Lost in the Tail: Addressing Geographic Imbalance in Urban Visual Place Recognition

Zhiyao Shu[1], Jiacheng Yang[2], Yang Lu[2], Waishan Qiu[3], Chuan Li[4], and Da Chen[5]✉

[1] George Mason University, Fairfax, USA [2] Xiamen University, Xiamen, China
[3] University of Hong Kong, Hong Kong, China
[4] Lambda, San Francisco, USA [5] University of Bath, Bath, UK
zshu2@gmu.edu, {jiachengyang,luyang}@xmu.edu.cn, waishanq@hku.hk, c@lambdal.com, da.chen@bath.edu

Abstract. Urban-scale Visual Place Recognition (VPR) aims to identify the geographic location of a query image by matching it against a geo-tagged database. While recent methods achieve impressive performance, they overlook a serious long-tailed problem hidden in urban-scale datasets, which biases the model towards locations with abundant images and ignores less-visited areas, causing models to systematically favor frequently photographed locations while failing in sparsely covered areas. In this paper, we systematically characterize this imbalance challenge and propose **Distribution-Aware Place Recognition (DAPR)**, a model-agnostic plug-in framework that rebalances gradient contributions across head and tail classes. Additionally, within classification-retrieval pipelines, DAPR applies a multi-scale distance search mechanism to compute per-class distributional compactness, providing complementary gains at the retrieval stage. On the large-scale SF-XL benchmark, our framework outperforms the previous classification-retrieval baseline by **18.3%** on test set v1, and **6.7%** on test set v2. As a plug-in module, it achieves consistent improvements across representative VPR methods on SF-XL, MSLS, and Pitts30k, demonstrating broad generalizability across different methods and benchmarks.

Keywords: Visual Place Recognition · Long-tailed · Image Retrieval

1 Introduction

Visual Place Recognition (VPR), also known as geo-localization, is a fundamental task in Computer Vision that aims to determine the geographic location of a query image by matching it against a geo-tagged reference database. Recent proposed VPR methods learn global descriptors for direct similarity search, ranging from lightweight CNN-based approaches [2, 5, 7] to transformer-based aggregation methods [3, 23, 34, 65]. Alternatively, some methods frame VPR as a classification problem, partitioning the map into discrete spatial cells to learn

✉ Corresponding author: Da Chen. da.chen@bath.edu

Fig. 1: Geographic classes in SF-XL [9] are ranked by sample count (blue curve) and partitioned into head (top 30%), middle (40%), and tail (bottom 30%). Intra-class feature diversity (green band) indicates richer and more varied visual content in sparsely sampled regions (widens toward tail classes). Per-class Recall@1 (orange curve), measured on the D&C [48] model, declines sharply toward the tail.

discriminative place descriptors [9, 12, 42], with D&C [48] further introducing a classification-retrieval pipeline (hereafter denoted as *mixed-pipeline*) to improve inference efficiency. These VPR approaches achieve strong accuracy on existing small-scale and large-scale benchmarks.

However, existing methods overlooked the severe long-tailed distribution challenge in the existing geographic datasets. Current VPR datasets are sourced from Google Street View [1, 5, 9] or crowd-sourced platforms [56], where image density is governed by vehicle traffic volume and photographer frequency rather than geographic feature richness. Consequently, environments that are visually distinctive and rich in diverse geographic features, such as residential streets, alleys and low-traffic urban corridors, accumulate only sparse training coverage. Figure 1 illustrates the extent of this imbalance on SF-XL [9]. We rank all geographic classes by image count and partition them into head (top 30% by frequency), middle (40%), and tail (bottom 30%), exposing a ∼300:1 imbalance ratio where head classes contain over 3,000 images while tail classes have as few as 12. Notably, intra-class feature diversity (green band) widens toward the tail, reflecting greater variance in pairwise feature similarity within sparsely sampled regions. This indicates that tail-class areas contain richer and more varied visual content. Per-class Recall@1 (orange curve, measured on D&C [48]) declines sharply toward the tail, revealing a compounding failure: the classes that are

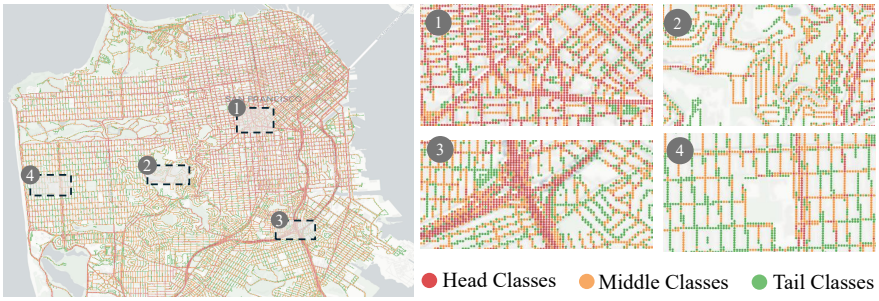

Fig. 2: The geographic classes distribution of SF_XL [9] on real-world map. Red regions (head classes) represent frequently photographed locations from SF_XL [9]. Orange regions (middle classes) show moderately sampled areas, and green regions (tail classes) indicate sparsely photographed locations. Four representative regions (1-4) are zoom-in spatial patterns of class imbalance: region 1 and 4 reveal predominantly head classes in major roads, regions 2-3 demonstrate mixed distributions.

inherently hardest to recognize due to their visual diversity receive the least training supervision.

This limitation fundamentally undermines the reliability of deployed real-world VPR systems. For example, urban safety applications such as autonomous patrol robots [61] are specifically deployed in low-traffic, GPS-degraded zones such as the residential and peripheral urban areas that constitute tail classes in existing datasets. A VPR model that has never adequately learned the distinctive features of these regions will suffer the worst localization failures at the locations of highest operational demand, leading to safety-critical risks where reliable spatial awareness is most needed.

Furthermore, direct similarity search over millions of images on a large-scale benchmark, such as SF-XL [9] containing 41.2 million images, introduces non-trivial computational overhead at inference. While mixed-pipeline methods [48] address this computational challenge by filtering candidate classes before similarity search and enabling practical deployment on large-scale datasets, their reliance on L2 distance during inference treats all feature distributions uniformly, implicitly assuming equal feature compactness across geographic classes. This causes head classes to dominate matching due to their dense and compact representations, while tail-class features are systematically underweighted despite their discriminative value.

To address these challenges, we propose a Distribution-Aware Place Recognition (DAPR) framework, which balances the feature distribution with a distribution-aware loss strategy (named as Low-visit Bias (LB) loss $\mathcal{L}_{lb}$) and a multi-scale distance search mechanism. In the learning phase, DAPR rebalances gradient contributions across classes through inverse-frequency weighting and corrects classifier bias via prior-based logit adjustment. Within mixed-pipelines, the multi-scale distance search mechanism introduces Characteristic Function [15] to adaptively weight amplitude (distributional compactness) and phase (directional similarity) components of feature distances, enabling fair comparison between head and tail

classes without requiring explicit labels at inference time. We summarize the contributions as follows:

- We identify and formalize the long-tailed geographic distribution problem in urban-scale VPR, revealing that class imbalance with spatial bias fundamentally limits existing VPR methods.
- We propose **DAPR**, a model-agnostic plug-in framework that addresses long-tailed geographic imbalance across different VPR methods. Within mixed pipelines, we further introduce a Multi-scale Distance Search Mechanism to provide complementary distribution-aware gains at the retrieval stage.
- Comprehensive experiments on SF-XL [9] demonstrate that DAPR achieves state-of-the-art performance, surpassing the previous best mixed-pipeline baseline by **+18.3%** R@1 on test v1 and **+6.7%** on test v2, while being over 60× faster than full-database retrieval methods. As a plug-in module, $\mathcal{L}_{lb}$ yields consistent improvements when integrated into representative retrieval methods (SALAD, BoQ) across MSLS and Pitts30k, confirming that long-tailed geographic imbalance is a universal bottleneck in urban-scale VPR.

2 Related Work

2.1 Visual Place Recognition

VPR research addresses localization at different spatial scales [21–23, 27, 35, 39, 48, 52, 64, 69]. At the global scale, systems recognize landmarks and determine geographic coordinates across vast areas by leveraging highly distinctive features within regional-level classification [18]. Beyond spatial scale, cross-view matching introduces additional complexity, where query and reference images are captured from fundamentally different viewpoints, requiring viewpoint-invariant representations [21,69]. Urban-scale VPR focuses on fine-grained localization within city environments [11,17], where the primary challenge lies in retrieving visually similar images based on robust feature matching. Recent feature extraction methods have shifted from CNN-based architectures [5, 41] to transformer-based models leveraging self-attention [52], hybrid-attention [20], and multi-attention mechanisms [39, 64]. In particular, DINOv2 has demonstrated strong capabilities for extracting semantically rich features that generalize well across diverse urban environments [22, 23, 27, 35].

Beyond feature extraction, aggregation strategies further improve descriptor quality [6]. VLAD-based methods [25, 28] pool local features into compact global representations, while attentional pyramid pooling approaches [40] enhance descriptor quality through multi-scale spatial encoding. Two-stage retrieval pipelines have become prevalent [35,68], first retrieving coarse candidates via global descriptors, then applying geometric verification or re-ranking with local features. Single-stage methods rely solely on global features, with recent DINOv2-based approaches such as SALAD [23], BoQ [3], CricaVPR [34], and EffoVPR [49] establishing strong baselines across benchmarks [16,44]. CliqueMining [22] mines visually similar place cliques to sharpen geographic distance

sensitivity, implicit aggregation [33] aggregates features within the transformer backbone itself. And MegaLoc [8] unifies training across datasets for strong cross-domain generalization. Methods that treat VPR as a geographic classification problem partition the map into discrete geo-cells and train a classifier over these place categories using large-margin cosine losses [51] for discriminative descriptor learning [9, 12, 42]. At inference, D&C [48] extend this trained classifier to filter the full database down to a small candidate pool, after which standard L2-based retrieval re-ranks the candidates.

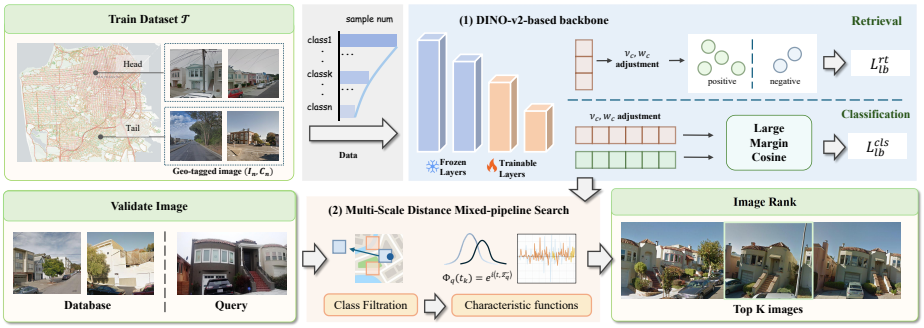

Fig. 3: Distribution-Aware Place Recognition Framework. The geo-tagged images are processed through (1) A DINOv2 backbone extracts features that are optimized under two paradigms: retrieval methods with $\mathcal{L}_{lb}^{rt}$, while classification methods with Large Margin Cosine producing $\mathcal{L}_{lb}^{cls}$. Both apply class-dependent weights w_c and logit adjustment ν_c to rebalance gradient contributions across head and tail classes. (2) Mixed-pipeline filters the database to candidate classes, which are then re-ranked using Characteristic Function Distance (CFD) in the frequency domain to return the top-K matched images.

2.2 Long-Tailed Distribution for Imbalanced Data

Real-world data are inherently imbalanced—a small number of head classes account for the majority of samples, whereas the more numerous tail classes contain only a few instances [4, 13, 45, 50, 62]. This long-tailed distribution biases empirical-risk-minimization models toward the head and leads to degraded performance on the tail [14, 54, 55, 59, 60]. Existing long-tailed learning methods can be broadly grouped into three families: data resampling [50, 62], training framework modifications [14, 45], and loss-function designs [59]. Among these, loss-based methods are widely adopted for downstream tasks because they require no extra training stages and are plug-and-play. For instance, [36] proposed the logit adjustment (LA) loss, which explicitly calibrates class confidences according to class frequency to mitigate head bias. [59] introduced the Distribution-Balanced (DB) loss, combining rebalanced weighting with negative-tolerant regularization to alleviate the suppression of tail classes by negatives. [67] incorporated the notion of "hard classes," assigning larger weights to harder-to-learn categories.

Moreover, [26] decoupled representation learning from classification and employed a cosine classifier to reduce the impact of imbalance. Building on these insights, in this paper, we design a tail-favoring loss, i.e., Low-visit Bias (LB) loss $\mathcal{L}_{lb}$, based on class frequency for VPR, which not only achieves better performance on tail classes, but also improves the performance of the head and middle classes.

3 Method

As illustrated in Fig. 3, the proposed DAPR framework consists of two plug-in modules. First, the Low-visit Bias (LB) loss addresses long-tailed geographic imbalance by correcting gradient contributions across head and tail classes (§3.2). Second, for the mixed-pipeline where the classifier filters the database to a candidate pool during inference, we propose a multi-scale distance search (§3.3) that operates in the frequency domain to provide multi-scale, distribution-aware similarity for retrieval within the filtered candidates.

3.1 Problem Formulation

Let $\mathcal{D} = \{(\mathbf{I}_i, \mathbf{p}_i)\}_{i=1}^{N_\mathcal{D}}$ denote a geo-tagged reference database, where $\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}$ is the i-th image and $\mathbf{p}_i = (x_i, y_i) \in \mathbb{R}^2$ its spatial coordinates. Given a query image $\mathbf{I}_q$ at an unknown location $\mathbf{p}_q$, VPR aims to retrieve the reference image(s) from $\mathcal{D}$ geographically closest to $\mathbf{p}_q$. A large-scale training set $\mathcal{T} = \{(\mathbf{I}_n, \mathbf{p}_n)\}_{n=1}^{N_\mathcal{T}}$ is used to learn a feature extractor $f\colon \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^d$ [5, 17, 24, 32].

Following the classification-based paradigm [9, 12, 48], which avoids the convergence difficulty of direct ranking optimization on large-scale datasets [5, 41], we partition the geographic space into C square grid cells of side length M meters:

$$\mathcal{C}_{i,j} = \{(x, y) \in \mathbb{R}^2 : \lfloor x/M \rfloor = i,\ \lfloor y/M \rfloor = j\}. \tag{1}$$

Each image $\mathbf{I}_n$ is assigned a class label C_n from its grid cell, yielding class frequencies $\hat{p}_c = n_c/N_\mathcal{T}$, where n_c denotes the number of samples in class c. At inference, the classifier filters the database to a candidate pool $\mathcal{D}_q \subset \mathcal{D}$ of predicted classes, within which similarity-based retrieval is performed. However, real-world urban datasets exhibit severely long-tailed class distributions, causing both training bias and retrieval degradation for under-represented tail classes.

3.2 Geographic Long-Tail Aware Training

To address the systematic under-representation of low-frequency geographic locations, we propose the **Low-visit Bias (LB) loss ($\mathcal{L}_{lb}$)**, a plug-and-play training module that corrects the systematic bias caused by long-tailed geographic distributions. As discussed, the long-tailed issue hidden in VPR datasets biases the model toward head classes, leading to degraded performance on tail classes. Specifically, this bias is mainly caused by the disproportionately larger gradients contributed by frequent areas during optimization [4, 30]. For this reason, we integrate three coupled mechanisms: class-distribution reweighting, prior-based

logit adjustment, and large-margin cosine similarity, instantiated as $\mathcal{L}_{lb}^{cls}$ for classification (Eq. 6) and $\mathcal{L}_{lb}^{rt}$ for retrieval (Eq. 7). First, to rebalance gradient contributions across head and tail areas, each class is assigned a weight w_c inversely proportional to its frequency $\hat{p}_c = n_c/N_{\mathcal{T}}$, where n_c is the number of samples in class c and N is the total number of samples:

$$w_c = \frac{C}{\sum_{c'=1}^{C}(\hat{p}_{c'} + \epsilon)^{-\beta}} \cdot (\hat{p}_c + \epsilon)^{-\beta}, \tag{2}$$

where $\beta \in [0, 1]$ controls the strength of reweighting, and ϵ helps avoid numerical instability. This weighting mechanism effectively amplifies the contribution of tail classes and prevents gradient explosion, while the normalization term ensures stable gradient magnitudes throughout training.

Second, to further mitigate bias caused by class prior encoding in decision boundaries, a prior-based logit adjustment is applied as:

$$\nu_c = -\kappa \cdot \log(\frac{\hat{p}_c}{1 - \hat{p}_c}), \tag{3}$$

where $\kappa \in [0, 1]$ controls the correction strength. Since $\nu_c > 0$ for tail classes, their logits are upweighted, promoting class-agnostic representation learning and reducing head-class bias.

Finally, after obtaining weighting and logit adjustment parameters, the Low-visit Bias loss function $\mathcal{L}_{lb}$ is designed to improve both retrieval and classification tasks. For the classification task, Large Margin Cosine Similarity d [51] is first computed to normalize both feature embeddings and classifier weights to the unit hypersphere as follows.

$$d = \frac{\mathbf{z}^\top W_j}{\|\mathbf{z}\|\|W_j\|}, \tag{4}$$

$$z_j = s \cdot (d - m \cdot \mathbb{I}[j = c]), \tag{5}$$

where $\mathbf{z}$ is the feature embedding for training sample $(\mathbf{I}_n, C_n)$, and W_j is the j-th classifier weight vector. s is a scale factor, m is the angular margin, and $\mathbb{I}[j = c]$ is the indicator function. This enforces minimum angular separation between classes in the embedding space. By doing this, the model can capture clear inter-class boundaries and learn compact, discriminative features across all classes. Then, w_c and ν_c are employed to enable robust and balanced representation learning across the long-tailed geographic distribution inherent in real-world VPR datasets by:

$$\mathcal{L}_{lb}^{cls} = -w_c \cdot \log\left(\frac{\exp(z_c - \nu_c)}{\sum_{j=1}^{C}\exp(z_j - \nu_j)}\right). \tag{6}$$

This function simultaneously addresses gradient imbalance through sample weighting w_c, removes classifier bias through logit adjustment ν_c, and enforces feature discriminability through the angular margin embedded in z_j.

For the retrieval-based task, feature similarity d_{ij}^f is first computed for sample i and j by $s_{ij} = \mathbf{f}_i^\top \mathbf{f}_j$, where $\mathbf{f}$ is the extracted feature, then combining widely used Multi-Similarity, the LB loss can be written as:

$$\begin{aligned}\mathcal{L}_{lb}^{rt} =& \frac{1}{N}\sum_{i=1}^{N} w_{y_i}\left[\frac{1}{\gamma_p}\log\left(1+\sum_{\substack{j\neq i\\ y_j=y_i}}\exp\left(-\gamma_p(d_{ij}^f-\tau)\right)\right)\right.\\ &\left.+\frac{1}{\gamma_n}\log\left(1+\sum_{\substack{j\\ y_j\neq y_i}}\exp\left(\gamma_n\left(d_{ij}^f-\nu_{y_j}-\tau\right)\right)\right)\right],\end{aligned} \tag{7}$$

where N denotes the mini-batch size, γ_p and γ_n are the positive and negative scaling factors, respectively, and τ is a similarity threshold. The index set $\{j \mid j \neq i,\, y_j = y_i\}$ corresponds to positive samples of anchor i, while $\{j \mid y_j \neq y_i\}$ denotes the set of negative samples. By doing this, the retrieval model can learn a more robust and balanced feature space for VPR datasets.

3.3 Multi-scale Distance Search Mechanism

While $\mathcal{L}_{lb}$ mitigates long-tail bias during training, distributional imbalance persists at retrieval time, particularly on large-scale datasets such as SF-XL. Within the classifier-filtered candidate pool $\mathcal{D}_q$, tail-class features remain sparse and follow complex, non-Gaussian distributions [66], whereas head-class features form tight, numerically dominant clusters. A static metric such as L_2 or cosine similarity treats both regimes identically, causing queries from tail classes to systematically match nearby head-class clusters.

To address this bias, we adopt the **Characteristic Function Distance (CFD)** as the retrieval metric, mapping features into the frequency domain [15, 53] for multi-scale similarity computation.

Let $\mathcal{S}_q$ denote the query feature set and $\mathcal{S}_j$ the descriptor set of candidate class j within the filtered pool $\mathcal{D}_q$, where $\mathcal{S}_q$ reduces to the single query descriptor. By definition, their empirical characteristic functions [15] are the sample averages over these sets, evaluated at K frequency points $\mathbf{T} = \{\mathbf{t}_k\}_{k=1}^K$:

$$\Phi_q(\mathbf{t}_k) = \frac{1}{|\mathcal{S}_q|}\sum_{\mathbf{z}\in\mathcal{S}_q} e^{i\langle \mathbf{t}_k, \mathbf{z}\rangle}, \quad \Phi_j(\mathbf{t}_k) = \frac{1}{|\mathcal{S}_j|}\sum_{\mathbf{z}\in\mathcal{S}_j} e^{i\langle \mathbf{t}_k, \mathbf{z}\rangle}, \tag{8}$$

where $i = \sqrt{-1}$, so that the modulus $|\Phi(\mathbf{t}_k)| \in [0, 1]$ is large when the set is compact and small when it is dispersed. To capture both fine-grained local patterns and global structure, we employ stratified multi-scale sampling for $\mathbf{T}$, drawing from Gaussian distributions at four logarithmically spaced scales. Please see the supplementary material for more details.

Each CF is then decomposed via Euler's formula into complementary components:

$$\Phi(\mathbf{t}_k) = |\Phi(\mathbf{t}_k)|e^{i\alpha(\mathbf{t}_k)}, \tag{9}$$

where $|\Phi(\mathbf{t}_k)|$ is the amplitude and $\alpha(\mathbf{t}_k) = \arg(\Phi(\mathbf{t}_k))$ is the phase. Head classes maintain more consistent amplitudes due to their dense feature clusters, while tail classes exhibit lower and more variable amplitudes, indicating dispersed, under-trained features. For this reason, to enable distribution-aware matching, we compute mean amplitudes for query and database features:

$$\bar{A}_q = \frac{1}{K}\sum_{k=1}^{K}|\Phi_q(\mathbf{t}_k)|, \quad \bar{A}_j = \frac{1}{K}\sum_{k=1}^{K}|\Phi_j(\mathbf{t}_k)|. \tag{10}$$

The reliability ratio $r^{(j)} = \frac{\bar{A}_q}{\bar{A}_j}$ captures relative compactness, and the adaptive weights follow $\alpha_w^{(j)} = \min(\alpha \cdot r^{(j)}, 1), \lambda_w^{(j)} = 1 - \alpha_w^{(j)}$, with $\alpha \in [0, 1]$ a base amplitude weight. At each frequency $\mathbf{t}_k$, the amplitude difference is:

$$D_{\text{amp}}^{(k)} = (|\Phi_q(\mathbf{t}_k)| - |\Phi_j(\mathbf{t}_k)|)^2. \tag{11}$$

and the phase difference $D_{\text{phase}}^{(k)}$ using circular distance with wrapping at 2π:

$$D_{\text{phase}}^{(k)} = \min(\delta^2, (2\pi - \delta)^2), \quad \delta = |\alpha_q(\mathbf{t}_k) - \alpha_j(\mathbf{t}_k)|. \tag{12}$$

Finally, the CFD averages the component differences across all frequencies with adaptive weighting by:

$$D_{\text{CFD}}(q, j) = \alpha_w^{(j)} \cdot \frac{1}{K}\sum_{k=1}^{K} D_{\text{amp}}^{(k)} + \lambda_w^{(j)} \cdot \frac{1}{K}\sum_{k=1}^{K} D_{\text{phase}}^{(k)}. \tag{13}$$

This distribution-aware adaptation enables fair similarity assessment without requiring explicit class labels at inference, and can be directly integrated with existing pre-trained VPR models. The amplitude ratio $r^{(j)}$ doubles as a per-query confidence estimate: a query whose descriptor distribution is compact relative to a candidate yields a high $r^{(j)}$ and a more reliable match. CFD thus reads an uncertainty signal from the descriptor distribution on the same retrieval pass, where existing methods obtain a comparable estimate by learning a stochastic embedding [57], measuring the spread of the top-K reference poses [63], or running a post-hoc local-feature matcher [43]. Note that CFD is applied only at the inference stage of the mixed pipeline since computing characteristic functions over all $C > 110{,}000$ class prototypes during training would scale as $\mathcal{O}(B \times C \times K \times D)$ per iteration, which is computationally prohibitive. We detail the uncertain experiment and CFD benefit to tail classes in the supplementary material.

4 Experiments

We conduct experiments to evaluate DAPR from three perspectives. First, we compare DAPR against state-of-the-art VPR methods on the large-scale SF-XL

benchmark (§4.2). Second, we verify the generalization of $\mathcal{L}_{lb}$ by integrating it as a plug-in module into representative classification-based and retrieval-only methods on SF-XL, MSLS, and Pitts30k (§4.3). Third, we conduct ablation studies to isolate the individual contributions of $\mathcal{L}_{lb}$ and the CFD retrieval module, and benchmark them against alternative long-tailed learning strategies (§4.4).

4.1 Experimental Setup

Datasets. We adopt the evaluation protocol from prior large-scale VPR works [9, 48] and conduct experiments on four benchmarks. The large-scale SF-XL benchmark [9] covers San Francisco with over 41M database images and two query sets (v1 and v2) captured under varied conditions. As shown in Figure 1, SF-XL exhibits a pronounced long-tail distribution. To demonstrate generalization beyond a single city, we evaluate the proposed module on MSLS [56], covering diverse global urban environments, and Pitts30k [47], a standard medium-scale benchmark built at Pittsburgh. We further include Nordland [46], to assess robustness under extreme seasonal appearance change. Long-tailed analysis for MSLS and Pitts30k is provided in the supplementary material.

Implementation Details. All experiments are conducted on a single NVIDIA RTX 3090 GPU. For optimal performance, we employ DINO-v2 [38] as the backbone, serving as the feature extractor. We train the model for 200 epochs using the Adam optimizer [29] with a batch size of 256 and a learning rate of 6×10^{-6}. More details are provided in the supplementary material.

Evaluation Metrics. In this paper, we adopt standard VPR evaluation protocols from prior work [5, 10, 17, 19]. For classification-only methods [37, 42], we report Localization Radius@N (LR@N) using a 25-meter localization threshold, following [48]. For measuring the search accuracy, we report Recall@N (R@N), defined as the percentage of queries where at least one of the top-N predictions falls within a 25-meter radius of the ground truth location. In fact, LR@N is equivalent to the R@N@25m used in retrieval according to [48]. For simplicity, we use R@N for all evaluations in this paper.

4.2 Comparison with VPR State-of-the-arts

We compare the proposed DAPR framework against a broad set of VPR methods on the SF-XL benchmark. As indicated in Table 1, DAPR-M (our mixed-pipeline framework) with DINOv2 achieves **89.7%** R@1 on test v1 and **94.3%** R@1 on test v2, surpassing all compared methods that do not incorporate distribution-aware training. Integrating $\mathcal{L}_{lb}$ as a plug-in to retrieval-only methods brings notable gains: SALAD* (SALAD [23] trained with LB Loss) improves from 87.6% to 88.0% on test v1 and from 93.5% to 94.5% on test v2, while BoQ* (BoQ [3] trained with LB Loss) improves from 83.7% to 88.8% and from 92.8% to 93.7%, respectively. These improvements confirm that long-tailed geographic imbalance is a shared bottleneck across VPR methods, not a weakness specific to any single model. DAPR-M further surpasses BoQ*, demonstrating that classifier-based

Method	Backbone	Infer. Time	R@1 v1	R@1 v2
Classification				
PlaNet [58]	EfficientNet	12 ms	24.5	53.1
HGE [37]	EfficientNet	15 ms	27.0	56.4
CPlaNet [42]	EfficientNet	17 ms	27.4	64.1
D&C [48]	EfficientNet	12 ms	61.0	79.1
DAPR - C	EfficientNet	–	67.2	84.1
DAPR - C	ResNet101	–	68.9	85.5
DAPR - C	Dinov2	–	86.4	91.1
Retrieval				
NetVLAD [5]	VGG16	12117 ms	40.0	71.1
SFRS [17]	VGG16	12117 ms	51.2	83.1
GeM [41]	VGG16	12117 ms	21.7	43.1
CosPlace [9]	VGG16	1514 ms	64.7	83.4
CosPlace [9]	ResNet101	1488 ms	70.9	81.9
SALAD [23]	DINOv2	4805 ms	87.6	93.5
SALAD*	DINOv2	4823 ms	88.0	94.5
BOQ [3]	DINOv2	21333 ms	83.7	92.8
BOQ*	DINOv2	21047 ms	88.8	93.7
Mixed pipeline				
D&C + CosPlace [48]	EfficientNet	**30** ms	71.4	87.6
DAPR - M	EfficientNet	51 ms	74.5	88.1
DAPR - M	ResNet101	54 ms	76.3	88.0
DAPR - M	DINOv2	74 ms	**89.7**	**94.3**

Table 1: Results on SF-XL [9] test v1 and v2 across three VPR types: classification-only (top), retrieval-only (middle), and mixed pipeline (bottom). SALAD* and BoQ* denote retraining with proposed LB loss $\mathcal{L}_{lb}$. **DAPR-M**: mixed pipeline with LB Loss and CFD.

candidate filtering offers additional gains by narrowing the search space to geographically relevant candidates, enabling more precise retrieval than searching the full database.

DAPR Across Pipelines and Backbones. Distribution-aware training provides consistent gains across all pipeline configurations. In the classification-only setting, DAPR-C (our classification-only framework) with DINOv2 reaches 86.4% R@1 on test v1, outperforming D&C [48] (61.0%) by +25.4%. As shown in Table 1, these improvements hold across EfficientNet, ResNet101, and DINOv2, indicating that $\mathcal{L}_{lb}$ addresses a data-level bottleneck largely orthogonal to backbone capacity.

Memory and Inference-Time Analysis. In the mixed-pipeline on SF-XL, when DAPR-M is applied, the classifier filters the database to ∼450 candidates per query, requiring only **1.35 MB** to store the 768-dim descriptors (FP32) for the following retrieval step. In contrast, retrieval-only methods must index the full database (∼2.8M images): SALAD [23] (8,448-dim) requires **94.8 GB** and BoQ [3] (8,192-dim) requires **91.9 GB**, over four orders of magnitude larger due to both higher-dimensional descriptors (8K+ vs. 768 dimensions) and full-database indexing (2.8M vs. 450 images). The compact candidate pool also accelerates inference, as D&C [48] achieves 30 ms and DAPR-M with DINOv2

Method	R@1		R@5		R@10	
	v1	v2	v1	v2	v1	v2
CosPlace	74.4	90.0	82.0	94.6	84.5	97.0
+ LB Loss	**76.1** (+1.7)	90.0	**84.2** (+2.2)	**95.8** (+1.2)	**86.3** (+1.8)	97.0
DC	44.1	64.0	57.2	84.9	63.6	90.1
+ LB Loss	**48.2** (+4.1)	**66.2** (+2.2)	**64.2** (+7.0)	**86.1** (+1.2)	**69.6** (+6.0)	**90.3** (+0.2)
DC (mixed)	55.1	81.1	62.2	88.6	64.7	90.0
+ LB Loss	**57.4** (+2.3)	**81.3** (+0.2)	**65.4** (+3.2)	**89.0** (+0.4)	**66.9** (+2.2)	**91.1** (+1.1)

Table 2: Generalization of LB loss across VPR frameworks on SF-XL test v1/v2. Recall@N (%) reported.

Method	Backbone	Dim.	MSLS			Pitts30k			Nordland		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CosPlace [9]	ResNet101	128	81.7	89.7	91.6	86.7	94.5	97.0	41.1	56.3	63.3
CosPlace* [9]	ResNet101	128	**82.2**	**89.8**	**92.4**	**89.4**	**96.6**	**97.9**	**44.6**	**59.4**	**65.7**
SALAD [23]	DINOv2	8448	91.9	96.4	96.5	92.3	96.3	97.3	76.0	89.2	92.0
SALAD* [23]	DINOv2	8448	**92.6**	**96.5**	**97.0**	**92.7**	**96.8**	**97.7**	**76.6**	**89.6**	**92.9**
BoQ [3]	DINOv2	8192	91.2	95.3	96.1	92.6	96.4	97.4	81.3	92.5	94.8
BoQ* [3]	DINOv2	8192	**93.7**	**96.2**	**96.8**	**92.9**	**96.6**	**97.5**	**83.7**	**94.6**	**96.5**

Table 3: Comparison with state-of-the-art retrieval methods on MSLS, Pitts30k, and Nordland. * denotes models retrained with $\mathcal{L}_{lb}$. **Bold** indicates the better value within each method pair.

achieves 74 ms per query, both over 60× faster than SALAD (4,805 ms), while delivering a 25.6% relative R@1 improvement over the best mixed-pipeline baseline on test v1.

Head-Tail Class Analysis. To pinpoint the source of improvement, we analyse performance across head, middle, and tail classes on SF-XL (Figure 4), using the same EfficientNet backbone in the mixed pipeline as D&C for a fair comparison. DAPR delivers consistent gains across all three groups: +0.98% R@1 on head, +1.30% on middle, and +1.47% on tail at R@1. This directly validates that our method targets the sparsely represented locations where existing models fail most. The gain pattern is more pronounced at R5, where tail classes improve by +7.35% compared to +1.30% for middle and +1.72% for head.

4.3 Generalization Across VPR Methods

To verify that the proposed DAPR framework addresses a fundamental problem rather than a method-specific weakness, we integrate the two plug-in modules independently into different VPR methods on both retrieval-only and mixed-pipeline settings.

Generalization on SF-XL. Table 2 presents results on the large-scale SF-XL [9] benchmark. Applying $\mathcal{L}_{lb}$ to CosPlace brings a gain of +1.7% R@1 on test v1. On the D&C baseline, the improvement is larger (+4.1% R@1, +7.0%

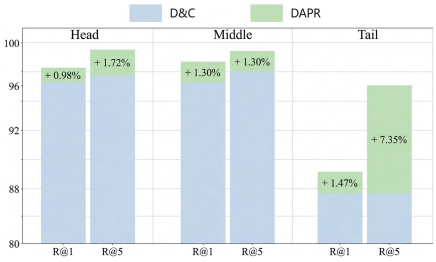

Fig. 4: Performance comparison across head, middle, and tail classes on the SF_XL dataset with Recall@N (%).

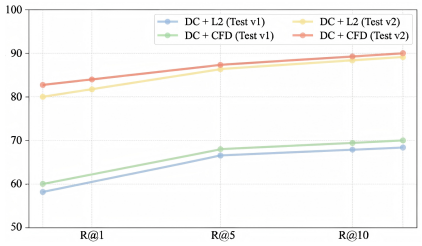

Fig. 5: Comparison of Retrieval Distance on D&C Method with SF-XL test v1 and test v2.

R@5 on test v1), as classification heads rely on per-class decision boundaries and are thus more sensitive to class imbalance than contrastive methods that inherently emphasize hard examples. Combining $\mathcal{L}_{lb}$ with the D&C mixed pipeline further adds +2.3% R@1, and gains are consistent across both test sets, confirming robustness to query-set variation. As shown in Figure 5, replacing L2 distance with CFD in both D&C and our mixed pipeline achieves consistent gains across R1, R5, on both mixed-pipeline methods (DAPR and D&C), indicating that CFD provides reliable distribution-aware matching without any additional training or label supervision. More baseline results including CliqueMining [22], Megaloc [8], CricaVPR [34], and ImAge [33] are indicated in the supplementary materials.

Generalization on Small-scale Benchmarks. We integrate DAPR as a plug-in module into two state-of-the-art contrastive retrieval-only methods, SALAD [23] and BoQ [3]. Both are trained on GSV-Cities [1] and evaluated on MSLS [56] and Pitts30k [47]. As indicated in Table 3, On MSLS, SALAD with $\mathcal{L}_{lb}$ (SALAD*), BoQ with $\mathcal{L}_{lb}$ (BoQ*) achieves +2.5% R@1, and SALAD* gains +0.7%. On Pitts30k, SALAD* and BoQ* improve by +0.4% and +0.3%, respectively. The improvements on both benchmarks, which cover diverse urban environments, confirm that $\mathcal{L}_{lb}$ reliably addresses geographic class imbalance as a lightweight, method-agnostic plug-in with different scales of datasets.

Tail-Class Analysis on Small-scale Benchmarks. Table 4 isolates gains on tail classes, the sparsely sampled zones identified in Figure 2 as most safety-critical. On MSLS, BoQ* improves tail-class R@1 by +2.70% and SALAD* by +0.90%. On Pitts30k, BoQ* and SALAD* gain +3.37% and +0.63%, respectively. The fact that tail-class gains consistently exceed overall gains confirms that $\mathcal{L}_{lb}$ primarily corrects the failure mode it was designed for, rather than providing uniform across-the-board improvements.

4.4 Ablation Study on DAPR

Ablation study on individual modules. We conduct comprehensive ablation studies to validate the individual contributions of the $\mathcal{L}_{lb}$ and the multi-scale distance retrieval module. Table 5 presents a cumulative study. Starting from

Method	MSLS			Pitts30k		
	R@1	R@5	R@10	R@1	R@5	R@10
SALAD	86.49	93.24	94.14	90.56	94.86	96.87
SALAD*	**87.39** (+0.90)	**94.14** (+0.90)	**95.05** (+0.91)	**91.19** (+0.63)	**95.50** (+0.64)	**97.36** (+0.49)
BoQ	89.64	93.69	94.59	89.19	93.69	94.59
BoQ*	**91.44** (+1.80)	**94.59** (+0.90)	**95.50** (+0.91)	**92.56** (+3.37)	**96.33** (+2.64)	**97.65** (+3.06)

Table 4: Tail-class performance on Pitts30k and MSLS. SALAD* and BoQ* denote training with $\mathcal{L}_{lb}$. Parentheses show absolute gain over original methods.

a CrossEntropy baseline with L_2 retrieval (75.8% R@1 on test v1), $\mathcal{L}_{lb}$ alone yields +12.2%, confirming that distribution-aware training is the primary driver of improvement. Replacing standard metrics with CFD yields performance gains in the proposed pipeline with a further +1.7%.

Long-Tailed Learning methods Analysis. We compare $\mathcal{L}_{lb}$ against Cross-Entropy (CE), Logit Adjustment (LA) [36], and Focal Loss (FL) [31] under the full mixed pipeline, as indicated in Table 6. $\mathcal{L}_{lb}$ outperforms CE by +13.9% R@1 on test v1 and +3.2% on test v2, and surpasses LA and FL by +1.7% R@1 and +3.3% R@1 on test v1, respectively. The advantage over LA and FL, which each address only one failure mode (prediction bias or gradient imbalance), supports the design choice of coupling logit adjustment and sample reweighting through the shared class prior $\hat{p}_c$. Sensitivity analyses for the hyperparameters of the two modules from DAPR, the Low-visit Bias Loss ($\mathcal{L}_{lb}$) and the CFD retrieval metric, are provided in the supplementary material.

Adaptive Partitioning vs $\mathcal{L}_{lb}$. Adaptive partitioning balances geographic classes by image count instead of at the loss level. We apply CPlaNet's [42] kd-tree partition to SF-XL, it evens the per-class counts but reshapes the imbalance into leaf area. As shown in Fig. 6, Per-class Recall@1 drops from 14.3 to 12.6. Adaptive partitioning rebalances the data layout, whereas $\mathcal{L}_{lb}$ rebalances the learning signal, so the two address different scopes of imbalance.

DAPR Classifier with SOTA Retrieval Head. We keep the DAPR-C classifier to filter the top classes from the full database, then re-rank the shortlist with SALAD* or BoQ*. Both bring only a marginal R@1 gain over DAPR-M,

VPR Method	**Test v1**			**Test v2**		
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
DINOv2 + CE	75.8	83.4	85.6	91.1	93.8	95.7
DINOv2 + LBLoss + L2	88.0	89.1	90.6	93.8	96.2	97.0
DINOv2 + LBLoss + CFD	**89.7**	**92.4**	**92.5**	**94.3**	**97.0**	**97.5**

Table 5: Ablation study on DAPR components. Baseline uses CrossEntropy loss with L2 retrieval. Results are reported as Recall@N (%).

Loss Type	R@1		R@5		R@10	
	v1	v2	v1	v2	v1	v2
CrossEntropy	75.8	91.1	83.4	93.8	85.6	95.7
LogitAdjust [36]	88.0 (+12.2)	93.3 (+2.2)	91.5 (+8.1)	**97.2** (+3.4)	92.2 (+6.6)	**97.7** (+2.0)
Focal Loss [31]	86.4 (+10.6)	94.0 (+2.9)	90.5 (+7.1)	96.3 (+2.5)	91.4 (+5.8)	97.3 (+1.6)
Ours (LB Loss)	**89.7** (+13.9)	**94.3** (+3.2)	**92.4** (+9.0)	97.0 (+3.2)	**92.5** (+6.9)	97.5 (+1.8)

Table 6: Comparison of LBLoss against CrossEntropy, LogitAdjust [36], and Focal Loss [31] on SF-XL test v1/v2. All methods use the same DINOv2 backbone with our mixed pipeline. Parentheses show absolute gain over the CrossEntropy baseline.

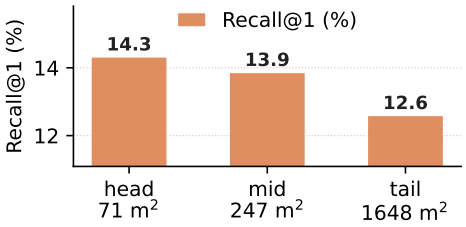

Fig. 6: Per-class Recall@1 across head, middle, and tail bins defined by leaf spatial extent (m^2) on CPlaNet's [42] of SF-XL (131,080 leaves).

while the inference time grows more than 10 times due to their feature size requirements. Moreover, SALAD and BoQ reshape DINOv2 features into cluster aggregators that the DAPR classifier cannot reuse, so the pipeline maintains two independent DINOv2 backbones at both the training and inference stages. DAPR-M reaches comparable accuracy with a single backbone at far lower cost.

5 Conclusion

In this paper, we first systematically reveal and address the fundamental long-tailed challenge arising from natural geographic distributions in the VPR task. We propose DAPR framework including two plug-and-play modules: 1) Low-visit Bias Loss rebalances gradient contributions; 2) CFD for multi-scale distance search in mixed pipeline. Extensive experiments on SF-XL, MSLS and Pitts30k demonstrate that DAPR can be independently integrated into existing VPR frameworks and generalizes across diverse backbone architectures, validating its robustness and broad applicability. Nevertheless, our current formulation defines the long-tail distribution based on *sample count per geographic class*, focusing on dataset-level imbalance. Another direction worth exploring is the *feature-level* long-tail phenomenon, where class difficulty is characterized by intra-class feature coherence rather than sample frequency.

Acknowledgments

We acknowledge Lambda for providing us with compute for this work.

References

1. Ali-bey, A., Chaib-draa, B., Giguère, P.: Gsv-cities: Toward appropriate supervised visual place recognition. Neurocomputing **513**, 194–203 (2022)
2. Ali-Bey, A., Chaib-Draa, B., Giguere, P.: Mixvpr: Feature mixing for visual place recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2998–3007 (2023)
3. Ali-Bey, A., Chaib-draa, B., Giguère, P.: Boq: A place is worth a bag of learnable queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17794–17803 (2024)
4. Anderson, C.: The long tail: why the future of business is selling less of more
5. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
6. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: Proceedings of the IEEE international conference on computer vision. pp. 1269–1277 (2015)
7. Barbarani, G., Mostafa, M., Bayramov, H., Trivigno, G., Berton, G., Masone, C., Caputo, B.: Are local features all you need for cross-domain visual place recognition? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6155–6165 (2023)
8. Berton, G., Masone, C.: Megaloc: One retrieval to place them all. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2861–2867 (2025)
9. Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4878–4888 (2022)
10. Berton, G., Masone, C., Paolicelli, V., Caputo, B.: Viewpoint invariant dense matching for visual geolocalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12169–12178 (2021)
11. Berton, G., Mereu, R., Trivigno, G., Masone, C., Csurka, G., Sattler, T., Caputo, B.: Deep visual geo-localization benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5396–5407 (2022)
12. Berton, G., Trivigno, G., Caputo, B., Masone, C.: Eigenplaces: Training viewpoint robust models for visual place recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11080–11090 (2023)
13. Chen, D., Chen, Y., Li, Y., Mao, F., He, Y., Xue, H.: Self-supervised learning for few-shot image classification. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1745–1749. IEEE (2021)
14. Chu, P., Bian, X., Liu, S., Ling, H.: Feature space augmentation for long-tailed data. In: European conference on computer vision. pp. 694–710. Springer (2020)
15. Feuerverger, A., Mureika, R.A.: The empirical characteristic function and its applications. The annals of Statistics pp. 88–97 (1977)
16. Garg, K., Puligilla, S.S., Kolathaya, S., Krishna, M., Garg, S.: Revisit anything: Visual place recognition via image segment retrieval. In: European Conference on Computer Vision. pp. 326–343. Springer (2024)
17. Ge, Y., Wang, H., Zhu, F., Zhao, R., Li, H.: Self-supervising fine-grained region similarities for large-scale image localization. In: European conference on computer vision. pp. 369–386. Springer (2020)

18. Haas, L., Skreta, M., Alberti, S., Finn, C.: Pigeon: Predicting image geolocations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12893–12902 (2024)
19. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14141–14152 (2021)
20. Hu, J., Nie, J., Ning, Z., Feng, C., Wang, L., Li, J., Cheng, S.: Enhancing visual place recognition with hybrid attention mechanisms in mixvpr. IEEE Access (2024)
21. Hu, S., Feng, M., Nguyen, R.M., Lee, G.H.: Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7258–7267 (2018)
22. Izquierdo, S., Civera, J.: Close, but not there: Boosting geographic distance sensitivity in visual place recognition. In: European Conference on Computer Vision. pp. 240–257. Springer (2024)
23. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17658–17668 (June 2024)
24. Jin Kim, H., Dunn, E., Frahm, J.M.: Learned contextual feature reweighting for image geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2136–2145 (2017)
25. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3304–3311 (2010). https://doi.org/10.1109/CVPR.2010.5540039
26. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: Eighth International Conference on Learning Representations (ICLR) (2020)
27. Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K.M., Scherer, S., Krishna, M., Garg, S.: Anyloc: Towards universal visual place recognition. IEEE Robotics and Automation Letters **9**(2), 1286–1293 (2023)
28. Khaliq, A., Xu, M., Hausler, S., Milford, M., Garg, S.: Vlad-buff: burst-aware fast feature aggregation for visual place recognition. In: European Conference on Computer Vision. pp. 447–466. Springer (2024)
29. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
30. Li, M., Cheung, Y.m., Lu, Y.: Long-tailed visual recognition via gaussian clouded logit adjustment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6929–6938 (2022)
31. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
32. Liu, L., Li, H., Dai, Y.: Stochastic attraction-repulsion embedding for large scale image localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2570–2579 (2019)
33. Lu, F., Jin, T., Ye, C., Lan, X., Liu, Y., Yuan, C.: Towards implicit aggregation: Robust image representation for place recognition in the transformer era. Advances in Neural Information Processing Systems **38**, 85096–85124 (2026)
34. Lu, F., Lan, X., Zhang, L., Jiang, D., Wang, Y., Yuan, C.: Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In: Proceed-

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16772–16782 (2024)
35. Lu, F., Zhang, L., Lan, X., Dong, S., Wang, Y., Yuan, C.: Towards seamless adaptation of pre-trained models for visual place recognition. In: The Twelfth International Conference on Learning Representations (2024)
36. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. In: International Conference on Learning Representations (2020)
37. Muller-Budack, E., Pustu-Iren, K., Ewerth, R.: Geolocation estimation of photos using a hierarchical model and scene classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 563–579 (2018)
38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
39. Paolicelli, V., Tavera, A., Masone, C., Berton, G., Caputo, B.: Learning semantics for visual place recognition through multi-scale attention. In: International Conference on Image Analysis and Processing. pp. 454–466. Springer (2022)
40. Peng, G., Zhang, J., Li, H., Wang, D.: Attentional pyramid pooling of salient visual residuals for place recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 885–894 (2021)
41. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1655–1668 (2018)
42. Seo, P.H., Weyand, T., Sim, J., Han, B.: Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 536–551 (2018)
43. Sferrazza, D., Berton, G., Trivigno, G., Masone, C.: To match or not to match: Revisiting image matching for reliable visual place recognition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2849–2860 (2025)
44. Shao, S., Chen, K., Karpur, A., Cui, Q., Araujo, A., Cao, B.: Global features are all you need for image retrieval and reranking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11036–11046 (2023)
45. Shi, J.X., Wei, T., Xiang, Y., Li, Y.F.: How re-sampling helps for long-tail learning? Advances in Neural Information Processing Systems **36**, 75669–75687 (2023)
46. Sünderhauf, N., Neubert, P., Protzel, P.: Are we there yet? challenging SeqSLAM on a 3000 km journey across all four seasons. In: Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA) (2013)
47. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1808–1817 (2015)
48. Trivigno, G., Berton, G., Aragon, J., Caputo, B., Masone, C.: Divide&classify: Fine-grained classification for city-wide visual geo-localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11142–11152 (October 2023)
49. Tzachor, I., Lerner, B., Levy, M., Green, M., Berkovitz Shalev, T., Habib, G., Samuel, D., Zailer, N., Shimshi, O., Darshan, N., et al.: Effovpr: Effective foundation model utilization for visual place recognition. In: International Conference on Learning Representations. vol. 2025, pp. 42817–42839 (2025)
50. Wang, B., Wang, P., Xu, W., Wang, X., Zhang, Y., Wang, K., Wang, Y.: Kill two birds with one stone: Rethinking data augmentation for deep long-tailed learn-

ing. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=RzY9qQHUXy
51. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5265–5274 (2018)
52. Wang, R., Shen, Y., Zuo, W., Zhou, S., Zheng, N.: Transvpr: Transformer-based place recognition with multi-level attention aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13648–13657 (2022)
53. Wang, S., Yang, Y., Liu, Z., Sun, C., Hu, X., He, C., Zhang, L.: Dataset distillation with neural characteristic function: A minmax perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25570–25580 (June 2025)
54. Wang, Z., Xu, Q., Yang, Z., He, Y., Cao, X., Huang, Q.: A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning. Advances in Neural Information Processing Systems **36**, 48417–48430 (2023)
55. Wang, Z., Xu, Q., Yang, Z., Xu, Z., Zhang, L., Cao, X., Huang, Q.: A unified perspective for loss-oriented imbalanced learning via localization. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
56. Warburg, F., Hauberg, S., Lopez-Antequera, M., Gargallo, P., Kuang, Y., Civera, J.: Mapillary street-level sequences: A dataset for lifelong place recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2626–2635 (2020)
57. Warburg, F., Jørgensen, M., Civera, J., Hauberg, S.: Bayesian triplet loss: Uncertainty quantification in image retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12158–12168 (2021)
58. Weyand, T., Kostrikov, I., Philbin, J.: Planet-photo geolocation with convolutional neural networks. In: European conference on computer vision. pp. 37–55. Springer (2016)
59. Wu, T., Huang, Q., Liu, Z., Wang, Y., Lin, D.: Distribution-balanced loss for multi-label classification in long-tailed datasets. In: European conference on computer vision. pp. 162–178. Springer (2020)
60. Yang, Z., Xu, Q., Wang, Z., Li, S., Han, B., Bao, S., Cao, X., Huang, Q.: Harnessing hierarchical label distribution variations in test agnostic long-tail recognition. In: Forty-first International Conference on Machine Learning (2024)
61. Ye, X., Robert, L.P.: A human–security robot interaction literature review. ACM Transactions on Human-Robot Interaction **14**(2), 1–36 (2024)
62. Yu, H., Du, Y., Wu, J.: Reviving undersampling for long-tailed learning. Pattern Recognition **161**, 111200 (2025)
63. Zaffar, M., Nan, L., Kooij, J.F.P.: On the estimation of image-matching uncertainty in visual place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17743–17753 (2024)
64. Zhang, H., Chen, X., Jing, H., Zheng, Y., Wu, Y., Jin, C.: Etr: An efficient transformer for re-ranking in visual place recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5665–5674 (2023)
65. Zhang, S., Mao, H., Chen, Q., Kim, Y.: Efficient visual place recognition through multimodal semantic knowledge integration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5601–5610 (2025)
66. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(9), 10795–10816 (2023)

67. Zhao, Y., Chen, W., Tan, X., Huang, K., Zhu, J.: Adaptive logit adjustment loss for long-tailed visual recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 3472–3480 (2022)
68. Zhu, S., Yang, L., Chen, C., Shah, M., Shen, X., Wang, H.: R2former: Unified retrieval and reranking transformer for place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19370–19380 (2023)
69. Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3640–3649 (2021)