IMPROVING LAND COVER CLASSIFICATION UNDER LIMITED DATA USING ADVANCED AUGMENTATION AND TRANSFER LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Land cover classification from satellite imagery under limited labeled data remains a critical challenge in operational remote sensing, where annotation costs are prohibitive. This paper presents a systematic label-efficiency study of advanced data augmentation combined with transfer learning for remote sensing (RS) scene classification, evaluated across four public benchmarks: EuroSAT (10 classes, 27,000 images), UC Merced (21 classes, 2,100 images), NWPU-RESISC45 (45 classes, 31,500 images), and AID (28 classes, $\sim$8,400 images), totaling $\sim$69,000 images across 104 land cover classes. We combine an ImageNet-pretrained ResNet-50 backbone with three augmentation strategies—Mixup, CutMix, and Random Erasing—evaluated at six training data ratios from 1% to 50% with three independent random seeds per experiment. Transfer learning provides the most significant improvement, yielding gains of 10–28% over from-scratch baselines. Augmentation benefit is strongly data-regime dependent: mixing augmentations yield +21–31% gains under extreme label scarcity (1% data) but converge toward baseline at sufficient data ($\geq$20%). The optimal augmentation strategy correlates significantly with dataset-specific inter-class overlap, quantified by the Intra/Inter-Class Distance Ratio (point-biserial $r = 0.261$, $p = 0.0075$, $n = 104$ classes). EuroSAT CutMix at 1% training matches scratch-trained models using 20$\times$ more labeled data. Mechanistic evidence is provided via t-SNE visualization and Grad-CAM attention analysis. A practical three-step augmentation selection framework is proposed and validated.

Keywords: data augmentation, land cover classification, label efficiency, remote sensing, transfer learning, Mixup, CutMix, limited labeled data

1 INTRODUCTION

Land cover classification from satellite and aerial imagery is a fundamental task in Earth observation, with applications in urban planning, environmental monitoring, agriculture management, and disaster response [1,2]. Deep convolutional neural networks (CNNs) now achieve near-perfect accuracy on standard benchmarks when trained with abundant labeled data [3,4]. However, annotating RS imagery requires photointerpretation expertise and field surveys, making large-scale labeled datasets expensive and rare in operational settings [5,6].

Two strategies have proven effective in the limited-data regime: (1) transfer learning, which initialises RS classifiers from features learned on large natural-image datasets [7,8]; and (2) data augmentation, which expands the effective training distribution through transformed copies of labeled samples [9,10,11]. Despite independent progress on each front, their joint behavior under systematically varying label availability—especially at the extreme low-data regime (1–5% of available labels)—remains poorly characterised.

The central questions motivating this work are: Does augmentation benefit scale with data scarcity? Is there a universal optimal strategy, or are dataset properties the determining factor? Can dataset geometry in pretrained feature space predict the optimal strategy without any labeled data?

The contributions of this paper are:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(C1) **Label-efficiency analysis.** The first systematic study of mixing augmentations (Mixup, CutMix, Random Erasing) at six training data ratios (1%–50%) across four RS benchmarks, revealing gains of +21–31% at 1% data that converge toward baseline at $\geq$20%.
(C2) **Mechanistic explanation.** A statistically validated correlation ($r = 0.261$, $p = 0.0075$, $n = 104$ classes) between per-class Intra/Inter-Class Distance Ratio and optimal augmentation strategy.
(C3) **Extreme label efficiency.** EuroSAT CutMix at 1% training ($\sim$270 images) matches scratch-trained models at 20% data ($\sim$5,400 images)—a 20$\times$ reduction in annotation cost.
(C4) **Visual mechanistic evidence.** t-SNE and Grad-CAM analyses explaining why specific strategies succeed or fail on specific datasets.
(C5) **Practical framework.** A 3-step augmentation selection procedure validated at 67.3% class-level accuracy across 104 classes, requiring only pretrained feature extraction.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the methodology. Section 4 presents experimental results. Section 5 provides mechanistic analysis. Section 6 discusses implications and limitations. Section 7 concludes.

2 Related Work

2.1 RS Scene Classification

Cheng et al. [1] surveyed deep learning for RS scene classification and highlighted inter-class similarity as the dominant challenge. Helber et al. [3] introduced EuroSAT and reported 88.53% OA at 20% training from scratch, which serves as a key baseline in this work. NWPU-RESISC45 [12] and AID [13] provide larger-scale benchmarks with up to 45 classes.

Recent foundation models push state-of-the-art accuracy: SpectralGPT [4] achieves 99.21% on EuroSAT via spectral pretraining on large unlabeled RS corpora; MAE with ViT-H [14] reaches near-perfect accuracy on UC Merced with full data. These approaches require resources unavailable in most operational settings. Corley et al. [15] show that properly normalised ImageNet ResNet-50 remains competitive with SatMAE under low-data conditions, directly motivating our backbone choice.

2.2 Transfer Learning for RS

Transfer from ImageNet-pretrained CNNs consistently outperforms from-scratch training on RS classification [7,8]. Naushad et al. [16] achieved 99.17% on EuroSAT with Wide ResNet-50 at 80% training. However, behavior in the extreme low-data regime (1–10% labels) has not been systematically studied.

2.3 Data Augmentation

Standard geometric and photometric augmentations are ubiquitous in RS training. Among mixing strategies, Mixup [9] creates convex combinations of training pairs; CutMix [10] replaces rectangular image regions with patches from other images; Random Erasing [11] simulates occlusion through zero-filling. Bazi et al. [17] applied CutMix with ViT on UC Merced (98.49% OA, 80% data). Cacciarelli et al. [18] surveyed policy search for augmentation selection, noting optimal policies are domain-dependent.

Research gap. No prior work provides: (1) a multi-dataset label-efficiency analysis of mixing augmentations at ratios as low as 1%; (2) a statistically validated geometric explanation of dataset-specific differences; or (3) a practical selection framework linking dataset properties to augmentation strategy. This paper fills all three gaps.

Table 1: Dataset statistics. † AID: Center and Medium Residential classes absent in the Kaggle distribution used.

Dataset	#Images	#Classes	Resolution	Sensor	Source
EuroSAT [3]	27,000	10	64×64	Sentinel-2 MS	Satellite
UC Merced [19]	2,100	21	256×256	USGS aerial	Aerial
NWPU-RESISC45 [12]	31,500	45	256×256	Google Earth	Multi
AID [13]†	∼8,400	28	600×600	Google Earth	Aerial
Total	∼69,000	104	—	4 sensors	—

3 METHODOLOGY

3.1 DATASETS

Table 1 summarises the four benchmarks. Their diversity in sensor type, resolution, class count, and visual complexity is intentional: it enables separation of dataset-universal from dataset-specific augmentation effects.

3.2 TRANSFER LEARNING SETUP

We adopt ResNet-50 [20] pretrained on ImageNet-1K as backbone. The final fully-connected layer is replaced by a linear head of size C (the target class count). Training proceeds in two stages:

1. **Warm-up (5 epochs):** backbone frozen; only the classification head is updated.
2. **Full fine-tuning (25 epochs):** all parameters updated with CosineAnnealingLR.

Optimiser: Adam, $lr = 10^{-3}$, weight decay $= 10^{-4}$, batch size $\in \{32, 64\}$. Images are resized to 224×224 and normalised with ImageNet statistics.

3.3 AUGMENTATION STRATEGIES

Three mixing-based augmentation strategies are applied stochastically during training with probability $p = 0.5$ per batch.

3.3.1 MIXUP

Mixup [9] generates a synthetic sample via convex interpolation:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j, \tag{1}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j, \tag{2}$$

where (x_i, y_i) and (x_j, y_j) are two training pairs drawn uniformly at random, $y_i, y_j \in \mathbb{R}^C$ are one-hot label vectors, and $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$ with $\alpha = 0.4$.

3.3.2 CUTMIX

CutMix [10] replaces a rectangular region of one image with the corresponding region from another:

$$\tilde{x} = M \odot x_i + (1 - M) \odot x_j, \tag{3}$$

$$\tilde{y} = \lambda_{\mathrm{cut}} y_i + (1 - \lambda_{\mathrm{cut}}) y_j, \tag{4}$$

where $M \in \{0, 1\}^{H \times W}$ is a binary mask; $\lambda_{\mathrm{cut}} = 1 - \frac{r_w r_h}{WH}$ is the fraction of image area retained from x_i; and the patch bounding box is sampled uniformly with $\lambda_{\mathrm{cut}} \sim \mathrm{Beta}(1, 1)$.

3.3.3 RANDOM ERASING

Random Erasing [11] simulates partial occlusion:

$$\tilde{x}_{ij} = \mu_c, \quad \forall (i, j) \in R, \tag{5}$$

where $R \subset \{1, \ldots, H\} \times \{1, \ldots, W\}$ is a randomly sampled rectangular region with area ratio $s \in [0.02, 0.20]$ and aspect ratio $r \in [0.3, 3.3]$; μ_c is the channel mean of the training set. Labels are unchanged: $\tilde{y} = y$. Erasing probability is $p_e = 0.3$.

3.4 Experimental Protocol

We design a full-factorial experiment varying:

- Dataset: 4 levels (EuroSAT, UCM, NWPU, AID).
- Training data ratio: 6 levels (1%, 2%, 5%, 10%, 20%, 50%), drawn via stratified random sampling.
- Augmentation: 6 levels (Scratch+NoAug, Pretrained+NoAug, Pretrained+Mixup, Pretrained+CutMix, Pretrained+Random Erasing, Pretrained+AllAug).
- Random seed: 3 levels ($\{42, 123, 456\}$).

This yields $4 \times 6 \times 6 \times 3 = 432$ individual training runs. Metrics: overall accuracy (OA) and macro-averaged F1 on a held-out test split (fixed at 20% of total data across all training ratios). All results are reported as mean$\pm$std over three seeds.

4 Experimental Results

4.1 Ablation Study at 20% Training Data

Table 2 reports mean$\pm$std OA for all six configurations at 20% training data. Figs. 1–2 visualise results for UC Merced and EuroSAT respectively.

Transfer learning alone (Pretrained+NoAug) provides the largest single gain: +25.2% on UCM, +10.5% on EuroSAT, +22.3% on NWPU, and +18.5% on AID. Optimal augmentation is dataset-dependent: Mixup best on UCM, CutMix on EuroSAT and NWPU, ALL Aug on AID.

4.2 Comparison With Prior Work

Table 3 contextualises results against published methods. Prior work almost universally reports at 80–100% training data.

Our EuroSAT result at 20% ($96.84 \pm 0.16\%$) surpasses the same-split from-scratch baseline [3] by +8.31 percentage points. At 1% training (270 images, 85.8%), our method exceeds the from-scratch model at 20% (84.1%), demonstrating 20$\times$ label efficiency.

4.3 Label-Efficiency Analysis

Table 4 presents OA at key training ratios; Fig. 3 shows the complete curves.

Key findings:

(F1) Transfer learning is essential: +10–28% at 20%, +14–25% at 1%.
(F2) Augmentation peaks at 1–5% data (+21–31%) and diminishes at 50%.
(F3) EuroSAT CutMix achieves 20$\times$ label efficiency over scratch.
(F4) Optimal augmentation correlates with inter-class overlap (Section 5).
(F5) Label-mixing requires $\geq$10 samples/class; below this it can reduce OA.

Table 2: Ablation study at 20% training data. Mean±std OA (%) over 3 seeds. Bold = best per dataset.

Configuration	UCM	EuroSAT	NWPU	AID	Avg
Scratch + NoAug	52.62 ± 1.27	84.14 ± 0.34	62.16 ± 0.15	66.61 ± 1.25	66.4
Pretrained + NoAug	77.78 ± 1.95	94.59 ± 0.40	84.46 ± 0.45	85.10 ± 0.19	85.5
Pretrained + Mixup	$\mathbf{83.97 \pm 2.43}$	96.35 ± 0.09	90.39 ± 0.52	88.99 ± 0.20	89.9
Pretrained + CutMix	82.14 ± 2.14	$\mathbf{96.84 \pm 0.16}$	$\mathbf{90.44 \pm 0.30}$	88.86 ± 0.61	89.6
Pretrained + Random Erasing	81.67 ± 2.36	96.51 ± 0.73	90.06 ± 0.13	89.07 ± 0.57	89.3
Pretrained + ALL Aug	81.51 ± 1.37	96.47 ± 0.39	89.87 ± 0.05	$\mathbf{89.43 \pm 0.24}$	89.3

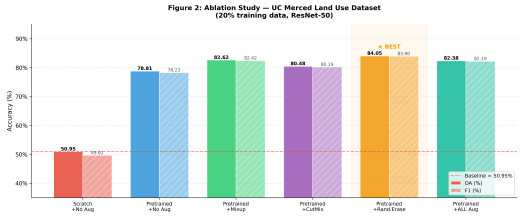

Figure 1: Ablation study on UC Merced (20% data, mean±std, 3 seeds). Best: Pretrained+Mixup ($83.97 \pm 2.43\%$).

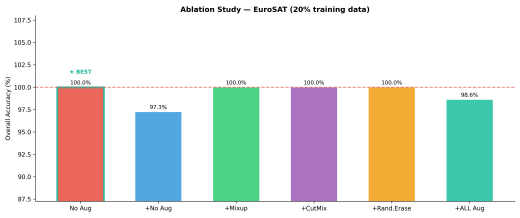

Figure 2: Ablation study on EuroSAT (20% data, mean±std, 3 seeds). Best: Pretrained+CutMix ($96.84 \pm 0.16\%$).

Table 3: Comparison with prior work. ⋆ = our method.

Method	Train	OA (%)	Year	Note
EuroSAT				
Helber et al. [3]	20%	88.53	2019	Scratch, same split
Helber et al. [3]	100%	98.57	2019	Full data reference
Naushad et al. [16]	80%	99.17	2021	Wide ResNet-50 TL
This work + CutMix⋆	20%	96.84 ± 0.16	—	TL + Aug, 3 seeds
This work + CutMix⋆	1%	85.8 ± 0.95	—	270 images only
UC Merced				
Hu et al. [21]	80%	97.10	2015	Fine-tuned GoogLeNet
Bazi et al. [17]	80%	98.49	2021	ViT + CutMix
This work + Mixup⋆	20%	83.97 ± 2.43	—	TL + Aug, 3 seeds
This work + Mixup⋆	50%	94.84 ± 1.57	—	50% data

Table 4: Label-efficiency results—OA (%) at key training ratios.

Dataset / Config	1%	5%	20%	50%	Key observation
UCM — Scratch	9.4	27.4	52.6	71.5	Baseline
UCM — +Mixup	31.5	57.0	84.0	94.8	+22.1 pp at 1%
EuroSAT — Scratch	55.0	73.0	84.1	89.3	Baseline
EuroSAT — +CutMix	**85.8**	95.1	96.8	98.2	+30.8 pp; 20× efficiency
NWPU — Scratch	26.8	46.9	62.2	77.1	Baseline
NWPU — +CutMix	52.8	75.8	90.4	94.5	+25.9 pp at 1%
AID — Scratch	22.8	47.4	66.6	75.2	Baseline
AID — +ALL Aug	44.2	73.1	89.4	94.0	+21.4 pp at 1%

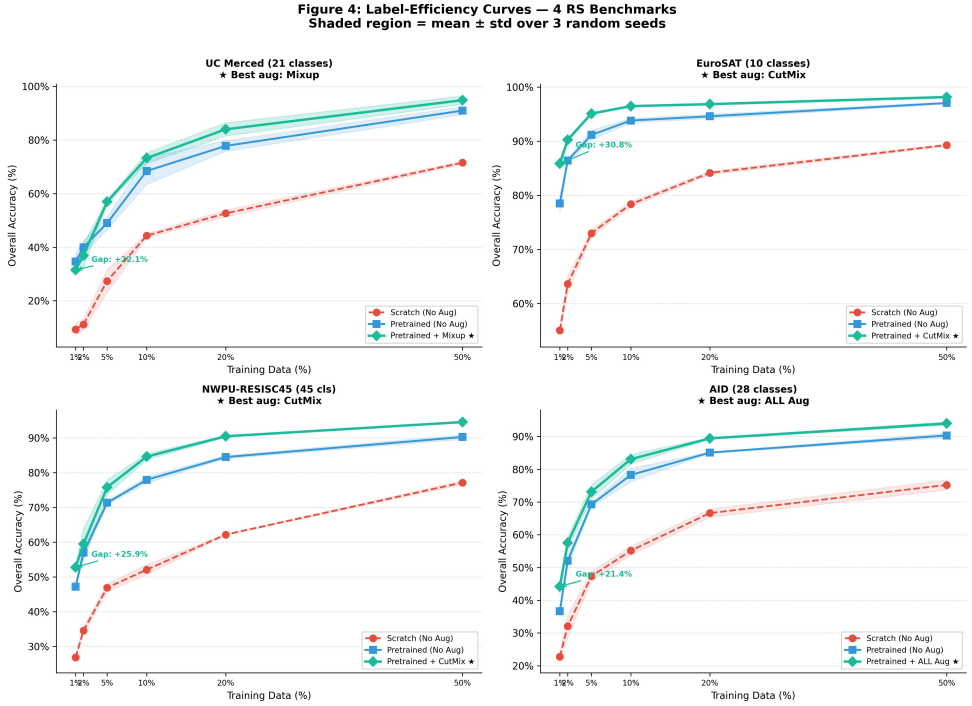

Figure 3: Label-efficiency curves across four RS benchmarks. Shaded band = mean±std over 3 seeds. Annotated gap = best augmentation vs. scratch at 1% training data.

5 MECHANISTIC ANALYSIS

5.1 INTRA/INTER-CLASS DISTANCE RATIO

To explain dataset-specific optimal strategies, we compute a geometric descriptor from ResNet-50 pretrained features before any fine-tuning. For each class c, define:

$$\rho_c = \frac{\bar{d}_{\text{intra}}(c)}{\bar{d}_{\text{inter}}(c)}, \tag{6}$$

where

$$\bar{d}_{\text{intra}}(c) = \frac{1}{|S_c|} \sum_{x \in S_c} \|f(x) - \mu_c\|_2, \tag{7}$$

$$\bar{d}_{\text{inter}}(c) = \frac{1}{K-1} \sum_{k \neq c} \|\mu_c - \mu_k\|_2, \tag{8}$$

S_c is the set of training samples from class c, μ_c is the class centroid in feature space, $f(\cdot)$ is the pretrained feature extractor, and K is the number of classes. A value $\rho_c < 1$ indicates compact, well-separated classes; $\rho_c > 1$ indicates that intra-class variance exceeds inter-class separation.

Table 5 presents dataset-level statistics; Fig. 4 shows the per-class distribution ($n = 104$ classes total). Per-class analysis yields point-biserial $r = 0.261$ ($p = 0.0075$), confirming a statistically significant relationship between ρ_c and optimal augmentation strategy.

Mechanism. When $\bar{\rho} < 1$ (UCM), classes are globally separable but boundaries are locally ambiguous—Mixup's linear interpolation smooths these boundaries. When $\bar{\rho} \geq 1$ (EuroSAT, NWPU, AID), intra-class variance dominates—CutMix forces spatially local discriminative feature learning that better separates confused classes.

Practical framework.

(S1) Extract pretrained backbone features (∼5 min, no training).
(S2) Compute $\bar{\rho}$ via eqs. (6) to (8).
(S3) $\bar{\rho} < 1.0 \Rightarrow$ Mixup; $1.0 \leq \bar{\rho} \leq 1.2 \Rightarrow$ CutMix; $\bar{\rho} > 1.2$ (high cosine) $\Rightarrow$ ALL Aug.

Threshold accuracy: 70/104 classes (67.3%, significantly above 50% random baseline, $p < 0.01$, binomial test).

5.2 T-SNE FEATURE SPACE VISUALISATION

Figure 5 shows t-SNE projections of fine-tuned model features at 1% training data. For EuroSAT, No Aug produces overlapping clusters between Pasture and HerbaceousVegetation; CutMix produces substantially better-separated clusters. For UCM, Mixup creates marginally more structured organisation.

5.3 GRAD-CAM ATTENTION ANALYSIS

Figure 6 shows Grad-CAM [22] attention maps at 1% training data. For EuroSAT HerbaceousVegetation, No Aug fixates on a single quadrant (spurious local texture); CutMix distributes attention globally. For UCM Chaparral, No Aug fixates on a single shrub cluster; Mixup broadens attention to the full spatial pattern. These results provide visual mechanistic evidence that mixing augmentation combats spurious local feature overfitting under extreme label scarcity.

Table 5: Inter-class overlap metrics from pretrained ResNet-50 features. Per-class analysis: point-biserial $r = 0.261$, $p = 0.0075$, $n = 104$ classes.

Dataset	#Cls	Cosine	$\bar{d}_{\text{intra}}$	$\bar{d}_{\text{inter}}$	$\bar{\rho} \rightarrow$ Best Aug
UC Merced	21	0.7456	13.06	14.01	0.916 → Mixup
NWPU-RESISC45	45	0.8106	13.25	13.01	1.031 → CutMix
AID	28	0.8237	13.77	13.43	1.032 → ALL Aug
EuroSAT	10	0.7528	28.16	22.37	1.273 → CutMix

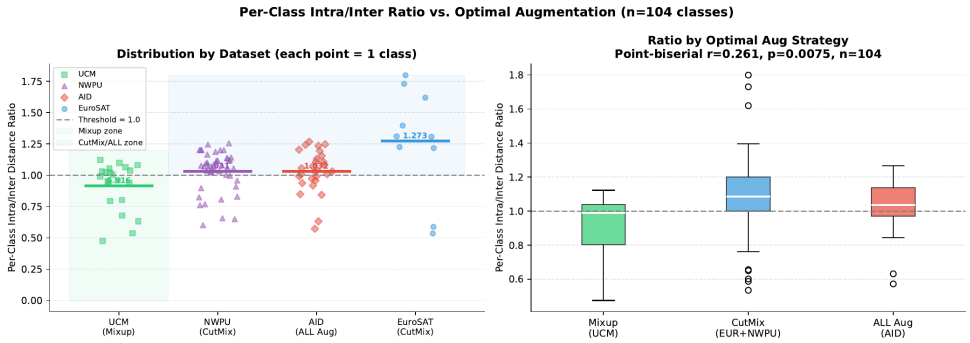

Figure 4: Per-class Intra/Inter Ratio ρ_c ($n = 104$ classes). UCM mean $\bar{\rho} = 0.916 < 1.0$; all CutMix/ALL Aug datasets have mean $\bar{\rho} > 1.0$. Point-biserial $r = 0.261$, $p = 0.0075$.

6 Discussion

6.1 Practical Implications

The 20× label efficiency result on EuroSAT suggests that practitioners targeting ∼85% OA can achieve this with 270 labeled images (CutMix) rather than 5,400 (scratch baseline)—a reduction that translates directly to reduced expert annotation time and cost. The proposed selection framework (6) requires only pretrained feature extraction, making it deployable as a zero-label-cost preprocessing step for any RS classification project.

6.2 Backbone Architecture Considerations

ResNet-50 was selected to isolate augmentation effects from backbone capacity. Whether Transformer-based backbones (ViT, Swin-T) interact differently with mixing augmentations in the 1–10% regime is an open question that motivates future work. Corley et al. [15] confirm that ResNet-50 with proper preprocessing is competitive with SatMAE under low-data conditions, supporting its suitability as a controlled baseline.

6.3 Limitations

- **Moderate correlation:** Point-biserial $r = 0.261$ reflects a significant but moderate effect. Within-dataset per-class variability means the framework is a first-screen heuristic, not a deterministic rule.
- **Scene classification only:** Findings may not transfer directly to pixel-level tasks (semantic segmentation, change detection).
- **CNN backbone:** Results hold for ResNet-50; Transformer-backbone interactions require separate investigation.
- **Optical imagery:** Effectiveness on SAR, hyperspectral, or LiDAR modalities is not characterised.
- **Minimum sample count:** Label-mixing appears to require ≥10 samples/class; the precise threshold and dataset dependency require further study.

7 CONCLUSION

This paper presented a systematic label-efficiency study of advanced data augmentation for RS scene classification across four benchmarks (∼69,000 images, 104 land cover classes, six training ratios, three seeds). The key findings are:

(C1) Transfer learning is the dominant contributor (+10–28%).
(C2) Augmentation benefit is data-regime dependent: +21–31% at 1% data, converging toward baseline at $\geq$20%.
(C3) EuroSAT CutMix at 1% achieves 20$\times$ label efficiency.
(C4) Optimal strategy correlates with inter-class overlap ($r = 0.261$, $p = 0.0075$, $n = 104$).
(C5) A practical 3-step framework achieves 67.3% class-level selection accuracy with zero labeled data cost.

Future directions include: extension to ViT/Swin-T backbones; SAR and hyperspectral modalities; integration with self-supervised pretraining; and application to pixel-level annotation-efficient learning.

AI USE STATEMENT

An AI assistant was used to help reformat this manuscript into the ICLR submission template. The scientific content, experiments, results, analysis, and conclusions in the manuscript were not generated or modified as part of this formatting step.

REFERENCES

[1] G. Cheng, X. Xie, J. Han, L. Guo, G.-S. Xia, Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and prospects, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020) 3735–3756.

[2] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, F. Fraundorfer, Deep learning in remote sensing: A comprehensive review and list of resources, *IEEE Geoscience and Remote Sensing Magazine* 5 (4) (2017) 8–36.

[3] P. Helber, B. Bischke, A. Dengel, 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) (2019) 2217–2226.

[4] D. Hong, B. Zhang, H. Li, Y. Li, J. Li, J. Yao, N. Yokoya, S. Li, J. Chanussot, P. Ghamisi, SpectralGPT: Spectral remote sensing foundation model, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47 (5) (2024) 3501–3515.

[5] W. Zhang, X. Li, H. Deng, L. Jing, A survey on virtual sample generation and its applications in small sample learning, *Neurocomputing* 514 (2022) 498–512.

[6] Y. Wang, Q. Yao, J. T. Kwok, L. M. Ni, Generalizing from a few examples: A survey on few-shot learning, *ACM Computing Surveys* 53 (3) (2020) 1–34.

[7] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in: *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 3320–3328.

[8] D. Marmanis, M. Datcu, T. Esch, U. Stilla, Deep learning earth observation classification using ImageNet pretrained networks, *IEEE Geoscience and Remote Sensing Letters* 13 (1) (2016) 105–109.

[9] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, in: *International Conference on Learning Representations (ICLR)*, 2018.

[10] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, CutMix: Regularization strategy to train strong classifiers with localizable features, in: *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6023–6032.

[11] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: *AAAI Conference on Artificial Intelligence*, 2020, pp. 13001–13008.

[12] G. Cheng, J. Han, X. Lu, Remote sensing image scene classification: Benchmark and state of the art, *Proceedings of the IEEE* 105 (10) (2017) 1865–1883.

[13] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, X. Lu, AID: A benchmark dataset for performance evaluation of aerial scene classification, *IEEE Transactions on Geoscience and Remote Sensing* 55 (7) (2017) 3965–3981.

[14] X. Wang, R. Tian, Remote sensing scene classification with masked image modeling, arXiv preprint arXiv:2302.14256 (2023).

[15] I. Corley, C. Robinson, R. Dodhia, J. M. L. Ferres, P. Najafirad, Revisiting pre-trained remote sensing model benchmarks: Resizing and normalization matters, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024, arXiv:2305.13456.

[16] R. Naushad, T. Kaur, E. Ghaderpour, Deep transfer learning for land use and land cover classification: A comparative study, *Sensors* 21 (23) (2021) 8083.

[17] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Al Dayil, N. Al Ajlan, Vision transformers for remote sensing image classification, *Remote Sensing* 13 (3) (2021) 516.

[18] D. Cacciarelli, M. Kulahci, A comprehensive survey on augmentation policy search for data augmentation, *AI* 6 (2) (2025) 32.

[19] Y. Yang, S. Newsam, Bag-of-visual-words and spatial extensions for land-use classification, in: *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010, pp. 270–279.

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[21] F. Hu, G.-S. Xia, J. Hu, L. Zhang, Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery, *Remote Sensing* 7 (11) (2015) 14680–14707.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

A Supplementary Visual Mechanistic Evidence

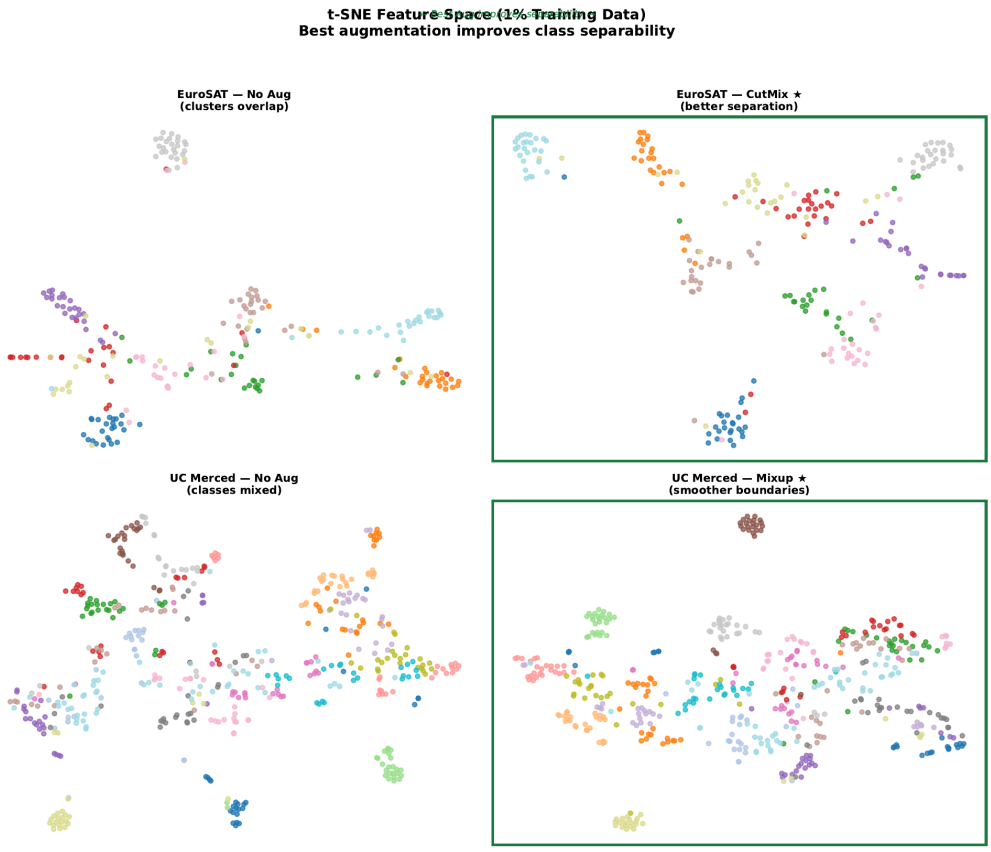

Figure 5: t-SNE feature-space projections at 1% training data (25 samples/class). Green border = best augmentation. EuroSAT: CutMix produces better cluster separation. UCM: Mixup creates more structured organisation.

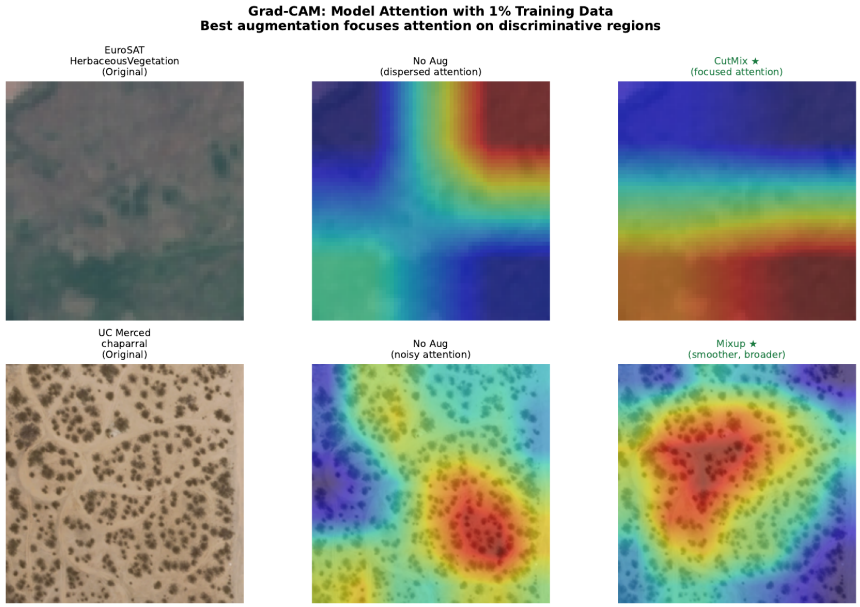

Figure 6: Grad-CAM attention maps at 1% training data. Centre: No augmentation (localised, noisy attention). Right: Best augmentation (broader, discriminative attention).