Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062834

ARTICLE

CerfeVPR: Cross-Environment Robust Feature Enhancement for Visual Place Recognition

Lingyun Xiang[1], **Hang Fu**[1] **and Chunfang Yang**[2,*]

[1]School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha, 410114, China
[2]Key Laboratory of Cyberspace Situation Awareness of Henan Province, Information Engineering University, Zhengzhou, 450001, China

*Corresponding Author: Chunfang Yang. Email: chunfangyang@126.com

Received: 28 December 2024; Accepted: 11 April 2025; Published: 09 June 2025

ABSTRACT: In the Visual Place Recognition (VPR) task, existing research has leveraged large-scale pre-trained models to improve the performance of place recognition. However, when there are significant environmental differences between query images and reference images, a large number of ineffective local features will interfere with the extraction of key landmark features, leading to the retrieval of visually similar but geographically different images. To address this perceptual aliasing problem caused by environmental condition changes, we propose a novel Visual Place Recognition method with Cross-Environment Robust Feature Enhancement (CerfeVPR). This method uses the GAN network to generate similar images of the original images under different environmental conditions, thereby enhancing the learning of robust features of the original images. This enables the global descriptor to effectively ignore appearance changes caused by environmental factors such as seasons and lighting, showing better place recognition accuracy than other methods. Meanwhile, we introduce a large kernel convolution adapter to fine tune the pre-trained model, obtaining a better image feature representation for subsequent robust feature learning. Then, we process the information of different local regions in the general features through a 3-layer pyramid scene parsing network and fuse it with a tag that retains global information to construct a multi-dimensional image feature representation. Based on this, we use the fused features of similar images to drive the robust feature learning of the original images and complete the feature matching between query images and retrieved images. Experiments on multiple commonly used datasets show that our method exhibits excellent performance. On average, CerfeVPR achieves the highest results, with all Recall@N values exceeding 90%. In particular, on the highly challenging Nordland dataset, the R@1 metric is improved by 4.6%, significantly outperforming other methods, which fully verifies the superiority of CerfeVPR in visual place recognition under complex environments.

KEYWORDS: Visual place recognition; cross-environment robustness; pre-trained model; feature learning

1 Introduction

With rapid advancements in fields such as panoramic navigation, autonomous driving, and virtual reality, the demand for precise recognition and localization of specific locations within extensive and diverse image datasets has become increasingly prominent. Consequently, Visual Place Recognition (VPR), also referred as visual geo-localization [1], has attracted increasingly widespread attention. It is a task that searches for the best matching result from an image database marked with geographical locations for a query image to identify the location described in that image, and it is usually regarded as an image retrieval problem [2]. Since each image in the database is associated with a geographical identifier, such as a place name or GPS

coordinates [3], the geographical location of a query image can be determined by representing the images with global or local descriptors, thereby enabling effective localization through result retrieval.

Traditional VPR methods typically rely on hand-crafted algorithms to extract local descriptors, which are subsequently aggregated into global descriptors using feature encoding techniques such as VLAD [4]. These global descriptors are then utilized to perform image-level queries through indexing mechanisms. With the significant success of deep learning, studies have shown that image features automatically learned by deep neural networks far surpass handcrafted local features [2,5–7], exhibiting greater robustness to environmental changes. This has led to the emergence of a series of VPR methods that utilize Convolutional Neural Networks (CNNs) or Transformers for local feature extraction in place recognition tasks [8–11]. Following the extraction of local features, these methods still necessitate the application of aggregation techniques, such as NetVLAD [12] or pooling methods like generalized mean pooling (GeM [13]) to generate global descriptors, which are subsequently utilized for place recognition via global descriptor retrieval.

In recent years, large-scale pre-trained models have achieved remarkable success in various downstream tasks, giving rise to a new technical paradigm for visual place recognition (VPR). These models, known for their strong representational capabilities, are used as backbones to extract local features, which are then aggregated into global descriptors for retrieval, enabling place recognition in images. For example, AnyLoc [14] leverages the pre-trained DINOv2 model as its backbone network, achieving significantly superior performance and efficiency compared to traditional VPR methods based on CNNs or Transformers. However, utilizing DINOv2 without fine-tuning may capture a considerable amount of irrelevant dynamic elements, such as pedestrians or vehicles. To mitigate this issue, SALAD [15] fine-tunes the pre-trained DINOv2 model on a VPR-specific dataset, concentrating on critical VPR features including architectural structures and scene layouts. Despite these advancements, full fine-tuning remains resource-intensive and poses the risk of catastrophic forgetting, which may undermine previously learned features. To overcome these limitations, CricaVPR [16] proposes the application of a parameter-efficient fine-tuning technique [17] to the pre-trained model, keeping the base model parameters frozen while integrating trainable adapters [18] into the base architecture, thereby enhancing the pre-trained model's adaptability to VPR tasks and bridging the gap between the pre-trained model and VPR task.

Although methods based on large-scale pre-trained models enhance VPR performance, they face several technical gaps in real-world location with diverse environmental conditions. On the one hand, in terms of feature extraction, existing models, even with fine-tuning, often struggle to disentangle the environmental-invariant features from the environmental-specific ones. For example, the pre-trained models may capture features that are too sensitive to lighting or weather changes, which are not essential for place recognition. This leads to sub-optimal performance when the query and reference images are taken under different environmental conditions. On the other hand, the training datasets used in current methods are limited. They rarely cover all possible environmental variations, such as different seasons, weather conditions, lighting, angles, and times [19]. As a result, the models trained on these datasets lack generalization ability to unseen environmental conditions. When faced with real-world scenarios where the environmental differences between query and reference images are significant, the models may misclassify or fail to recognize the correct location, as illustrated in Columns 1 & 2 of Fig. 1.

To overcome this, we propose generating environment-specific variations of query and reference images, and leveraging their commonality to learn robust features for the original images. This mitigates the impact of environmental changes on feature learning. Consequently, the robust features enable the visual place recognition framework to retain strong spatial relationship analysis capabilities under drastically changing environmental conditions, thereby improving place recognition accuracy. Building on this foundation, we propose CerfeVPR, a method specifically designed to enhance cross-environment robustness

and improve matching precision for images under diverse conditions. First, CerfeVPR employs Generative Adversarial Networks (GAN) [20] to generate similar images under different environments for the original images, which are processed by the pre-trained DINOv2 model to extract general feature maps. To help DINOv2 better focus on global features and reduce interference from less discriminative local features, a novel large-kernel adapter is introduced to fine-tune DINOv2. Next, CerfeVPR's 3-layer pyramid scene parsing network further refines and fuses local and global features for comprehensive representation. Finally, a Transformer-based robust feature learning network processes the fused features of original and synthesized images to corporately learn robust features for orignal images [21], significantly enhancing perceptual capabilities and resilience to environmental variations.

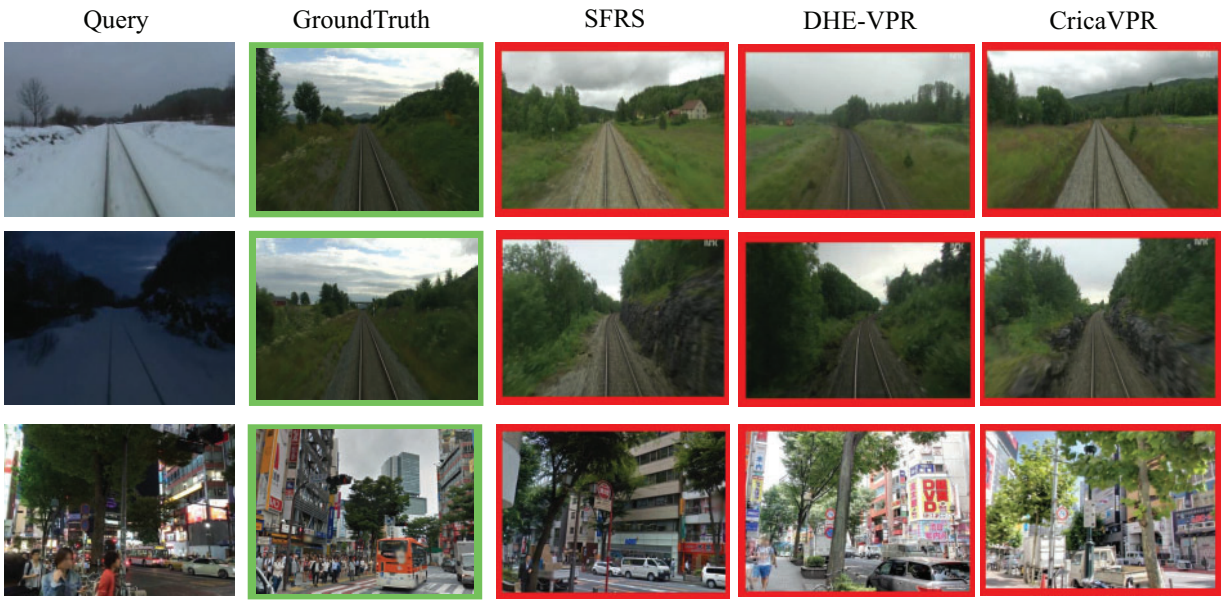

Figure 1: Examples of challenging queries with significant environmental differences from their corresponding GroundTruth, causing existing VPR methods to fail in correctly recognizing places

Extensive experiments conducted on five widely used datasets demonstrate that CerfeVPR outperforms existing benchmark methods for visual place recognition.

The main contributions are as follows:

1) We propose a novel visual place recognition method CerfeVPR that leverages GAN-generated similar variants of an image to enhance the robustness of its global descriptor. By focusing on invariant landmark features, CerfeVPR effectively overcomes the challenges posed by diverse environmental conditions and discrepancies between query and reference images of the same location.
2) We introduce a large-kernel adapter to fine-tune the pre-trained DINOv2 model, enabling it to learn more effective general representations for images. To further optimize this representation for the visual place recognition task, its local and global information are further fined and fused to produce a comprehensive feature representation.
3) Experimental results demonstrate that the proposed method surpasses existing baseline methods on multiple public datasets. Notably, it achieves a significant performance improvement of up to 4.6% in Recall@1 on datasets, where references images share similar environmental conditions with the generated variants of query images.

The rest of the paper is organized as follows. Section 2 reviews the related work in visual place recognition. Section 3 presents our proposed CerfeVPR method, detailing its design and components. Section 4 provides a comparative analysis of different approaches. Finally, we conclude this paper in Section 5.

2 Related Work

2.1 Visual Place Recognition Based on Convolutional Neural Networks

Chen et al. [6] were the first to employ Convolutional Neural Networks (CNNs) for VPR, demonstrating that features extracted by CNNs can achieve robust performance under varying environments. Sunderhauf et al. [5] evaluated the performance of features from each layer of the neural network and found that features extracted from the third convolutional layer exhibited high robustness to changes in environmental appearance. Arandjelović et al. [12] designed the NetVLAD layer, which not only allows for end-to-end training but can also be applied to any CNN architecture. However, NetVLAD directly aggregates global features without considering more granular features. To remedy this limitation, Patch-NetVLAD [10] extracts locally-global descriptors by incorporating patch-level features from the feature space, effectively combining the advantages of both local and global descriptor methods, thereby achieving superior performance. In addition, the optimization of training strategies has garnered significant attention. For instance, SFRS [22] achieved noticeable improvements through self-supervised iterative training. CosPlace [1] reformulated the training as a classification problem, eliminating the necessity for hard negative mining. Additionally, MixVPR [8] innovatively employs a fully MLP-based aggregation technique, iteratively incorporating global relationships into each feature map extracted from the pre-trained backbone, resulting in outstanding performance.

Table 1: Advantages and disadvantages of recent representative works in the VPR field

Year	Method	Advantages	Disadvantages
2014	Chen [6]	It pioneered the use of pre-trained CNN for VPR.	When faced with severe enviromental changes, the model's performance drops.
2018	NetVLAD [12]	It can be trained end-to-end and is applicable to any CNN architecture.	It aggregates the global features directly without considering detailed features.
2020	SFRS [22]	It achieved significant improvement through self-supervised iterative training.	There is high computational complexity and poor robustness to lighting changes.
2021	Patch-NetVLAD [10]	It combines the advantages of local and global descriptor methods and can capture more detailed spatial information.	This algorithm is sensitive to image noise and has relatively high memory consumption.
2022	CosPlace [23]	It transforms the training process into a classification problem and eliminates the need for mining negative examples.	This algorithm is less accurate in low-texture environments.
2022	TransVPR [11]	It can adaptively extract robust image representations of the image, and re-order through the RANSAC algorithm to improve accuracy.	This model has high computational cost in some complex and variable environments.

(Continued)

Table 1 (continued)

Year	Method	Advantages	Disadvantages
2023	MixVPR [8]	It uses full MLP aggregation to iteratively incorporate global relationships into the feature maps. This method can effectively capture the long-range dependencies in the image.	This algorithm has relatively low recognition accuracy for scenes with drastic changes in illumination and perspective.
2023	R2former [24]	It adopts Transformer to represent deep features, and considers multiple factors in the re-ordering module.	This model has high running time and memory consumption.
2023	AnyLoc [14]	It adopts the large pre-trained model DINOv2 combined with unsupervised aggregation methods. The use of DINOv2 can leverage the pre-learned knowledge from large-scale data.	It is sensitive to dynamic objects during image representation generation and overlooks static discriminative backgrounds.
2024	SALAD [15]	It also utilizes the DINOv2 to enhance the performance of VPR.	Environmental interferences may impact the model's accuracy.
2024	DHE-VPR [25]	It constructs a compact convolutional Transformer backbone network to extract dense feature maps and determines image similarity through regression analysis.	This algorithm is prone to misrecognition in scenes with similar global features.
2024	PlaceFormer [26]	It generates multi-scale patches, uses the self-attention mechanism to select relevant patches for geometric verification and re-ordering.	The model shows poor robustness in the face of dramatic changes in the environment.
2024	SelaVPR [27]	It adds an adapter module to the pre-trained model framework, bridges the gap between pre-training and VPR tasks, and improves the performance of the base model.	The model has a relatively complex model structure that requires more computing resources and longer training time.
2024	CricaVPR [16]	The cross-image correlation aware encoder in CricaVPR enables the model to comprehensively capture and analyze correlations across different images.	The model has relatively high GPU requirements due to the additional module.

2.2 Visual Place Recognition Based on Transformer

With the remarkable success of Transformer in computer vision, several representative works have emerged based on visual Transformer. For example, TransVPR [11] utilizes an adaptive strategy to extract robust image representations from different regions of an image, which are subsequently utilized as global features for candidate sequence retrieval. Then, the RANSAC algorithm is employed for spatial matching to refine the ranking. R2former [24] employs Transformer to enhance image location representation by

incorporating depth features, in the image re-ranking module, the characteristics of features, attention values, and xy coordinates are factored in to determine whether two images originate from the same location. The DHE-VPR [25] employs the dense feature map extracted by the backbone network as input, and then performs a regression analysis to achieve a single response matrix that assesses the internal similarity between matching local features. PlaceFormer [26] initially generates multiple scale blocks and then employs self-attention mechanisms to select relevant blocks for geometric verification based on the task at hand. Subsequently, a similarity score is computed for resorting. Currently, transformer-based methods predominantly follow a two-stage process that involves two rounds of ranking to improve the precision and robustness of localization. However, this method often incurs substantial computational and memory overhead.

2.3 Visual Place Recognition Based on a Large Pre-Trained Model

Large Vision Models (LVM) can produce powerful feature representations, outperforming many existing models in general visual localization, which brings both new challenges and opportunities for VPR. The use of large visual pre-trained models in place recognition tasks has become increasingly popular, with examples like AnyLoc [14] and SALAD [15] both using the large pre-trained model DINOv2 [28] as their feature extraction backbone, combined with unsupervised aggregation methods to extract image pixel features, significantly improving the accuracy and efficiency of VPR.

While using pre-trained models can provide powerful general learning representation capabilities on large-scale datasets, directly applying them to VPR tasks still has limitations, particularly in being susceptible to dynamic objects when generating image representations and tending to ignore some static discriminative backgrounds (such as buildings and vegetation). SelaVPR [27] and CricaVPR [16] bridge the gap between model pre-training and VPR tasks by adding adapter modules to the pre-trained model framework and then training or fine-tuning on VPR datasets, greatly improving the performance of foundation models in VPR tasks.

We have comprehensively summarized the advantages and disadvantages of the methods introduced in the above-mentioned related works and systematically organized them into Table 1.

3 Proposed Method

In this section, we introduce CerfeVPR, a novel method designed to mitigate the environmental impact on feature representations of place images by leveraging their similar variants to learn inherently robust and invariant feature representations. CerfeVPR begins with a well-designed global descriptor learning process to generate robust global descriptors for each query and reference image. Then, we utilize cosine similarity to estimate the similarity between the global descriptors of a query and a reference image. The location of the query image is approximately determined by identifying the most similar global descriptor in the retrieval database. This method significantly improves the matching accuracy of global descriptors, thereby enhancing overall place recognition performance. The core of CerfeVPR lies in its global descriptor learning process. As illustrated in Fig. 2, this process is mainly composed of the following four modules:

1. Similar Image Generation: Using the collaborative efforts of generators and discriminators in the GAN model, this module generates new similar images that retain the original key landmark features while incorporating new environmental conditions. These generated images simulate the same location under varying environmental conditions, enriching the information for robust feature learning.

2. Adapted ViT Backbone Network: It leverages the large-scale pre-trained Vision Transformer(ViT), DINOv2, as the backbone network to learn general feature representation for each image. To effectively

extract spatial information from the image's sub-regions at various scales, a novel large-kernel adapter is integrated, enabling efficient fine-tuning of DINOv2. This adaptation significantly enhances the model's ability to learn discriminative general feature representations.

3. Local-Global Feature Fusion: This module separates global and multi-level local features from the feature representations learned by the fine-tuned DINOv2. It utilizes a 3-layer pyramid scene parsing network to capture contextual relationships across various scales and sub-regions. An adaptive average pooling (AAP) algorithm is then applied to reduce redundant local information. Finally, it integrates the global and multi-scale local features to create a compact global representation.

4. Robust Feature Learning: This module utilizes a Transformer-based architecture to construct a robust feature learning network, designed to generate a reliable global descriptor for each original query or reference image. By taking the fused features of an original image and its GAN-generated similar variants as input, the network captures shared features among images of the same location under varying environmental conditions. This process facilitates the learning of discriminative landmark features as the global descriptor, ensuring robust and accurate place recognition despite environmental changes.

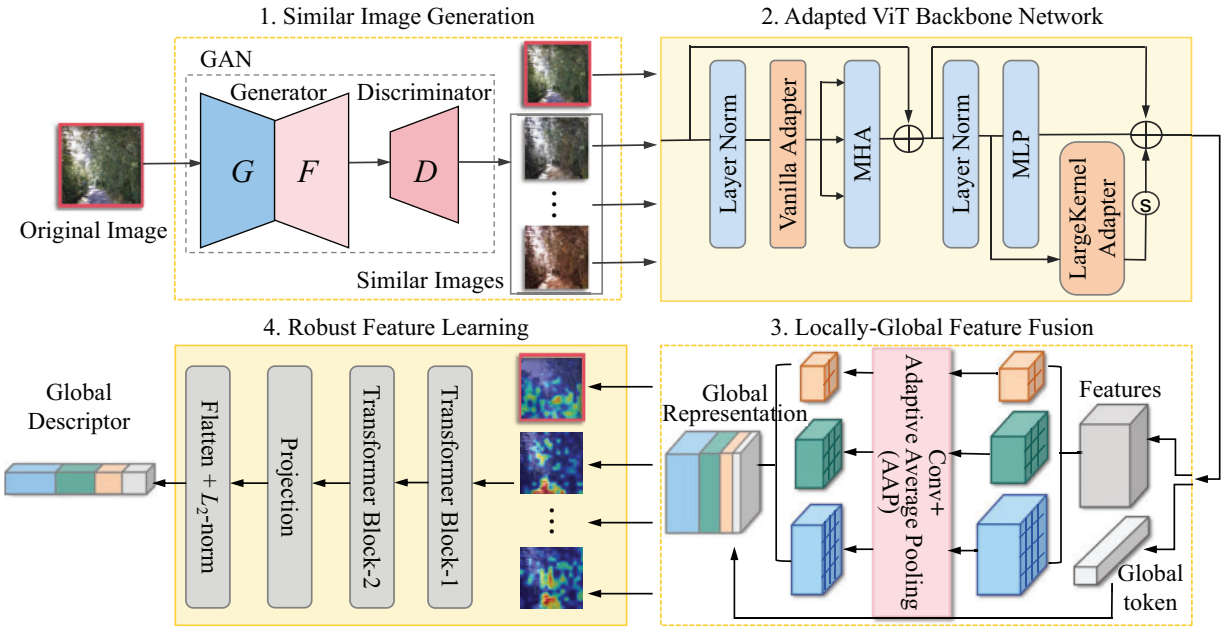

Figure 2: Overview of the robust global descriptor learning pipeline in the proposed CerfeVPR. To enhance the learning of robust features across different environments, it initiates by generating similar images under diverse environmental conditions. Subsequently, utilizing shared inherent features of these images, it learns robust feature representations, reducing environmental impacts on extraction and recognition

3.1 Similar Image Generation

Since a query or reference image in the dataset can only represent the state of a location at a specific moment, this state is heavily influenced by the environment. This variability makes it challenging to accurately match query images with their corresponding reference images under different environmental conditions. To overcome this, we propose generating similar images under diverse environmental conditions, which are utilized to facilitate the learning of robust feature representations for query and reference images.

Inspired by existing neural network models [29–31], we adopt the CycleGAN [30] model to generate high-quality images that are very similar to the original images, which transfers the environmental type while keeping the content unchanged.

Fig. 3 shows some examples of similar images generated by the CycleGAN.

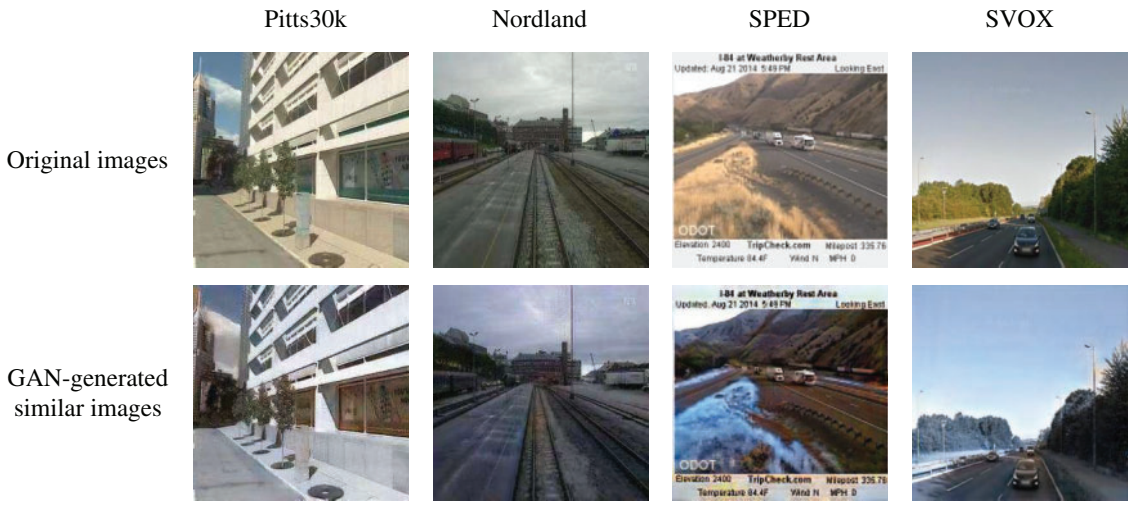

Figure 3: A simple illustration of similar images generated using the CycleGAN. The displayed generated images have the source domain as summer and the target domain as winter in different datasets

The cycle consistency loss mechanism of CycleGAN enables it to generate paired images in the target domain (i.e., under different environmental conditions) while maintaining a high level of visual similarity to real-world images. This characteristic is particularly crucial for our research. Therefore we adopt the CycleGAN model to transfer environmental types while maintaining content unchanged, generate similar variants under specified environmental conditions for original query images or reference images. CycleGAN employs two generators and two discriminators to learn mapping functions between two different domains M and N, where images from these domains are captured in varying environments (such as winter vs. summer, day vs. night). Specifically, for a source domain image $m \in M$, an image generator $G : M \to N$ is trained to generate a corresponding target domain image n', such that $G(m) = n'$. Conversely, a reverse generator $F : N \to M$ is trained to generate source domain image m' corresponding to the target domain image $n \in N$, where $F(n) = m'$. Meanwhile, two discriminators, D_M and D_N, are trained to distinguish between real images and generated images within their respective domains. Specifically, D_M differentiates between real images $\{m\}$ from domain M and generated images $\{F(n)\}$ produced by the generator F. Similarly, D_N distinguishes between real images $\{n\}$ and generated images $\{G(m)\}$ produced by the generator G. These discriminators are trained to maximize their ability to correctly classify real and generated images, while the generators aim to produce images that can deceive the discriminators, thereby improving the quality of the generated images.

The objective function combines two components: adversarial loss L_{GAN} and cycle consistency loss L_{cyc}, which are jointly optimized to enable high-quality image generation. The overall objective is defined as:

$$L(G, F, D_M, D_N) = L_{GAN}(G, D, M, N) + L_{GAN}(F, D, N, M) + \lambda L_{cyc}(G, F), \tag{1}$$

where λ controls the relative importance of the adversarial and cycle consistency losses. The adversarial loss L_{GAN} is designed to ensure that the generated images are indistinguishable from real images within their

respective target domains. It is expressed as:

$$L_{GAN}(G, D_N, M, N) = E_{n\sim P_{data(n)}}[\log D_N(n)] + E_{m\sim P_{data(m)}}[\log(1 - D_N(G(m)))],$$
$$L_{GAN}(F, D_M, N, M) = E_{m\sim P_{data(m)}}[\log D_M(m)] + E_{n\sim P_{data(n)}}[\log(1 - D_M(G(n)))], \quad (2)$$

where $E_{m\sim P_{data(m)}}$ and $E_{n\sim P_{data(n)}}$ represent the expectation over real images in domains M and N, respectively.

Additionally, the cycle consistency loss L_{cyc} ensures that an image translated to the target domain can be accurately reconstructed back to its original form in the source domain. This loss guarantees the accuracy of the mapping function and enhances the quality of generated images. It is formulated as:

$$L_{cyc}(G, F) = E_{m\sim P_{data(m)}}[\| F(G(m)) - m\|_1] + E_{n\sim P_{data(n)}}[\| F(G(n)) - n\|_1], \quad (3)$$

where $\|\cdot\|_1$ represents the $L1$ norm, which is used to calculate the distance between two image representations.

To handle diverse environmental conditions, we train CycleGAN using various datasets representing a set of common conditions: $Ec = \{c_0, ec_1, \ldots, ec_{r-1}\}$, where r is the number of environmental conditions (e.g., spring, summer, night, $\cdots$). For an arbitrary query or reference image I, different trained models are applied under these conditions to produce r similar images $I' = \{I'_0, I'_1, \cdots, I'_{r-1}\}$, each preserving the original structural information while reflecting different environmental variations. The collection of the original image I and similar images is denoted as $\tilde{I} = \{I'_0, I'_1, \cdots, I'_r\}$, where $I'_r = I$.

3.2 Backbone with Large Kernel Adapters

To achieve robust place recognition, efficient image representations must be learned. Considering that the latest VPR models AnyLoc [14] and SALAD [15] have demonstrated that pre-trained large-scale model DINOv2 possesses powerful representation capability and can effectively enhance VPR task performance, we also adopt DINOv2 to learn universal feature maps from original images and their similar images.

As shown in Fig. 4a, DINOv2 utilizes a Vision Transformer (ViT) as its backbone. However, directly applying DINOv2 to downstream tasks underutilizes its potential, while full-volume fine-tuning on a VPR dataset imposes significant computational and storage demands. Therefore, during the feature learning process, most layers of the model are typically frozen, with only some parameters being efficiently fine-tuned. For example, SelaVPR [27] adapts the pre-trained model to VPR tasks by adding an efficient fine-tuning adapter called Vanilla adapter, as shown in Fig. 4b. Vanilla adapter has a simple structure. It uses fully connected layers to project the input down to a lower dimension, activates the reduced representation with a non-linear activation function ReLU, and then projects it back up to the original dimension. This design efficiently captures VPR task-specific knowledge while ensuring computational efficiency.

To capture features that differ at the visual perception level but remain consistent in visual cognition, we propose a new adapter: LargeKernelAdapter, whose structure is illustrated in Fig. 4c. This adapter leverages multi-scale convolution kernels to capture cognitive-level visual features. It first employs the down-projection layer W_{down} to reduce the input's dimension. A 1×1 convolution is then applied to compress the features and provide an appropriate number of channels for subsequent operations. Next, LargeKernelAdapter constructs three parallel convolution paths with convolution kernels of different scales (3×3, 5×5, and 7×7) to capture multi-scale features [32]. The outputs of these paths are concatenated and combined with the residual connection after downsampling, ensuring efficient reuse of the original features and minimizing feature loss during convolution. This integration enables residual transfer while achieving multi-scale feature fusion.

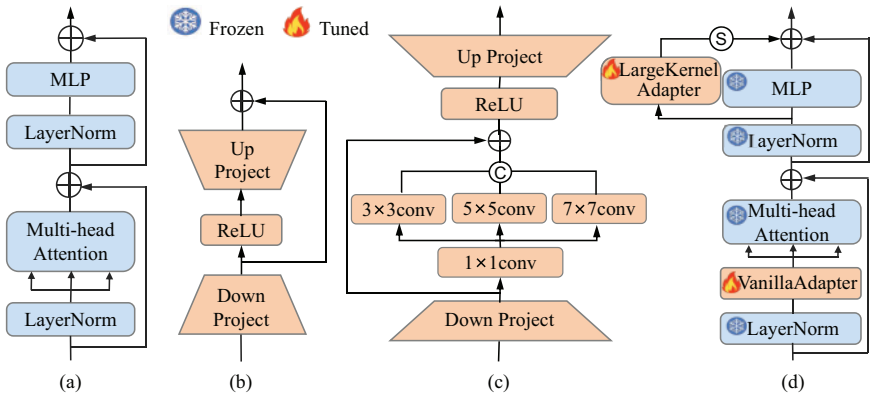

Figure 4: Illustration of the adapters and backbone network optimization. **(a)** Transformer block in ViT, **(b)** Vanilla adapter, **(c)** Proposed LargeKernel adapter, **(d)** Transformer with Vanilla and LargeKernel adapters

Finally, the processed features pass through ReLU, and the up-projection layer W_{up} projects the features back to the original dimension. The entire process of LargeKernelAdapter is expressed as following:

$$LargeKernelAdapter = ReLU(Conv(X' \cdot W_{down})) \cdot W_{up}, \tag{4}$$

where ReLU is the activation function, X' represents the features after Layer Normalization of the output of the Multi-Head Attention (MHA) layer. As depicted in the structure shown in Fig. 4c, the $Conv(\cdot)$ operation represents a series of convolutional operations. It encompasses a residual connection and a 1×1 convolution, along with three parallel convolutional paths with kernel sizes of 3×3, 5×5, and 7×7, respectively. Unlike SelaVPR [27], which cascades the adapter after MHA, we integrates VanillaAdapter in series between LayerNorm and MHA to refine feature extraction, while connecting the LargeKernelAdapter in parallel with the Multi-Layer Perceptron (MLP) layer. This design, depicted in Fig. 4d, constructs a new Transformer module that combines serial and parallel adapters, which does not disrupt the features of the original independent branches, better preserving and utilizing the data information from the original branches. By incorporating these lightweight adapters into the Transformer, the optimized backbone network DINOv2 enhances its feature representation capability with only a small number of additional parameters, allowing it to focus more on discriminative landmark feature representations for the VPR task.

The optimized backbone DINOv2 is employed to learn the general feature map of each query, reference image and their generated similar images. An image is first divided into N patches of size $p \times p$, and then transformed into D-dimensional vector representation $x_1, \ldots, x_N$ via linear projection, and positional embeddings P is added to retain spatial information. To capture the global semantic information of the entire patches, a learnable class token x_{cls} is prepended, forming the serialized input representation $X_0 = [x_{cls}, x_1, \ldots, x_N] + P$.

The sequence X_0 is processed through the above optimized Transformer, which integrates the LargeKernelAdapter for multi-scale attention computation and feature learning. The output from the prior layer first undergoes LayerNorm, followed by the VanillaAdapter and MHA layer, as described in Eq. (5):

$$X_t' = MHA(VanillaAdapter(LN(X_t - 1))), \tag{5}$$

where X_{t-1} represents the output of the $(t-1)$-th layer of the Transformer, and $X_t^{'}$ denotes the output of the MHA in the current layer. Subsequently, $X_t^{'}$ is normalized and processed by two parallel branches:

LargeKernelAdapter and the MLP layer. The features from these branches are fused via residual connections, with the fully connected feedforward layer incorporating the LargeKernelAdapter expressed as:

$$X_t = MLP(LN(X_t^{'})) + s \cdot LargeKernelAdapter(LN(X_t^{'})) + X_t^{'}, \tag{6}$$

where s is a scaling factor. Finally, the output of the last Transformer is used as the extracted feature map F, representing the final general feature representation of the input image.

3.3 Local-Global Feature Fusion

Effective representation of VPR scene requires a balanced integration of global and local features. While global features capture the overall semantic context of an image, local features provide the fine-grained details essential for distinguishing between visually similar places. To achieve this balance, we combines global and local information in the learned feature map F to enhance the feature representation. The feature map F is first separated into two components: a global token summarizing image content ($F_c \in \mathbb{R}^{H\times W\times 1}$, derived from the class token x_{cls}), and patch tokens containing local information ($F_p \in \mathbb{R}^{H\times W\times (D-1)}$). Each component is processed through specialized paths to maximize their respective contributions. The architecture of this fusion process is illustrated in Part 3 of Fig. 2.

We introduce a 3-layer pyramid scene parsing network to integrate local contextual information under different scales and subregions. It divides the patch tokens F_p into three different scales (2×2, 3×3, and 4×4, respectively), with each region forming a compact feature representation through feature aggregation algorithm. The aggregation algorithm includes convolution and average pooling, which are used to get compact global feature representations. The three aggregated layers of features will be combined in sequence and, the global information F_c will be concatenated through skip connections at the end, form the global feature representation. In the VPR field, feature aggregation algorithm chosen determines how the model combines local features to describe images and environments. Common feature aggregation methods can be divided into aggregation layers based on feature extractors (such as NetVLAD [12], MixVPR [8], etc.) and unsupervised lightweight pooling layers (such as GeM [13], VLAD [4], etc.), which can effectively perform global feature representation for retrieval. Unlike these methods, we use a combination of 1×1 convolution and adaptive average pooling [23] as the feature aggregation module to pool hierarchical feature maps. Firstly, F_p linearly combines the features of different channels through a 1×1 convolution to realize the information integration across channels. This allows feature maps of different scales to not only retain their original spatial position information but also integrate global position information from different channels. Then, the spatial dimension of F_p is reduced by adaptive average pooling, and the feature map space is divided into sub-regions of the same size as $s_1 \times s_2$, and the average value is taken to perform effective feature aggregation to obtain a feature map of size $s_1 \times s_2 \times d$. Then, F_c containing global information is merged into local features with three levels to obtain the global representation Z, which is formally described as:

$$Z = F_c \oplus AAP_{s_1\times s_2}(Conv_{1\times 1}(F_p)), \tag{7}$$

where $AAP(\cdot)$ represents adaptive average pooling operation and $Conv_{1\times 1}(\cdot)$ represents the 1×1 convolution operation. By leveraging convolution and feature aggregation within this module, this model enhances its ability to sense environmental context and produce comprehensive global feature representations. As a result, compact global features that encode relationships across different scales and sub-regions are obtained, enabling superior recognition performance in visual place analysis.

3.4 Robust Feature Learning

Achieving reliable place recognition in different environments requires feature representations that not only capture general visual patterns but also exhibit robustness across varying conditions. The fused features often lack the resilience necessary to handle perceptual aliasing and cross-environment variability inherent to VPR. Therefore, we construct a robust feature learning network, as shown in Part 4 of Fig. 2, to learn a more robust global descriptor for each query or reference image. It utilizes the attention mechanism in the Transformer block to conduct continuous identification and analysis on the global feature representation sequence $\tilde{Z}$, thereby effectively enhancing the robustness of the global features. $\tilde{Z}$ is generated after the feature extraction and fusion operations on the similar image set $\tilde{I} = \{I'_0, I'_1, \cdots, I'_r\}$, and is represented as:

$$\tilde{Z} = \{Z_i\}, i \in \{0, \cdots, r\}, \tag{8}$$

where Z_i corresponds to the feature representation of the i-th image. This sequence already contains abundant feature information, but to produce the same correct results for different transformations of images from the same location, we need to further focus on features that remain unchanged despite varying conditions. Thus, we feed global representation sequence composed of images from the same location under different conditions into a Transformer encoder, calculating feature correlations at the same positions in sequence and learning dependencies of similar images. First, global representation sequence $\tilde{Z}$ is input into Transformer encoder layer for normalization, then feature $Z^{'}$ obtained through the MHA layer and a residual connection:

$$Z^{'} = MHA(LN(\tilde{Z})) + \tilde{Z}. \tag{9}$$

After the MHA layer, feature Z' undergoes normalization and pass through two Linear Layers, with a ReLU activation function applied between the two Linear Layers to help regularize the network, while residual connections are used to preserve information flow. This process can be described as follows:

$$Z'' = Linear(ReLU(Linear(LN(Z^{'})))) + Z^{'}, \tag{10}$$

where $Linear(\cdot)$ is a linear transformation function. The global representation sequence obtains feature representation Z'' after passing through two encoder layers mentioned above. Since a large amount of local feature information is aggregated in the feature fusion part, at this point, Z'' is typically high-dimensional. In order to fully utilize these features with strong robustness across different environments while reducing feature dimension, we use a fully connected layer $W_{proj}(\cdot)$ to control the size of the global descriptor, followed by a $Flatten(\cdot)$ dimensionality reduction and a normalization operation $LN(\cdot)$ to generate the final global descriptor O of original image I:

$$O = LN(Flatten(W_{proj}(Z''))). \tag{11}$$

This structural setup allows the robust feature learning module to not only focus on features of individual images, but also comprehensively consider the feature dependencies between the original image and all similar images. This process results in robust, high-quality global descriptors that facilitate accurate VPR under complex real-world scenarios [33]. By learning highly discriminative global features during training, the model achieves improved place recognition performance even in challenging environments.

3.5 Model Training

During model training, to reduce interference from similar yet location-different negative samples to query images, we adopt the online hard example mining (OHEM) [34] algorithm. In each iteration, it

prioritizes samples with larger loss functions (i.e., hard examples hard for the neural network to distinguish). By mining and integrating these representative samples into the training set, the model can focus on difficult samples and strengthen its robust feature representation for negative samples. Moreover, efficiently and accurately adopt and weight important sample pairs, can be expressed as: OHEM addresses the issue of excessive simple samples in the dataset by eliminating easily distinguishable examples based on loss value rankings. This not only reduces computational overhead but also enables efficient training on large-scale datasets. We employ Multiple Similarity Loss (MS-Loss) [35] for computation as it has proven to perform best in VPR [23]. This loss comprehensively considers self-similarity and relative similarity, optimizing the weighting of sample pairs to more efficiently and accurately adopt and weight important sample pairs:

$$\mathscr{L}_{MS} = \frac{1}{B}\sum_{q=1}^{B}\left\{\frac{1}{\alpha}\log\left[1+\sum_{p\in P_q} e^{-\alpha\left(S_{qp}-\lambda\right)}\right]+\frac{1}{\beta}\log\left[1+\sum_{p\in N_q} e^{\beta\left(S_{qp}-\lambda\right)}\right]\right\}, \tag{12}$$

where B represents the number of image samples in a batch, and q represents the index of each query image. The index sets of the positive sample pairs and negative sample pairs are P_q and N_q, respectively. S_{qp} is the cosine similarities of positive and negative pairs, α, β and λ are three constant hyperparameters.

4 Experiments and Results

4.1 Training Strategies and Implementation Details

Training was conducted on the GSV-Cities [23] dataset, which contains 560 k images captured at 62 k locations around the world, featuring highly accurate labels, extensive geographical coverage, and highly precise ground truth. All experiments in this work were conducted on two NVIDIA GeForce RTX 3090 GPUs. In the similar image generation module, two environmental conditions, summer and winter, were designated to generate similar images. Each original image has two similar variants under the conditions of summer and winter generated by the trained CycleGAN model. The resolution of all input images was unified to 224 × 224, and the patch size p was set to 14. Thus, an image was segmented into 256 patches. In the adapted ViT backbone network, the basic backbone network adopted ViT-B/14, the scale factor s in the parallel adapter was set to 0.5, and the dimension of the input of the adapter was limited to 768. The number of channels was reduced to 192 through 1 × 1 convolution, and then it was further reduced to 128 through convolutions with kernel sizes of 3 × 3, 5 × 5, and 7 × 7. For all convolutional components, we fix the stride to 1, use a single convolutional layer, and set the output channel dimension to 3. The outputs of the convolutions at three scales were concatenated and then added to the output through a skip connection, so that the final channel output still maintained 768 dimensions. In the loss function, the hyperparameters were set as α = 1, β = 50 and λ = 0. The margin in online mining was set to 0.1. The initial learning rate was set to 0.0001 and multiplied by 0.5 after every 3 epochs. The total number of parameters in our model is 107.3 M, and the number of tunable parameters is 21.2 M. The inference time for each query image is 0.0082 s.

4.2 Test Datasets

To thoroughly evaluate the capabilities of different methods, we selected the following five widely-used real-world datasets for test and analysis. These datasets cover a variety of real-world scenarios, namely images under different scales, viewpoints, illuminations, weather conditions, seasons, and camera types.

1. Mapillary Street-Level Sequences (MSLS)

The Mapillary Street-Level Sequences (MSLS) [36] is a large-scale long-term place recognition dataset that contains over 1.6 million street-level images collected in urban, suburban, and natural scenes over a

period of seven years. It encompasses various challenging visual variations, such as changes in lighting, weather, seasons, viewpoints, as well as dynamic objects.

2. Pittsburgh30k

The Pittsburgh [37] dataset was collected from Google Street View maps and provides images collected at different locations and poses in the city center. The images in this dataset have significant viewpoint variations and condition changes in various geometric structures. We used the test dataset consisting of 10,000 database images and 6,816 query images.

3. Nordland Dataset

The Nordland [38] dataset shows severe appearance changes images in four different seasons recorded from the perspective of the front of the train. It exhibits no viewpoint variations, thus allowing the testing of the algorithm's robustness on isolated condition variations. Note that two versions of Nordland are used for VPR benchmarking in this paper: one uses the entire summer image sequence as query images and the entire winter image sequence as reference images, while the other uses the entire winter image sequence as query images and the entire summer image sequence as reference images (accompanied by the marker*).

4. SPED Dataset

The SPED [39] (Specific PlacEs Dataset) dataset contains 2.5 million pictures collected from over 1,000 places, with hundreds of pictures for each place. These images encounter constantly changing conditions, including weather, season, and day-night variations.

5. SVOX-Night Dataset

The SVOX [40] is a cross-domain VPR dataset sourced from Google Street View images across the city of Oxford. It contains a large amount of data under various weather conditions such as sunny, rainy, snowy, and from a night-time perspective, and is divided into multiple subsets according to different conditions. The test set SVOX-Night we used, is a challenging subset of night-time image queries within the dataset.

4.3 Comparison with Previous Works

We conducted comparative experiments on six place recognition datasets (including variants of Nordland) between the method proposed in this paper and eight of the most representative VPR methods. The evaluations follow the standard VPR evaluation framework [41], using Recall@N [42] as the evaluation metric. Recall@N (R@N) measures the percentage of queries where at least one of the top-*N* retrieved reference images falls within a specific threshold of the ground truth. The proposed CerfeVPR is compared with eight promising VPR methods. These compared methods include CNN-based VPR methods: NetVLAD [12], SFRS [22], MixVPR [8], CosPlace [1], EigenPlaces [9]; a Transformer-based VPR method: DHE-VPR [25]; and VPR methods based on large pre-trained models: SelaVPR [27], CricaVPR [16]. Each method in our experiments used its original author-released parameters. The comparison results for Recall@N across six datasets, with N set to 1, 5, 10, and 20, are listed in Table 2.

Table 2: Comparison on the Recall@N. The best is highlighted in **bold** and the second is underlined

Datasets	Method	NetVLAD	SFRS	MixVPR	CosPlace	EigenPlaces	DHE	SelaVPR	CricaVPR	CerfeVPR
MSLS-val	R@1	54.0	64.9	83.3	79.4	85.0	80.5	87.3	**88.4**	88.1
	R@5	65.4	74.5	90.1	86.8	91.4	89.1	93.9	94.1	**94.3**
	R@10	70.0	78.3	92.0	89.4	93.0	91.6	**95.6**	95.5	95.5
	R@20	74.3	81.6	93.5	91.4	94.3	93.2	**96.9**	96.3	96.4
Pitts30k	R@1	85.0	89.1	91.6	88.4	91.9	89.4	92.8	94.5	**95.0**
	R@5	92.1	94.6	95.6	94.6	96.4	94.9	97.0	97.4	**97.5**

(Continued)

Table 2 (continued)

Datasets	Method	NetVLAD	SFRS	MixVPR	CosPlace	EigenPlaces	DHE	SelaVPR	CricaVPR	CerfeVPR
	R@10	94.4	96.0	96.4	95.7	97.4	96.1	97.9	**98.0**	**98.0**
	R@20	95.9	97.0	97.4	96.5	97.9	97.1	98.5	**98.7**	**98.7**
Nordland	R@1	13.1	16.0	76.4	58.5	67.9	45.8	87.0	89.1	**93.7**
	R@5	21.1	24.1	87.1	73.6	81.1	54.1	94.1	95.3	**97.4**
	R@10	26.1	28.7	90.6	79.4	85.6	56.5	95.8	96.7	**98.1**
	R@20	32.0	34.4	93.6	84.8	89.6	58.4	97.1	97.9	**98.8**
Nordland*	R@1	9.3	11.3	59.5	49.3	56.4	40.1	79.0	82.7	**85.3**
	R@5	15.0	18.1	72.3	64.4	70.4	49.0	86.5	90.6	**92.7**
	R@10	18.9	22.1	77.5	71.1	76.1	67.6	87.9	93.2	**94.9**
	R@20	23.7	27.6	82.4	77.5	81.7	70.3	89.2	95.3	**96.6**
SPED	R@1	65.7	72.3	85.7	74.6	72.0	75.6	88.8	91.8	**92.6**
	R@5	83.0	88.1	93.1	87.3	85.2	81.4	95.2	95.7	**96.2**
	R@10	88.0	92.3	95.2	90.3	89.6	82.9	96.5	96.7	**97.7**
	R@20	92.4	94.9	96.9	93.6	92.4	83.5	97.7	97.5	**98.6**
SVOX-Night	R@1	8.0	28.7	63.1	51.6	51.6	34.9	79.2	85.1	**87.0**
	R@5	17.4	40.6	79.8	70.8	70.8	50.2	92.6	**95.0**	94.8
	R@10	23.1	46.4	84.1	78.4	78.4	56.7	95.0	96.7	**96.8**
	R@20	29.4	52.1	88.1	84.1	84.1	62.0	96.6	97.1	**97.8**
Average	R@1	39.2	47.1	76.6	67.0	70.8	61.1	85.7	88.6	**90.3**
	R@5	49.0	56.7	86.3	79.6	82.6	69.8	93.2	94.7	**95.5**
	R@10	53.4	60.6	89.3	84.1	86.7	75.2	94.8	96.1	**96.8**
	R@20	58.0	64.6	92.0	88.0	90.0	77.4	96.0	97.1	**97.8**

The proposed CerfeVPR exhibits outstanding overall performance. On average, CerfeVPR achieves the highest results, with all Recall@N values exceeding 90%. Furthermore, it achieves state-of-the-art performance across all metrics on Pitts30k, Nordland, Nordland*, and SPED. In particular, CerfeVPR shows a significant advantage on Nordland and its variant Nordland*, achieving a remarkable 4.6% improvement in Recall@1. This improvement may be attributed to the alignment of query images' similar variants with reference images under similar environmental conditions. Unlike other datasets, Nordland, comprising video frames captured during ten-hour train journeys across four distinct seasons, presents drastic seasonal appearance changes, often causing a season gap between query and reference images. CerfeVPR bridges this gap by generating seasonally aligned variants for summer and winter using CycleGAN. These variants enhance the global descriptors of query and reference images, substantially improving VPR performance and resulting in a marked increase in Recall@N values on Nordland and Nordland*.

Fig. 5 presents the qualitative results under five highly challenging query cases. From top to bottom, it covers a variety of complex situations such as significant seasonal changes, rainy and snowy weather, perceptual aliasing, illumination changes, and viewpoint variations, all of which are typical difficult problems in VPR tasks. Especially in the first two query images, where heavy snow in winter covered many local features and need to match them with the reference images taken in summer, our method was still able to accurately locate the images, while other methods generally returned images that were visually similar to the query images but had incorrect geographical locations.

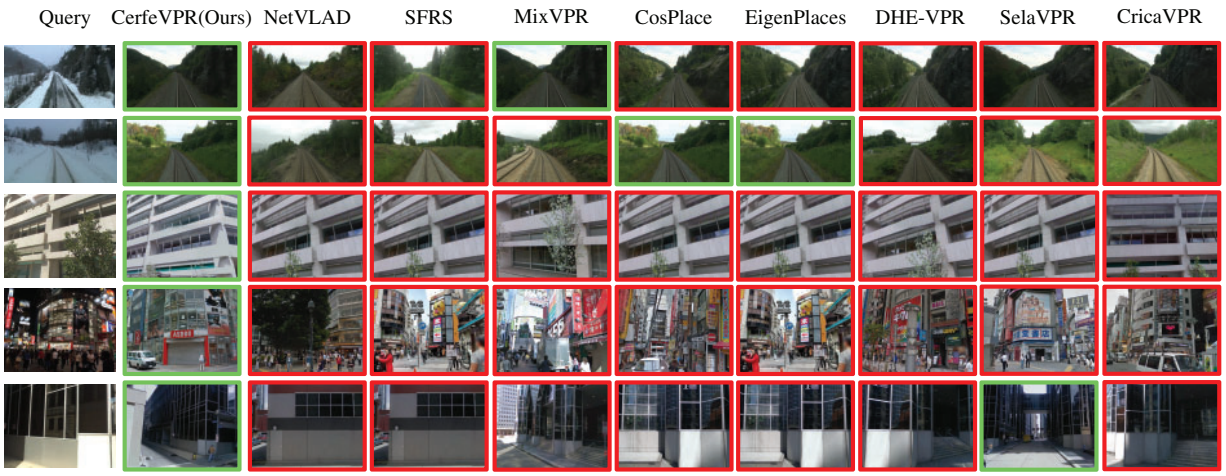

Figure 5: Qualitative results. In these queries, our method accurately returned the correct retrieval results, while most of the other comparative methods returned incorrect matching images

4.4 Ablation Study

To deeply explore the effectiveness of each component module in the model and test its contribution to the overall performance, we removed the corresponding modules from CerfeVPR to conduct ablation experiments. We denote the version without the similar image generation module as "-w/o SIG". "-w/o PSPN" represents the removal of the 3-layer pyramid scene parsing network used to obtain the initial global feature representation of the image. "-w/o RFL" represents the removal of the robust feature learning module used to enhance the robustness of global features. This ablation experiment verified the improvement of model performance brought by different modules, and the results are shown in Table 3.

Table 3: Ablation experiments. The best results are highlighted in **bold** and the second is underlined

Versions	NordLand			NordLand*			SPED			SVOX-Night		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
-w/o SIG	93.3	97.2	98.0	85.0	92.5	94.4	92.4	95.9	97.5	86.8	93.3	95.3
-w/o PSPN	92.5	96.7	97.8	81.4	88.9	91.5	89.0	94.1	96.3	84.3	90.0	92.4
-w/o RFL	89.2	95.7	96.5	80.9	88.3	90.2	88.1	93.6	94.9	83.8	89.5	91.9
CerfeVPR	**93.7**	**97.4**	**98.1**	**85.3**	**92.7**	**94.9**	**92.6**	**96.2**	**97.7**	**87.0**	**94.8**	**96.8**

The similar image generation module generates retrieval and reference images with new backgrounds but still retaining the original features by changing the seasons of the images without changing the original location labels of the data. Experimental data shows that the similar image generation module expands the diversity of images, enabling the model to learn features that are more crucial for the place recognition task. The 3-layer pyramid scene parsing network aggregates features at different scales, enabling the model to better adapt to the scale changes of images. The robust feature learning module improves the model's robustness to interferences such as illumination and seasons by learning the features beneficial for recognition in similar images. The combined effect of each module in the model makes the model perform better when facing complex inputs in the real world.

4.5 Impact Analysis of Adapters

We investigated the structural design of adapters and the impact of their insertion positions on retrieval performance. Considering the ease of training residual structures and the superior performance of superimposing trainable weights on frozen weights compared to splicing them, we devised four adapter configurations to fine-tune and optimize DINOv2. 1) Serial LKA: two LargeKernel adapters serially connected before the MHA and the MLP layer in the Transformer modul of DINOv2. 2) Serial + Parallel LKA: A modification of Serial LKA where the LargeKernel adapter connected before MLP is arranged in parallel. 3) Serial + Parallel VA: similar to Serial + Parallel LKA, but the LargeKernel adapters are replaced with Vanilla adapters. This configuration, employed in SelaVPR, integrates both serial and parallel Vanilla adapters. 4) Serial VA + Parallel LKA: This configuration is adopted in the proposed CerfeVPR (Fig. 4d), with a Vanilla adapter in series and a LargeKernel adapter in parallel with the MLP layer. We fine-tuned DINOv2 under these four configurations, followed by GeM pooling to generate image descriptors, ensuring experimental consistency across all evaluations.

The retrieval results presented in Table 4 show that Serial + Parallel LKA outperforms standalone Serial LKA, as its non-intrusive residual connections preserve discriminative features. Nevertheless, the Serial + Parallel LKA configuration performs slightly lower performance than Serial + Parallel VA. This difference may stem from the relatively complex structure of the LargeKernel adapter and its serial integration into the backbone network, which can adversely affect the model's performance. In contrast, the Serial VA + Parallel LKA configuration, which is employed in CerfeVPR, achieves the best performance by striking an optimal balance. Specifically, Serial VA allows for a sequential extraction of features, capturing long-range dependencies effectively. Parallel LKA, on the other hand, can capture local and global features simultaneously in parallel, enhancing the representational power of the model. The combination of these two structures leverages their respective advantages, leading to the best performance in VPR tasks.

Table 4: Ablation studies of 4 different structures on the SVOX-Night and SPED datasets. The best is highlighted in **bold** and the second is underlined

Method	SPED				SVOX-Night			
	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
Serial LKA	87.0	94.2	95.7	**97.5**	78.3	89.1	91.9	94.2
Serial + Parallel LKA	88.3	94.1	95.9	97.2	79.8	88.9	92.0	94.2
Serial + Parallel VA	88.6	94.9	95.6	**97.5**	80.6	90.6	93.0	94.9
Serial VA + Parallel LKA	**88.8**	**95.2**	**96.6**	**97.5**	**85.9**	**94.4**	**96.1**	**97.3**

4.6 Impact Analysis of Descriptor Dimensions

We utilized Principal Component Analysis (PCA) to reduce the dimensionality of the original high-dimensional global descriptors of each query and reference image, aiming to evaluate the impact of descriptor dimensions on VPR performance. VPR tests were conducted using these reduced-dimension global descriptors on three datasets: MSLS-val, Pitts30k, and Nordland. The place recognition results for descriptor with various dimensions are illustrated in Table 5. The experimental results reveal that dimensionality reduction has minimal impact on place recognition performance for the Pitts30k dataset. In contrast, performance on the MSLS-val dataset decreases significantly when descriptors are reduced to 512 or 1024 dimensions. The decline is even more pronounced for the Nordland dataset, with reductions exceeding 1% when dimensions

drop below 4096. This degradation is primarily because that severe environmental changes and perceptual aliasing present in the MSLS-val and Nordland datasets. High-dimensional descriptors retain more effective information captured by the model, enabling precise differentiation between images of different locations and enhancing retrieval accuracy.

Table 5: Experimental results for descriptors with varying dimensions. The best is highlighted in **bold** and the second is underlined

Dim	MSLS-val				Pitts30k				Nordland			
	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
512	82.4	90.4	92.4	93.7	94.1	96.9	97.7	98.1	81.5	91.0	93.8	95.8
1024	85.4	92.8	94.2	95.3	94.6	97.2	97.9	98.3	87.4	94.4	96.1	97.6
2048	87.0	93.8	95.0	96.0	94.9	97.4	97.9	98.5	90.8	96.0	97.2	98.3
4096	87.7	94.2	95.3	96.3	94.9	**97.5**	**98.0**	**98.7**	92.5	96.8	97.8	98.6
10752	**88.1**	**94.3**	**95.5**	**96.4**	**95.0**	**97.5**	**98.0**	**98.7**	**93.7**	**97.4**	**98.1**	**98.8**

5 Conclusions

This work addresses the problem of perceptual aliasing caused by environmental differences in the field of VPR and proposes CerfeVPR, an innovative cross-environment enhanced robustness VPR method. It uses a Generative Adversarial Network to generate images with similar environmental conditions and introduces large-kernel convolutional adapters and a 3-layer pyramid scene parsing network to analyze real-world place scenes. The accuracy of this method has been improved to a certain extent on multiple datasets, demonstrating the model's ability to learn the features of key landmarks in complex environments and achieving an ideal trade-off between global and local features. However, some datasets may have significant architectural appearance shifts and extreme viewpoint changes. When there's a lack of distinctive features, the model can produce wrong results. Also, the similar-image generation module, though generating diverse scene-based similar images, has a marginal-efficiency-decline issue and limited impact on retrieval accuracy. Future research can focus on the currently emerging real-time VPR systems. Their fast and accurate positioning capabilities are crucial for scenarios such as autonomous driving and drone navigation. In-depth exploration of the environmental alignment between query images and reference images, and the excavation of more efficient feature extraction and similarity measurement models will help reduce the dependence on expensive computing power, and promote the VPR system to achieve higher robustness, accuracy, and practicability in dynamic environments, serving scenarios such as autonomous driving and robot navigation.

Acknowledgement: The authors gratefully acknowledge the helpful comments and suggestions of the reviewers and editors, which have improved the presentation.

Funding Statement: This project is supported by Postgraduate Scientific Research Innovation Project of Hunan Province CX20230915, National Natural Science Foundation of China under Grant 62472440.

Author Contributions: The authors confirm contribution to the paper as follows: Lingyun Xiang: Conceptualization, Methodology, Writing—Original Draft, Investigation. Hang Fu: Methodology, Software, Writing—Reviewing & Editing. Chunfang Yang: Writing—Reviewing & Editing, Analysis and Interpretation of Results, Visualization. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All relevant data are within the paper. The data are available from the corresponding author on reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. 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 (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 4878–88.
2. Cao B, Araujo A, Sim J. Unifying deep local and global features for image search. In: Computer Vision-ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK. p. 726–43.
3. Masone C, Caputo B. A survey on deep visual place recognition. IEEE Access. 2021;9:19516–47. doi:10.1109/ACCESS.2021.3054937.
4. 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; 2010 Jun 13–18; San Francisco, CA, USA. p. 3304–11.
5. Sunderhauf N, Shirazi S, Dayoub F, Upcroft B, Milford M. On the performance of ConvNet features for place recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2015 Sep 28–Oct 2; Hamburg, Germany. p. 4297–304.
6. Chen Z, Lam O, Jacobson A, Milford M. Convolutional neural network-based place recognition. In: Proceedings of the 16th Australasian Conference on Robotics and Automation; 2014 Dec 2–4; Melbourne, Australia. p. 1–8.
7. Hou Y, Zhang H, Zhou S. Convolutional neural network-based image representation for visual loop closure detection. In: IEEE International Conference on Information and Automation; 2015 Aug 8–10; Lijiang, China. p. 2238–45.
8. Ali-Bey A, Chaib-draa B, Giguere P. MixVPR: feature mixing for visual place recognition. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2–7; Waikoloa, HI, USA. p. 2997–3006.
9. Berton G, Trivigno G, Caputo B, Masone C, di Torino P. EigenPlaces: training viewpoint robust models for visual place recognition. In: IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 2–6; Paris, France. p. 11046–56.
10. Hausler S, Garg S, Xu M, Milford M, Fischer T. Patch-NetVLAD: multi-scale fusion of locally-global descriptors for place recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 14136–47.
11. 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; 2022 Jun 18–24; New Orleans, LA, USA. p. 13648–57.
12. Arandjelović R, Gronat P, Torii A, Pajdla T, Sivic J. NetVLAD: CNN architecture for weakly supervised place recognition. IEEE Trans Pattern Anal Mach Intell. 2018;40(6):1437–51. doi:10.1109/TPAMI.2017.2711011.
13. Radenović F, Tolias G, Chum O. Fine-tuning CNN image retrieval with no human annotation. IEEE Trans Pattern Anal Mach Intell. 2018;41(7):1655–68. doi:10.1109/TPAMI.2018.2846566.
14. Keetha N, Mishra A, Karhade J, Jatavallabhula KM, Scherer S, Krishna M, et al. AnyLoc: towards universal visual place recognition. IEEE Robot Autom Lett. 2024;9(2):1286–93. doi:10.1109/LRA.2023.3343602.
15. Izquierdo S, Civera J. Optimal transport aggregation for visual place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 17–21; Seattle, WA, USA. p. 17658–68.
16. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 17–21; Seattle, WA, USA. p. 16772–82.

17. Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, de Laroussilhe Q, Gesmundo A, et al. Parameter-efficient transfer learning for NLP. In: International Conference on Machine Learning, PMLR; Jun 9–15; Long Beach, CA, USA. p. 2790–9.
18. Chen S, Ge C, Tong Z, Wang J, Song Y, Wang J, et al. AdaptFormer: adapting vision transformers for scalable visual recognition. Adv Neural Inf Process Syst. 2022;35:16664–78.
19. Shen X, Wu W, Wang X, Zheng Y. Multiple riemannian kernel hashing for large-scale image set classification and retrieval. IEEE Trans Image Process. 2024;33(11):4261–73. doi:10.1109/TIP.2024.3419414.
20. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM. 2020;63(11):139–44. doi:10.1145/3422622.
21. Hao W, Wang J, Lu H. A real-time semantic segmentation method based on transformer for autonomous driving. Comput Mater Contin. 2024;81(3):4419–33. doi:10.32604/cmc.2024.055478.
22. Ge Y, Wang H, Zhu F, Zhao R, Li H. Self-supervising fine-grained region similarities for large-scale image localization. In: Computer Vision-ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK. p. 369–86.
23. Ali-bey A, Chaib-draa B, Giguere P. Gsv-cities: toward appropriate supervised visual place recognition. Neurocomputing. 2022;513:194–203. doi:10.1016/j.neucom.2022.09.127.
24. 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; 2023 Jun 18–22; Vancouver, BC, Canada. p. 19370–80.
25. Lu F, Dong S, Zhang L, Liu B, Lan X, Jiang D, et al. Deep homography estimation for visual place recognition. Proc AAAI Conf Artif Intell. 2024;38(9):10341–9. doi:10.1609/aaai.v38i9.28901.
26. Kannan SS, Min BC. PlaceFormer: transformer-based visual place recognition using multi-scale patch selection and fusion. IEEE Robot Autom Lett. 2024;9(7):6552–9. doi:10.1109/LRA.2024.3408075.
27. 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; May 7–11; Vienna, Austria. [cited 2025 Jan 1]. Available from: http://OpenReview.net.
28. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. Dinov2: learning robust visual features without supervision. arXiv:230407193. 2023.
29. Shaham TR, Dekel T, Michaeli T. Singan: learning a generative model from a single natural image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 4570–80.
30. Zhu J, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice, italy. p. 2223–32.
31. Qin J, Wang J, Tan Y, Huang H, Xiang X, He Z. Coverless image steganography based on generative adversarial network. Mathematics. 2020;8(9):1394. doi:10.3390/math8091394.
32. Gao P, Qin J, Xiang X, Tan Y. Robust and privacy-preserving feature extractor for perturbed images. Pattern Recognit. 2025;161(2):111202. doi:10.1016/j.patcog.2024.111202.
33. He G, Zhang X, Wang F, Fu Z. A novel copy-move detection and location technique based on tamper detection and similarity feature fusion. Int J Auton Adapt Commun Syst. 2024;17(6):514–29. doi:10.1504/IJAACS.2024.142523.
34. Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2016 Jun 26–Jul 1; Las Vegas, NV, USA. p. 761–9.
35. Wang X, Han X, Huang W, Dong D, Scott MR. Multi-similarity loss with general pair weighting for deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA, USA. p. 5022–30.
36. 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; 2020 Jun 14–19; Seattle, WA, USA. p. 2623–32.

37. Torii A, Sivic J, Pajdla T, Okutomi M. Visual place recognition with repetitive structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2013 Jun 23–28; Portland, OR, USA. p. 883–90.
38. Sünderhauf N, Neubert P, Protzel P. Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In: IEEE International Conference on Robotics and Automation (ICRA); May 6–10; Karlsruhe, Germany.
39. Chen Z, Jacobson A, Sunderhauf N, Upcroft B, Liu L, Shen C, et al. Deep learning features at scale for visual place recognition. In: 2017 IEEE International Conference On Robotics And Automation (ICRA); 2017 May 29–Jun 3; Singapore. p. 3223–30.
40. Berton G, Paolicelli V, Masone C, Caputo B. Adaptive-attentive geolocalization from few queries: a hybrid approach. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2021 Jan 3–8; Waikoloa, HI, USA. p. 2918–27.
41. Berton G, Mereu R, Trivigno G, Masone C, Csurka G, Sattler T. Deep visual geo-localization benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 5396–407.
42. Zaffar M, Garg S, Milford M, Kooij J, Flynn D, McDonald-Maier K, et al. VPR-bench: an open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. Int J Comput Vis. 2021;129(7):2136–74. doi:10.1007/s11263-021-01469-5.