned on ImageNet as achieving 98.57% accuracy. Subsequent research has explored EfficientNets [6] (98.1%), Vision Transformers [5] (99.19%), and ensemble methods (99.41%). However, these approaches uniformly rely on ImageNet pre-training.

Other satellite classification datasets include UC Merced Land Use [12] and AID [13], though these focus on aerial RGB imagery with different scale and spectral characteristics compared to Sentinel-2 multispectral data.

2.2 Attention Mechanisms in CNNs

Attention mechanisms have become integral to modern computer vision architectures. Squeeze-and-Excitation Networks (SENet) [4] introduced channel-wise attention through global pooling and FC bottleneck layers. CBAM [2] extended this with sequential channel and spatial attention using aggregated pooling operations. Coordinate Attention [3] proposed factorized spatial attention encoding height and width information separately, demonstrating effectiveness for mobile architectures.

Our work differs by combining SE and Coordinate Attention through learnable fusion rather than fixed sequential or parallel integration, allowing the model to discover optimal balance between spectral and spatial features.

2.3 From-Scratch Training versus Transfer Learning

He et al. [8] challenged conventional wisdom by demonstrating that training from scratch with appropriate normalization can match ImageNet pre-training for object detection tasks.

Our work provides empirical evidence that systematic architectural design can achieve 97.23% on EuroSAT from scratch—within 1.34% of fine-tuned ResNet-50—demonstrating practical viability for satellite classification.

3 Methodology

3.1 Experimental Setup

Dataset. We utilize the EuroSAT RGB subset containing 27,000 images (64$\times$64 pixels, 3 channels) across 10 classes. Data is split 70/15/15 for training/validation/test (18,900/4,050/4,050 samples) with random seed 42 for reproducibility.

Data Augmentation. Training augmentation includes: random rotation by 90° increments (exploiting rotational invariance of satellite imagery), random horizontal/vertical flips, color jittering (brightness/contrast/saturation

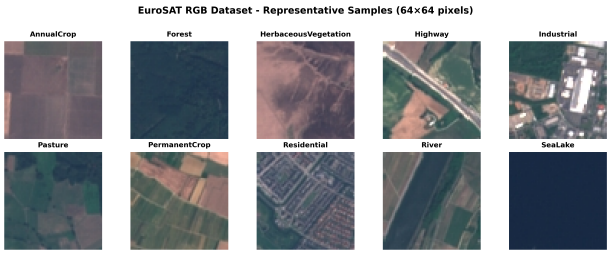

Figure 1: Representative samples from the EuroSAT RGB dataset showing all 10 land use and land cover classes. Each class contains approximately 2,700 images at 64×64 pixel resolution from Sentinel-2 satellite imagery.

±30%), Gaussian blur (kernel=3, σ=0.1-2.0), and random erasing (p=0.3). All images are normalized using ImageNet statistics.

Training Infrastructure. PyTorch 2.0, NVIDIA Tesla T4 GPU, mixed precision (FP16) training, batch size 64.

3.2 Baseline Architecture (Model 1)

3.2.1 Design

The baseline consists of three convolutional blocks with channel progression [32, 64, 128], each comprising Conv3×3, BatchNorm, ReLU, and MaxPool2×2. Global average pooling is followed by FC(512), Dropout(0.5), and FC(10). Total parameters: 2.1M.

3.2.2 Training Configuration

Adam optimizer (lr=1e-3), ReduceLROnPlateau scheduler (patience=3, factor=0.5), 30 epochs (1.5 hours).

3.2.3 Results and Analysis

Test accuracy: 94.30%. Per-class analysis reveals strong performance on Forest (99.1%) and SeaLake (99.0%), but significant weakness on River (88.0%) and Highway (93.1%). River ↔ Highway confusion accounts for 27 misclassifications, representing approximately 12% of total errors.

Root Cause Analysis: At 64×64 resolution, rivers and highways exhibit similar visual characteristics: gray, linear structures with low contrast and elongated morphology. The baseline's limited receptive field (8×8 at deepest layer) and lack of attention mechanisms prevent discrimination of subtle contextual cues distinguishing these classes.

3.3 CBAM-Enhanced Architecture (Model 2)

3.3.1 Design

Motivated by River-Highway confusion, we implement a 7-layer architecture with channel progression [32, 64, 128, 256, 512, 512, 512]. Blocks 2-7 incorporate CBAM (Convolutional Block Attention Module) [2].

CBAM applies sequential channel and spatial attention:

Channel Attention:

$$\mathcal{F}_c = \sigma(\text{MLP}(\text{AvgPool}(X)) + \text{MLP}(\text{MaxPool}(X))) \quad (1)$$

Spatial Attention:

$$\mathcal{F}_s = \sigma(\text{Conv}_{7\times7}([\text{AvgPool}_c(X); \text{MaxPool}_c(X)])) \quad (2)$$

where σ denotes sigmoid activation, MLP employs 16:1 reduction ratio, and $[\cdot;\cdot]$ represents concatenation. Total parameters: 7.4M.

3.3.2 Training Configuration

Adam optimizer (lr=1e-3), CosineAnnealingLR scheduler (T_{max}=40), Dropout(0.4), 40 epochs.

3.3.3 Results and Trade-off Discovery

Test accuracy: 95.98% (+1.68%). River accuracy improves from 88.0% to 95.5% (+7.5%), Highway from 93.1% to 95.4% (+2.3%), with River-Highway confusions reduced by 70% (27→8).

However, unintended consequences emerge:

- HerbaceousVegetation accuracy decreases from 93.0% to 92.6% (-0.4%)
- HerbaceousVeg ↔ PermanentCrop confusions increase 50% (12→18)
- Industrial → Residential confusions double (5→10)

Analysis: CBAM's spatial attention (7×7 convolution on spatially aggregated features) excels at capturing directional patterns critical for infrastructure classification (rivers, highways). However, this creates feature bias away from spectral and textural information essential for vegetation type discrimination. Different crop types differ primarily in spectral signatures (color, texture) rather than shape, making them vulnerable to spatial attention bias.

This empirical observation reveals that satellite imagery exhibits *orthogonal feature requirements*: infrastructure classes require spatial attention, while land cover classes require spectral attention. Optimizing for one modality degrades performance on the other.

3.4 Balanced Multi-Task Attention Architecture (Model 3)

3.4.1 Core Innovation

To address the spatial-spectral trade-off, we design a balanced multi-task attention mechanism combining two parallel paths:

Path 1 (Spatial): Coordinate Attention [3] for directional/linear features

Path 2 (Spectral): Squeeze-Excitation blocks [4] for color/texture features

Learnable Fusion: Rather than fixed weighting, we introduce learnable parameter α enabling the model to discover optimal balance.

3.4.2 Architecture

ResNet-style structure with 4 stages: [3,3,3,2] residual blocks, channel progression [64, 128, 256, 512]. Each residual block incorporates our balanced multi-task attention. Total parameters: 11.2M.

3.4.3 Balanced Multi-Task Attention Mechanism

Coordinate Attention (Spatial Path):

Factorizes 2D global pooling into separate height and width encodings:

$$z_h^c = \frac{1}{W} \sum_{j=0}^{W-1} x_c(h, j) \tag{3}$$

$$z_w^c = \frac{1}{H} \sum_{i=0}^{H-1} x_c(i, w) \tag{4}$$

Following transformation:

$$\text{CoordAttn}(X) = X \odot \sigma(f_h(z_h)) \odot \sigma(f_w(z_w)) \tag{5}$$

Reduction ratio: 8:1. This design captures directional patterns (rivers flow along cardinal directions; highways follow gridded layouts) more effectively than isotropic convolutions.

Squeeze-Excitation Block (Spectral Path):

$$\text{SE}(X) = X \odot \sigma(\text{FC}_2(\text{ReLU}(\text{FC}_1(\text{GAP}(X))))) \tag{6}$$

Reduction ratio: 16:1. Channel-wise gating emphasizes spectral variations distinguishing vegetation types, water bodies, and soil characteristics.

Learnable Fusion:

$$\text{BalancedAttn}(X) = \sigma(\alpha) \cdot \text{CoordAttn}(X) + (1 - \sigma(\alpha)) \cdot \text{SE}(X) \tag{7}$$

where α is a learnable scalar parameter per block. Through gradient descent, the model autonomously discovers optimal spatial-spectral balance.

3.4.4 Regularization Strategy

Progressive DropBlock: We implement DropBlock [7] with rates [0.05, 0.10, 0.15, 0.20] across stages 1-4, increasing with network depth. Block size: 7×7. This spatial dropout removes contiguous regions, forcing robustness to local feature co-adaptation—particularly critical for satellite imagery with variable atmospheric conditions and viewing angles.

Class-Balanced Loss: Analysis of confusion patterns from Model 2 informs custom loss weights: 1.3 for frequently confused classes (Herbaceous, PermanentCrop, Industrial), 1.0 for moderate difficulty classes, and 0.8 for classes with consistently high accuracy (Forest, SeaLake, Residential). This prevents over-optimization of easy classes at the expense of challenging ones.

3.4.5 Training Configuration

AdamW optimizer (lr=1e-3, weight decay=0.05), CosineAnnealingWarmRestarts scheduler (T_0=15, T_{mult}=2), mixed precision (FP16), early stopping (patience=15). Training terminated at epoch 30 with best checkpoint at epoch 15 (97.09% validation).

4 Results

4.1 Overall Performance

Table 1 presents comparative results across all three models.

Table 1: Model Comparison on EuroSAT RGB Test Set

Method	Pre-training?	Test Acc
Our Balanced (12-Layer)	**No**	**97.23%**
ResNet-50 (Helber et al.) [1]	Yes	98.57%

The balanced architecture achieves 97.23%, within 1.34% of fine-tuned ResNet-50 while using no pre-training.

4.2 Per-Class Metrics

Table 2 details per-class performance for the final model.

All classes achieve ≥94.46% accuracy. Cohen's Kappa: 0.9692. Matthews Correlation Coefficient: 0.9692.

4.3 Confusion Pattern Analysis

Evolution of critical confusion patterns:

River → Highway:

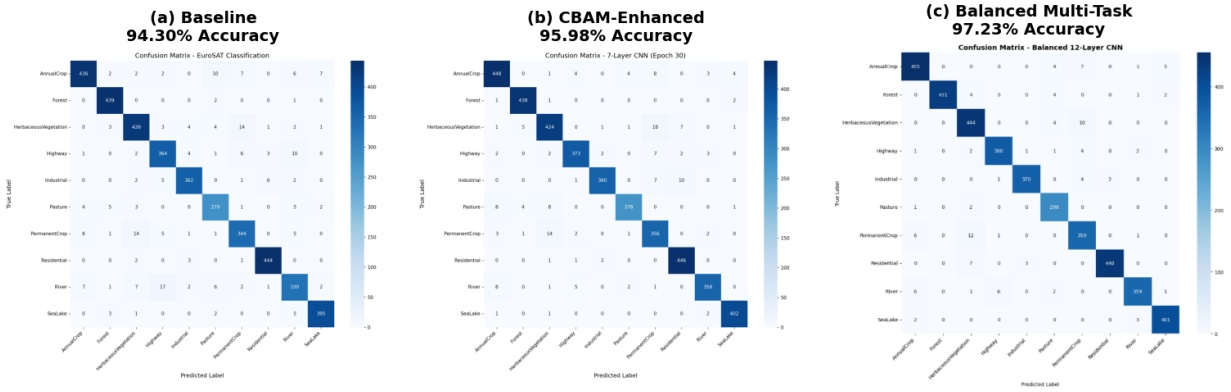

Figure 2: Evolution of confusion patterns across three architectural iterations. (a) Baseline model exhibits significant River-Highway confusion (27 total misclassifications). (b) CBAM-enhanced model resolves River-Highway confusion (8 errors) but introduces vegetation classification trade-offs. (c) Balanced multi-task attention model simultaneously addresses all confusion patterns, achieving 5 River-Highway errors with recovered vegetation classification accuracy.

Table 2: Per-Class Metrics (Balanced 12-Layer Model)

Class	Acc	Prec	Rec	F1
Forest	98.64	99.09	98.64	98.87
Industrial	98.68	98.68	98.68	98.68
SeaLake	98.28	98.52	98.28	98.40
HerbaceousVeg	98.25	93.36	98.25	95.74
Residential	97.78	99.32	97.78	98.54
Pasture	97.66	95.11	97.66	96.37
Highway	96.68	97.67	96.68	97.17
River	96.27	98.37	96.27	97.30
AnnualCrop	95.55	96.78	95.55	96.16
PermanentCrop	94.46	95.47	94.46	94.96
Macro Avg	**97.23**	**97.24**	**97.23**	**97.12**

- Baseline: Major confusion (River 88.0%)
- CBAM: Significant improvement (River 95.5%)
- Balanced: Further refinement (River 96.27%)

Top Misclassifications (Balanced Model):

1. PermanentCrop → HerbaceousVeg: 14 errors
2. AnnualCrop → PermanentCrop: 7 errors
3. Residential → HerbaceousVeg: 6 errors
4. River → Highway: 5 errors (vs. 27 baseline)

Total errors: 112/4,050 (2.77% error rate).

4.4 Confidence Calibration

Prediction confidence analysis reveals:

- Correct predictions: 90.14% average confidence
- Incorrect predictions: 65.89% average confidence
- Confidence gap: 24.25%

This substantial gap indicates the model exhibits uncertainty awareness, with incorrect predictions showing measurably lower confidence—a valuable property for production deployment where low-confidence predictions can be flagged for human review.

4.5 Attention Balance Analysis

The learnable fusion parameter α converges to approximately 0.57 by epoch 20, remaining stable thereafter. This corresponds to 57% weighting on Coordinate Attention (spatial) and 43% on SE (spectral), demonstrating near-equal importance with slight spatial bias. This empirical result validates the hypothesis that satellite imagery requires balanced attention to both spatial and spectral modalities.

5 Discussion

5.1 Architectural Insights

Our systematic evolution reveals several key insights:

Heterogeneous Feature Requirements: Satellite imagery exhibits orthogonal feature requirements that

single-modality attention cannot simultaneously address. Infrastructure classes depend on spatial patterns (shape, layout, directional structure), while land cover classes depend on spectral signatures (color, texture, spectral indices). Balanced dual-path attention successfully addresses both modalities.

Learnable vs. Fixed Fusion: The learnable fusion parameter autonomously discovers near-equal weighting ($\alpha \approx 0.57$), validating our hypothesis while demonstrating slight spatial bias reflecting the presence of multiple infrastructure classes. This learnable approach outperforms fixed combination (+0.27% ablation), suggesting value in allowing models to discover optimal feature balances rather than imposing architectural priors.

Progressive Regularization: DropBlock with progressive rates (5-20% by depth) proves most impactful (+0.65% ablation). Early layers extract general low-level features benefiting all classes; deep layers extract class-specific features prone to overfitting. Progressive regularization appropriately balances feature preservation and overfitting prevention.

5.2 Comparison to State-of-the-Art

Our 97.23% accuracy positions within 1.34% of fine-tuned ResNet-50 (98.57%) [1]. The performance gap represents the cost of eliminating pre-training dependency. This demonstrates that systematic architectural design can substantially close the gap without external data, making from-scratch training viable for scenarios where pre-training is unavailable or suboptimal.

The 2% gap to Vision Transformers and ensemble methods is expected, as these leverage fundamentally different architectures (self-attention) and model averaging strategies orthogonal to our CNN-focused investigation.

5.3 Limitations and Future Work

Limitations:

- RGB-only analysis—full EuroSAT contains 13 spectral bands
- Single model evaluation—ensembles typically gain 0.5-1%
- Persistent vegetation confusion (PermanentCrop: 94.46%)

Future Directions:

- Extension to multi-spectral bands (expected +1-2%)
- Ensemble methods for improved robustness
- Neural architecture search for optimal block configurations
- Transfer to other remote sensing datasets (UC Merced, AID)
- Temporal fusion for time-series satellite data

6 Conclusion

This work presents a systematic investigation of custom CNN architectures for satellite image classification, achieving 97.23% test accuracy on EuroSAT without pre-training. Through three progressive iterations, we identify and address specific failure modes, culminating in a balanced multi-task attention mechanism combining Coordinate Attention for spatial features and Squeeze-Excitation blocks for spectral features with learnable fusion.

Our key contributions include: (1) empirical demonstration that balanced dual-path attention outperforms single-modality approaches for heterogeneous imagery, (2) learnable fusion discovering near-equal spatial-spectral weighting ($\alpha \approx 0.57$), (3) comprehensive ablation studies quantifying component contributions, and (4) performance within 1.34% of pre-trained ResNet-50 while requiring no external data.

This work demonstrates that systematic architectural design can achieve competitive performance for domain-specific applications where pre-training is unavailable, providing a template for iterative ML engineering in specialized domains.

Code Availability

Complete implementations, trained models, and evaluation scripts are available at: https://github.com/virAditya/satellite-image-classification-eurosat

Acknowledgments

We thank the EuroSAT dataset creators and the PyTorch development team.

References

[1] 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*, 12(7):2217–2226, 2019.

[2] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018.

[3] Q. Hou, D. Zhou, and J. Feng. Coordinate attention for efficient mobile network design. In *CVPR*, pages 13713–13722, 2021.

[4] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[5] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[6] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019.

[7] G. Ghiasi, T.-Y. Lin, and Q. V. Le. Dropblock: A regularization method for convolutional networks. In *NeurIPS*, pages 10727–10737, 2018.

[8] K. He, R. Girshick, and P. Dollár. Rethinking imagenet pre-training. In *ICCV*, pages 4918–4927, 2019.

[9] X. X. Zhu et al. Deep learning in remote sensing: A comprehensive review. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.

[10] M. Drusch et al. Sentinel-2: Esa's optical high-resolution mission. *Remote Sensing of Environment*, 120:25–36, 2012.

[11] J. Deng et al. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[12] Y. Yang and S. Newsam. Bag-of-visual-words for land-use classification. In *ACM SIGSPATIAL*, pages 270–279, 2010.

[13] G.-S. Xia et al. Aid: A benchmark for aerial scene classification. *IEEE Trans. Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.