applied sciences

MDPI

Article

CriSALAD: Robust Visual Place Recognition Using Cross-Image Information and Optimal Transport Aggregation

Jinyi Xu [1], Yuhang Ming [1,2,*], Minyang Xu [1,2], Yaqi Fan [1], Yuan Zhang [1] and Wanzeng Kong [1,2]

[1] School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China; 22050505@hdu.edu.cn (J.X.); minyangxu@hdu.edu.cn (M.X.); 22150605@hdu.edu.cn (Y.F.); 22062107@hdu.edu.cn (Y.Z.); kongwanzeng@hdu.edu.cn (W.K.)
[2] Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province, Hangzhou 310018, China
* Correspondence: yuhang.ming@hdu.edu.cn

Abstract: Visual place recognition (VPR) is crucial for enabling autonomous agents to accurately localize themselves within a known environment. While existing methods leverage neural networks to enhance performance and robustness, they often suffer from the limited representation power of local feature extractors. To address this limitation, we propose CriSALAD, a novel VPR model that integrates visual foundation models (VFMs) and cross-image information to improve feature extraction robustness. Specifically, we adapt pre-trained VFMs for VPR by incorporating a parameter-efficient adapter inspired by Xception, ensuring effective task adaptation while preserving computational efficiency. Additionally, we employ the Sinkhorn Algorithm for Locally Aggregated Descriptors (SALAD) as a global descriptor to enhance place recognition accuracy. Furthermore, we introduce a transformer-like cross-image encoder that facilitates information sharing between neighboring images, thus enhancing feature representations. We evaluate CriSALAD on multiple publicly available place recognition datasets, achieving promising performance with a recall@1 of 89.3% on the Nordland dataset, while the closest rival achieves only 76.2%. CriSALAD outperforms both baseline models and advanced VFM-based VPR approaches.

Keywords: visual place recognition; visual localization; transfer learning; feature aggregation

check for updates

Academic Editor: Atsushi Mase

Received: 13 March 2025
Revised: 25 April 2025
Accepted: 7 May 2025
Published: 9 May 2025

Citation: Xu, J.; Ming, Y.; Xu, M.; Fan, Y.; Zhang, Y.; Kong, W. CriSALAD: Robust Visual Place Recognition Using Cross-Image Information and Optimal Transport Aggregation. *Appl. Sci.* **2025**, *15*, 5287. https://doi.org/10.3390/app15105287

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Visual place recognition (VPR) is a key research area in computer vision and robotics that aims to localize an agent by matching query images against place-tagged reference datasets [1,2]. It plays a crucial role in applications such as mobile robotic localization [3], augmented reality [4], and autonomous navigation [5], where accurate place recognition is essential for state estimation in complex environments. Despite substantial advancements, VPR remains a challenging task due to factors such as appearance variations [6,7], viewpoint discrepancies [8,9], and perceptual aliasing [10,11], where visually similar locations lead to incorrect matches.

Typically, VPR is formulated as an image retrieval problem, where a query image is represented as a compact feature descriptor and compared against a database of place-tagged reference images to retrieve the closest match. The process of generating a global descriptor generally involves two key steps: local feature extraction and global descriptor aggregation. Conventional VPR methods commonly rely on combinations of handcrafted local features with global descriptors to generate global image representations. Examples include SURF [12] with Vector of Locally Aggregated Descriptors (VLAD) [13], and ORB [14] with Bag of Words (BoW) [15]. However, these approaches fail to adequately

preserve geometric and structural information, making it challenging to handle lighting and appearance variations. Targeting these limitations, advanced deep learning models are leveraged to enhance VPR performance by improving both local feature extraction and global descriptor aggregation.

In terms of local feature extraction, early approaches replaced hand-crafted features with Convolutional Neural Networks (CNNs) [16], demonstrating improved robustness and scalability. In contrast, Vision Transformer (ViT) [17] leverages a self-attention mechanism to effectively capture global dependencies in images, thus excelling at understanding the overall scene structure and long-range relationships. In the context of VPR, ViT precisely models relationships between regions, such as the positions of buildings or the interaction between landmarks and backgrounds, thereby generating unique global descriptors. This makes ViT well suited for tasks demanding holistic scene comprehension. Consequently, visual foundation models (VFMs), such as AnyLoc [18] and CricaVPR [19], which are based on ViT, have gained popularity due to their strong generalization capabilities.

Regarding global descriptor aggregation, NetVLAD [20] stands as a milestone, introducing a differentiable modification of the VLAD layer within a CNN for end-to-end optimization. Additionally, MixVPR [21] enhances feature aggregation by incorporating a multilayer perceptron (MLP), thus improving feature expressiveness. In contrast, generalized mean pooling (GeM) [22] prioritizes efficiency, utilizing a simple yet effective pooling mechanism for global descriptor generation.

More recently, a promising development is the Sinkhorn Algorithm for Locally Aggregated Descriptors (SALAD) [23], which formulates feature aggregation as an optimal transport problem. By selectively discarding uninformative local features, SALAD improves the discriminative power of global descriptors. However, its reliance on fine-tuning DiNOv2, a VFM, increases computational costs, as it requires unfreezing and retraining the final layers of the model. Additionally, SALAD processes each query image independently, without leveraging contextual relationships between images in the database—an aspect that could further enhance recognition robustness.

To address these limitations, we propose CriSALAD, a novel visual place recognition framework that integrates cross-image information with optimal transport-based feature aggregation. Inspired by CricaVPR [19], we employ parameter-efficient transfer learning, incorporating a lightweight depthwise separable convolutional adapter into the frozen pretrained backbone. This design enhances inter-channel information fusion while preserving spatial structure, thus improving adaptability in various environments. Furthermore, we introduce a cross-image correlation-aware representation mechanism, which computes relationships between feature embeddings across multiple images within a training batch, thus enhancing global descriptor robustness.

Our key contributions can be summarized as follows:

1. Global descriptor generation with cross-image information—We introduce a novel global descriptor generation approach that integrates cross-image information with optimal transport-based feature aggregation, thus enhancing recognition robustness.
2. Parameter-efficient adaptation strategy—We propose a lightweight adaptation mechanism that decouples spatial filtering via depthwise convolutions from channel-wise interactions via pointwise convolutions. This design reduces the computational overhead (FLOPs) compared to other conventional adapters.
3. State-of-the-art performance—Extensive experiments conducted on multiple VPR benchmarks demonstrate that CriSALAD outperforms existing approaches on several datasets, such as Pitts30k and Nordland.

The remainder of this paper is organized as follows. Section 2 provides a review of related works in the fields of VPR and parameter-efficient transfer learning. Section 3

begins with a brief overview of transformer blocks, followed by a detailed presentation of our proposed method, CriSALAD. This includes the design of the VPR framework, which integrates depthwise separable convolution adaptation, global descriptor aggregation, and a cross-image method. The experimental details and results are discussed in Section 4. Finally, this paper concludes in Section 5.

2. Related Works

2.1. Visual Place Recognition

Following the milestone work NetVLAD [20], which first enabled end-to-end training for VPR by integrating CNN-based feature extraction with a differentiable VLAD layer, many subsequent methods have been proposed. In image-based place recognition, Patch-NetVLAD [24] derives patch-level features and combines local and global descriptor methods. Ref. [25] proposed a hardness-aware metric learning method with cluster-guided attention. VIPeR [26] balances performance in a single environment with generality across multiple environments. In the domain of point cloud-based place recognition, Indoor DH3D from FD-SLAM [27] introduces color into 3D point clouds, CGiS-Net [10] identifies indoor places from color point clouds by aggregating low-level color and geometric features with high-level implicit semantic features, and AEGIS-Net [11] uses a self-attention module to select the most important local features.

In particular, to address the issue of perceptual aliasing, Oh et al. proposed a variational autoencoder that effectively extracts condition-invariant features by leveraging probabilistic modeling to separate environment-dependent and environment-independent features [28]. Meanwhile, Peng et al. demonstrated that the integration of attention mechanisms can effectively identify salient regions [6]. Ye et al. proposed the use of global features from scan-context descriptors and dictionary-based coding [29], enabling the model to handle challenging perceptual aliasing scenarios. Focusing on the viewpoint variant problem, Chen et al. introduced a coarse-to-fine hierarchical VPR pipeline that integrates semantic aggregation and scene understanding [30].

Apart from using a NetVLAD [20] layer for global aggregation, there are many other popular aggregation methods, such as pooling layer-based approaches [22], learning global aggregation [21], and optimal transport aggregation [23]. Moreover, instead of formulating VPR as a retrieval problem, CosPlace [31] and EigenPlaces [9] innovatively treat the VPR task as a classification problem and have achieved impressive VPR performance, with reduced GPU memory consumption and smaller descriptor size.

Considering the powerful feature extraction capabilities of ViT, many methods have incorporated ViT into the VPR task. AnyLoc [18] directly utilizes a pre-trained DiNOv2 [32] model as a feature extractor, achieving promising results. Following this, many other methods have been proposed to fine-tune the pre-trained DiNOv2 model to not only leverage its general feature extraction ability but also fit the model better for the VPR task. In particular, DiNO-MIX [33] and DiNO-SALAD [23] use a straightforward approach, fine-tuning the last few blocks of the DiNOv2 model. On the other hand, CricaVPR [19] and SelaVPR [34] use parameter-efficient fine-tuning, also known as parameter-efficient transfer learning, to freeze the weights of DiNOv2, utilizing different adapter architectures to enable fine-tuning. Our proposed CriSALAD model also utilizes the parameter-efficient fine-tuning approach and achieves better performance with fewer fine-tuning parameters.

2.2. Parameter-Efficient Transfer Learning

Pre-trained ViTs have demonstrated outstanding performance across various visual tasks. However, deploying ViTs for visual recognition presents significant challenges due to their high computational and storage demands. Furthermore, each ViT model typically

requires extensive task-specific fine-tuning, limiting its transferability and efficiency. To address these challenges, parameter-efficient transfer learning (PETL) [35,36] adapts large pre-trained models to specific downstream tasks with minimal computational overhead. Originally introduced in natural language processing (NLP) to mitigate catastrophic forgetting and reduce training costs, PETL encompasses three primary strategies: task-specific adapters [37], prompt tuning [38], and low-rank adaptation (LoRA) [35]. In this work, we adopt the task-specific adapter approach, which enables the adaptation of foundation models by optimizing only a small subset of additional parameters, thus reducing overfitting risks while preserving generalization capabilities.

Recent works have applied PETL to visual tasks, enabling better adaptation of large models to specific tasks. AdaptFormer [39] efficiently adapts pre-trained ViT models with a lightweight AdaptMLP module. The Side Adapter Network [40] generates mask proposals and attention biases via side networks, allowing efficient fine-tuning of pre-trained CLIP models for open-vocabulary semantic segmentation. In the visual place recognition task, CricaVPR [19] improves feature representation through cross-image correlation modeling and multi-scale convolutions. SelaVPR [34] achieves efficient hybrid global–local adaptation for visual place recognition using lightweight adapters. Inspired by Xception [41], we design a depthwise separable convolutional adapter to enhance the model's feature decoupling capability.

3. Methodology

In this paper, we present CriSALAD, a VPR model that extracts local features using visual foundation models and aggregates global descriptors using optimal transport aggregation. Our CriSALAD model also leverages cross-image information to boost performance. An overview of CriSALAD is shown in Figure 1. In the following, we first provide a concise overview of the ViT architecture, which is the core of the DiNOv2 model [32]. The adapted DiNOv2 model consists of L layers of transformer blocks. In each block, we follow the original transformer design [42], which is composed of multi-head attention (MHA) and multi-layer perceptron (MLP). Then, we dive into our adapted DiNOv2 model and present the depthwise separable convolution-based adapter. Subsequently, we discuss the global descriptor aggregation method used to establish global matching relationships between features. Finally, we detail the cross-image correlation-aware place representation, which significantly improves robustness in challenging environments and effectively addresses the perceptual aliasing problem through cross-image correlation modeling.

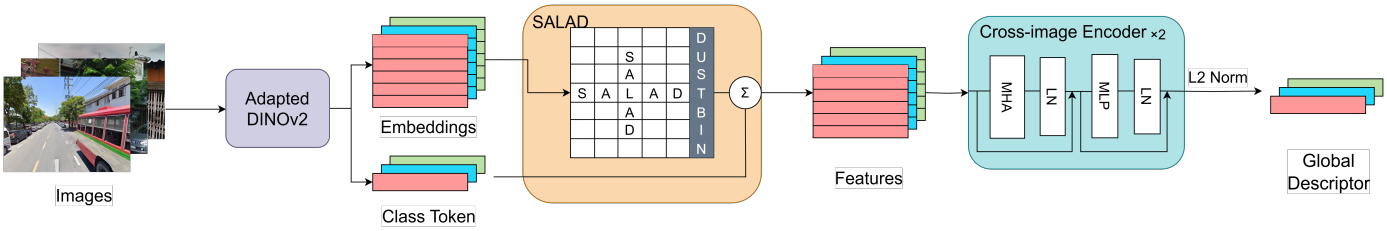


Figure 1. Network architecture of CriSALAD. It has three main components: an adapted DiNOv2 model for local feature extraction, a SALAD module to aggregate global descriptors, and a cross-image encoder to share information across neighboring images.

3.1. Preliminaries

ViT [17] initially partitions the input image $x \in \mathbb{R}^{H\times W\times C}$ into N patches, where $N = HW/P^2$, $P = 14$, which are subsequently linearly projected onto D-dimensional patch embeddings $x_p \in \mathbb{R}^{N\times D}$. A learnable class token embedding x_{class} is prefixed to the patch embeddings x_p, forming the initial sequence $X_0 = \left[X_{class}; X_p^1, X_p^2, ..., X_p^N\right]$. Positional

embeddings $E_{pos} \in \mathbb{R}^{(N+1)\times D}$ are then added to the patch embeddings to preserve spatial information between patches, resulting in $z_0 = X_0 + E_{\text{pos}}$, which is then passed to the transformer block. The transformer block comprises alternating MHA and MLP blocks, with layer normalization (LN) applied before each block and residual connections after each block. The forward propagation of input z_{l-1}, where l denotes the layer index, through a transformer encoder layer to produce the output can be formally expressed as

$$\begin{aligned} z'_l &= \text{MHA}(\text{LN}(z_{l-1})) + z_{l-1} \\ z_l &= \text{MLP}(\text{LN}(z'_l)) + z'_l \end{aligned} \tag{1}$$

3.2. Depthwise Separable Convolutional Adapter

As demonstrated in the pioneering work AnyLoc [18], DiNOv2 [32] exhibits promising performance not only in common visual tasks but also in VPR tasks. With proper fine-tuning, as shown in DiNO-SALAD [23], the model can achieve state-of-the-art (SoTA) performance in VPR tasks. However, due to the high complexity of the DiNOv2 model, even fine-tuning the last few blocks incurs significant computational cost. However, we take inspiration from the parameter-efficient transfer learning approach used in CricaVPR [19] and SelaVPR [34] and present an efficient adapter for the DiNOv2 model.

Inspired by Xception [41], we design a depthwise separable convolutional adapter. As shown on the left side of Figure 2, our adapter first reduces the dimensionality through a fully connected layer to decrease computational load and enhance generalization ability. Subsequently, it employs a separable convolution to achieve efficient feature extraction with a minimal number of parameters. Additionally, a skip connection is utilized to prevent performance degradation. Finally, a fully connected layer is applied to up-project the features back to the original dimension. As shown in Figure 2, we follow the idea of parallel adaptation and attach the depthwise separable convolutional adapter in parallel to the MLP of each adapted transformer block. With the pre-trained weights loaded into the DiNOv2 backbone, we keep all these weights frozen and only update the trainable depthwise separable convolutional adapter. The forward operations within each adapted transformer module are formally defined below, with α being a pre-defined weighting factor:

$$\begin{aligned} z'_l &= \text{MHA}(\text{LN}(z_{l-1})) + z_{l-1} \\ z_l &= \text{MLP}(\text{LN}(z'_l)) + \alpha \cdot \text{Adapter}(\text{LN}(z'_l)) + z'_l \end{aligned} \tag{2}$$

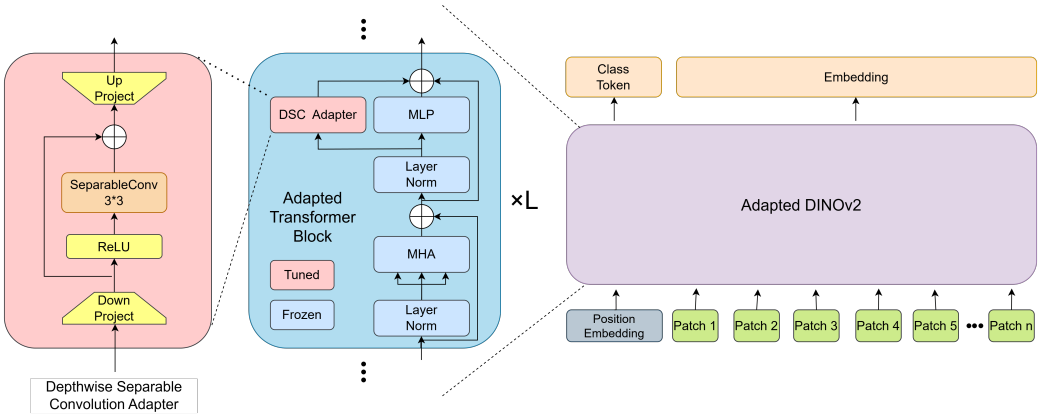


Figure 2. Detailed architecture of our adapter and adapted DiNOv2 model. The adapted DiNOv2 model is composed of L layers of transformer blocks, and we freeze the parameters within these blocks, training only the parameters in the adapter.

3.3. Global Descriptor Aggregation

In our work, we adapt the Sinkhorn algorithm from the SALAD method to generate global image descriptors, reformulating feature-to-cluster assignment as an optimal transport problem. Unlike NetVLAD, which relies on k-means cluster centers, SALAD employs a learnable, data-driven assignment strategy. Instead of introducing prior biases, it utilizes a randomly initialized two-layer fully connected network to compute the feature-to-cluster assignment score matrix $S \in \mathbb{R}^{N \times m}$, where m is the number of cluster centroids, dynamically generating per-row assignment scores. For each local feature $t_i \in \mathbb{R}^D$, where $i \in \{1, 2, \ldots, N\}$ indexes the i-th local feature, the corresponding score vector $s_i \in \mathbb{R}^m$ is computed as

$$s_i = W_{s_2}(\sigma(W_{s_1}(t_i) + b_{s_1})) + b_{s_2} \tag{3}$$

where W_{s_1} and W_{s_2}, and b_{s_1} and b_{s_2} are learnable weights and biases, and σ is a non-linear activation function. This formulation allows dynamic learning of cluster structures, making assignments more adaptive.

To ensure robustness, SALAD introduces a "dustbin" cluster, allowing non-informative features to be discarded rather than forced into predefined clusters. This is achieved by extending the assignment matrix:

$$\overline{S} = [S, \bar{s}_{i,m+1}] \in \mathbb{R}_{>0}^{N \times m+1} \tag{4}$$

where $\bar{s}_{i,m+1} = \gamma \cdot 1_N$ is a constant column vector, $\gamma \in \mathbb{R}$ is a learnable scalar, and $\mathbf{1}_N \in \mathbb{R}^N$ is an all-one vector.

By considering feature-to-cluster and cluster-to-feature assignments, the method formulates clustering as an optimal transport problem. The Sinkhorn algorithm solves this constrained optimization, ensuring effective feature distribution:

$$\overline{P} 1_{m+1} = \mu \quad \text{and} \quad \overline{P}^\top 1_n = \kappa \tag{5}$$

where $\mu = 1_N$ and $\boldsymbol{\kappa}$ defines the mass distribution constraints. The matrix $\overline{p}$ is the complete assignment matrix containing the dustbin column. Once optimized, the dustbin column is removed, yielding the refined assignment matrix P, where $P_{i,j}$ represents the assignment weight of feature i to cluster j.

With the optimal assignments, local features are aggregated to form the global descriptor. Traditional methods use residual aggregation (e.g., VLAD and NetVLAD) or statistical aggregation (e.g., GeM). Instead, SALAD dynamically assigns features to clusters and pools them based on learned weights.

Each local feature t_i is then projected onto a lower-dimensional space via a two-layer MLP to obtain $f_i \in \mathbb{R}^l$, thus improving efficiency while retaining task-relevant information. The final global descriptor $V \in \mathbb{R}^{m \times l}$ is constructed via a weighted sum of features as

$$V_{j,k} = \sum_{i=1}^{n} P_{i,k} \cdot f_{i,k} \tag{6}$$

where $f_{i,k}$ is the k-th dimension of the reduced feature f_i.

This eliminates reliance on residual encoding while enabling adaptive feature selection. By integrating optimal transport principles with learned clustering, SALAD enhances descriptor flexibility and robustness, thus improving adaptability in complex environments.

3.4. Cross-Image Correlation-Aware Place Representation

Traditional VPR methods employ MHA to analyze individual images, capturing internal region relationships. However, this approach is inherently limited, as it cannot

directly model interactions between multiple images or fully exploit their contextual dependencies. As a result, it provides an incomplete representation of complex scenes. In contrast, a cross-image encoder extends MHA to process multiple images simultaneously, enabling it to focus on region correspondences across images. When handling a batch of images, this method learns both image similarities and differences, making it particularly effective for image matching, retrieval, and multi-view understanding. By integrating features from multiple images, the model becomes more robust to noise and variations, improving its ability to handle scene complexity.

Following CricaVPR [19], our approach treats the features described in Section 3.3 as sequential data and processes them through a cross-image encoder composed of two transformer encoder layers. Each layer consists of an MHA module, an MLP module, an LN module, and skip connections. This architecture effectively models region-level correlations among images in a batch, enhancing feature invariance.

During inference, the cross-image encoder operates on single images due to the lack of contextual batches, which limits its ability to leverage cross-image correlations. While this does have some impact, its effect on retrieval accuracy is relatively minor. Finally, the regional features are concatenated and L2-normalized to generate the global descriptor, ensuring a compact and discriminative representation.

4. Experiments

4.1. Datasets

We trained our proposed CriSALAD model on the GSV-Cities dataset [43], which is a large-scale VPR dataset containing images taken with Google Street View Time Machine. With the images collected from 40 major cities across the world, the dataset covers about 67,000 places spanning 2000 km^2 geographically and 14 years in time. In addition, it provides a wide variety of viewpoints, illumination changes, weather conditions, etc. For evaluation, we followed the recent literature [19,23,34] and adopted several widely used benchmark datasets for visual place recognition, including Pitts30k [44], Tokyo24/7 [45], Nordland [46], and Mapillary Street-Level Sequences (MSLS) [47].

Pitts30k is a smaller subset of Pitts250k [44], which is a large-scale dataset collected in Pittsburgh, encompassing urban environments with significant viewpoint variations. It contains approximately 30,000 database images. Moreover, Tokyo24/7 [45], which is also an urban environment dataset containing thousands of day-time and night-time images taken in the city of Tokyo, was used to evaluate the model's robustness against illumination changes. In contrast, Nordland [46] was designed for long-term VPR evaluation. It comprises approximately 1 million images captured during a single train journey across northern Norway, spanning all four seasons, to assess place recognition under extreme seasonal variations. In particular, images from the winter sequence are designated as the query set, with the summer counterparts constituting the reference database. Finally, MSLS is one of the largest and most diverse street-level datasets for VPR tasks. It contains millions of images taken from over 30 cities across the world, covering different viewpoints, cities, times of day, and weather conditions.

4.2. Implementation Details

We implemented our proposed CriSALAD model using PyTorch v2.0.0 and fine-tuned it under the parameter-efficient setup with a single NVIDIA GeForce RTX 4090D GPU. For the adapted DiNOv2 feature extractor, we used ViT-B/14 as our backbone. We set the input image resolution at 224×224 and the feature tokens with dimensions of 768. Regarding the adapter, we used a bottleneck ratio of 0.5 and a kernel size of 3×3 for the separable convolution, and the weighting factor in Equation (2) for α was set to 0.2.

For the SALAD global aggregator, we set all hidden fully connected layers to 512 neurons, activated by ReLU functions. Following the SALAD default setup, we designed 64 clusters, each with 128 dimensions, and set the global token dimension to 256. The final global descriptor of our model combined the global token and all cluster tokens, resulting in a dimension of 8448.

For training, we used the GSV-Cities dataset for fully supervised training. The dataset contains 560,000 images from 67,000 locations and provides highly accurate label information. The richness of the dataset and the accuracy of the labels make it an important benchmark for urban visual tasks. We followed the standard framework of the dataset and used the multi-similarity (MS) loss function for training. To further enhance training performance, we also incorporated the online hard example mining (OHEM) strategy to more effectively learn from challenging samples, thereby enhancing the model's robustness and generalization. For inference, our method was evaluated with an inference batch size of 16.

To implement these strategies effectively, we fine-tuned the model using the AdamW [48] optimizer with an initial learning rate of 6×10^{-5} and applied a linear decay to improve learning effectiveness. We used a training batch of 48 places, each with four images. The model was trained on the GSV-Cities dataset until the *Recall*@5 on the Pitts30k [44] dataset stopped improving for three consecutive epochs. In our experiments, the CriSALAD model achieved convergence within five epochs.

4.3. Evaluation Metrics

In line with common practice in VPR research, we adopted the average recall rate over all queries for the top-K retrievals, denoted as *Recall*@K, which is computed as

$$Recall@K = \frac{|\{q \in Q | \exists i \leq K, d_i \in \mathcal{G}(q)\}|}{|Q|}, \tag{7}$$

where q is a query image from the set Q, d_i represents the i-th retrieved database image (also referred to as a place image) for the query image q, $\mathcal{G}(q)$ denotes the set of ground-truth place images corresponding to the query image q, K is the number of top retrieved results considered, and $|\cdot|$ stands for the counting operation to obtain the cardinality of the set.

It is worth mentioning that various ground-truth criteria were used to evaluate different datasets. Specifically, we followed the evaluation procedures outlined in [44,45,47] and used 25 m for the Pitts30k [44] and Tokyo24/7 [45] datasets, 10 frames for the Nordland [46] dataset, and 25 m and 40° for the MSLS [47] dataset.

4.4. Comparison Methods

We first established baseline performance with the milestone VPR models NetVLAD [20] and GeM [22]. For more recent and robust learning-based models, we experimented with Conv-AP [43], Patch-VPR [24], TransVPR [49], CosPlace [31], MixVPR [21], and EigenPlaces [9]. Finally, we compared our proposed CriSALAD model with AnyLoc [18], Crica-VPR [19], and DiNO-SALAD [23], three recent VPR models that leverage visual foundation models like DiNOv2. To ensure a fair comparison, we re-trained all models with the same training dataset, i.e., GSV-Cities [43], except for CosPlace and EigenPlaces, which were trained on a different large-scale dataset (SF-XL [31]), and AnyLoc, which did not require re-training. For the training batch size, CricaVPR followed the original paper by using a training batch size of 288, while the other methods maintained consistency with our current training batch size of 192. Moreover, we disabled the principal component analysis (PCA) dimension reduction step in all models and used the output dimension of the global aggregator directly.

4.5. Results and Discussion

The quantitative and qualitative results, presented in Table 1 and Figure 3, demonstrate that CriSALAD significantly outperformed DiNO-SALAD and other baseline and SoTA methods across most tested scenarios. CricaVPR(16 bs) refers to the results of CricaVPR using an inference batch size of 16, which matched our settings for a fair comparison. Notably, in terms of the average Top-1 recall, CriSALAD achieved 94.7% on the Pitts30k [44] dataset and 89.3% on the Nordland [46] dataset, surpassing the closest competing methods, which reached only 93.7% and 76.2%, respectively. This performance demonstrates the superiority of our proposed CriSALAD model in structured urban environments and challenging long-term place recognition scenarios.

Despite the fact that CriSALAD did not achieve the highest performance on Tokyo24/7 [45] and MSLS [47], it remained highly competitive. On the one hand, our method did not outperform other methods on Tokyo24/7 and MSLS due to the lack of generalization in illumination and dynamic issues, which need to be addressed in future work. On the other hand, our method outperformed other methods on Nordland and Pitts30k, demonstrating its strength in handling weather and viewpoint changes. These results highlight the robustness of CriSALAD, demonstrating its strong generalization ability across diverse datasets while maintaining top-tier accuracy in challenging scenarios. Our CriSALAD model can be further optimized to deal with illuminance changes, as in the Tokyo24/7 dataset, and occlusions, as in the MSLS dataset. In terms of inference speed, CricaVPR, using the multi-resolution convolutional adapter, achieved 148.96 FPS on a single NVIDIA GeForce RTX 4090D GPU, whereas our method achieved 427.36 FPS on the same hardware.

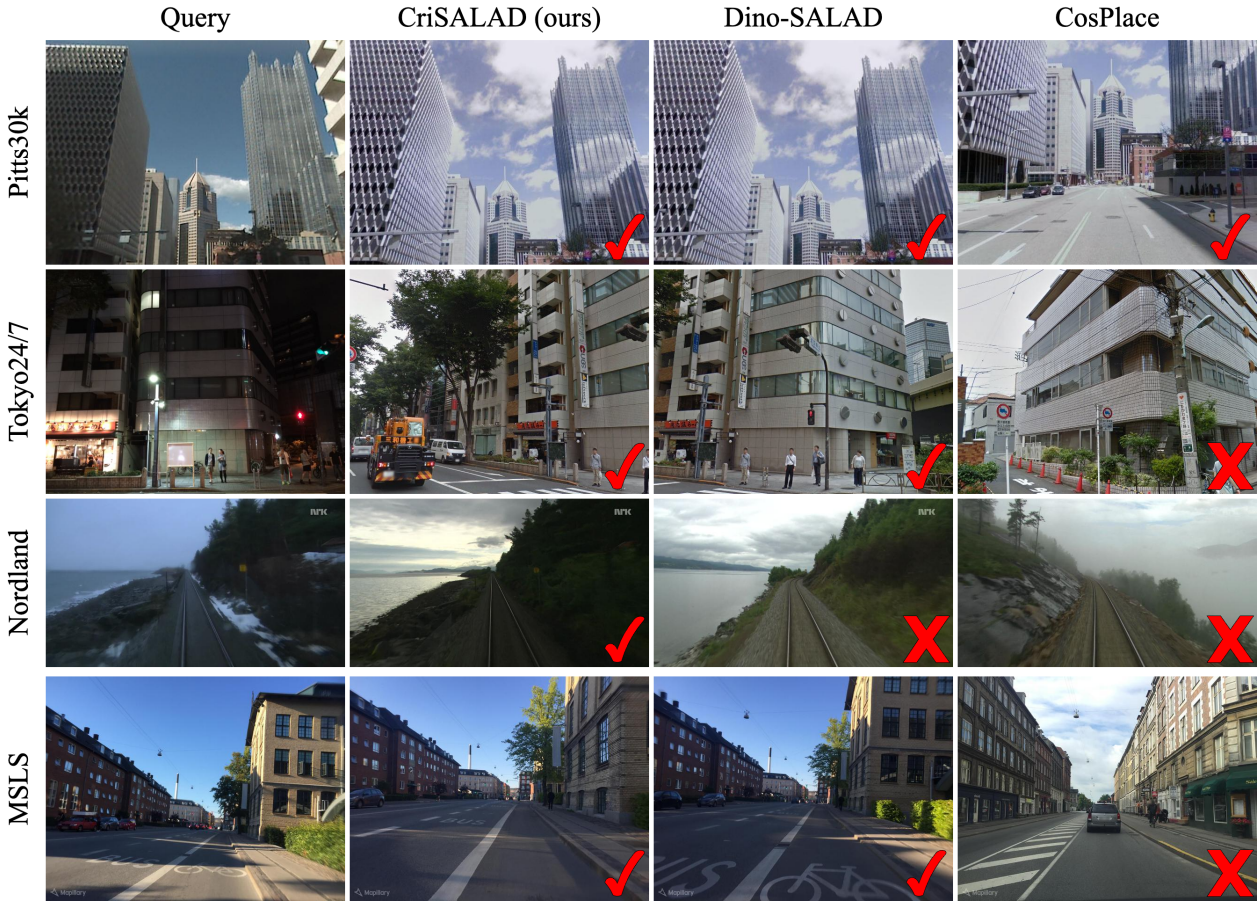


Figure 3. Examples of Top-1 retrieval from the datasets. Check and fork marks stand for correct and incorrect retrievals.

Table 1. Comparison with state-of-the-art methods on benchmark datasets. The best results are indicated in **bold**, and the second-best results are underlined.

Method	Dim.	Pitts30k			Tokyo24/7			Nordland			MSLS-Val		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
NetVLAD [20]	32,768	81.9	91.2	93.7	60.6	68.9	74.6	-	-	-	53.1	66.5	71.1
Conv-AP [43]	4096	89.4	94.7	95.9	81.0	88.3	92.4	-	-	-	69.2	80.3	83.1
Patch-NetVLAD [24]	/	88.7	94.5	95.9	86.0	88.6	90.5	-	-	-	79.5	86.2	87.7
TransVPR [49]	/	89.0	94.9	96.2	79.0	82.2	85.1	-	-	-	86.8	91.2	92.4
CosPlace [31]	512	88.4	94.5	95.7	81.9	90.2	92.7	58.5	73.7	79.4	82.8	89.7	92.0
MixVPR [21]	4096	91.5	95.5	96.3	85.1	91.7	94.3	76.2	86.9	90.3	88.0	92.7	94.6
EigenPlaces [9]	2048	92.5	96.8	97.6	**93.0**	96.2	97.5	71.2	83.8	88.1	89.1	93.8	95.0
CricaVPR(16 bs) [19]	4096	93.7	97.0	**98.1**	**93.0**	**97.1**	**97.8**	-	-	-	89.9	95.4	96.2
DiNO-SALAD [23]	8448	91.9	96.3	97.4	92.1	96.2	96.5	76.0	89.2	92.0	**92.2**	**96.4**	**97.0**
CriSALAD (ours)	8448	**94.7**	**97.2**	97.8	90.8	95.6	97.1	**89.3**	**95.7**	**97.2**	87.4	93.8	95.0

4.6. Ablation Studies

To validate the design choices of our CriSALAD model, we conducted a series of ablation studies on the Nordland dataset [46].

Learning rate and scheduler: We experimented with two different learning rate and scheduler configurations. Specifically, we compared performance using a learning rate of 6×10^{-5} with a linear scheduler, as adopted in SALAD [23], against a learning rate of 0.0001 with a step scheduler, as used in CricaVPR [19]. The results, presented in the first three rows of Table 2, indicate a significant degradation in place recognition performance when transitioning from 6×10^{-5} (linear scheduler) to 0.0001 (linear scheduler) and 6×10^{-5} (step scheduler). In particular, the average Top-1 recall rate dropped from 89.3% to 80.9% and 85.0%, respectively. These findings highlight the importance of selecting an appropriate learning rate and scheduler in optimizing model performance for visual place recognition.

Table 2. Ablation studies on the model's design choices. CIE stands for the cross-image encoder, lr stands for the learning rate, DSConv stands for the depthwise separable convolutional adapter, and MultiConv stands for the multi-resolution convolutional adapter.

Model Configuration				Nordland		
Adapter	CIE	lr	Scheduler	R@1	R@5	R@10
DSConv-DiNOv2	✓	6×10^{-5}	Linear	**89.3**	**95.7**	**97.2**
DSConv-DiNOv2	✓	0.0001	Linear	80.9	90.4	93.4
DSConv-DiNOv2	✓	6×10^{-5}	Step	85.0	93.1	95.3
MultiConv-DiNOv2	✓	6×10^{-5}	Linear	84.8	93.1	95.3
MLP-DiNOv2	✓	6×10^{-5}	Linear	86.4	94.1	95.2
DSConv-DiNOv2	×	6×10^{-5}	Step	69.4	80.9	85.4

Depthwise Separable Convolutional Adapter: To assess the effectiveness of our adapter, we performed an ablation study by replacing the depthwise separable convolutional adapter with the multi-resolution convolutional adapter from CricaVPR [19] and a simple MLP adapter. All adapters were placed in the same position within the DiNOv2 backbone to ensure a fair comparison. The results, shown in the fourth and fifth rows of Table 2, indicate a notable drop in performance. The average Top-1 recall rate dropped from 89.3% to 84.8% when using the multi-resolution convolutional adapter and to 86.4% when using the MLP adapter. These findings suggest that the depthwise separable convolutional adapter achieves superior results and improves computational efficiency.

Although the MLP adapter incurred lower computational cost, its lack of convolutional operations limited its capacity to learn expressive features. As a result, its performance remained suboptimal, as evidenced by the results in Table 2. Our adapter replaced the traditional convolution operations with depthwise separable convolutions. As shown in Table 3, our depthwise separable convolutional adapter outperformed CricaVPR's multi-resolution convolutional adapter in the Multiply-Accumulate Operations (MACs), Floating Point Operations (FLOPs), and parameter count, demonstrating that our adapter achieved higher performance and significantly enhanced computational efficiency and model compactness.

Table 3. Comparison of adapter efficiency: FLOPs, MACs, and parameter counts, all calculated on a per-sample basis for the forward pass during inference.

	MACs (M)	FLOPs (M)	Params (K)
MLP adapter	1.181	2.362	2356.08
MultiConv adapter	44.236	88.472	761.472
Our adapter	39.223	78.446	741.504

Cross-Image Encoder: To assess the contribution of our cross-image encoder, we conducted an ablation experiment by removing it from the architecture. The results, presented in the last row of Table 2, showed a significant performance drop. Without the cross-image encoder to facilitate information sharing between neighboring images, the average Top-1 recall dropped from 89.3% to just 69.4%. This underscores the critical role of the cross-image encoder in enhancing feature contextualization and improving retrieval accuracy by leveraging multi-image relationships.

5. Conclusions

In this paper, we introduced CriSALAD, a novel VPR model that fully leverages the generalization capabilities of extensively pre-trained visual foundation models, allowing it to extract environment-invariant features. To improve task-specific adaptability, we incorporated a depthwise separable convolutional adapter into the pre-trained backbone, enabling efficient fine-tuning by updating only the adapter's weights.

Furthermore, we integrated the SALAD descriptor to both optimize global feature aggregation and suppress dynamic interference and leveraged cross-image correlations to strengthen contextual representations, thereby facilitating more robust scene understanding. Our CriSALAD model achieved promising performance on several datasets. Particularly, on the Nordland dataset, CriSALAD achieved a Recall@1 of 89.3%, surpassing the closest competitor by 13.1 percentage points (76.2%).

In future work, we aim to enhance the model's generalization ability to effectively address challenges such as illumination variations and occlusions in visual place recognition tasks. Additionally, we plan to expand our experimental framework to include a broader range of unstructured environments, such as indoor settings and underwater scenarios.

Author Contributions: Conceptualization, J.X., Y.M., M.X., Y.F., Y.Z. and W.K.; methodology, J.X., Y.M. and M.X.; software, J.X., Y.M., Y.F., Y.Z. and M.X.; writing—original draft preparation, J.X., Y.M., Y.F., Y.Z., M.X. and W.K.; writing—review and editing, J.X., Y.M., Y.F., Y.Z., M.X. and W.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Zhejiang Provincial Natural Science Foundation of China under grant No. LQN25F030015.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This research utilizes publicly available data for training and evaluation. The GSV-Cities dataset is available at https://www.kaggle.com/datasets/amaralibey/gsv-cities (accessed on 4 January 2025) , the Pitts30k dataset is available at https://github.com/Relja/netvlad/issues/37 (accessed on 4 January 2025), the Tokyo247 dataset is available at http://www.ok.ctrl.titech.ac.jp/~torii/project/247/ (accessed on 4 January 2025), the Nordland dataset is available at https://github.com/jmfacil/single-view-place-recognition (accessed on 4 January 2025), and the MSLS dataset can be found at https://www.mapillary.com/dataset/places (accessed on 4 January 2025). All the datasets are formatted using the VPR-Datasets-downloader repo, available at https://github.com/gmberton/VPR-datasets-downloader (accessed on 4 January 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VPR	Visual Place Recognition
VFMs	Visual Foundation Models
SALAD	Sinkhorn Algorithm for Locally Aggregated Descriptors
SURF	Speeded-Up Robust Features
VLAD	Vector of Locally Aggregated Descriptors
ORB	Oriented FAST and Rotated BRIEF
BoW	Bag of Words
CNNs	Convolutional Neural Networks
ViT	Vision Transformer
MLP	Multi-Layer Perceptron
GeM	Generalized Mean Pooling
DiNOv2	Distillation with NO labels (Version 2)
3D	Three-dimensional
GPU	Graphical processing unit
PETL	Parameter-efficient transfer learning
NLP	Natural Language Processing
LoRA	Low-Rank Adaptation
CLIP	Contrastive Language–Image Pre-training
MHA	Multi-head attention
LN	Layer normalization
SoTA	State-of-the-art
PCA	Principal Component Analysis
FLOPs	Floating Point Operations
MACs	Multiply–Accumulate Operations
GSV-Cities	Google Street View-Cities
MSLS	Mapillary Street-Level Sequences
MS	Multi-Similarity
OHEM	Online Hard Example Mining
SF-XL	San Francisco eXtra Large Dataset
DSConv	Depthwise Separable Convolution

References

1. Masone, C.; Caputo, B. A survey on deep visual place recognition. *IEEE Access* **2021**, *9*, 19516–19547. [CrossRef]
2. Schubert, S.; Neubert, P.; Garg, S.; Milford, M.; Fischer, T. Visual place recognition: A tutorial. *arXiv* **2023**, arXiv:2303.03281.
3. Chen, Z.; Jacobson, A.; Sünderhauf, N.; Upcroft, B.; Liu, L.; Shen, C.; Reid, I.; Milford, M. Deep learning features at scale for visual place recognition. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 3223–3230.

4. Middelberg, S.; Sattler, T.; Untzelmann, O.; Kobbelt, L. Scalable 6-dof localization on mobile devices. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part II 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 268–283.
5. Qian, J.; Cheng, Y.; Ying, R.; Liu, P. A Novel Indoor Localization Method Based on Image Retrieval and Dead Reckoning. *Appl. Sci.* **2020**, *10*, 3803. [CrossRef]
6. Peng, Z.; Song, R.; Yang, H.; Li, Y.; Lin, J.; Xiao, Z.; Yan, B. Adaptive Feature Refinement and Weighted Similarity for Deep Loop Closure Detection in Appearance Variation. *Appl. Sci.* **2024**, *14*, 6276. [CrossRef]
7. Garg, S.; Sünderhauf, N.; Milford, M. LoST? Appearance-Invariant Place Recognition for Opposite Viewpoints using Visual Semantics. In Proceedings of the Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, PA, USA, 26–30 June 2018. [CrossRef]
8. Ming, Y.; Yang, X.; Calway, A. Object-Augmented RGB-D SLAM for Wide-Disparity Relocalisation. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 2203–2209. [CrossRef]
9. Berton, G.; Trivigno, G.; Caputo, B.; Masone, C. EigenPlaces: Training Viewpoint Robust Models for Visual Place Recognition. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 11046–11056. [CrossRef]
10. Ming, Y.; Yang, X.; Zhang, G.; Calway, A. CGiS-Net: Aggregating Colour, Geometry and Implicit Semantic Features for Indoor Place Recognition. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; pp. 6991–6997. [CrossRef]
11. Ming, Y.; Ma, J.; Yang, X.; Dai, W.; Peng, Y.; Kong, W. AEGIS-Net: Attention-Guided Multi-Level Feature Aggregation for Indoor Place Recognition. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 4030–4034. [CrossRef]
12. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]
13. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
14. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571. [CrossRef]
15. Angeli, A.; Filliat, D.; Doncieux, S.; Meyer, J.A. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Trans. Robot.* **2008**, *24*, 1027–1037. [CrossRef]
16. A, V.; Garg, H.; Anand, A.; Nigam, R.; Gupta, A.; Murthy, K.N.B.; Natarajan, S. Aggregation of Deep Local Features using VLAD and Classification using R2 Forest. *Procedia Comput. Sci.* **2018**, *143*, 998–1006. . [CrossRef]
17. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
18. Keetha, N.; Mishra, A.; Karhade, J.; Jatavallabhula, K.M.; Scherer, S.; Krishna, M.; Garg, S. AnyLoc: Towards Universal Visual Place Recognition. *IEEE Robot. Autom. Lett.* **2024**, *9*, 1286–1293. [CrossRef]
19. Lu, F.; Lan, X.; Zhang, L.; Jiang, D.; Wang, Y.; Yuan, C. CricaVPR: Cross-Image Correlation-Aware Representation Learning for Visual Place Recognition. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16772–16782. [CrossRef]
20. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 5297–5307.
21. 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, Waikoloa, HI, USA, 2–7 January 2023; pp. 2998–3007.
22. Radenović, F.; Tolias, G.; Chum, O. Fine-tuning CNN image retrieval with no human annotation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1655–1668. [CrossRef]
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, Seattle, WA, USA, 16–22 June 2024; pp. 17658–17668.
24. 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 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 14136–14147. [CrossRef]
25. Guan, P.; Cao, Z.; Fan, S.; Yang, Y.; Yu, J.; Wang, S. Hardness-Aware Metric Learning With Cluster-Guided Attention for Visual Place Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 367–379. [CrossRef]
26. Ming, Y.; Xu, M.; Yang, X.; Ye, W.; Wang, W.; Peng, Y.; Dai, W.; Kong, W. VIPeR: Visual Incremental Place Recognition With Adaptive Mining and Continual Learning. *IEEE Robot. Autom. Lett.* **2025**, *10*, 3038–3045. [CrossRef]

27. Yang, X.; Ming, Y.; Cui, Z.; Calway, A. FD-SLAM: 3-D Reconstruction Using Features and Dense Matching. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; pp. 8040–8046. [CrossRef]
28. Oh, J.; Eoh, G. Variational Bayesian Approach to Condition-Invariant Feature Extraction for Visual Place Recognition. *Appl. Sci.* **2021**, *11*, 8976. [CrossRef]
29. Ye, M.; Tanaka, K. Visual Place Recognition of Robots via Global Features of Scan-Context Descriptors with Dictionary-Based Coding. *Appl. Sci.* **2023**, *13*, 9040. [CrossRef]
30. Chen, B.; Song, X.; Shen, H.; Lu, T. Hierarchical Visual Place Recognition Based on Semantic-Aggregation. *Appl. Sci.* **2021**, *11*, 9540. [CrossRef]
31. Berton, G.; Masone, C.; Caputo, B. Rethinking Visual Geo-localization for Large-Scale Applications. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4868–4878. [CrossRef]
32. 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* **2023**, arXiv:2304.07193.
33. Huang, G.; Zhou, Y.; Hu, X.; Zhang, C.; Zhao, L.; Gan, W. DINO-Mix enhancing visual place recognition with foundational vision model and feature mixing. *Sci. Rep.* **2024**, *14*, 22100. [CrossRef]
34. Lu, F.; Zhang, L.; Lan, X.; Dong, S.; Wang, Y.; Yuan, C. Towards Seamless Adaptation of Pre-trained Models for Visual Place Recognition. In Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
35. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv* **2021**, arXiv:2106.09685.
36. He, J.; Zhou, C.; Ma, X.; Berg-Kirkpatrick, T.; Neubig, G. Towards a unified view of parameter-efficient transfer learning. *arXiv* **2021**, arXiv:2110.04366.
37. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 2790–2799.
38. Lester, B.; Al-Rfou, R.; Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv* **2021**, arXiv:2104.08691.
39. Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; Luo, P. Adaptformer: Adapting vision transformers for scalable visual recognition. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16664–16678.
40. Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; Bai, X. Side adapter network for open-vocabulary semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2945–2954.
41. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
43. Ali-bey, A.; Chaib-draa, B.; Giguère, P. GSV-Cities: Toward appropriate supervised visual place recognition. *Neurocomputing* **2022**, *513*, 194–203. [CrossRef]
44. Torii, A.; Sivic, J.; Pajdla, T.; Okutomi, M. Visual Place Recognition with Repetitive Structures. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 883–890. [CrossRef]
45. Torii, A.; Arandjelović, R.; Sivic, J.; Okutomi, M.; Pajdla, T. 24/7 Place Recognition by View Synthesis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 257–271. [CrossRef] [PubMed]
46. Olid, D.; Fácil, J.M.; Civera, J. Single-View Place Recognition under Seasonal Changes. *arXiv* **2018**, arXiv:1808.06516.
47. Warburg, F.; Hauberg, S.; López-Antequera, M.; Gargallo, P.; Kuang, Y.; Civera, J. Mapillary Street-Level Sequences: A Dataset for Lifelong Place Recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 2623–2632. [CrossRef]
48. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
49. Wang, R.; Shen, Y.; Zuo, W.; Zhou, S.; Zheng, N. TransVPR: Transformer-Based Place Recognition with Multi-Level Attention Aggregation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 13638–13647. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.