12-2024

Learning de-biased representations for remote-sensing imagery

Zichen TIAN
Singapore Management University, zichen.tian.2023@phdcs.smu.edu.sg

Zhaozheng CHEN
Singapore Management University, zhaozhengc@smu.edu.sg

Qianru SUN
Singapore Management University, qianrusun@smu.edu.sg

Citation

TIAN, Zichen; CHEN, Zhaozheng; and SUN, Qianru. Learning de-biased representations for remote-sensing imagery. (2024). *Proceedings of the 38th Conference on Neural Information Processing Systems, Vancouver, Canada, 2024 December 10-15*.
Available at: https://ink.library.smu.edu.sg/sis_research/9400

Learning De-Biased Representations for Remote-Sensing Imagery

Zichen Tian Zhaozheng Chen Qianru Sun
School of Computing and Information Systems
Singapore Management University
{zichen.tian.2023,zzchen.2019}@phdcs.smu.edu.sg, qianrusun@smu.edu.sg

Abstract

Remote sensing (RS) imagery, requiring specialized satellites to collect and being difficult to annotate, suffers from data scarcity and class imbalance in certain spectrums. Due to data scarcity, training any large-scale RS models from scratch is unrealistic, and the alternative is to transfer pre-trained models by fine-tuning or a more data-efficient method LoRA [22]. Due to class imbalance, transferred models exhibit strong bias, where features of the major class dominate over those of the minor class. In this paper, we propose debLoRA—a generic training approach that works with any LoRA variants to yield debiased features. It is an unsupervised learning approach that can diversify minor class features based on the shared attributes with major classes, where the attributes are obtained by a simple step of clustering. To evaluate it, we conduct extensive experiments in two transfer learning scenarios in the RS domain: from natural to optical RS images, and from optical RS to multi-spectrum RS images. We perform object classification and oriented object detection tasks on the optical RS dataset DOTA and the SAR dataset FUSRS. Results show that our debLoRA consistently surpasses prior arts across these RS adaptation settings, yielding up to 3.3 and 4.7 percentage points gains on the tail classes for natural → optical RS and optical RS → multi-spectrum RS adaptations, respectively, while preserving the performance on head classes, substantiating its efficacy and adaptability [1].

1 Introduction

Remote sensing (RS) is crucial in various applications such as environmental monitoring, resource management, and disaster response [70, 36]. RS data is collected by various sensors and has multiple spectrums, including optical RS imagery (dubbed as ORS, 400–700nm) [32], multi-spectral RS imagery (MSRS, 400–2500nm) [8], and synthetic aperture radar imagery (SAR, 1mm-1m) [48, 13]. These spectrums differ significantly in imaging mechanisms, leading to distinct data characteristics and processing pipelines [71]. Given this diversity, learning robust and generic representation models for such data is desirable to reduce processing costs and complexities.

Recently, in natural image domains, large-scale pre-trained visual foundation models (*e.g.*, CLIP [45], Stable Diffusion [47], and DINO [4]) have shown great advances in robustness and generalization ability. The zero-shot features extracted from the models show impressive performance in downstream tasks such as object classification, detection and semantic segmentation [66], even outperforming the supervised models trained on the specific datasets of those tasks. However, in the RS domain, training such foundation models from scratch remains challenging. Even though some trials have been made in past years [8, 16], their works have clear limitations. First, they require large-scale RS data for effective training, which are available for only ORS but not other spectrums such as SAR

[1] Code: https://github.com/doem97/deblora

and MSRS [43, 10, 17]. Collecting and annotating images in "other" spectrums is difficult due to many factors such as military restrictions, sensor availability, and high acquisition costs, so the data scarcity is unlikely to be alleviated in the near future [71]. Second, their works are constrained in small- or medium-scale models, *i.e.*, they use ViT-L (300M) in [8] and Swin-L (197M) in [16], while the foundation models in the natural image domain are much larger (*e.g.*, Latent Diffusion has 860M, and OpenCLIP-H/14 has 986M). Third, their training-from-scratch approaches are computationally inefficient, requiring a huge amount of GPU memory (VRAM). For instance, [16] reported the need of 80 * A100 GPU with 80GB VRAM each, totaling 6.4TB.

Instead of learning a foundation model from scratch, we propose to transfer existing foundation models to RS domains. This approach is both data-efficient and computation-efficient. We answer two questions: 1) Which foundation models to transfer? 2) Which transfer learning methods to use?

For the first question, we consider foundation models pre-trained on natural images (*e.g.*, CLIP [45], Stable Diffusion [47]) as well as the models from remote sensing (RS) images (*e.g.*, SkySense [16]). A positive aspect of these models is that they contain the semantic knowledge necessary for learning a new RS domain. However, a great challenge is the large domain gap between natural images and RS domains, or between different RS spectrums. In our preliminary study, we conduct validation experiments. Fortunately, we observe successful transfer results both from natural to ORS in Figure 1 and between different RS spectrums in Table 3, when compared to the method of TRS-Res101 [65] which does not perform any transfer learning. The success of natural→ORS is due to the shared underlying visual elements like edges, textures, and contours, which are intrinsic to both natural and RS images. The success of ORS→other RS is due to the shared spatial structures, *e.g.*, urban areas, buildings, and object outlines, in different RS spectrums.

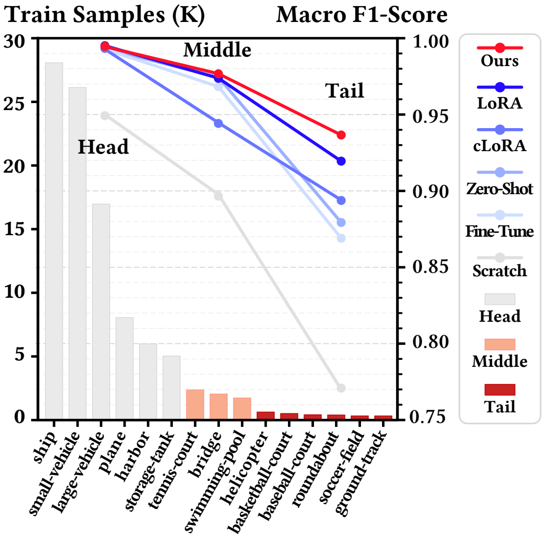
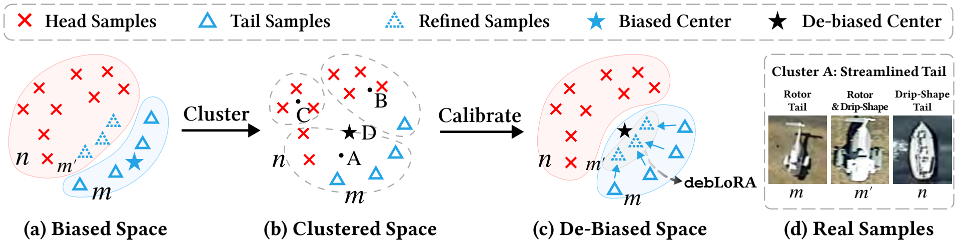


Figure 1: **Long-tailed Problems.** This figure shows 1) ORS datasets (take DOTA [59] as an example) have the long-tailed distribution issue. 2) Model adaptation methods suffer from weak performance in tail classes.

For the second question, we found that data-efficient transfer learning methods on foundation models exhibit a strong bias towards major classes. As shown in Fig. 1, both Fine-Tune and LoRA have significantly lower F1 scores for tail classes. This is because their learned feature space is biased towards the discriminative features of head classes while neglecting the tail [62]. Taking the head class `ship` (which takes 28.35%) and tail class `helicopter` (0.64%) as examples on the DOTA dataset [59]. Fig. 2(a) shows biased LoRA features of "oval tail" in the `ship` sample n and "rotor tail" in the `helicopter` sample m. We say biased because the LoRA fails to understand the "oval tail with a rotor" in another `helicopter` sample m' and embeds m' wrongly as a `ship` sample in the feature space. Please note that the real feature distribution is shown in Figure 3 to support the illustration of Figure 2. This long-tail issue is particularly severe for transfer learning in the RS domain due to two reasons. *First, RS datasets suffer from more severe data imbalance than natural image datasets.* For instance, the imbalance ratios[2] of RS datasets DOTA and ShipRSImageNet reach 86 and 112, respectively, while CIFAR100-LT [2], a natural image dataset with a similar data scale, has a ratio of only 50. This is because annotating under-represented tail class samples in RS, *e.g.*, identifying a rare naval vessel, such as the "Nimitz", from SAR image, requires a high level of domain expertise. *Second, the data scarcity in RS domains determines that RS adaptation methods must be data-efficient*, such as LoRA. However, as shown in Table 2, using fewer parameters in LoRA (being more data-efficient) exacerbates long-tail issues. The reason is that this restricts the model capacity and forces the model to prioritize a limited number of features—usually from head classes.

[2] The imbalance ratio is measured by n_1/n_k, where 1 and k are the largest and smallest categories. It reflects the severity of data imbalance [69].



(a) Biased Space (b) Clustered Space (c) De-Biased Space (d) Real Samples

Figure 2: **Two key steps of debLoRA: feature clustering and calibration.** (a) The baseline LoRA feature space is biased towards head classes. Red crosses ✗ represent head class samples, and blue triangles △ represent tail class samples. The blue star ★ indicates the center of tail class samples. Dashed blue triangles △ show the validation samples of the tail class wrongly embedded in the head class region, indicating the model bias towards head classes. (b) We cluster all features (clusters denoted by gray dotted boundaries) regardless of class labels. A, B and C are cluster centers used to generate a de-biased center D, as in Eq. 2. (c) We calibrate the tail class features by "moving" them closer to D, as in Eq. 3. After these steps, we train the debLoRA module on the calibrated features of tail classes (together with the original head class features).

To mitigate this bias without needing more data or labels in tail classes, we propose an unsupervised learning approach, debiased LoRA, dubbed debLoRA. debLoRA is based on the features extracted from LoRA (or a LoRA variant) and is generic to LoRA variants. To be concise, we use LoRA in the following to represent itself and its variants. Given the LoRA features, debLoRA has three steps: clustering, calibration, and training. First, it clusters all the features regardless of class labels by K-means. Each obtained cluster center represents an attribute from one or shared by multiple classes. Second, these cluster centers are used to calibrate the LoRA features of tail classes and enhance the territory of tail classes in the feature space. We illustrate these two steps in Figure 2. Last, the calibrated features are used as the learning objectives to train a debLoRA module with a similar network architecture to LoRA. The learned debLoRA is thus a de-biased feature extractor.

We observe that after K-means clustering, each cluster center captures a general visual attribute shared across different classes. For instance, in Figure 2(b), cluster A corresponds to the general vehicle attribute "streamlined tail", which includes both head class sample n and tail class sample m. Such clusters can thus yield a balanced representation base, making the tail more robust by integrating common attributes with the head.

One may ask "what if some attributes are dominated by the attribute features of head classes?" We address this question by proposing a weighting scheme, in the step of calibration. In specific, for each tail class sample (*e.g.*, m in Fig. 2(c)), we calibrate it by forcing its feature closer to the de-biased center (D)—the weighted average of all cluster centers. The weights are determined by the number of samples in each cluster, ensuring that this center is not dominated by clusters with mostly head class samples. This calibration process results in de-biased representations that capture a more comprehensive range of visual attributes shared across classes, leading to improved features of tail classes (*e.g.*, m'). Lastly, we re-train a LoRA module to map biased representations towards these debiased centers. Please find more details of justifications in Sec. 4.4. Our method significantly improves the features of tail classes. Moreover, it is efficient as it learns only a lightweight low-rank module while keeping the original foundation model frozen.

Our contributions can be concluded three-fold: 1) We demonstrate the effectiveness of adapting foundation models for data-scarce RS domains. 2) We propose Incremental LoRA, a novel method that de-biases category-specific representations for long-tailed RS adaptation. 3) We conduct extensive experiments to validate our approach on multiple RS adaptation settings and downstream tasks.

2 Related Works

Representation Learning for RS Images. Self-supervised representation learning in RS image domains mainly includes contrastive- and generative-based methods. **Contrastive-based** methods, such as Tile2vec [27], Seasonal contrast [37] and SauMoCo [30], heavily rely on rich temporal data or high-resolution samples, which are often unavailable for data-scarce RS spectrums [56]. **Generative-based** methods, such as RR-SSL [67] and SGSAGANs [15], reconstruct inputs to capture the global data distribution and learn fine-grained patterns. However, they require large-scale data to form robust latent space [14]. Recently, **foundation models** in the RS domain, such as

SatMAE [8], SpectralGPT [20], and SkySense [16], have shown promising results for ORS tasks. SpectralGPT [20] tackles spectrum diversity by pre-training separate tokenizers for each spectrum, which still needs large amounts of data. Another problem is that existing RS foundation models are much smaller than those in the natural image domain (*e.g.*, SatMAE-L [8] has 300M parameters *v.s.* 986M of OpenCLIP-H/14 [5]). Instead of learning RS foundation models from scratch, we propose to adapt them from pre-trained models to RS. Our approach 1) greatly reduces the computational cost, 2) can be easily adapted to various data-scarce RS spectrums, and 3) benefit from the strong representation power of large-scale foundation models in other domains.

Long-tailed Data Distribution and its Bias Problem. Long-tailed data distribution, where a few head classes cover most of the samples, is prevalent in both natural and RS image domains [54, 69]. This imbalance leads to biased feature representations, where the model focuses on discriminative features for head classes while neglecting subtle but crucial features for tail classes [69, 29]. Zhang et al. [69] observed that such a feature space is usually broader for head classes than tail classes, and the decision boundary tends to be biased towards head classes, *i.e.*, many false positive predictions for head classes. Existing solutions include sample-level, meta-learning, and representation-level approaches [69]: **Sample-level** methods, such as re-sampling [49] and data augmentation [7], aim to directly balance the sample distribution. However, they require sample annotations [2, 49] or rely on data diversity [7], both of which are unrealistic in the data-scarce RS spectrums such as SAR [13] and MSRS [8]. **Meta-learning** methods [26, 57] formulate the problem as "learning to learn" and adapt the model to a balanced meta-test set. They depend on the data diversity of the training sets and the availability of balanced validation sets, and therefore, are less applicable for data-scarce RS domains. The **representation-level** methods enhance the learned representation space, including metric learning losses [23], margin-based losses [2], and feature transfer from head to tail classes [33, 63]. However, they are designed for supervised single-domain settings and do not address the challenges of model adaptation to RS: 1) handling multiple downstream tasks (*e.g.*, small object detection, scene segmentation, change detection), and 2) multiple spectrums (such as ORS and SAR). In contrast, we propose an *unsupervised adaptation* method to tackle these challenges in this paper.

Transfer Learning in Remote Sensing. Transfer learning in remote sensing primarily focuses on adaptation within the optical imagery domain. They can be categorized into supervised and unsupervised methods. Supervised methods [12, 35, 46, 44, 39] align distributions using target labels. However, they require task-specific annotations, which are scarce in SAR and multispectral domains and limit the applicability of the obtained models to multiple downstream tasks. Unsupervised DA (UDA) methods aim to learn domain-invariant features without requiring labeled data in the target domain, including transfer component analysis [42, 40], manifold alignment [53, 60, 61], and adversarial learning [1, 11, 51]. However, they are designed for single-source, single-target adaptation within the same spectrum [41, 38]. Besides, the manifold alignment and adversarial methods require significant computational resources, often involving the training of several copies of the source model, while component analysis methods involve complex pipelines. These factors make them unsuitable for foundation models, which are already computationally intensive. In contrast, our method tackles multi-spectrum adaptation without requiring extra labels. It is also computationally efficient.

3 LoRA and cLoRA

Our debLoRA is based on the LoRA [22] or its variants [64], but is orthogonal and generic to them.

LoRA. LoRA was initially proposed to adapt a pre-trained large-scale language model to downstream tasks. It assumes adapted parameters are sparse during model training when the data is limited. It introduces a low-rank factorization of the difference between original and adapted parameters, *i.e.*, $\Delta\theta := B{\cdot}A$. Here, $\theta \in \mathbb{R}^{d\times k}$ represents the parameters of pre-trained model, and $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ denote low-rank factors, with $r \ll \min(d, k)$. The updated parameters $\hat{\theta}$ are thus given by $\hat{\theta} = \theta{+}\Delta\theta = \theta{+}B{\cdot}A$. During inference, the obtained LoRA modules could be combined through a weighted sum, $\hat{\theta} = \theta + \sum_i w_i\Delta\theta_i$, where w_i denotes combination weights.

cLoRA. To tackle the long-tailed issue of LoRA, we also explore its variant cLoRA [64]. The key idea of cLoRA is to learn a separate LoRA module for each class, denoted as $\Delta\theta_c$ for class c, to ensure that the learned representations of one class do not interfere with those of other classes. Formally, the adapted parameters for class c are given by $\hat{\theta}_c = \theta + \Delta\theta_c = \theta + B_c \cdot A_c$, where $B_c \in \mathbb{R}^{d\times r}$ and $A_c \in \mathbb{R}^{r\times k}$ are the low-rank factors specific to class c. During training, each cLoRA module $\Delta\theta_c$ is optimized using only the data from class c, allowing it to capture class-specific features. During inference, as there is no class label available, we use all the cLoRA modules to extract features for the input. Specifically, for an input x, we obtain the features $z_c = \hat{\theta}_c(x)$ using each cLoRA module $\hat{\theta}_c$. The final feature representation is then obtained by concatenating the features from all the cLoRA: $z = [z_1; z_2; \ldots; z_C]$, where C is the total number of classes.

4 De-biased LoRA (debLoRA)

The algorithm of debLoRA consists of two steps: generating debiased features, and then using them to train a debLoRA module. In the first step, we perform unsupervised clustering on biased feature space $\mathcal{Z}$ (*i.e.*, composed by original LoRA features biased to head classes) to obtain debiased features $\hat{\mathcal{Z}}$. In the second step, we use $\hat{\mathcal{Z}}$ as the learning target to train a debLoRA module. The debLoRA learns the mapping between biased and de-biased features. We justify the feasibility of learning such a mapping in Section 4.4.

4.1 Problem Formulation

Given a pre-trained feature extractor $f : \mathcal{X} \rightarrow \mathcal{Z}$ and a long-tailed RS dataset $\mathcal{D} = (x, y)$, where $x \in \mathcal{X}$ is an RS image, $y \in \mathcal{Y}$ is its annotation and $\mathcal{Z}$ is the biased feature space[3], our goal is to adapt f to the target dataset $\mathcal{D}$ while yielding a de-biased feature space $\hat{\mathcal{Z}}$, *i.e.*, adapted encoder is $\hat{f} : \mathcal{X} \rightarrow \hat{\mathcal{Z}}$. The de-biased feature representation $\hat{\mathcal{Z}}$ should improve downstream task performance on tail classes without sacrificing the performance on head classes.

4.2 Stage 1: Representation De-biasing

Feature Clustering. Given a pre-trained encoder $f_\theta : \mathcal{X} \rightarrow \mathcal{Z}$ that maps input images to a biased representation space, where f_θ is parameterized by θ, we first extract features for each sample in the dataset: $z_i = f_\theta(x_i)$, $i \in N$. We then apply K-means clustering on $\{z_i\}$ to obtain K clusters. To mitigate imbalanced clusters, we impose a constraint that each cluster should contain at least $\frac{N}{K \cdot \rho}$ samples, where ρ is a pre-defined constant. The clustering objective is:

$$\min_{\mu_k} \sum_{i=1}^{N} \min_k \|z_i - \mu_k\|^2, \quad \text{s.t. } \forall k,\ n_k \geq \frac{N}{K \cdot \rho}, \tag{1}$$

where μ_k and n_k denote the center and size of the k-th cluster, respectively.

De-biased Cluster Centers. For each tail class c, we calculate its de-biased representation center $\hat{\mu}_c$ by weighted averaging all the cluster centers:

$$\hat{\mu}_c = \sum_k w_k \cdot \mu_k, \quad \text{where } \ w_k = \frac{n_k}{n_c}. \tag{2}$$

Here n_k denotes the number of samples from class c in the k-th cluster, and n_c is the total number of samples in class c. The weight w_k is proportional to the fraction of class c samples in the k-th cluster. This ensures that the de-biased center $\hat{\mu}$ is not dominated by head classes.

4.3 Stage 2: De-Biased Low Rank Adaptation (debLoRA)

Tail Class Calibration. For each tail class sample x with representation z, we calibrate z by moving it closer to the de-biased center $\hat{\mu}$:

$$\tilde{z} = \alpha z + (1 - \alpha)\hat{\mu}, \tag{3}$$

where $\alpha \in [0, 1]$ is a hyper-parameter controlling the degree of calibration. We empirically set α based on the imbalance ratio γ of each tail class: $\alpha = \min(1, \frac{10}{\gamma})$. For tail classes with larger imbalance ratio, a higher α encourages the calibrated representation $\tilde{z}$ to be closer to the de-biased center $\hat{\mu}$, as the original representation z is less reliable due to its learning from limited samples. While for classes with smaller γ, a lower α is used to retain the discriminative information of z. For instance, the DOTA dataset's tail class helicopter has high $\gamma = 45.45$, so its α reaches 0.22.

[3] We define feature space $\mathcal{Z}$ as biased if Vol$(\mathcal{Z}_h) \gg$ Vol$(\mathcal{Z}_t)$, and $\exists z_t \in \mathcal{Z}_t : P(z_t \in \mathcal{Z}_h) > P(z_t \in \mathcal{Z}_t)$, where $\mathcal{Z}_h$ and $\mathcal{Z}_t$ denotes the feature spaces of head and tail classes respectively, Vol$(\cdot)$ denotes feature space volume, and $P(\cdot)$ denotes the probability predicted by the model.

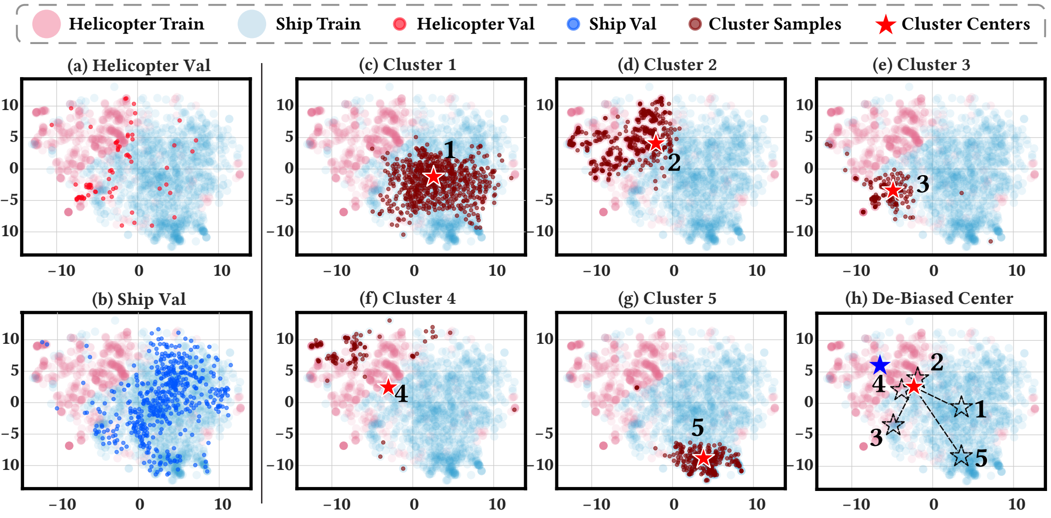


Figure 3: **t-SNE visualization of validation samples and clusters.** The first column shows the distribution of `helicopter` (tail) and `ship` (head) validation samples. Subfigures (c)-(g) are the clusters and their centers when K=5 in K-means. In (h), the dotted lines and stars indicate that we compute a de-biased center for the tail class (`helicopter`) by weighted averaging the five cluster centers, and the blue star is the original biased center of `helicopter` training samples.

Learning debLoRA. With the pre-trained encoder f_θ frozen, we learn a LoRA module $g_\phi : \mathcal{Z} \rightarrow \hat{\mathcal{Z}}$ parameterized by ϕ to map the biased representations to the calibrated ones. The training objective is:

$$\min_{\phi} \frac{1}{|\mathcal{D}_t|} \sum_{x \in \mathcal{D}_t} \| g_\phi(f_\theta(x)) - \tilde{z} \|^2, \tag{4}$$

where $\mathcal{D}_t$ is the set of tail class samples. During inference, we apply the learned LoRA module to extract the de-biased representations $z = g_\phi(f_\theta(x))$ for an input image x. The complete algorithm of debLoRA is summarized in Algorithm 1.

Algorithm 1 debLoRA

```
Require: Long-tailed training set $\mathcal{D} = \{(x, y)\}$, pre-trained encoder $f_\theta : \mathcal{X} \rightarrow \mathcal{Z}$, number of
    clusters $K$, balance factor $\rho$
Ensure: A LoRA module $g_\phi$ that de-biases $f_\theta$
Extract biased representations $z = f_\theta(x)$ for each sample $x \in \mathcal{D}$ using pre-trained $f_\theta$
Perform constrained $K$-means clustering on $\{z\}$ (equation 1) to obtain cluster centers $\{\mu_k\}_{k=1}^K$,
    where each cluster has at least $\frac{N}{K \cdot \rho}$ samples
for each tail class $c$ do
    Calculate its de-biased representation center $\hat{\mu}_c$ by weighted averaging all cluster centers
    $\{\mu_k\}_{k=1}^K$ (equation 2)
    for each sample $x \in \mathcal{D}_c$ do
        Extract biased representation $z = f_\theta(x)$
        Calibrate $z$ to $\tilde{z}$ by moving it closer to $\hat{\mu}_c$ with factor $\alpha = 10/\gamma$ (equation 3)
    end for
end for
Learn a LoRA module $g_\phi : \mathcal{Z} \rightarrow \hat{\mathcal{Z}}$ to map biased representations to calibrated ones
return $g_\phi$
```

4.4 Justification

We discuss the biased representation space of LoRA, and then justify the effectiveness of our three critical operations in debLoRA: **clustering**, **weighting**, and **calibration**. We show the real sample distribution in Figure 3 and an illustrative example in Figure 2.

LoRA is Biased. The feature space learned by LoRA is biased towards head classes [62], evidenced by two observations. 1) The head class representations over-expand their territory into the tail class space. As shown in Figure 3, most of the `ship` (head) validation samples are distributed within its own representation space, while many `helicopter` (tail) validation samples are wrongly distributed in the ship's space. 2) The center of the entire space is biased towards head class, as the ship training samples significantly overlap with the helicopter training samples. This bias occurs because, during training, the encoder is exposed to significantly more diverse samples of head class.

Clustering. By feature clustering, we obtain a set of cluster centers that benefit the tail classes in two ways. 1) *Improved robustness.* The obtained cluster centers, shown as red stars in Figure 3(c)-(g), represent visual prototypes[3], *i.e.*, general visual attributes common to both head and tail classes, such as "streamlined tail" or "with wooden deck". These cluster centers are more robust than the original tail class representations because they leverage the diversity of head class samples. 2) *Reduced imbalance.* Certain clusters exhibit reduced long-tail issues. The clusters in Figure 3(d)-(f) contain more samples from helicopter than ship. This is because the clusters are formed based on intrinsic visual similarities among images, regardless of their imbalanced class labels. Using these cluster centers avoids the risk of tail class attributes (*e.g.*, "rotor tail" and its variants in helicopter) being overwhelmed by head class attributes (*e.g.*, "oval tail" and its variants in ship).

Weighting and Calibration. One might ask, "Are the data imbalances within each cluster or among different clusters still issues?" *E.g.*, the 5-th cluster in Figure 3 contains only ship samples and seems irrelevant to helicopter. To answer this, we perform the *weighted averaging* over cluster centers, and the *calibration* over tail class samples: 1) *Weighted averaging.* When calculating the de-biased representation center for each tail class (equation 2), we assign higher weights to clusters containing a larger fraction of that particular tail class. The de-biased center (red star in Figure 3(h)) better captures the true distribution of the validation samples of helicopter, compared to the original biased center (blue star in Figure 3(h)). 2) *Calibration.* We calibrate the representation of each tail class sample by moving it closer to the class's de-biased center (equation 3). The calibration factor α is inversely proportional to the imbalance ratio of the tail class. This design ensures severely underrepresented tail classes like helicopter receive stronger calibration.

5 Experiments and Analyses

We evaluate our debLoRA on two settings: 1) adapting natural image foundation models to RS, and 2) adapting ORS foundation models to SAR. For the first setting, we conduct experiments on two representative RS tasks: object classification and oriented object detection. For the second setting, we conduct experiments on a representative SAR task—fine-grained ship classification.

Natural → RS adaptation. 1) *Foundation model.* We use two state-of-the-art foundation models: Stable Diffusion v1.5 (SD) [47] and OpenCLIP [25]. Both models have shown impressive generalization ability on various tasks when adapted to domains like medical images [58]. However, their transferability from natural images to the RS domain remains under-explored. 2) *RS dataset.* We use the DOTA dataset [10], a large-scale benchmark for RS object recognition. DOTA contains 188,282 instances from 15 categories, covering various scales, orientations, and shapes. We define the long-tail split as follows: 6 classes with <1% instances as tail, 3 classes with 1%-5% instances as middle, and the remaining 6 classes (each with >5% instances) as head. This split exhibits a clear long-tail distribution, evidenced by the performance gap between head and tail classes for the baseline methods (see Table 1 row 1). 3) *Tasks.* For the classification task, we obtain features from the adapted foundation models and train a linear classifier. We report the macro F1-score that fairly evaluate the performance across all classes. For detection, we train a FCOS detector head [52] on obtained representations and evaluate using the mAP.

ORS → SAR adaptation. 1) *Foundation model.* We use SatMAE-L [8], the state-of-the-art open-sourced foundation model for RS. SatMAE-L is pre-trained on large ORS datasets using self-supervised learning. It has 307M parameters and requires 6,144 GPU hours to train from scratch. 2) *SAR dataset.* We evaluate our method on the fine-grained ship classification task of SAR. Existing SAR ship datasets have insufficient samples to evaluate the model performance reliably, *e.g.*, only 2 samples in test set for tail class "WingInGrnd" on the FUSAR-Ship dataset. We thus create a new dataset by combining two high-resolution ($<$10m/pixel) SAR datasets: FUSAR-Ship [21] and SRSDD [31]. Details of this combined dataset are provided in the Appendix. 3) *Ship classification task.* We follow the same setup as in the natural → RS setting for this SAR task.

Implementation Details. 1) *Fine-tuning baseline.* We fine-tune the foundation models until the training loss stabilizes. During inference, we use null prompts as no ground truth is available. For SD, we extract features from the U-Net after applying one denoising step [50]. For OpenCLIP, we extract features from its visual encoder's final layer before the projection head. 2) *LoRA and variants.*

We apply LoRA modules to all linear layers in the foundation models. We use a rank of 8 for LoRA, as it suffers from the most severe long-tail issues. We also validate our method with higher ranks (*e.g.*, 64) in Table 2. During inference, we extract features from the U-Net encoder output followed by global average pooling (GAP). For cLoRA, we concatenate the category-specific features after GAP. 3) *debLoRA*. The debLoRA involves two hyperparameters: the calibration factor α, and the number of clusters K. We set α as inversely proportional to the imbalance ratio of the tail class, as described in Section 4.4. We empirically set K=32 (ablation study on K are provided in Appendix).

Evaluation Metrics. 1) *Classification.* We use linear probing (i.e., train a linear classifier on the top of frozen features) to evaluate the learned representations [18, 19]. It is simple and avoids introducing additional learning operations. We apply GAP and ReLU on the extracted features before linear probing. We report the macro F1-score, which assigns equal importance to all classes—more suitable for evaluating imbalanced datasets. We report scores for head, middle, and tail classes separately, as well as the overall score averaged across all categories. 2) *Detection.* We use the lightweight FCOS [52], an anchor-free detector head, to avoid potential interference from pre-defined anchors. We extract high-resolution feature maps from the SD U-Net output. During feature clustering and re-training, we use per-instance features for each category. During inference, we extract features from the entire image and feed them to the detector head. We report the mAP metric.

Ablation study. In Table 1, rows 1 and 2 show the results of using zero-shot features of SD or fine-tuned SD features on DOTA to train RS object recognizers. Recognizers' performances are strongly biased to head classes—around 12 percentage points drop for tail classes. From rows 3 and 5, we can see such issues get resolved a bit when using LoRA methods. Rows 4 and 6 show that debLoRA significantly outperforms LoRA methods on tail classes—by 4.2 points and 2.7 points, respectively. Specifically, compared to cLoRA, debLoRA does not even sacrifice the performance for head classes. To quantitatively validate its working mechanism, we analyzed feature discrimination. Results show that debLoRA enlarges inter-class distances and reduces intra-class distances for tail classes (see Appendix). In addition, debLoRA needs just the same amount of parameters as LoRA (0.08M), which is appealing for computation.

Table 1: **Ablation study of debLoRA.** We apply our debLoRA based on LoRA and cLoRA. Results are reported for the adaptation from SD → DOTA recognizer. Params (M) refers to the number of updated parameters during the adaptation. Our results are marked in gray.

Method	Macro F1 Score (%)				Params (M)
	Head	Middle	Tail	Overall	
Zero-Shot	99.2	97.3	87.8	94.3	—
Fine-Tune	99.1	96.7	86.8	93.7	860
cLoRA	99.1	94.3	89.3	94.2	0.08
w/ debLoRA	**99.3**	**97.5**	**93.5**	**96.6**	0.08
LoRA	**99.4**	97.2	91.8	95.9	0.08
w/ debLoRA	99.1	**98.7**	**94.5**	**97.1**	0.08

LoRA Ranks. We investigate the impact of different LoRA ranks on the long-tailed classification performance in Table 2. We have two key observations. 1) As the LoRA rank decreases, the performance on tail classes drops more significantly than on head classes. For example, when the rank is reduced from 64 to 8, the F1-score of tail classes decreases by 2.2 percentage points, while that of head classes even increases by 0.3 percentage. This supports our hypothesis that the limited parameter capacity of low-rank LoRA forces it to prioritize learning the head classes, exacerbating the long-tail problem. 2) debLoRA consistently improves the performance on middle and tail classes across different LoRA ranks. Notably, with rank 64, debLoRA achieves a 2.2 percentage points improvement on tail classes while maintaining the performance on head classes.

Table 2: **Compare LoRA ranks.** The table compares different ranks of the LoRA module. Our results are marked in gray.

Method	Macro F1 Score (%)				Params (M)
	Head	Middle	Tail	Overall	
Rank 8	99.4	97.2	91.8	96.1	0.08
w/ debLoRA	99.1	98.7	94.5	97.1	0.08
Rank 16	99.0	95.9	92.4	95.8	0.16
Rank 32	99.4	96.9	93.0	96.4	0.32
Rank 64	99.1	96.9	94.0	96.7	0.64
w/ debLoRA	99.1	98.7	96.2	98.0	0.64

Compare with SOTA. 1) *Object Classification.* Table 3 compares our debLoRA with state-of-the-art methods under three adaptation tasks. We draw four key observations from the table. 1) debLoRA consistently outperforms LoRA on tail classes across all adaptation tasks, with the largest gain of 4.7 percentage points for ORS → SAR (*i.e.*, SatMAE → FUSRS). This shows the consistent efficiency of our approach in tackling the long-tail problem of RS domains. 2) Compared to SD → DOTA setting, cLoRA performs exceptionally well under OpenCLIP → DOTA setting, slightly surpassing LoRA. We hypothesize that OpenCLIP's feature space aligns particularly well with cLoRA's class-specific

Table 3: **State-of-the-art comparison under different adaptation settings.** The experiments are conducted on two RS adaptation settings: 1) Natural→ORS, where we adopt Stable Diffusion (SD) and OpenCLIP as foundation models and DOTA as the target dataset. 2) ORS→SAR, where we adopt SatMAE as the foundation model and FUSRS (SAR imagery dataset) as the target dataset. Results are evaluated by linear probing and reported in macro F1-Score (%). The highest result in each position is highlighted by **bold**. Our results are marked in gray.

Method	**SD → DOTA**			**OpenCLIP → DOTA**			**SatMAE → FUSRS**		**Mean**		
	Head	Middle	Tail	Head	Middle	Tail	Head	Tail	Head	Middle	Tail
Zero-Shot	99.2	97.3	87.9	93.1	92.7	91.7	78.3	67.8	90.2	95.0	82.5
Fine-Tune	99.1	96.7	86.8	93.1	91.1	89.2	88.6	73.6	93.6	93.9	83.2
cLoRA	99.1	94.3	89.3	97.3	93.3	92.2	89.9	82.0	95.5	93.8	87.9
w/ debLoRA	99.3	97.5	93.5	**97.6**	**95.8**	**95.0**	**92.5**	**86.1**	**96.5**	**96.7**	**91.5**
LoRA	99.4	97.2	91.8	96.6	92.7	91.6	87.1	76.3	94.4	95.0	86.6
w/ ResLT [9]	**99.4**	97.7	93.0	97.7	94.1	93.8	86.6	75.4	94.6	95.9	87.4
w/ SADE [68]	99.1	97.3	92.4	97.3	93.0	92.5	89.6	78.4	95.3	95.2	87.8
w/ debLoRA	99.3	**97.7**	**95.1**	97.2	95.6	94.8	90.1	81.0	95.5	**96.7**	90.3

Table 4: **Evaluation on the oriented object detection task.** We implement debLoRA for long-tailed detection tasks. Our results are marked in gray.

Method	**mAP** (%)↑			**Average**
	Head	Middle	Tail	(%)↑
Zero-Shot	71.0	73.7	55.9	66.9
Fine-Tune	76.3	84.9	64.3	75.2
LoRA	77.5	86.3	66.5	76.7
w/ Reweight [28]	74.3	86.8	66.9	76.0
w/ ECM [24]	78.1	87.4	68.5	78.0
w/ debLoRA	**79.4**	**88.5**	**73.2**	**80.4**

design. However, debLoRA remains robust across both foundation models. 3) The performance gains of debLoRA are most significant for SatMAE → FUSRS (+4.7 points) compared to SD → DOTA and OpenCLIP → DOTA (+3.3 and +3.2 points, respectively). This suggests that our method can leverage domain similarity more effectively when adapting between related image domains (SatMAE and FUSRS are RS datasets). We think this is because debLoRA's clustering step captures and utilizes the shared domain-specific visual patterns (*e.g.*, spatial structures and textures) when the source and target domains are closely related. 4) debLoRA consistently outperforms long-tailed recognition methods, ResLT [9] and SADE [68] (2.5 and 2.9 points by average). ResLT and SADE mainly introduce re-weighting strategies to balance the learning of different classes, but they do not directly rectify the bias in the feature space. In contrast, debLoRA explicitly learns a de-biased representation center for tail classes. 5) We further validate the generalizability of our method by conducting experiments on additional long-tailed datasets Places365-LT [34], iNaturalist [55], and fMoW-S2 [6, 8]. Our debLoRA consistently outperforms baselines, achieving up to 7.2% improvement on tail classes (see Appendix). 2) *Oriented Object Detection.* We validate our method's generalization ability on the oriented object detection task in Table 4. We have two key findings. 1) Our debLoRA achieves the highest mAP scores across all positions. Notably, debLoRA outperforms vanilla LoRA by an impressive 6.7 percentage points. 2) Notably, all methods performed better in the middle classes than in the head. This might be attributed to the greater intra-class variation in head classes, whereas middle classes have more distinct and compact features.

6 Conclusion

In this paper, we propose debLoRA, a novel approach for adapting foundation models to data-scarce and long-tailed remote sensing domains while mitigating representation bias. Our method introduces unsupervised clustering to capture robust visual attributes shared across classes, and feature calibration to rectify the bias in tail class representations. We validate the effectiveness of debLoRA through extensive experiments on multiple RS adaptation settings and downstream tasks, where it consistently outperforms vanilla LoRA and other long-tailed recognition methods. Notably, debLoRA achieves significant performance gains on tail classes without sacrificing the performance on head classes, highlighting its ability to learn debiased feature representations.

Acknowledgments and Disclosure of Funding

The author gratefully acknowledges the support from the DSO research grant awarded by DSO National Laboratories, Singapore, and the Lee Kong Chian Fellow grant awarded to Dr. Qianru Sun by Singapore Management University.

References

[1] Mesay Belete Bejiga, Farid Melgani, and Pietro Beraldini. Domain adversarial neural networks for large-scale land cover classification. *Remote Sensing*, 11(10):1153, 2019.

[2] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[5] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[6] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

[7] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 694–710. Springer, 2020.

[8] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35: 197–211, 2022.

[9] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45 (3):3695–3706, 2022.

[10] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7778–7796, 2021.

[11] Ahmed Elshamli, Graham W Taylor, Aaron Berg, and Shawki Areibi. Domain adaptation using representation learning for the classification of remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(9):4198–4209, 2017.

[12] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013.

[13] Africa Ixmuca Flores-Anderson, Kelsey E Herndon, Rajesh Bahadur Thapa, and Emil Cherrington. The sar handbook: comprehensive methodologies for forest monitoring and biomass estimation. Technical report, NASA SERVIR Global Program, 2019.

[14] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE transactions on knowledge and data engineering*, 35(4):3313–3332, 2021.

[15] Dongen Guo, Ying Xia, and Xiaobo Luo. Self-supervised gans with similarity loss for remote sensing image scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2508–2521, 2021.

[16] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. *arXiv preprint arXiv:2312.10115*, 2023.

[17] Yue Guo, Hengchao Li, Wen-Shuai Hu, and Wei-Ye Wang. Sar image data augmentation via residual and attention-based generative adversarial network for ship detection. *IEEE International Geoscience and Remote Sensing Symposium*, pages 439–442, 2022. URL `https://api.semanticscholar.org/CorpusID:252589527`.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[20] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Xiuping Jia, Antonio Plaza, et al. Spectralgpt: Spectral foundation model. *arXiv preprint arXiv:2311.07113*, 2023.

[21] Xiyue Hou, Wei Ao, Qian Song, Jian Lai, Haipeng Wang, and Feng Xu. Fusar-ship: Building a high-resolution sar-ais matchup dataset of gaofen-3 for ship detection and recognition. *Science China Information Sciences*, 63:1–19, 2020.

[22] 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. *arXiv preprint arXiv:2106.09685*, 2021.

[23] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.

[24] Jang Hyun Cho and Philipp Krähenbühl. Long-tail detection with effective class-margins. In *European Conference on Computer Vision*, pages 698–714. Springer, 2022.

[25] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`. If you use this software, please cite it as below.

[26] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7610–7619, 2020.

[27] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019.

[28] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8420–8429, 2019.

[29] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020.

[30] Jian Kang, Ruben Fernandez-Beltran, Puhong Duan, Sicong Liu, and Antonio J Plaza. Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3):2598–2610, 2020.

[31] Songlin Lei, Dongdong Lu, Xiaolan Qiu, and Chibiao Ding. Srsdd-v1. 0: A high-resolution sar rotation ship detection dataset. *Remote Sensing*, 13(24):5104, 2021.

[32] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020.

[33] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2970–2979, 2020.

[34] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019.

[35] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2200–2207, 2013.

[36] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177, 2019.

[37] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021.

[38] Mauro Martini, Vittorio Mazzia, Aleem Khaliq, and Marcello Chiaberge. Domain-adversarial training of self-attention-based networks for land cover classification using multi-temporal sentinel-2 satellite imagery. *Remote Sensing*, 13(13):2564, 2021.

[39] Giona Matasci, Devis Tuia, and Mikhail Kanevski. Svm-based boosting of active learning strategies for efficient domain adaptation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(5):1335–1343, 2012.

[40] Giona Matasci, Michele Volpi, Mikhail Kanevski, Lorenzo Bruzzone, and Devis Tuia. Semisupervised transfer component analysis for domain adaptation in remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7):3550–3564, 2015.

[41] Fabio Pacifici, Jocelyn Chanussot, and Qian Du. 2011 grss data fusion contest: Exploiting worldview-2 multi-angular acquisitions. In *2011 IEEE International Geoscience and Remote Sensing Symposium*, pages 1163–1166. IEEE, 2011.

[42] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2):199–210, 2010.

[43] Fernando Paolo, Tsu-ting Tim Lin, Ritwik Gupta, Bryce Goodman, Nirav Patel, Daniel Kuster, David Kroodsma, and Jared Dunnmon. xview3-sar: Detecting dark fishing activity using synthetic aperture radar imagery. *Advances in Neural Information Processing Systems*, 35: 37604–37616, 2022.

[44] Claudio Persello and Lorenzo Bruzzone. Active learning for domain adaptation in the supervised classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4468–4483, 2012.

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

[46] Suju Rajan, Joydeep Ghosh, and Melba M Crawford. An active learning approach to hyperspectral data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(4): 1231–1242, 2008.

[47] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10674–10685, 2021. URL `https://api.semanticscholar.org/CorpusID:245335280`.

[48] Gary A Shaw and Hsiaohua K Burke. Spectral imaging for remote sensing. *Lincoln laboratory journal*, 14(1):3–28, 2003.

[49] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 467–482. Springer, 2016.

[50] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=ypOiXjdfnU`.

[51] Onur Tasar, Yuliya Tarabalka, Alain Giros, Pierre Alliez, and Sébastien Clerc. Standardgan: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 192–193, 2020.

[52] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1922–1933, 2020.

[53] Devis Tuia, Michele Volpi, Maxime Trolliet, and Gustau Camps-Valls. Semisupervised manifold alignment of multimodal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 52(12):7708–7720, 2014.

[54] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

[55] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[56] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):213–247, 2022.

[57] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. *Advances in neural information processing systems*, 30, 2017.

[58] Bram De Wilde, Anindo Saha, Richard P. G. ten Broek, and Henkjan J. Huisman. Medical diffusion on a budget: textual inversion for medical image generation. *ArXiv*, abs/2303.13430, 2023. URL `https://api.semanticscholar.org/CorpusID:257687585`.

[59] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018.

[60] Hsiuhan Lexie Yang and Melba M Crawford. Domain adaptation with preservation of manifold geometry for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(2):543–555, 2015.

[61] Hsiuhan Lexie Yang and Melba M Crawford. Spectral and spatial proximity-based manifold alignment for multitemporal hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 54(1):51–64, 2015.

[62] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301, 2020.

[63] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5704–5713, 2019.

[64] Zhongqi Yue, Pan Zhou, Richang Hong, Hanwang Zhang, and Qianru Sun. Few-shot learner parameterization by diffusion time-steps. *arXiv preprint arXiv:2403.02649*, 2024.

[65] Jianrong Zhang, Hongwei Zhao, and Jiao Li. Trs: Transformers for remote sensing scene classification. *Remote Sensing*, 13(20):4143, 2021.

[66] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[67] Shuai Zhang, Zaidao Wen, Zhunga Liu, and Quan Pan. Rotation awareness based self-supervised learning for sar target recognition. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 1378–1381. IEEE, 2019.

[68] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. *Advances in Neural Information Processing Systems*, 35:34077–34090, 2022.

[69] 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*, 2023.

[70] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017.

[71] Xiao Xiang Zhu, Sina Montazeri, Mohsin Ali, Yuansheng Hua, Yuanyuan Wang, Lichao Mou, Yilei Shi, Feng Xu, and Richard Bamler. Deep learning meets sar: Concepts, models, pitfalls, and perspectives. *IEEE Geoscience and Remote Sensing Magazine*, 9(4):143–172, 2021.

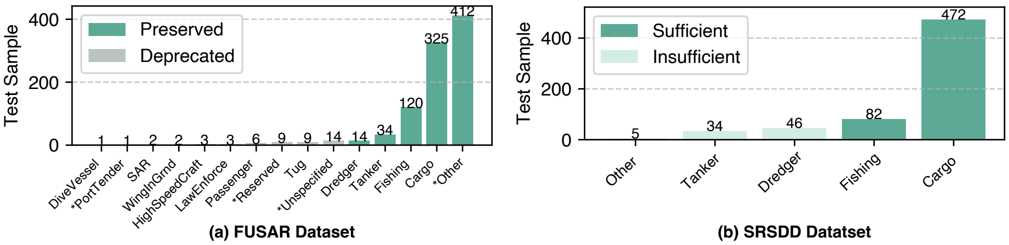
A Appendix

This appendix contains the following supplementary information:

1. Section A.1 details on the customized SAR ship dataset used in the ORS → SAR setting, complementing the experiments in Section 5.
2. Section A.2 presents experiments on additional datasets, including natural image datasets and a multi-spectral remote sensing dataset, to demonstrate the generalizability of our method.
3. Section A.3 provides ablation studies and additional analyses, including quantitative feature analysis, sensitivity to cluster number K, and statistical analysis with error bars.
4. Section A.4 discusses the limitations of our work.

A.1 Details of the customized SAR ship dataset

We selected the FUSAR-Ship [21] and SRSDD [31] datasets as our source datasets due to their high resolution ($\leq$ 10m) and fine-grained ship subcategories, as shown in Figure A1. However, both datasets have limitations. Figure A1(a) shows that the FUSAR-Ship dataset has insufficient test samples (i.e., certain categories have only ≤ 15 test samples) and unclear category definitions (e.g., "Reserved" or "Unspecified" categories). Figure A1(b) reveals that the SRSDD dataset also suffers from insufficient test samples. To address these issues and establish a robust benchmark, we combined the ship categories from both datasets, merging those with fewer than 10 test samples into an "others" category.



(a) FUSAR Dataset

(b) SRSDD Datatset

Figure A1: **Constraints of the SAR datasets' test sets.** This figure illustrates the per-category test sample distribution of (a) the FUSAR dataset and (b) the SRSD dataset. The FUSAR dataset suffers from insufficient test samples and vaguely defined classes (indicated by "$*$"). Similarly, the SRSDD dataset also has the issue of insufficient test samples.

A.2 Experiments on Additional Datasets

To demonstrate the generalizability of our debLoRA method, we conducted experiments on three additional datasets: two from the natural image domain (Places365-LT [34] and iNaturalist 2018 [55]) and one multi-spectral remote sensing dataset (fMoW-S2 [6, 8]). These datasets were chosen for their unique properties: 1) Places365-LT exhibits a substantial domain gap from Stable Diffusion's pre-training data, allowing us to evaluate the performance of our domain adaptation model. 2) iNaturalist 2018 has a high imbalance ratio of 500, enabling us to assess our model's performance under severe class imbalance conditions. 3) fMoW-S2 contains multi-spectral data, including visible, near-infrared, and shortwave infrared bands, complementing our existing experiments on optical (DOTA) and SAR (FUSRS) imagery. The results are given in Table A1 and Table A2.

1) On Places365-LT and iNaturalist 2018 (Table A1), debLoRA consistently outperforms LoRA, especially for tail classes. We observe improvements of 4.3% and 7.2% for Places365-LT and iNaturalist 2018 tail classes, respectively.

Table A1: **Comparison on Places365-LT and iNaturalist2018 datasets**. Results reported in top-1 accuracy (%). Our results are marked in gray .

Method	Places365-LT			iNaturalist 2018			Mean		
	Head	Middle	Tail	Head	Middle	Tail	Head	Middle	Tail
Zero-Shot	40.3	36.9	24.9	36.2	29.4	8.9	38.3	33.2	16.9
Fine-Tune	43.2	31.1	39.0	66.5	69.2	67.5	54.9	50.2	53.3
LoRA	48.2	42.0	44.9	71.9	74.6	71.2	60.1	58.3	58.1
w/ debLoRA	**50.9**	**51.2**	**49.2**	**72.6**	**79.8**	**78.4**	**61.8**	**65.5**	**63.8**

2) For the fMoW-S2 dataset (Table A2), we adapted Stable Diffusion (SD) to the scene recognition task. The dataset was manually divided into "Head" (34 classes comprising 80% of the

samples) and "Tail" (28 classes comprising 20% of the samples). Results were evaluated by linear probing. debLoRA achieves the highest overall accuracy (46.8%) and tail class accuracy (41.2%), surpassing the second-best method (ResLT) by 0.3 and 2.6 percentage points, respectively.

These results confirm that our method effectively addresses the long-tailed distribution problem across various domains, including natural images and multi-spectral remote sensing data. The consistent improvements, particularly for tail classes, highlight the robustness of debLoRA in handling class imbalance and domain adaptation challenges.

Table A2: **Results on the fMoW-S2 dataset.**

Method	**SD → fMoW-S2**		
	Head	Tail	Overall
Fine-Tune	46.2	34.6	44.9
LoRA	46.5	38.1	46.2
w/ ResLT	46.8	38.6	46.5
w/ debLoRA	**46.8**	**41.2**	**46.8**

A.3 Ablation Studies and Additional Analyses

To provide a comprehensive evaluation of our debLoRA method, we conducted several ablation studies and additional analyses. These experiments aim to validate the effectiveness of our approach, investigate its sensitivity to key hyperparameters (*i.e.*, cluster number K), and demonstrate the statistical significance.

Quantitative Feature Analysis. To further validate the effectiveness of our debLoRA method, we present a quantitative analysis of the learned features, focusing on inter-class and intra-class distances. Table A3 shows the results on the DOTA dataset.

Our analysis reveals several key observations about debLoRA. First, it enlarged the inter-class distance between tail and head classes, with the average cosine distance increasing from 0.702 to 0.719. This indicates improved separation between these class groups. Second, debLoRA reduced the intra-class distance for tail classes, as evidenced by the decrease in average cosine distance from 0.182 to 0.146. This suggests a tighter clustering of tail samples. Finally, we observed an increase in inter-class distance among tail classes, with the average cosine distance rising from 0.607 to 0.632. This demonstrates better separation among different tail classes. These findings support the effectiveness of debLoRA in improving feature separation for tail classes.

Table A3: **Quantitative feature analysis on the DOTA dataset.** Inter-class distance is measured as the average cosine distance between class centers, while intra-class distance is the average cosine distance between samples and their corresponding class centers.

Method	**Inter-class**		**Intra-class**
	Head-Tail	Tail-Tail	Tail
Fine-tuning	0.674	0.621	0.170
LoRA	0.702	0.607	0.182
w/ debLoRA	0.719	0.632	0.146

Sensitivity to Cluster Number K. We conducted an ablation study to investigate the sensitivity of our method to the number of clusters (K) used in the de-biasing process. Table A4 shows the results on the SD → DOTA adaptation.

From the table we can observe that performance generally improves as K increases, with the most significant gains observed for tail classes. For instance, when K increases from 16 to 32, the F1 score for tail classes improves by 4.7%. The performance peak around K=32 suggests a good default value for our method. These findings indicate that our method is sensitive to K but remains effective across different values.

Table A4: **Ablation study on the number of clusters (K) in debLoRA.** Our default value is marked in gray.

K	**Macro F1 Score (%)**		
	Head	Middle	Tail
16	99.1	96.9	90.4
32	99.3	97.7	95.1
64	99.3	97.4	94.8

Statistical Analysis with Error Bars. To demonstrate the statistical significance of our results, we report the results of three runs with random initializations on the SD → DOTA experiment. Table A5 shows the results.

Table A5: **Error Bar Analysis.** Reported in mean ± std. Our results are marked in gray.

Method	**Macro F1 Score (%)**		
	Head	Middle	Tail
Zero-Shot	99.2 ± 0.1	97.4 ± 0.3	87.6 ± 0.6
Fine-Tune	99.1 ± 0.1	96.7 ± 0.1	86.8 ± 0.2
LoRA	99.3 ± 0.1	97.2 ± 0.1	91.8 ± 0.2
w/ ResLT	99.3 ± 0.1	97.5 ± 0.3	92.9 ± 0.3
w/ debLoRA	**99.3 ± 0.1**	**97.5 ± 0.2**	**94.8 ± 0.3**

These results demonstrate that our debLoRA method consistently outperforms other approaches, especially for tail classes, with statistically stable improvements. The small std across all methods indicate the stability of the results. Notably, debLoRA shows the most substantial improvement for tail classes, with a mean

F1 score of 94.8% and a standard deviation of only 0.3%, highlighting both the effectiveness and consistency of our approach in addressing the long-tailed distribution problem.

A.4 Limitations

While our proposed debLoRA method has proven effective in adapting foundation models to remote sensing domains with limited data and long-tailed distributions, we acknowledge three key limitations: **Assumption of shared visual attributes.** Our method assumes that visual attributes are shared across classes, enabling robust representation learning through clustering. However, if the visual attributes are highly class-specific or there is significant intra-class variation, the effectiveness of our approach may be reduced.

Sensitivity to hyperparameters. The performance of debLoRA depends on the selection of hyperparameters, such as the number of clusters K. The optimal value of K may differ depending on the specific dataset and adaptation setting.

Limited evaluation on SAR datasets. Due to the scarcity of large-scale SAR datasets with sufficient samples for reliable evaluation, we created a customized dataset by combining two existing SAR datasets. Further investigation is needed to assess the performance of our method on a broader range of SAR datasets and tasks.

By acknowledging these limitations, we aim to provide a transparent and objective assessment of our work and to encourage future research addressing these challenges to further improve long-tailed adaptation in remote sensing domains.