ONTOLOGY-GUIDED SEMANTIC MIXING WITH ADAPTIVE LORA EXPERTS FOR LONG-TAILED REMOTE SENSING RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning from long-tailed data remains an obstacle for recognition systems, because rare categories carry the greatest decision value but provide the weakest training signal. However, correcting class imbalance alone does not ensure that augmented samples preserve domain-specific semantics. Mixing-based augmentation guided by generic vision-language embeddings selects partners by linguistic co-occurrence rather than domain semantics, injecting semantic noise where data are scarcest. We propose a knowledge-guided learning framework that couples an explicit symbolic knowledge source with parameter-efficient adaptation. Ontology-Guided Semantic Mixing (OGSM) transforms frozen CLIP text embeddings into ontology-consistent representations through triplet metric learning over expertcurated relations, producing a similarity matrix that restricts augmentation to semantically admissible class pairs. Bag of LoRAs (BAGL) then supplies the discriminative capacity the enriched distribution demands. The number of low-rank experts is derived in closed form from the class count and the imbalance factor, which removes per-dataset hyperparameter search, and per-expert logit tempering is optimized by a secondary objective that jointly minimizes mean task loss and inter-expert loss variance. Because the knowledge component is precomputed and architecture-agnostic, it transfers across tasks without modification. On four long-tailed land-cover benchmarks the framework attains 91.1% average accuracy with consistent tail-class gains over the strongest baseline, and OGSM alone improves tail mIoU on iSAID segmentation by 14.5 points over CutMix.

1 INTRODUCTION

Visual recognition in RS confronts a persistent statistical challenge. The long-tailed distribution (LTD) means that a few dominant "head" classes account for most training samples while the majority of "tail" classes are critically underrepresented Sharma & Gosain (2025); Zhang et al. (2023). Models trained on such distributions systematically overfit head categories and generalize poorly to rare yet operationally important land-cover types, a problem amplified by the high intra-class variability and inter-class similarity inherent to RS imagery Zhou et al. (2023); Cheng et al. (2017).

Generic augmentation strategies such as MixUp Zhang et al. (2017) and CutMix Yun et al. (2019) blend arbitrary class pairs and can produce semantically implausible composites in specialized domains Qin et al. (2024). Recent methods address this by using CLIP's cosine similarity to select compatible mixing partners Wang et al. (2025); Franklin et al. (2024), but CLIP's embedding space is shaped by web-scale co-occurrence rather than domain ontology. As Figure 1 shows, domain-specific pairs such as *stadium* $\leftrightarrow$ *tennis court* receive poor CLIP similarity ranks despite being ontologically related, while our OGSM-adapted embeddings correctly promote them.

We propose Ontology-Guided Semantic Mixing as a lightweight and architecture-agnostic module that adapts frozen CLIP text embeddings to a formal remote sensing ontology through triplet metric learning Schroff et al. (2015). The resulting semantic similarity matrix ensures that the mixing process is constrained exclusively to ontologically compatible classes. This targeted selection produces high-fidelity augmented samples that effectively enrich tail-class representations. As a purely data-level module, our approach is compatible with any downstream architecture.

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.

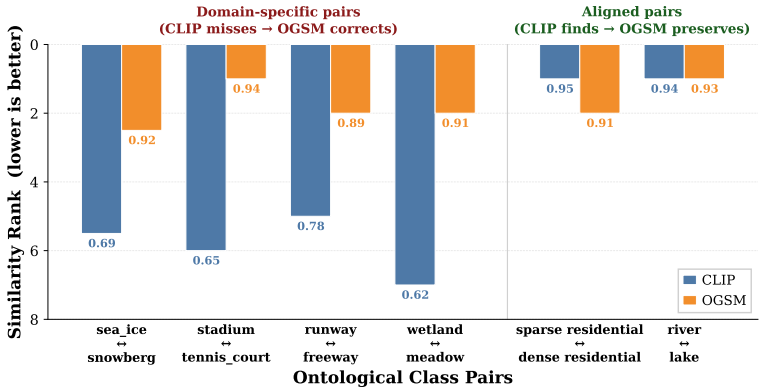


Figure 1: Similarity rank comparison between raw CLIP Radford et al. (2021) text embeddings (**blue**) and OGSM-adapted embeddings (**orange**) for six ontological class pairs from NWPU-RESISC45 Cheng et al. (2017). Lower rank indicates higher similarity. Left group shows domain-specific pairs that CLIP underestimates but OGSM corrects. Right group shows aligned pairs that both methods rank correctly.

Enriching the data distribution alone is still insufficient because the classifier itself must also possess adequate specialized capacity. Mixture-of-Experts architectures offer a principled solution for this capacity issue Wang et al. (2020); Zhao et al. (2023); Ru et al. (2024), yet two practical limitations persist: (i) the expert count is typically fixed through costly empirical tuning Yang et al. (2025) and (ii) the tempering strategies remain entirely static. We address both of these limitations within our Bag of LoRAs framework, which is built on a frozen CLIP Vision Transformer equipped with lightweight LoRA adapters Hu et al. (2021). We introduce an optimal expert allocation scheme that deterministically derives the ideal expert count directly from dataset statistics. We also implement an adaptive tempering mechanism that dynamically adjusts per-expert logit rebalancing. This adjustment is guided by a secondary objective that jointly minimizes the task loss and inter-expert performance variance. We validate our framework on four classification benchmarks and establish new state-of-the-art results with consistent tail-class gains.

Our contributions are as follows: **1.** We propose OGSM, a lightweight, architecture-agnostic module that adapts frozen CLIP embeddings to a formal RS ontology via triplet metric learning, ensuring semantically plausible tail-class augmentation. **2.** We introduce an optimal expert allocation formula that deterministically derives the LoRA expert count from dataset statistics, replacing empirical tuning in existing MoE frameworks. **3.** We design an adaptive tempering mechanism, a secondary objective jointly minimizing task loss and inter-expert performance variance, enforcing balanced specialization during training. **4.** We conduct comprehensive evaluation achieving state-of-the-art results on four classification benchmarks, and validate on a severely imbalanced semantic segmentation benchmark, where OGSM alone yields substantial tail-class gains without any architectural modification, confirming generality to diverse RS tasks and imbalance regimes.

2 Related Work

Semantic-Aware Augmentation for Long-Tailed Learning: Simple interpolation strategies such as MixUp Zhang et al. (2017) and CutMix Yun et al. (2019) blend arbitrary class pairs and can produce semantically implausible composites in specialized domains Qin et al. (2024). A parallel direction incorporates external knowledge from VLMs, DMRS Wang et al. (2025) and Text-Guided Mixup Franklin et al. (2024) use the cosine similarity of a frozen CLIP text encoder Radford et al. (2021) to select compatible mixing partners. However, these methods operate on generic VLM embeddings that lack domain-specific semantics. Our OGSM differs by adapting CLIP embeddings to a formal RS ontology via triplet metric learning Schroff et al. (2015), injecting domain-specific semantic structure rather than relying on generic linguistic similarity, while operating as a lightweight, architecture-agnostic pre-processing step applicable to both classification and dense prediction tasks.

Multi-Expert Long-Tailed Recognition: Multi-expert frameworks partition the class space across specialized sub-models for complementary head–tail specialization. RIDE Wang et al. (2020) es-

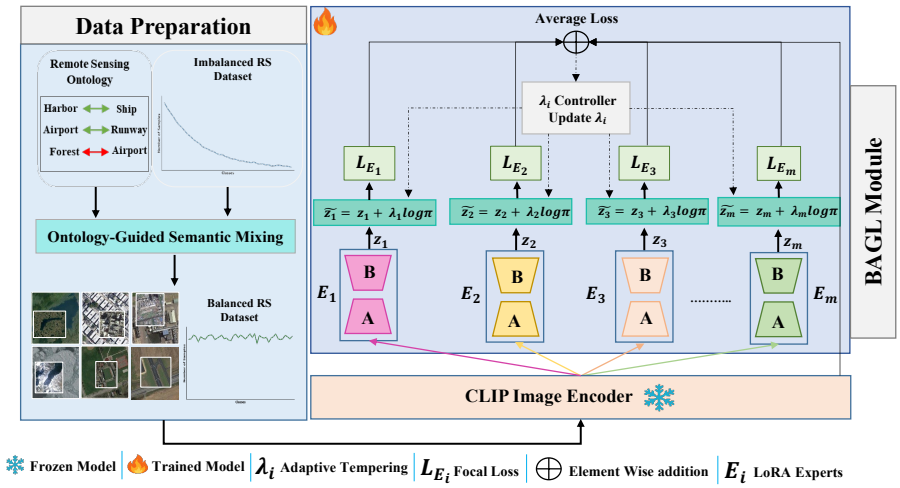


Figure 2: Architectural overview of the proposed framework. Stage 1 (left) shows OGSM, where an RS ontology $\mathcal{O}$ supervises the training of an MLP adapter that transforms CLIP's generic text embeddings into RS-aware representations. The semantic similarity matrix $\mathbf{S}$ guides CutMix-style augmentation. Stage 2 (right) shows the BAGL Module, where E lightweight LoRA experts are integrated into a frozen CLIP ViT-B/16 image encoder.

tablished this paradigm with distribution-aware routing, SADE Zhang et al. (2022) added test-time self-supervised expert selection, and MDCS Zhao et al. (2023) promoted diversity via consistency self-distillation. Across these methods, the expert count is universally treated as a manually tuned hyperparameter, almost always fixed at three Wang et al. (2020); Zhao et al. (2023); Ru et al. (2024); Wang et al. (2025), and tempering strategies remain static Wang et al. (2020); Zhao et al. (2023); Wang et al. (2025); Menon et al. (2021), applying globally fixed rebalancing factors that ignore the evolving specialization of experts during training Tao et al. (2023); Chen et al. (2024). The fusion of LoRA with MoE has become a dominant paradigm for parameter-efficient capacity scaling Wu et al. (2024); Gou et al. (2024); Dou et al. (2024); Buehler & Buehler (2024); Li et al. (2024). In remote sensing, DMRS Wang et al. (2025) integrates D-LoRA modules into a fixed three-expert architecture with learned routing that introduces gating hyperparameters and reduces interpretability. BAGL addresses these limitations by activating all experts for every input and aggregating via simple averaging. The expert count is derived analytically from dataset statistics, and expert-specific tempering factors are adapted dynamically through a secondary optimization objective, replacing the static rebalancing strategies that dominate prior work.

3 Proposed Methods

Let an imbalanced remote sensing dataset be denoted as $\mathcal{I} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{H\times W\times 3}$ is an RGB image and $y_i \in \{1, \ldots, C\}$ is its class label. The class distribution follows a long-tailed structure with imbalance factor $\rho = n_{\min}/n_{\max}$. We propose a two-stage framework (Figure 2). In Stage 1, a lightweight MLP adapter $\mathcal{A}_\theta$ is trained offline to correct CLIP's domain-agnostic embedding space using supervision from a formal RS ontology $\mathcal{O}$. The resulting $C \times C$ semantic similarity matrix $\mathbf{S}$ guides CutMix-style augmentation that enriches the imbalanced dataset with semantically plausible samples. In Stage 2, E lightweight LoRA adapters are inserted into a frozen CLIP ViT-B/16 image encoder, where E is derived analytically from the class count C and ρ. Training employs Focal Loss with an adaptive tempering mechanism to coordinate inter-expert specialization.

3.1 Ontology-Guided Semantic Mixing (OGSM)

3.1.1 Embedding Transformation.

OGSM is implemented as a lightweight two-layer MLP adapter $\mathcal{A}_\theta$ (hidden dimension 256, ReLU activation) that operates entirely in the text embedding space. For a frozen base embedding $e_{\text{base}} = \text{CLIP-Text}(c)$ of class c, the adapter produces a domain-aware embedding via a residual formulation:

$$\mathbf{e}_{\text{RS}} = \frac{\mathbf{e}_{\text{base}} + \alpha A_\theta(\mathbf{e}_{\text{base}})}{\|\mathbf{e}_{\text{base}} + \alpha A_\theta(\mathbf{e}_{\text{base}})\|_2}. \quad (1)$$

where α is a learnable residual scaling factor. The residual connection preserves CLIP's general-purpose representations while allowing targeted corrections toward RS-specific semantics.

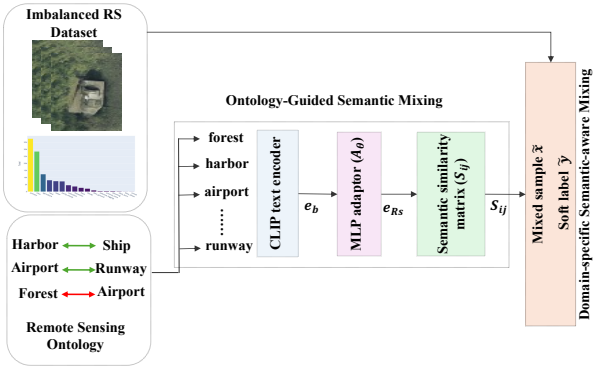


Figure 3: OGSM pipeline.

The adapter is trained once as an independent preprocessing step via a triplet loss that directly optimizes for ontological alignment. For each triplet (a, p, n) sampled from $\mathcal{O}$, where p is a class ontologically related to anchor a and n is a distinct class, the loss enforces:

$$\mathcal{L}_{\text{OGSM}} = \mathbb{E}_{(a,p,n)\sim\mathcal{O}}[\max(0,\ \|e_a - e_p\|_2 - \|e_a - e_n\|_2 + m)]\,, \quad (2)$$

with margin m. This objective pulls together embeddings of ontologically related classes while pushing apart distinct ones, directly correcting the failure modes documented in Figure 1. A critical design choice is that CLIP's own similarities cannot serve as training supervision, as they encode precisely the domain-agnostic relationships that OGSM aims to correct. The RS ontology $\mathcal{O}$ therefore provides essential external supervision independent of CLIP's learned biases. Upon convergence, the adapter parameters are frozen at $\mathcal{A}_{\theta^*}$, and the full $C \times C$ cosine similarity matrix is computed once and cached:

$$\mathbf{S}_{ij} = \frac{e_{\text{RS},i}^\top e_{\text{RS},j}}{\|e_{\text{RS},i}\|_2\,\|e_{\text{RS},j}\|_2}. \quad (3)$$

3.1.2 Semantic-Constrained Augmentation.

The cached similarity matrix $\mathbf{S}$ governs all augmentation decisions during downstream training. For a training sample x_i belonging to class y_i, a mixing partner x_j is selected by sampling its class y_j proportionally to the similarity row $\mathbf{S}_{y_i,\cdot}$:

$$P(y_j = c \mid y_i) = \frac{\exp(\mathbf{S}_{y_i,c}/\tau)}{\sum_{c'=1}^{C}\exp(\mathbf{S}_{y_i,c'}/\tau)}, \quad c \neq y_i, \quad (4)$$

where τ is a temperature parameter controlling selection sharpness. Given the selected pair (x_i, x_j), the mixed sample is produced via a CutMix-style Yun et al. (2019) binary mask $\mathbf{M} \in \{0,1\}^{H\times W}$:

$$\tilde{x}_i = \mathbf{M} \odot x_i + (1 - \mathbf{M}) \odot x_j, \quad (5)$$

with soft label:

$$\tilde{y}_i = \lambda\,\mathbf{e}_{y_i} + (1-\lambda)\,\mathbf{e}_{y_j}, \quad \text{where} \quad \lambda = \frac{1}{HW}\sum_{h,w}\mathbf{M}_{h,w}, \quad (6)$$

and $\mathbf{e}_{y_i}$ denotes the one-hot encoding of class y_i. The mask area ratio λ is drawn from $\text{Beta}(1.0, 1.0)$ following standard CutMix practice Yun et al. (2019). Augmentation is applied preferentially to tail classes, with the probability of augmenting a sample of class c modulated by inverse frequency:

$$P_{\text{aug}}(c) = \min\left(1,\ p_0 \cdot \frac{n_{\max}}{n_c}\right), \quad (7)$$

where p_0 is the base mixing probability and n_c is the sample count for class c.

3.2 BAGL Module

The BAGL Module operates on the enriched dataset produced by OGSM. A frozen CLIP ViT-B/16 image encoder serves as a shared feature extractor, with all trainable parameters residing in E lightweight LoRA adapters Hu et al. (2021) inserted into the self-attention layers.

3.2.1 LoRA Expert Architecture.

For each expert $k \in \{1, \ldots, E\}$, the adapted weight matrix in each self-attention layer is:

$$W_k = W_0 + \frac{\alpha_{\text{LoRA}}}{r} B_k A_k, \quad k \in \{1, \ldots, E\} \tag{8}$$

where W_0 is the frozen pretrained weight, $A_k \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B_k \in \mathbb{R}^{d_{\text{out}} \times r}$ are the low-rank factors (rank r=16, scaling α_{LoRA}=32). All E experts share the same frozen backbone but develop distinct specializations through the adaptive tempering mechanism described below.

Optimal Expert Allocation. Existing MoE frameworks treat the expert count as a manually tuned hyperparameter Zhao et al. (2023); Ru et al. (2024) or rely on complex learned routing dynamics Mercurius et al. (2024); Yang et al. (2025). We replace this with an analytically derived allocation that ties model capacity directly to the long-tailed structure. Given C classes and imbalance factor ρ, the optimal expert count is:

$$E^* = \left\lfloor \sqrt{\frac{C \cdot \log(1/\rho)}{2}} + 1 \right\rfloor. \tag{9}$$

The formula reflects that representational demand scales with both class count and imbalance severity, measured by $\log(1/\rho)$, while the square root enforces sublinear growth since expert representations overlap and share structure. The additive constant ensures a minimum of one expert and handles the balanced case ($\rho = 1$) gracefully. Figure 5 empirically validates that these predicted optima coincide with observed accuracy peaks across all benchmarks.

3.2.2 Loss Function and Adaptive Tempering.

Intra-expert rebalancing: Each expert k addresses residual class imbalance via Focal Loss Lin et al. (2017). The raw logit $z_{k,i,c}$ for sample i and class c is modulated by a trainable adaptive tempering factor λ_k:

$$\tilde{z}_{k,i,c} = z_{k,i,c} + \lambda_k \log(\pi_c), \tag{10}$$

where $\pi_c = n_c / \sum_k n_k$ is the class prior. The expert-specific loss is:

$$\mathcal{L}_k = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c}\, \alpha_{\text{FL}}\, (1 - \tilde{p}_{k,i,c})^{\gamma} \log(\tilde{p}_{k,i,c}), \tag{11}$$

where $\tilde{p}_{k,i,c} = \text{softmax}(\tilde{z}_{k,i})_c$ is the post-softmax probability, α_{FL} is the balancing factor, and γ is the focusing parameter.

Inter-expert coordination: The tempering factors $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_E]$ are trainable parameters governed by a dedicated secondary objective:

$$\mathcal{J}(\boldsymbol{\lambda}) = \underbrace{\frac{1}{E} \sum_{k=1}^{E} \mathcal{L}_k(\lambda_k)}_{\text{Mean Task Loss}} + \beta \cdot \underbrace{\text{Var}(\mathcal{L}_1(\lambda_1), \ldots, \mathcal{L}_E(\lambda_E))}_{\text{Loss Balancing Regularizer}}. \tag{12}$$

The variance regularizer penalizes disparity among expert losses, encouraging a collaborative equilibrium. The λ_k values are updated via gradient descent with learning rate $\eta_\lambda = 0.01$:

$$\boldsymbol{\lambda}^{(t+1)} = \boldsymbol{\lambda}^{(t)} - \eta_\lambda \nabla_{\boldsymbol{\lambda}} \mathcal{J}(\boldsymbol{\lambda}^{(t)}). \tag{13}$$

The factors are initialized with evenly spaced values in $[-0.9, 2.5]$ to seed differentiated expert specialization from the outset.

Overall training objective: The total BAGL loss is the mean of all expert losses:

$$\mathcal{L}_{\text{BAGL}} = \frac{1}{E} \sum_{k=1}^{E} \mathcal{L}_k. \tag{14}$$

Only the LoRA parameters $\{A_k, B_k\}_{k=1}^{E}$ and the tempering factors $\boldsymbol{\lambda}$ are updated during training. At inference, the tempering factors are deactivated and final predictions are computed by averaging softmax outputs across all experts: $P(y \mid x) = \frac{1}{E} \sum_{k=1}^{E} \text{softmax}(z_k(x))$.

Table 1: Comprehensive comparison across four benchmarks. Numbers in **blue** represent the relative error rate reduction (%) of the proposed BAGL compared to the corresponding DMRS Wang et al. (2025) SOTA method. Avg. is the unweighted mean of the four dataset overalls.

	RSD46-WHU			NWPU-RESISC45			EuroSAT			UCMerced			
Method	Head	Med	Tail	Head	Med	Tail	Head	Med	Tail	Head	Med	Tail	**Avg.**
MDCS Zhao et al. (2023)	81.7	78.3	75.2	**94.6**	82.7	63.2	95.2	84.5	78.8	95.3	87.4	81.6	84.5
SADE Zhang et al. (2022)	80.4	76.1	75.8	92.0	78.1	61.4	93.5	82.0	77.6	94.8	86.2	80.5	81.5
RIDE Wang et al. (2020)	83.9	72.1	58.8	93.7	79.9	53.6	94.8	81.5	75.2	**98.5**	84.3	79.8	79.7
BALMS Ren et al. (2020)	81.7	75.9	69.4	93.2	78.8	61.3	93.8	82.8	74.1	92.1	88.5	79.2	80.9
CLIP-rsicd Radford et al. (2021)	37.7	28.6	32.5	70.5	66.9	57.7	62.5	55.3	48.1	76.4	73.2	65.9	56.5
RemoteCLIP liu (2024)	41.8	30.4	27.3	77.4	65.8	63.7	65.8	58.2	52.7	78.5	74.1	64.6	58.4
DMRS Wang et al. (2025)	**90.1**	90.6	81.9	92.8	86.8	84.3	**96.3**	87.2	84.8	94.4	92.6	86.5	88.7
BAGL (Ours)	89.8	**92.4**	**84.7**	92.9	**90.6**	**86.5**	95.8	**90.5**	**87.3**	98.3	**94.9**	**87.9**	**91.1**
	(-3.0)	(19.1)	(15.5)	(4.39)	(28.8)	(14.0)	(13.5)	(25.8)	(16.4)	(69.6)	(8.9)	(10.4)	(21.2)

4 Experimental Results

4.1 Experimental Settings

Datasets: We evaluate on four RS benchmarks spanning diverse imbalance regimes. RSD46-WHU Long et al. (2017) contains 117,000 images across 46 scene categories with extreme natural imbalance ($\rho = 0.01$). NWPU-RESISC45 Cheng et al. (2017) comprises 31,500 images across 45 classes, imbalanced to $\rho = 0.01$ following established protocols Wang et al. (2025). UCMerced Yang & Newsam (2010) consists of 2,100 images across 21 land-use categories, imbalanced to $\rho = 0.1$. EuroSAT Helber et al. (2019) contains 27,000 Sentinel-2 patches across 10 land-cover classes, imbalanced to $\rho = 0.01$, providing a distinct multi-spectral evaluation regime.

Implementation Details: We use a frozen CLIP ViT-B/16 backbone with trainable LoRA adapters (r=16, α=32). The expert count E is derived via Equation equation 9, yielding E=7 for RSD46-WHU and NWPU-RESISC45, E=4 for UCMerced and EuroSAT. Tempering factors $\boldsymbol{\lambda}$ are initialized uniformly over $[-0.9, 2.5]$ and updated with η_λ=0.01.

For the OGSM triplet loss (Eq. 2), we set the margin m=0.2 following Schroff et al. (2015) and the partner sampling temperature τ=0.5. Focal Loss (Eq. 11) uses balancing factor α_{FL}=0.25 and focusing parameter γ=2.0, which are the default parameters from Lin et al. (2017). The variance regularizer weight in Eq. 12 is set to β=0.5.

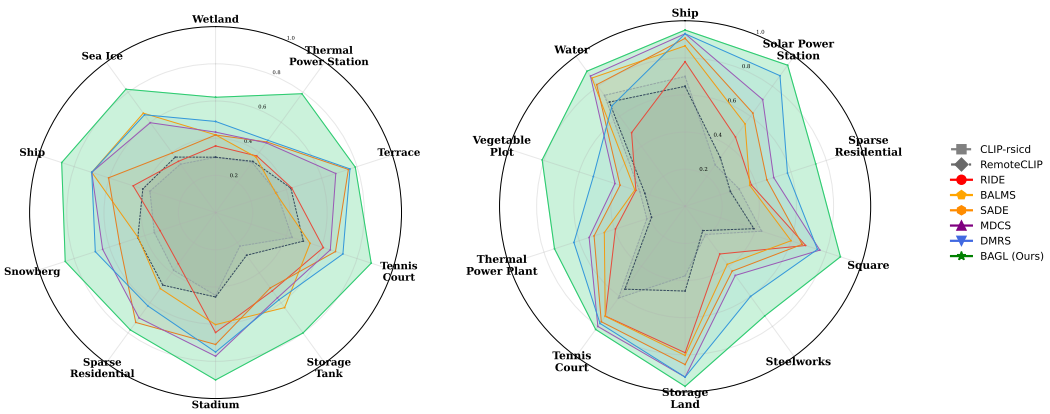
Training: All models train for 50 epochs with AdamW (weight decay 0.01), cosine annealing (lr : $5\times10^{-2} \rightarrow 1\times10^{-4}$), and batch size 64.

Evaluation: We report overall accuracy and group-wise accuracy over equal-sized Head, Medium, and Tail splits defined by training frequency. Group accuracy is $\mathrm{Acc}_G = \left(\sum_{c\in G}\sum_{i=1}^{N_c} \mathbb{1}(\hat{y}_i{=}y_i)\right) / \left(\sum_{c\in G} N_c\right) \times 100\%$.

Baselines: We compare against multi-expert methods (MDCS Zhao et al. (2023), RIDE Wang et al. (2020), SADE Zhang et al. (2022)), a reweighting method (BALMS Ren et al. (2020)), and RS-specific frameworks (DMRS Wang et al. (2025), RemoteCLIP liu (2024), CLIP-rsicd Radford et al. (2021)). All baselines are reimplemented under identical training conditions.

4.2 Quantitative Results across Benchmarks

Table 1 summarizes quantitative results on four long-tailed remote-sensing benchmarks. Across all datasets, **BAGL** consistently outperforms prior SOTA methods, achieving the highest Tail-class accuracy on every benchmark. BAGL improves Tail accuracy over DMRS Wang et al. (2025) by +2.8, +2.2, +2.5, and +1.4 pp on RSD46-WHU, NWPU-RESISC45, EuroSAT, and UCMerced respectively. Averaged across all four datasets, BAGL reaches 91.1% (+2.4 pp over DMRS), establishing a new SOTA. Notably, BAGL exhibits minor declines in some Head-class accuracies. This trade-off arises from its adaptive multi-expert LoRA ensemble and trainable λ parameters, which dynamically reallocate gradient emphasis toward underrepresented classes during optimization, intention-



(a) NWPU-RESISC45 (b) RSD46-WHU

Figure 4: Per-class accuracy comparison on the 10 most underrepresented tail classes across (a) NWPU-RESISC45 and (b) RSD46-WHU benchmarks.

ally softening overfitting to dominant categories to promote a more balanced representation across Head, Medium, and Tail distributions.

4.3 Tail Class Performance Analysis

To provide deeper insights into BAGL's effectiveness on minority classes, Figure 4 presents radar charts visualizing per-class accuracy for the 10 most underrepresented tail classes. These visualizations reveal the granular performance characteristics that underlie the aggregate tail-group improvements reported in Table 1. The radar charts demonstrate BAGL's consistent superiority (green polygons) in achieving broader coverage across challenging tail categories compared to all baseline methods. On NWPU-RESISC45 (Figure 4 (a)), BAGL exhibits particularly strong performance on environmentally linked classes such as Wetland, Sea Ice, and Snowberg, where semantic relationships captured by OGSM prove especially beneficial. The method achieves near-perfect accuracy on Ship and Thermal Power Station while maintaining robust performance ($>$80%) across all displayed tail classes, significantly outperforming the previous SOTA DMRS (blue polygon), which shows considerable performance degradation on classes like Sparse Residential and Stadium.

On RSD46-WHU (Figure 4 (b)), the most challenging dataset with ρ=0.01, BAGL maintains competitive performance despite extreme imbalance. While certain industrial classes like Steelworks and Storage Land present challenges across all methods, BAGL achieves substantial gains on natural scene categories (Water, Vegetable Plot) and infrastructure classes (Solar Power Station, Ship), validating the effectiveness of our adaptive expert allocation strategy in capturing diverse tail-class representations. To better understand the source of our tail-class performance gains, we visualize the feature activation maps using GradCAM on rare classes from the NWPU-RESISC45 dataset (detailed visualizations and analysis are provided in Appendix D).

4.4 Generalization on RS Semantic Segmentation

To validate that OGSM generalizes beyond scene classification to dense prediction tasks, we evaluate on semantic segmentation using iSAID Zamir et al. (2019), a large-scale aerial dataset comprising 2,806 high-resolution satellite images with dense annotations across 15 object categories (655,451 instances). Following standard practice, we merge per-instance masks into per-pixel semantic labels and split into 1,411 training and 458 validation images. The dataset exhibits severe natural imbalance: *small vehicle* contains roughly 200$\times$ more instances than *ground track field*, making iSAID a compelling testbed for assessing whether OGSM can improve tail-class segmentation without architectural modifications.

We adopt SegFormer-B2 Xie et al. (2021) as the segmentation backbone, trained for 80k iterations with AdamW (lr 6×10^{-5}, weight decay 0.01) and a batch size of 16. Input patches are cropped to 512×512. OGSM's precomputed similarity matrix $\mathbf{S}$ is applied identically to the classification

Table 2: Per-class IoU (%) on the iSAID validation set. Classes are ordered from most to least frequent by instance count Zamir et al. (2019). Shaded rows denote the five tail classes. **Bold**: best result per row.

Class	Freq.	IoU (%)		
		No Aug	*CutMix*	*OGSM (Ours)*
Small Vehicle	Head	59.9	60.0	**61.2**
Large Vehicle	Head	81.3	82.4	**84.9**
Ship	Head	48.6	52.7	**54.4**
Plane	Head	82.5	80.7	**81.1**
Storage Tank	Head	62.8	62.2	**68.0**
Harbor	Medium	62.9	61.4	**62.6**
Swimming Pool	Medium	49.7	53.1	**56.3**
Tennis Court	Medium	89.9	91.7	**92.3**
Bridge	Medium	37.2	31.4	**45.9**
Baseball Diamond	Medium	67.6	72.8	**78.9**
Basketball Court	Tail	42.2	57.6	**73.6**
Soccer-Ball Field	Tail	72.7	74.0	**84.3**
Roundabout	Tail	55.5	64.7	**81.7**
Helicopter	Tail	13.3	8.1	**23.1**
Ground Track Field	Tail	45.7	52.0	**66.2**
mIoU	—	58.1	60.3	**72.1**
Tail mIoU	—	45.9	51.3	**65.8**

setting, guiding CutMix partner selection during training without any segmentation-specific modification.

Table 2 reports per-class IoU on the iSAID validation set. OGSM achieves 72.1% mIoU, outperforming CutMix by +11.8 pp and the no-augmentation baseline by +14.0 pp. The gains are most pronounced on the five tail classes, where OGSM reaches 65.8% tail mIoU versus 51.3% for CutMix (+14.5 pp). Importantly, OGSM also improves every head class, confirming that domain-aware mixing produces globally better augmentation partners rather than trading head-class accuracy for tail-class gains. *Helicopter* remains the most challenging category due to extreme rarity, though OGSM still nearly triples the CutMix result (8.1%→23.1%).

4.5 Ablation Study

4.5.1 Impact of LoRA expert allocation

Figure 5 empirically validates our expert allocation formula across three benchmarks. The observed optima precisely match the predictions of Equation equation 9: RSD46-WHU ($C = 46$, $\rho = 0.01$) and NWPU-RESISC45 ($C = 45$, $\rho = 0.01$) peak at $E = 7$ experts, while UCMerced ($C = 21$, $\rho = 0.1$) peaks at $E = 4$. All datasets show monotonic improvement from $E = 2$ up to the predicted optimum, with substantial gains over the fixed three-expert DMRS baseline.

Beyond the optimum, the datasets exhibit divergent sensitivity to over-parameterization. RSD46-WHU and NWPU-RESISC45 degrade gradually beyond $E = 7$, whereas UCMerced suffers catastrophic collapse past

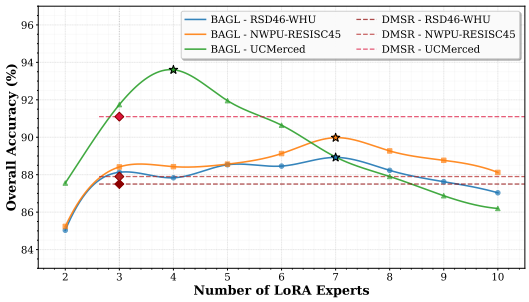


Figure 5: Impact of LoRA expert allocation on classification accuracy. Solid lines: BAGL with varying expert count. Dashed lines: fixed three-expert DMRS baselines. ★: analytically predicted optima from Eq. equation 9. ♦: DMRS operating points at $E = 3$.

$E = 5$, dropping below the DMRS baseline - confirming that smaller datasets with limited class diversity cannot sustain excess capacity. These results validate that our closed-form allocation provides theoretically grounded guidance and prevents performance degradation from suboptimal expert assignment.

4.5.2 Component Contributions

To isolate the specific improvements yielded by our proposed modules, we conduct a component-wise ablation on the NWPU-RESISC45 and RSD46-WHU datasets, detailed in Table 3. The baseline model consists of a single LoRA adapter trained with standard CutMix augmentation.

Table 3: Component ablation study on the NWPU-RESISC45 and RSD46-WHU benchmarks. We report Overall Accuracy (OA) and Tail-class Accuracy (Tail) to evaluate the individual and synergistic contributions of Ontology-Guided Semantic Mixing (OGSM) and Bag of LoRAs (BAGL). The baseline employs a single LoRA adapter with standard CutMix augmentation.

Components		NWPU-RESISC45		RSD46-WHU	
OGSM	BAGL	OA (%)	Tail (%)	OA (%)	Tail (%)
–	–	85.2	62.1	81.4	58.3
✓		88.5	75.4	84.7	69.2
	✓	89.1	75.0	85.9	68.1
✓	✓	**90.3**	**86.5**	**88.9**	**84.7**

Replacing standard CutMix with our Ontology-Guided Semantic Mixing (OGSM) yields substantial improvements in tail-class accuracy across both datasets. This confirms that preventing semantically implausible interpolations, which plague generic MixUp/CutMix strategies in remote sensing, directly benefits the representation of rare classes. Conversely, replacing the single adapter with our Bag of LoRAs (BAGL) architecture while retaining standard CutMix provides notable gains in Overall Accuracy, demonstrating the superior representational capacity of the adaptive multi-expert system. Finally, the full system combining both components achieves the highest performance. This highlights a critical synergy. OGSM provides the semantically coherent training signals necessary to prevent expert collapse, allowing the specialized experts within BAGL to effectively discriminate among the newly enriched tail-class variations.

5 Conclusion

We presented OGSM and BAGL for long-tailed remote sensing recognition. OGSM corrects CLIP's domain-agnostic embedding space via triplet metric learning under RS ontological constraints, producing a semantic similarity matrix that restricts augmentation to ontologically compatible class pairs. BAGL complements this enriched training signal with an analytically derived expert count and dynamically adapted tempering factors, replacing the fixed configurations and static rebalancing strategies prevalent in prior work.

Experiments across four classification benchmarks demonstrate consistent state-of-the-art performance, with an average accuracy of 91.1% (+2.4 pp over DMRS) and systematic tail-class improvements on every dataset. On iSAID semantic segmentation, OGSM alone yields +14.5 pp tail mIoU over CutMix without any architectural modification, validating its generality to dense prediction tasks. Future work will explore automated ontology construction via large language models or existing taxonomic systems, and extend the framework to additional downstream tasks.

References

Remoteclip: A vision language foundation model for remote sensing, 2024. URL https://arxiv.org/abs/2306.11029.

Eric L. Buehler and Markus J. Buehler. X-LoRA: Mixture of low-rank adapter experts, a flexible framework for large language models with applications in protein mechanics and design. *APL Machine Learning*, 2(2):026119, 2024.

Jiawei Chen, Jingchao Wang, and Qiulong Yuan. Dynamic learnable logit adjustment for long-tailed visual recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34 (10):9803–9815, 2024.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. doi: 10.1109/JPROC.2017.2675998.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhicheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1932–1945, 2024.

Richard Franklin, Jiawei Yao, Deyang Zhong, Qi Qian, and Juhua Hu. Text-guided mixup towards long-tailed image categorization. *arXiv preprint arXiv:2409.03583*, 2024.

Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T. Kwok, and Yu Zhang. MoCLE: Mixture of cluster-conditional LoRA experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*, 2024.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2019. URL `https://arxiv.org/abs/1709.00029`.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021.

Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. MixLoRA: Enhancing large language models fine-tuning with LoRA-based mixture of experts. *arXiv preprint arXiv:2404.15159*, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Yang Long, Yiming Gong, Zhifeng Xiao, and Qiang Liu. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2486–2498, May 2017. doi: 10.1109/TGRS.2016.2645610.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Ray Coden Mercurius, Ehsan Ahmadi, Soheil Mohamad Alizadeh Shabestary, and Amir Rasouli. Amend: A mixture of experts framework for long-tailed trajectory prediction. *arXiv preprint arXiv:2402.08698*, 2024.

Huafeng Qin, Xin Jin, Hongyu Zhu, Hongchao Liao, Mounîm A El-Yacoubi, and Xinbo Gao. Sumix: Mixup with semantic and uncertain information. In *European Conference on Computer Vision*, pp. 70–88. Springer, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *NeurIPS*, 33:4175–4186, 2020.

Lixiang Ru, Xin Guo, Lei Yu, Yingying Zhang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yansheng Li, and Ming Yang. Parameter-efficient complementary expert learning for long-tailed visual recognition. In *ACM MM*, pp. 5393–5402, 2024.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shweta Sharma and Anjana Gosain. Addressing class imbalance in remote sensing using deep learning approaches: A systematic literature review. *Evolutionary Intelligence*, 18(1):23, 2025.

Jian Tao, Bingchen Zhao, and Hao Yang. Local and global logit adjustments for long-tailed learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11773–11783, 2023.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. Long-tailed recognition by routing diverse distribution-aware experts, 2020. arXiv:2010.01809.

Yifan Wang, Fan Zhang, Qihao Zhao, Wei Hu, and Fei Ma. Dmrs: Long-tailed remote sensing recognition via semantic-aware mixing and diversity experts. *Int. J. Appl. Earth Obs. Geoinf.*, 141:104623, 2025.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of LoRA experts. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, pp. 12077–12090, 2021.

Minghao Yang, Guang Li, Guofeng Cui, Yuzhen Niu, and Haowei Huang. Adaptive shared experts with lora-based mixture of experts for multi-task learning. *arXiv preprint arXiv:2510.00570*, 2025.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS)*, 2010.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Hisham Cholakkal, Ling Shao, Gui-Song Xia, Xiang Bai, and Fan Yang. iSAID: A large-scale dataset for instance segmentation in aerial images. In *CVPR Workshops*, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization, 2017. arXiv:1710.09412.

Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. *NeurIPS*, 35:34077–34090, 2022.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023.

Qihao Zhao, Chen Jiang, Wei Hu, Fan Zhang, and Jun Liu. Mdcs: More diverse experts with consistency self-distillation for long-tailed recognition. In *ICCV*, pp. 11563–11574, 2023.

Zheng Zhou, Change Zheng, Xiaodong Liu, Ye Tian, Xiaoyi Chen, Xuexue Chen, and Zixun Dong. A dynamic effective class balanced approach for remote sensing imagery semantic segmentation of imbalanced data. *Remote Sensing*, 15(7):1768, 2023.

A Implementation details

Our experiments were conducted on an NVIDIA RTX A6000 GPU using CUDA 12.9 and PyTorch 2.2.1. All classification models were trained for 50 epochs with the AdamW optimizer, employing a weight decay of 0.01, a batch size of 64, and a cosine annealing learning rate schedule from 5e-2 to 1e-4. The architecture utilized a frozen CLIP ViT-B/16 backbone equipped with trainable LoRA adapters (rank r=16, scaling alpha=32). For the semantic segmentation tasks on iSAID, we adopted a SegFormer-B2 backbone, which was trained for 80k iterations using the AdamW optimizer with a learning rate of 6e-5, a weight decay of 0.01, and a batch size of 16.

B Dataset Imbalance Configuration

To evaluate BAGL under realistic long-tailed scenarios, we introduce controlled class imbalance during training following an exponential decay function. The number of training samples for class i (sorted by original frequency) is:

$$n_i = n_{\max} \times \rho^{i/(C-1)}, \tag{15}$$

where $n_{\max}$ is the sample count of the most frequent class, ρ is the imbalance factor, C is the total number of classes, and $i \in \{0, \ldots, C-1\}$. Table 4 summarises the resulting configurations.

Table 4: Training set statistics under the applied imbalance.

Dataset	C	ρ	Total	$n_{\max}$	$n_{\min}$
RSD46-WHU	46	0.01	3,847	380	3
NWPU-RESISC45	45	0.01	6,960	700	7
UCMerced	21	0.1	829	100	10

C Per-Class Performance Comparison

This section provides a comprehensive per-class comparison between BAGL and DMRS Wang et al. (2025) in terms of Precision, Recall, and F1-score across all three benchmarks. The full results are reported in Tables 5, 6, and 7.

C.1 Results on RSD46-WHU

Table 5 presents the per-class results on RSD46-WHU Long et al. (2017). BAGL achieves a macro-averaged F1 of 0.90 compared to 0.86 for DMRS, and a weighted-averaged F1 of 0.92 versus 0.83, reflecting consistent gains across both frequent and rare categories.

The most pronounced improvements appear in classes where DMRS suffers from low recall. Railway classification improves from an F1 of 0.51 to 0.94, driven almost entirely by a recall increase from 0.35 to 0.96, indicating that DMRS frequently misses this class while BAGL recovers it reliably. A similar pattern holds for medium density scattered building (F1: 0.61 $\rightarrow$ 0.74), natural dense forest land (F1: 0.70 $\rightarrow$ 0.86), solar power station (F1: 0.81 $\rightarrow$ 0.95), and water (F1: 0.72 $\rightarrow$ 0.94). Tail classes such as ship, vegetable plot, and sparse residential area also show consistent improvements, confirming that BAGL's augmentation and expert allocation strategies benefit underrepresented categories broadly rather than isolated cases.

A small number of classes remain challenging for both methods. Refinery (F1: 0.60) and steelworks (F1: 0.68) exhibit high visual similarity with other industrial categories, suggesting that further refinement of the ontology or additional semantic cues may be needed for fine-grained industrial scene discrimination.

C.2 Results on NWPU-RESISC45

Table 6 reports the per-class results on NWPU-RESISC45 Cheng et al. (2017), a 45-class benchmark with 700 images per class. BAGL achieves a macro-averaged F1 of 0.89 versus 0.84 for DMRS,

Table 5: Detailed per-class performance of BAGL and DMRS on RSD46-WHU ($\rho = 0.01$).

Class Name	BAGL (Ours)			DMRS		
	Precision	Recall	F1	Precision	Recall	F1
airplane	0.94	0.99	0.96	0.91	0.99	0.95
airport	0.96	0.92	0.94	0.84	0.95	0.89
artificial dense forest land	0.82	0.60	0.69	0.61	0.79	0.69
artificial sparse forest land	0.85	0.80	0.83	0.80	0.83	0.81
bare land	0.88	0.97	0.92	0.89	0.97	0.93
basketball court	0.81	0.63	0.71	0.81	0.54	0.65
blue structured factory building	0.83	0.97	0.90	0.89	0.96	0.92
building	0.81	0.60	0.69	0.44	0.93	0.60
construction site	0.61	0.85	0.71	0.63	0.86	0.73
cross river bridge	1.00	0.99	1.00	0.96	1.00	0.98
crossroads	1.00	0.92	0.96	0.98	0.95	0.96
dense tall building	0.85	0.58	0.69	0.88	0.86	0.87
dock	0.96	0.86	0.91	0.82	1.00	0.90
fish pond	0.84	1.00	0.91	0.79	1.00	0.88
footbridge	1.00	0.89	0.94	0.98	0.94	0.96
graff	0.97	0.72	0.82	0.75	0.83	0.79
grassland	0.90	0.92	0.91	0.76	0.76	0.76
irregular farmland	0.92	0.93	0.93	0.91	0.87	0.89
low scattered building	0.93	0.97	0.95	0.71	0.99	0.83
medium density scattered building	0.75	0.73	0.74	0.82	0.38	0.61
medium density structured building	0.61	0.81	0.69	0.50	0.94	0.65
natural dense forest land	0.77	0.97	0.86	0.55	0.97	0.70
natural sparse forest land	0.88	0.91	0.90	0.89	0.84	0.86
oil tank	0.83	0.88	0.85	0.83	0.86	0.84
overpass	0.93	0.98	0.96	0.91	1.00	0.95
parking lot	0.92	0.89	0.90	0.87	0.88	0.88
plastic greenhouse	0.98	1.00	0.99	0.94	0.96	0.95
playground	0.93	0.91	0.92	0.75	0.91	0.82
railway	0.92	0.96	0.94	0.96	0.35	0.51
red structured factory building	0.89	0.96	0.92	0.90	0.82	0.85
refinery	0.57	0.62	0.60	0.55	0.86	0.67
regular farmland	0.82	0.94	0.87	0.83	0.76	0.79
scattered blue roof factory building	0.83	0.85	0.84	0.78	0.80	0.79
scattered red roof factory building	0.83	0.72	0.77	0.88	0.51	0.64
sewage plant-type-one	0.95	0.79	0.87	0.96	0.76	0.86
sewage plant-type-two	0.90	1.00	0.95	0.90	1.00	0.95
ship	0.99	0.94	0.96	0.87	1.00	0.93
solar power station	0.99	0.92	0.95	0.73	0.92	0.81
sparse residential area	0.81	0.73	0.77	0.76	0.59	0.67
square	0.91	0.83	0.87	0.94	0.69	0.80
steelworks	0.70	0.65	0.68	0.72	0.59	0.65
storage land	0.83	0.95	0.89	0.98	0.85	0.91
tennis court	0.78	0.95	0.86	0.74	0.82	0.77
thermal power plant	0.81	0.56	0.66	0.79	0.50	0.64
vegetable plot	0.86	0.81	0.83	0.72	0.74	0.73
water	0.84	0.85	0.94	0.77	0.65	0.72
Macro Average	0.91	0.89	0.90	0.88	0.86	0.86
Weighted Average	0.92	0.91	0.92	0.85	0.82	0.83

while the weighted-averaged F1 reaches 0.90 compared to 0.81, indicating substantial improvements across all class frequency regimes.

The largest gains are concentrated in infrastructure and environmental classes where DMRS exhibits severe recall deficiencies. Railway improves from 0.81 to 0.93 in F1, ship from 0.71 to 0.95, thermal power station from 0.54 to 0.78, and runway from 0.68 to 0.89. In each case, the improvement is predominantly recall-driven: DMRS achieves high precision on these classes but detects only a fraction of the true positives, whereas BAGL recovers the majority of instances without sacrificing

precision. Beach (F1: 0.82 $\rightarrow$ 0.97) and terrace (F1: 0.69 $\rightarrow$ 0.83) follow a similar pattern, with BAGL substantially improving recall on categories that DMRS underdetects.

Certain ambiguous categories remain difficult for both methods. Palace (F1: 0.71), railway station (F1: 0.69), and wetland (F1: 0.69) involve high visual overlap with neighbouring classes, posing a shared challenge that may benefit from finer-grained ontological distinctions in future work.

Table 6: Detailed per-class performance of BAGL and DMRS on NWPU-RESISC45 ($\rho = 0.01$).

Class Name	**BAGL (Ours)**			**DMRS**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
airplane	0.97	0.99	0.98	0.94	0.97	0.95
airport	0.92	0.93	0.92	0.87	0.94	0.90
baseball diamond	0.83	0.98	0.90	0.89	0.90	0.90
basketball court	0.80	0.94	0.86	0.81	0.90	0.85
beach	0.94	1.00	0.97	0.71	0.96	0.82
bridge	0.96	0.95	0.95	0.77	0.88	0.82
chaparral	0.98	0.98	0.98	0.90	0.99	0.94
church	0.76	0.88	0.82	0.75	0.85	0.80
circular farmland	0.95	0.95	0.95	0.98	0.97	0.97
cloud	0.99	0.92	0.95	0.83	0.99	0.90
commercial area	0.85	0.92	0.88	0.88	0.94	0.91
dense residential	0.76	0.96	0.85	0.69	0.88	0.76
desert	0.87	0.97	0.92	0.83	0.89	0.86
forest	0.82	0.98	0.89	0.73	0.99	0.84
freeway	0.94	0.82	0.88	0.72	0.79	0.76
golf course	0.97	0.98	0.97	0.94	0.99	0.96
ground track field	0.93	0.91	0.92	0.75	0.99	0.85
harbor	0.98	0.98	0.98	0.89	0.98	0.93
industrial area	0.71	0.84	0.77	0.57	0.96	0.71
intersection	0.96	0.84	0.90	0.75	0.90	0.82
island	0.93	0.97	0.95	0.98	0.86	0.92
lake	0.78	0.94	0.85	0.81	0.81	0.81
meadow	0.89	0.91	0.90	0.88	0.87	0.88
medium residential	0.74	0.90	0.81	0.75	0.90	0.82
mobile home park	0.95	0.95	0.95	0.94	0.92	0.93
mountain	0.80	0.97	0.88	0.84	0.79	0.82
overpass	0.93	0.94	0.93	0.76	0.95	0.84
palace	0.84	0.62	0.71	0.80	0.58	0.67
parking lot	0.96	0.93	0.94	0.97	0.91	0.94
railway	0.94	0.92	0.93	0.90	0.69	0.81
railway station	0.85	0.59	0.69	0.63	0.95	0.75
rectangular farmland	0.81	0.92	0.86	0.75	0.97	0.85
river	0.90	0.79	0.84	0.72	0.83	0.77
roundabout	0.94	0.84	0.89	0.88	0.67	0.79
runway	0.97	0.83	0.89	0.91	0.51	0.68
sea ice	0.94	0.97	0.96	0.91	0.97	0.94
ship	0.98	0.93	0.95	0.97	0.55	0.71
snowberg	0.93	0.96	0.95	0.89	0.82	0.86
sparse residential	0.94	0.79	0.86	0.91	0.81	0.86
stadium	0.81	0.93	0.87	0.78	0.87	0.82
storage tank	0.88	0.83	0.85	0.82	0.74	0.79
tennis court	0.89	0.65	0.75	0.83	0.77	0.80
terrace	0.90	0.77	0.83	0.87	0.54	0.69
thermal power station	0.86	0.71	0.78	0.79	0.41	0.54
wetland	0.88	0.54	0.67	0.83	0.47	0.60
Macro Average	0.90	0.89	0.89	0.86	0.83	0.84
Weighted Average	0.91	0.89	0.90	0.82	0.81	0.81

Table 7: Detailed per-class performance of BAGL and DMRS on UCMerced Land Use ($\rho = 0.1$).

Class Name	BAGL (Ours)			DMRS		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
agricultural	0.95	1.00	0.97	1.00	0.62	0.77
airplane	0.93	1.00	0.97	1.00	0.94	0.97
baseball diamond	1.00	1.00	1.00	0.95	0.95	0.95
beach	0.95	0.98	0.96	0.96	0.98	0.97
buildings	0.93	0.88	0.90	0.92	0.65	0.76
chaparral	1.00	1.00	1.00	0.92	0.96	0.94
dense residential	1.00	0.71	0.83	0.64	0.89	0.74
forest	0.95	1.00	0.98	0.70	1.00	0.82
freeway	0.75	1.00	0.86	0.79	0.90	0.84
golf course	1.00	1.00	1.00	1.00	0.95	0.97
harbor	1.00	1.00	1.00	1.00	1.00	1.00
intersection	0.90	0.95	0.93	0.54	1.00	0.70
medium residential	0.75	1.00	0.86	0.93	0.76	0.84
mobile home park	1.00	1.00	1.00	0.95	0.91	0.93
overpass	1.00	0.75	0.86	0.80	0.47	0.59
parking lot	1.00	1.00	1.00	1.00	1.00	1.00
river	0.90	0.93	0.92	0.56	0.59	0.57
runway	1.00	1.00	1.00	0.94	0.89	0.92
sparse residential	0.95	0.93	0.94	0.93	0.93	0.93
storage tanks	0.90	0.86	0.88	0.89	0.59	0.71
tennis court	1.00	0.89	0.94	0.95	1.00	0.98
Macro Average	0.95	0.94	0.94	0.87	0.86	0.85
Weighted Average	0.95	0.95	0.95	0.88	0.86	0.86

C.3 Results on UCMerced Land Use

Table 7 reports results on UCMerced Land Use Yang & Newsam (2010), a 21-class benchmark with 100 images per class. BAGL achieves a macro-averaged F1 of 0.94 versus 0.85 for DMRS, and a weighted-averaged F1 of 0.95 compared to 0.86, representing the largest per-metric gains among all three benchmarks.

BAGL attains perfect F1 scores on six categories – baseball diamond, beach, chaparral, golf course, mobile home park, and runway – while matching DMRS's perfect scores on harbour and parking lot. The most striking improvements occur in classes where DMRS suffers severe confusion: river (F1: 0.57 $\rightarrow$ 0.92), overpass (F1: 0.59 $\rightarrow$ 0.86), and intersection (F1: 0.70 $\rightarrow$ 0.93). These are categories with limited training data and high visual overlap with neighbouring classes, precisely the setting where OGSM's domain-aware augmentation and BAGL's adaptive expert allocation are designed to help.

An informative comparison concerns the three residential density levels. DMRS exhibits uneven performance across dense (F1: 0.74), medium (F1: 0.84), and sparse (F1: 0.93) residential classes, suggesting difficulty in capturing fine-grained density distinctions. BAGL yields more balanced scores across the three levels (0.83, 0.86, and 0.94 respectively), indicating that the ontology-guided mixing helps the model learn the subtle structural differences among these visually similar categories.

D VISUAL EXPLANATION ANALYSIS: GRADCAM VISUALIZATIONS

D.1 ATTENTION PATTERNS ON TAIL CLASSES

To complement the quantitative results in the main paper, we employ Gradient-weighted Class Activation Mapping (GradCAM) Selvaraju et al. (2017) to examine the internal representations learned by BAGL under extreme data scarcity. Figure 6 presents side-by-side visualizations for five representative tail classes from NWPU-RESISC45, each containing fewer than 30 training samples.Comparing BAGL with DMRS Wang et al. (2025) highlights notable differences in where each model focuses its attention.

D.2 SEMANTIC COHERENCE AND FEATURE DISCRIMINATION

Two consistent patterns emerge from the activation maps. First, BAGL exhibits *semantically coherent* attention: activations concentrate on task-relevant structures rather than diffuse background regions. For instance, in the Railway example (26 training samples), BAGL strongly activates along the track and platform, whereas DMRS produces scattered, low-intensity activations across the scene, yielding a substantially lower confidence (65% vs. 87%). Similarly, for Runway (13 samples), BAGL attends to the linear pavement markings and runway surface, while DMRS focuses on surrounding grass texture and misclassifies the image as Freeway (68% confidence).

Second, BAGL demonstrates superior *inter-class discrimination* between visually similar categories. The Stadium example (9 training samples) is particularly illustrative: BAGL correctly identifies the oval structure and seating arrangement with 85% confidence, whereas DMRS confuses the scene with Tennis Court at only 58% confidence. This failure mode aligns precisely with the motivation for OGSM – without domain-aware semantic guidance, generic embeddings conflate structurally related but ontologically distinct classes. The same pattern recurs for Wetland (7 samples), where DMRS attends to generic water features and predicts Lake (53%), while BAGL captures the characteristic combination of water channels and vegetation, correctly predicting Wetland with 79% confidence.

D.3 LINKING VISUAL EVIDENCE TO ARCHITECTURAL DESIGN

These qualitative observations connect directly to the three architectural components of BAGL. The improved attention coherence is consistent with the multi-expert design, because different LoRA experts specialize in complementary feature subspaces, the ensemble avoids collapsing onto head-class patterns even when tail-class training data are scarce. The adaptive tempering mechanism further ensures that experts assigned to underrepresented classes retain sufficient gradient signal throughout training, preventing the attention degradation visible in the DMRS maps for extreme tail classes such as Wetland and Stadium. Finally, the improved ability to distinguish between visually similar classes supports the effectiveness of OGSM. By adapting the embedding space to reflect domain-specific class relationships, OGSM encourages the model to focus on the features that truly differentiate one class from another.

Across all five examples in Figure 6, BAGL correctly classifies every tail-class sample, while DMRS fails on three. The confidence margins are consistently wider for BAGL (averaging 87% vs. 67% for DMRS), indicating not only more accurate but also more decisive predictions. These visualizations confirm that the tail-class gains reported in Table 1 of the main paper stem from more meaningful learned features, not merely from score recalibration.

E BROADER IMPACTS

Our proposed framework (OGSM and BAGL) significantly improves the identification of rare but operationally critical land-cover categories. In environmental monitoring and climate change research, the accurate recognition of tail classes such as sea ice, snowbergs, and wetlands is vital for tracking ecological degradation and managing natural resources. Furthermore, in urban planning and disaster management, the ability to reliably identify specialized infrastructure (e.g., solar power stations, bridges, or storage tanks) despite severe data imbalance can enhance emergency response times and support sustainable development initiatives. By preventing the model from ignoring rare

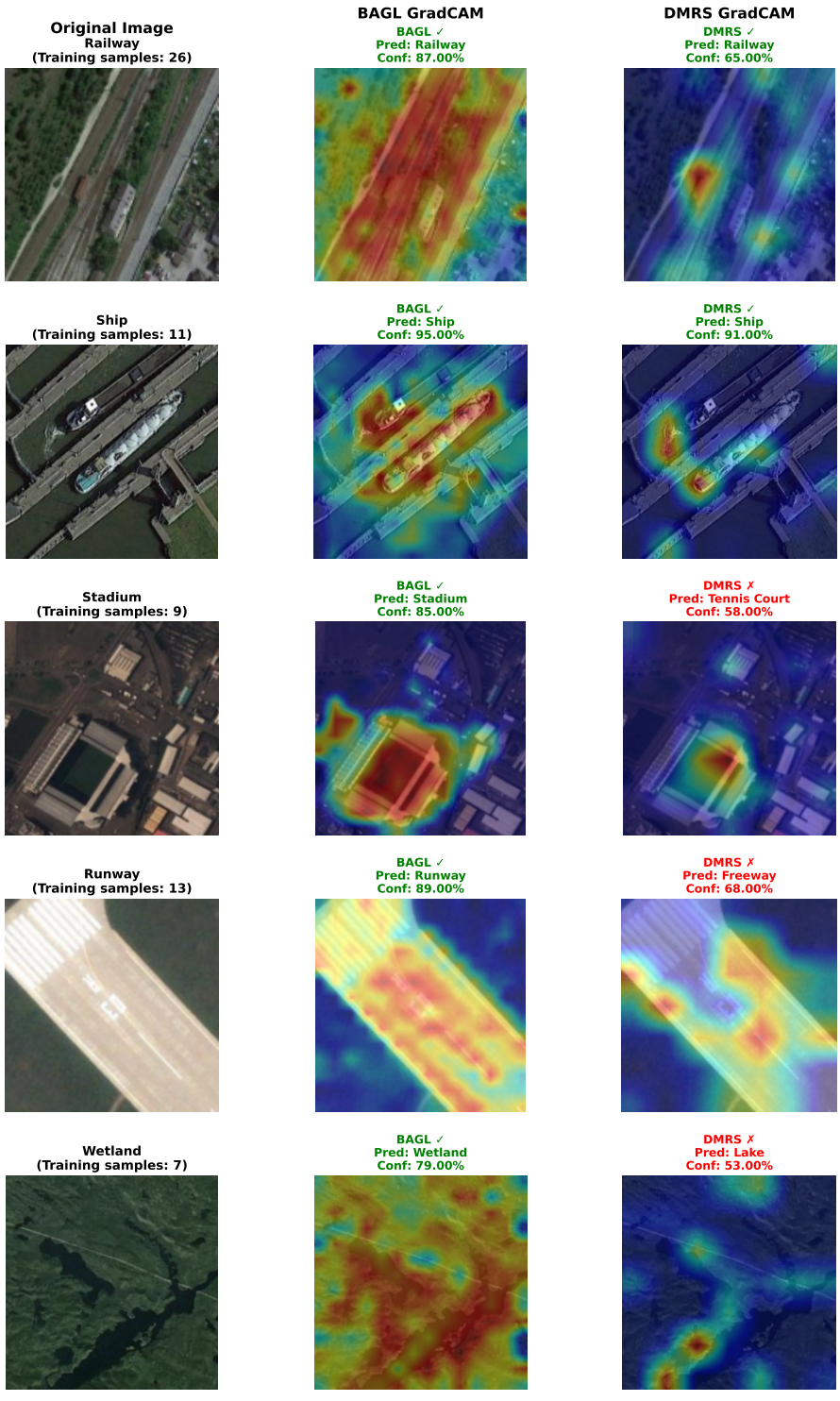


Figure 6: GradCAM visualization comparison between BAGL and DMRS on tail-class samples from NWPU-RESISC45. Each row shows a different tail class with the number of training samples indicated. BAGL produces focused, semantically coherent activation maps and correctly classifies all five examples, whereas DMRS misclassifies three out of five, attending to non-discriminative or ambiguous regions. **Green** headers denote correct predictions, while **red** headers denote incorrect predictions.

classes in favor of dominant ones, our method promotes fairer and more comprehensive geographical analyses.