Open access Original scientific paper

https://doi.org/10.5552/crojfe.2026.2924

ForestsNet: Mixer Feature and Binary Neural Networks towards Robust and Efficient Visual Place Recognition in Forest

Junshuai Wang, Junyu Han, Ruifang Dong, Jiangming Kan

Abstract

*Visual Place Recognition (VPR) enables robots to determine current location by comparing input image against previously stored reference images. It is essential in autonomous location and simultaneous localization and mapping (SLAM). A key task of VPR is evaluating similarity between images, as state-of-the-art deep learning-based approaches have achieved outstanding performance in standard indoor/outdoor scenes. However, the SOTA deep learning-based methods underperform in forestry robotic owing to two challenges, constrained computational capabilities and appearance variation due to seasonal shifts, weather/light/viewpoint variations, which substantially impair visual similarity computation. Consequently, this work proposes ForestsNet, a novel lightweight VPR network, to resolve this issue. First, a Binary Neural Network (*BNN*) was constructed to achieve considerable memory reduction. A novel binarization function, Leaky Sign, is proposed; it adaptively applies quantization factors to input activations, it retains richer feature information during binarization while significantly reducing accuracy degradation of place recognition. Second, Mixer Forests, a novel multilayer perceptron-based aggregation method is introduced to integrate global context into feature maps, substantially enhancing the robustness against appearance variation. In addition, two novel evaluation metrics, Memory Allocation Efficiency and Balance Compression Recall, are designed to quantify the trade-off between memory efficiency and place recognition accuracy. Experimental results demonstrate that ForestsNet achieves substantially higher memory usage efficiency than full-precision networks. Compared to state-of-the-art BNNs, it presents superior performance in both memory efficiency and place recognition accuracy, establishing itself as a robust VPR solution for resource-constrained forestry robots.*

Keywords: Visual Place Recognition, Forest Scene, Memory usage efficiency, Aggregate Feature, Binary Neural Networks

1. Introduction

Place recognition has gained increasing attention in robotics community to help robots understand the spatial characteristics of the world. It refers to the problem of deciding whether a place has been visited before, and if it has been visited before, which place it is. It is essential in autonomous location and SLAM. In SLAM, it is always used in loop closure or relocalization work to correct results when the position tracking fails or drifts due to accumulated errors. Visual place recognition concerns the ability to recognize previously seen places in the world based solely on images (Lowery et al. 2016, Fan et al. 2022). VPR is usually defined as an image retrieval problem, where it determines whether a query image has been seen by evaluating the similarity between this query image and previously stored reference images. Many recent studies have achieved great progress, especially the state-of-the-art deep learning-based approaches have achieved outstanding performance in standard indoor/outdoor scenes. However, VPR in forest is still extremely challenging which retards the development of forestry robots. Two crucial challenges include the constrained computational capabilities of robots, and serious environmental appearance variations due to seasonal shifts, weather/light/viewpoint variations (as

Fig. 1 The same place in forest shows different appearance due to weather, season, and illumination variations

shown in Fig. 1), which substantially impair VPR computation (Chen et al. 2022, Li et al. 2024).

The success of deep learning-based VPR methods relies heavily on high computational costs and large parameter sizes (Chen et al. 2023, Zhao et al. 2023), which conflicts with the limited computing and storage resources of mobile robots (Lee et al. 2022), especially the robots operating in wild places, such as forestry robots. Reducing model size is therefore critical for mobile robots (Ferrarini et al. 2019, Fan et al. 2022). Fortunately, Binary Neural Networks (*BNNs*) can reduce memory consumption by converting weights and activations from 32 bits to 1 bit. For example, FloppyNet (Ferrarini et al. 2022) first applied *BNNs* to reduce model size using the Straight-through Estimator (STE) as binarization function (Courbariaux et al. 2016). However, its limited expressiveness causes significant loss of information during binarization, leading to serious accuracy degradation in *BNNs*. Therefore, it is necessary to conduct more in-depth research on the

binarization function to decrease accuracy loss, and this work explored a novel binarization function to preserve more information.

On the other hand, in order to tackle the challenges from environmental appearance variations, researchers proposed global and robust image representation methods, such as NetVLAD (Arandjelović et al. 2016), GeM (Radenović et al. 2018), CosPlace (Berton et al. 2022), and MixVPR (Ali-Bey et al. 2023). These methods describe the entire image and they are called aggregation methods. However, they still underperform in forest scene due to the aforementioned extreme appearance variations, while local features can provide fine-grained details and demonstrate strong robustness to occlusions and illumination variations. However, their high computational and storage demands hinder deployment in resource-constrained forestry robot. Thus, this work proposed Mixer Forests aggregation, which aims to combine global and local information to deal with this challenge.

Therefore, ForestsNet was constructed in this work to address VPR challenges in forest environments. In addition, VPR algorithms are designed to achieve high recognition accuracy with low memory consumption. The evaluation metrics that can simultaneously account for both the place recognition accuracy and memory usage efficiency of VPR algorithms are essential. However, the current evaluation metrics cannot meet this requirement. The contributions of this work are as follows:

- ⇒ a novel binarization activation function, Leaky Sign, was presented to preserve more information by adapting binarization factors to input activation values, minimizing loss and enabling efficient storage/computation. It addresses *BNNs* accuracy degradation from STE limited expressiveness, which is critical for resource-constrained forestry robots
- ⇒ a novel aggregation method, Mixer Forests, was designed for changing environments. It uses feature maps from a binary backbone network, iteratively integrates global relations into local feature maps to obtain global-local information, thereby enhancing the robustness of forest image representation
- ⇒ two evaluation metrics, Memory Allocation Efficiency (MAE) and Balance Compression Recall (*BCR*), were constructed to evaluate the performance of VPR algorithms by balancing compression and accuracy trade-offs
- ⇒ ForestsNet was constructed for VPR, primarily consisting of a binarized backbone network and the Mixer Forests module. For the binarized backbone network, Leaky Sign and DoReFaNet were applied to binarized activation and weight function. Mixer Forests aggregate features from the binarized backbone to generate robust forest image representations. Experiments showed that ForestsNet achieves higher Recall than current typical aggregation methods and superior comprehensive performance compared to prevailing *BNNs*.

The remainder of the paper is structured as follows: Section 2 provides a review of related work. Section 3 presents the proposed ForestsNet in detail. Section 4 introduces the evaluation metrics essential to our study. Section 5 discusses the experimental results and analysis. Finally, Section 6 discusses and concludes the paper.

2. Related Work

2.1 Deep Learning for VPR

Visual Place Recognition has been regarded as an image retrieval task, where the goal is to localize a query image by matching it against the most similar reference image in a pre-built database. Deep learning techniques have achieved remarkable success in visual community (Yan et al. 2015, Heidari et al. 2018, Proto et al. 2020, Daou et al. 2023, Zhang 2025), demonstrating strong performance across diverse environments, including both indoor and outdoor urban scenes (Petrakis et al. 2023, Yu et al. 2024). A typical deep learning-based VPR pipeline consists of two main stages, feature extraction which employs a backbone network that encodes input images into high-dimensional features, and feature aggregation which uses a trainable aggregation layer that transforms these features into robust and compact representations for efficient matching. Depending on the approaches, deep learning-based VPR methods may leverage either local features or global features to encode place information.

Local feature-based methods focus on extracting and aggregating features from key regions in feature map. One of the earliest approaches is MAC (Maximum Activation of Convolutions) (Babenko et al. 2015), which employed max-pooling to aggregate the most activated neurons across each feature map. Building upon MAC, R-MAC (Regional Maximum Activation of Convolutions) (Tolias et al. 2015) enhanced robustness by extracting multiple Regions of Interest (RoIs) from convolutional feature maps to form a compact representation. More recently, Patch-NetVLAD (Häusler et

al. 2021) extended this idea by leveraging NetVLAD to aggregate multi-scale patch features, enabling effective matching of deep-learned local features across a spatially distributed feature-space grid.

In contrast, global feature-based methods encode the entire image into a single high-dimensional vector for place representation. GeM (Generalized Mean pooling) is a widely adopted aggregation technique (Radenović et al. 2018), a learnable global pooling method that generalizes max and average pooling. Further improving efficiency, CosPlace (Berton et al. 2022) introduced a lightweight aggregation layer by combining GeM with a linear projection. Another prominent approach, NetVLAD (Arandjelović et al. 2016), adapts the VLAD descriptor into a trainable deep learning framework, where local features are softly assigned to learned cluster centers to form a discriminative global representation.

One of the latest trends in VPR is the adoption of a two-stage retrieval strategy. In the first stage, a global descriptor-based k-NN search efficiently retrieves the top k candidate images from the reference database. The second stage re-ranks these candidates using local feature matching, significantly improving retrieval accuracy. Representative methods in this paradigm include Patch-NetVLAD, Contextual Patch-NetVLAD (Sun et al. 2024), and TransVPR (Yang et al. 2021), etc.

Recently, MixVPR (Ali-Bey et al. 2023) introduced a novel feature aggregation technique that diverges from conventional approaches. Unlike TransVPR and Patch-NetVLAD – which rely on self-attention mechanisms or regional feature pooling – MixVPR generated global descriptors without re-ranking, yet achieved superior performance over two-stage methods. The Focus on Local (FoL) (Wang et al. 2025) was introduced to explicitly model critical image regions via specialized losses and a pseudo-supervised training scheme, simultaneously improving retrieval and re-ranking performance and setting a new state-of-the-art on multiple benchmarks.

For forest environments, local feature-based methods excel at capturing fine-grained details and demonstrate strong robustness to occlusions and illumination variations. However, their high computational and storage demands hinder deployment in resource-constrained forestry robots. Conversely, global feature methods are computationally efficient but suffer from sensitivity to environmental changes. To bridge this gap, local information was integrated into global feature map, enhancing robustness to environmental variations while minimizing computational and storage overhead.

2.2 *BNNs* for VPR

Although deep learning-based methods demonstrate strong performance in VPR, their computational efficiency is hindered by high parameter counts and large model sizes. Over the past decade, numerous techniques have been developed to reduce inference latency. These techniques primarily focus on reducing redundancy and non-essential weights, such as Optimal Brain Damage and Optimal Brain Surgeon, which leverage Hessian matrices to prune model connections; weight pruning (Han et al. 2015) was shown to reduce parameters in state-of-the-art networks by an order of magnitude. However, aggressive compression trades off accuracy, binary networks (1-bit precision) achieve maximal compression and acceleration post-training but incur significant accuracy loss (Courbariaux et al. 2014).

Binary Neural Networks (*BNNs*) achieved competitive performance while operating with low-precision models (Simons et al. 2019). Courbariaux and Bengio initially (Courbariaux et al. 2014) trained a fully binary network using a straight-through estimator (STE). STE approach maintains real-valued weights in memory during training while performing forward passes with binarized activations, enabling effective gradient propagation during backpropagation. Subsequent research has further advanced *BNN* performance; XNOR-Net (Rastegari et al. 2016) introduced optimized convolutional block architectures to enhance classification accuracy. Standard implementations typically apply batch normalization (BatchNorm) between convolutional operations and activation functions to improve training stability. However, XNOR-Net adopted a unique configuration where BatchNorm and binary activation precede the convolutional operation, causing pooling prior to binarization. While Bi-RealNet (Liu et al. 2018) improved ImageNet top-1 accuracy through feature map shortcuts, it maintains 32-bit floating-point operations for batch normalization and addition processes. DoReFa-Net (Zhou et al. 2016) accelerated training by employing bitwise operations for weight-input dot products. Esser et al. (2020) advanced this approach by jointly learning binarized thresholds and weights to narrow the accuracy gap with full-precision models. Real-to-Bin-Net further enhanced performance by aligning binary and real-valued convolutions through spatial attention maps and an auxiliary loss function.

FloppyNet (Ferrarini et al. 2022) first demonstrated *BNNs* applicability to VPR by introducing a compact binary architecture. Ternary networks (Li et al. 2016, Zhu et al. 2017) represented weights using three discrete values. Although they offer reduced memory

usage (2-bit storage) and computational complexity compared to binary networks (Hubara et al. 2017, Lin et al. 2017, Phan et al. 2020), their performance still lags behind full-precision counterparts. The STBLLM (Dong et al. 2025) represents a milestone as the first work to achieve sub-1-bit structural binarization for LLMs. By integrating N:M sparsity with a layer-wise and group-wise binarization strategy, it effectively compresses models to 0.55 bits/weight.

For *BNNs*, weights and activations exhibit distinct quantization behaviors. Activations, being input-dependent, display wide dynamic ranges and asymmetric distributions. To address these properties, a novel adaptive binarization approach was proposed that employs input-dependent scaling factors for activation quantization. Unlike activations, weights are static parameters optimized during training. DoReFa-Net weight binarization function replaces floating-point computations with bitwise operations, significantly accelerating computation. Following DoReFa-Net approach, we present the first integration of *BNN* principles with Mixer techniques for VPR in forest environments, which includes a novel binarized activation function with adaptive thresholds and an efficient feature aggregation technique.

2.3 Memory Usage Efficiency for VPR

Forestry robots are equipped with limited computing resources, so VPR algorithm is desired to achieve high recognition accuracy with low memory consumption. Traditional metrics evaluate either memory footprint or recognition performance in isolation, which fails to comprehensively assess this trade-off.

Ferrarini et al. (2019) introduced a metric based on the ratio of model size to s_{p100}, s_{p100} referring to the ratio of correct number of retrieved images to all retrieved images. Both factors were considered, but this approach emphasizes model compactness and overlooks place recognition accuracy as a critical performance indicator, leading to an incomplete evaluation. In this study, *BNNs* are employed to compact model size, which typically comes at the cost of place recognition accuracy. It is necessary to systematically consider both the reduction in model memory and the place recognition performance loss. However, current VPR methods use model size as the main evaluation metric to assess memory usage efficiency, thus reflecting storage requirements, but it fails to capture the crucial relationship between memory usage and recognition performance.

Therefore, in order to enhance the rationality of the evaluation metrics, this work proposed Balanced Compression Ratio (*BCR*) comprehensively quantifying the trade-off between model compression and accuracy. Meanwhile, Memory Allocation Efficiency (MAE) was designed to quantify the recognition performance achieved per unit of memory allocation.

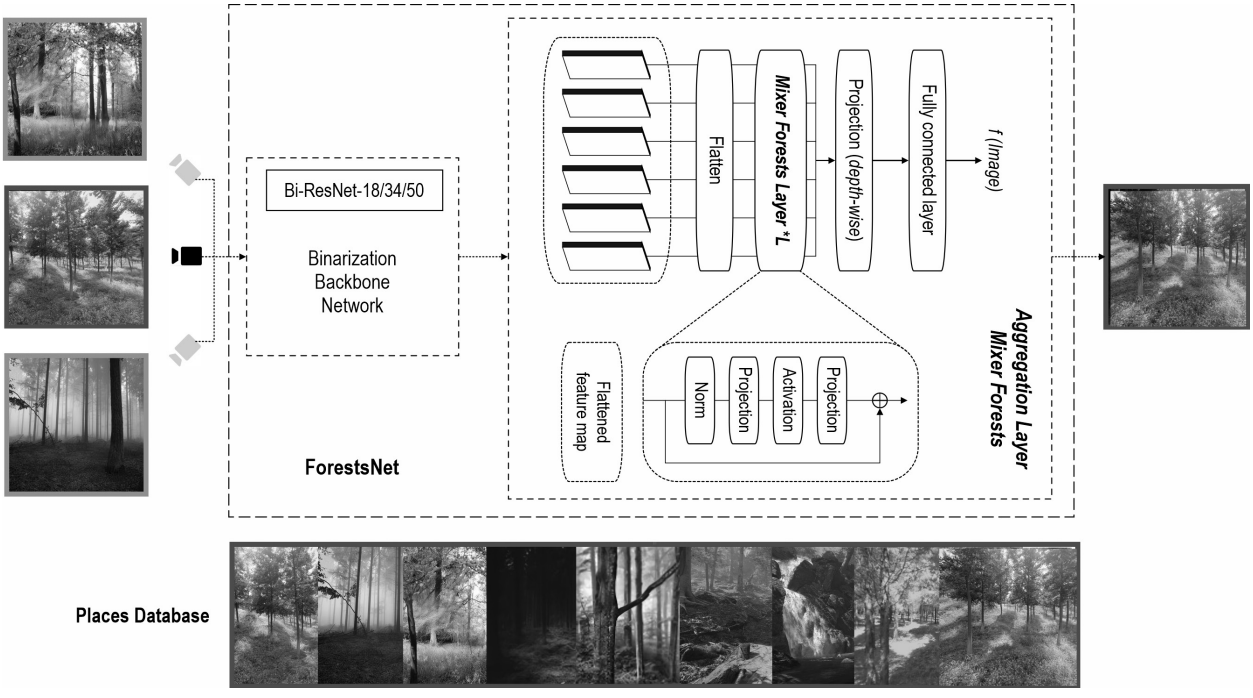

Fig. 2 ForestsNet network structure

3. Methods

This section introduces our ForestsNet (as shown in Fig. 2), which is specifically designed for Visual Place Recognition (VPR) in forest environments. ForestsNet comprises two key components: a backbone network constructed using Binary Neural Networks (*BNNs*), and Mixer Forests which serves as the network top layer. When inputting an image captured by forestry robot, ForestsNet retrieves the most similar place from the stored map if this place has been visited before; if it has not been visited, the map is updated.

It consists of two core components: Binarization Backbone Network and Aggregation layer.

3.1 Binarization Backbone Network

The binarization backbone network employs a standard ResNet18/34/50 architecture (He et al. 2016) with two key modifications:

⇒ activations binarized using our proposed Leaky Sign function

⇒ weights quantized via DoReFaNet binarization scheme.

As illustrated in Fig. 3, each residual block executes operations in the sequence: BatchNorm → Binarization → Binary Convolution → Pool. This configuration has been empirically validated as optimal for gradient propagation. Crucially, the BatchNorm layer incorporates an affine transformation that enables its bias term to function as a learnable binarization threshold. Notably, during downsampling operations, input data bypasses binarization and weight quantization. Instead, standard full-precision convolutional layers perform downsampling operations that lead to spatial resolution reduction.

Fig. 3 Convolutional blocks in a CNN (left) and in a *BNN* (right). The execution order of the standard convolutional structure has been modified. Specifically, Batch Normalization (Batch Norm) and activation operations have been placed prior to the convolutional layer, with binarization applied to both the activation function and the convolutional layer

Binarization compresses memory usage by converting model input activations and weights from 32-bit to 1-bit (Ding et al. 2022). It greatly reduces memory requirement to store the model. However, standard backpropagation proves be unsuitable for training *BNNs* due to its dependence on high-precision gradient accumulation (Hubara et al. 2017). Courbariaux and Bengio (Courbariaux et al. 2016) employed STE to solve this problem during forward propagation. They proposed a sign function as follows:

$$Forword: r_0 = \mathrm{sign}(r_\mathrm{i}) = \begin{cases} 1, & \text{if } x \geq 0 \\ -1, & \text{otherwise} \end{cases} \quad (1)$$

$$Backward: \frac{\partial l}{\partial r_\mathrm{i}} = \frac{\partial l}{\partial r_\mathrm{o}} 1_{|r_\mathrm{i}| \leq t_\mathrm{clip}} \quad (2)$$

Where:

l denotes loss function

r_i is a real-value input

$r_\mathrm{o} \in$ {–1, +1} is a binary output.

T_clip is a threshold for clipping gradients, and is originally set to 1. Here Sign returns one of the possible values of {–1, +1}.

Paradoxically, while the STE enables binary networks, the binarization process itself – which compresses continuous inputs into discrete values – causes a substantial degradation in model accuracy. To mitigate this limitation, Leaky Sign activation function is proposed, as shown in Eq. (3). This piecewise function dynamically adapts binarization factors based on input activation value.

$$a_\mathrm{o} = \begin{cases} 1 + (a_\mathrm{i} - 1) \cdot \alpha, & if\ a_\mathrm{i} > 1, \\ a_\mathrm{i}, & if -1 \leq a_\mathrm{i} \leq 1, \\ -1 + (a_\mathrm{i} + 1) \cdot \alpha, & otherwise \end{cases} \quad (3)$$

Where:

a_o output activation values

a_i input activation values.

Binarization factor α is 0.2 when $\alpha_\mathrm{i} > 1$, and 0.1 when $\alpha_\mathrm{i} > 2$.

In terms of weights, standard gradient descent fails to optimize binary weights directly due to vanishingly small gradients that cannot update discrete parameters. To solve this problem, Rastegari et al. (2016) introduced real weights W and sign functions during the training process. Thus, the binary weight scalar can be considered as the output of the sign function. In the weight function binarization method, DoReFaNet (Zhou et al. 2016) scales all filters uniformly using a constant scalar. The STE is adopted for all binarized weights as Eq. (4)

and Eq.(5). In Eq. 4, k>1. STE f_w^k is used for the weights as follows:

$$Forword: w_o = f_w^k = 2Q_k\left(\frac{tan\ h(w_i)}{2max(|tan\ h(w_i)|)} + \frac{1}{2}\right) \quad (4)$$

$$Backward: \frac{\partial l}{\partial w_i} = \frac{\partial w_o}{\partial w_i}\frac{\partial l}{\partial w_o} \quad (5)$$

Where:

Q_k quantizes a real-value w_i to a k-bit value w_o in [0, 1], where $tanh(\cdot)$ restricts the weight to [–1, 1].

Thus, Eq. (1) constrains weights f_w^k to [0, 1]. An affine transformation is then used to adjust $f_w^k(r_i)$ to [–1, 1]. Eq. (5) is the backpropagation with respect to DoReFaNet, the partial derivatives with respect to w_o and w_i, and l is the loss function.

3.2 Mixer Forests

This section presents the proposed aggregation method – Mixer Forests. As shown in Fig. 2, at network's top, Mixer Forests aggregates features to reduce performance loss caused by binarization operations and generate a robust feature representation of forest environments.

In the feature aggregation Layer, Mixer Forests uses the top-level flat feature maps from the backbone network as input. It incorporates global relationships into each feature map by using multiple Mixer Forests. The resulting output is then projected onto a compact representation space, which is utilized as a global descriptor.

Given an image I, feature map $F \in R^{c \times h \times w}$ can be extracted from the binary backbone network, where $F = BNN(I)$. (c, h, and w, respectively, represent the channel, height, and width of a feature map). In this study, the 3D tensor F is considered as a set of 2D features of size $h \times w$ as shown in Eq.(6):

$$F = X^i, i = \{1, .., c\} \quad (6)$$

Where:

X^i denotes the $i^{\text{-th}}$ activation map in F across the entire image (each feature map contains information about the entire image).

Each 2D feature X^i is flattened to obtain a single feature map $F \in R^{c \times n}$, where $n = h \times w$. These feature maps are fed into mixer forests with the same MLP structure. The mixer forests accept a set of flattened feature maps as input and add the following global relations to each $X^i \in F$:

$$X^i \leftarrow W_2\left(\sigma(W_1 X^i)\right) + X^i, i = \{1, ..., c\} \quad (7)$$

In our study, W^1 and W^2 are weights of the fully connected layers in MLP, while σ denotes the nonlinearity (ReLU). The inputs are connected back to generated projection via skip connections. This design helps the gradient flow and further improves its performance. For mixer forests, the overall capability of fully connected layers can be leveraged to automatically aggregate features, rather than just focus on local features, to achieve an effect similar to the attention mechanism. Instead of hierarchical (pyramidal) aggregation, the whole receptive field of the feature mixer means that each neuron sees the entire input image. A cascading block of mixer forests is used to incorporate spatial features into each feature map in iterative relationships.

For a given input $F \in R^{c \times n}$, the Mixer Forests (MF) produce an output of the same shape $Z \in R^{c \times n}$ (Z is of the same dimension as F). Since it has an isotropic structure, this output is used as an input to continue feeding it into next Mixer Forests (MF) until L consecutive blocks are produced, as shown in Eq.(8):

$$Z = MF_L(MF_{L \circledast 1}(...MF_1 F)) \quad (8)$$

The dimension of Z can be reduced by adding a fully connected layer to reduce its dimensional depth-wise (channel-wise). This layer acts as a weighted pooling operation that allows us to control the size of final global descriptor. Specifically, a depth-wise projection was applied that maps Z from $R^{c \times n}$ to $R^{d \times n}$ using Eq.(9):

$$Z = W_d(Transpose(Z)) \quad (9)$$

Where:

W_d weights of a fully-connected layer.

4. Evaluation Metrics

This section introduces the evaluation metrics to our VPR performance. These metrics are designed in terms of three perspectives: recognition capability, Memory Allocation Efficiency and quantitative effect of model binarization.

4.1 Recognition Capability

Recognition capability is evaluated by Recall and Recall@N. Both of them are the most used evaluation metrics in VPR community.

Recall (also known as true positive rate) can be used to evaluate the robustness of VPR against environmental

changes (e.g., seasons, weather, lighting, etc.). Its formula is as follows:

$$R = \frac{TP}{TP + FN} \tag{10}$$

Where:

R Recall

TP True-Positive, it represents a correctly identified place

FN False-Negative, it represents an incorrectly retrieved image based on ground-truth information.

A recall of 1.0 indicates that the VPR method has identified all positive instances correctly, while a recall of 0.0 means that the VPR method failed to identify any positive ones.

Recall@N tries to model the fact that a correctly retrieved reference image (as per the ground-truth) does not necessarily has to be the top-most retrieved image, but only needs to be among the Top-N retrieved images. The primary motivation behind this is that subsequent filtering steps, e.g. geometric consistency or weak GPS-prior, can be used to re-arrange the ranking of the retrieved images and avoid false-positives.

4.2 Memory Allocation Efficiency

VPR performance is also evaluated concerning memory requirements. Ferrarini (2022) proposed memory usage (η_m) by calculating the ratio of model size M_{size} to S_{p100}, as shown in Eq.(11), which indicates the memory costs of each S_{p100}. S_{p100} refers to the ratio of correct recognition places to ground-truth.

$$\eta_m = \frac{M_{size}}{S_{p100}} \tag{11}$$

Parameter η_m quantifies the memory cost per S-point under a fixed M_{size} (where M_{size} is held constant). A key observation is its inverse relationship with S_{p100}: as S_{p100} decreases, η_m increases proportionally, and conversely, larger S_{p100} values yield smaller η_m results. Notably, however, this inverse correlation fails to accurately capture the true memory utilization efficiency, as it neglects critical system-level factors such as memory fragmentation and cache coherence overhead that dynamically affect resource allocation.

Therefore, Memory Allocation Efficiency (MAE) $\eta_{m_{new}}$ is defined as the ratio of Recall@1 to model size:

$$\eta_{m_{new}} = \frac{Recall@1}{M_{size}} * 100 \tag{12}$$

Where:

$\eta_{m_{new}}$ is a comprehensive metric used to quantify the recall achieved by the model in unit of memory resource, where a larger value indicates a higher memory utilization efficiency. Specifically, at fixed Recall@1, larger models yield lower MAE; at fixed model size, higher Recall@1 increases MAE.

MAE takes memory resources as a core, which can intuitively quantify recognition performance level achieved by the model per unit of memory. By directly correlating memory consumption with recognition efficiency, this index clearly reflects the Memory Allocation Efficiency of the model for limited memory resources, and provides an intuitive quantitative basis for evaluating performance of a model for memory-sensitive tasks.

4.3 Quantitative Effect of Model Binarization

Existing metrics for quantitative effect of model binarization suffer from fragmented evaluation: model compression ratios measure memory savings, while accuracy loss quantifies performance degradation. To holistically assess binarization effect, the Balance Compression and Recall (*BCR*) metric was proposed, as shown in Eq. (13). It defines binarization effect by the proportion of model size reduction and the proportion of recall reduction.

$$BCR = \frac{\frac{M^F_{size} - M^B_{size}}{M^F_{size}}}{\frac{R^F@1 - R^B@1}{R^F@1}} \tag{13}$$

M^F_{size} is full-precision network model size. M^B_{size} is model size of *BNNs*. $R^F@1$ is Recall@1 of full precision network. $R^B@1$ is Recall@1 of *BNNs*. The *BCR* metric provides a superior quantitative measure of binarization effect, where higher values indicate more favorable binarization outcomes.

According to the strict definition of Eq. (13), *BCR* score classifies the binarization effect into four typical states by quantifying the relationship between model reduction and accuracy loss in the model binarization process:

⇒ when *BCR*=1, the model does not achieve effective improvement, showing a balance between model reduction and accuracy loss

⇒ when *BCR*<1, the model shows a negative improvement effect (accuracy loss more than model reduction gains)

$\Rightarrow$ when *BCR*>1, the model approach achieves positive improvement

$\Rightarrow$ when $BCR \rightarrow \infty$, the model achieves the theoretical optimal solution of the improvement effect.

BCR score comprehensively evaluates binarization effects on full-precision networks for *BNNs* rather than solely assessing model reduction or accuracy loss.

5. Experiments

A series of experiments were carried out to evaluate the performance of the proposed ForestsNet. First, the constructed binary neural networks was compared with some prevailing binary models, i.e. BinaryNet, FloppyNet, ShallowNet, DoReFaNet and XNOR-Net, to verify the recognition capability and binarization efficacy. Second, the proposed Mixer Forests was compared with a few typical conventional aggregation methods, i.e. AVG, GeM, CosPlaces, to validate the robustness of ForestsNet that resists environment changes. Then, multiple ablation experiments were conducted to further validate the design of ForestsNet. Finally, qualitative evaluation was presented. ForestsNet was built using TensorFlow and Larq (Bannink et al. 2020, Geiger et al. 2020,). Larq is a TensorFlow-based *BNNs* framework. In ForesetsNet, we found that when *L*=4 in Mixer Forests, it balances the performance of baseline network for ResNet18/34/50 (*L* is the number of stacked feature mixer blocks). Experimental datasets and baselines are detailed below.

5.1 Experimental Details

5.1.1 Datasets

We used four datasets for evaluation: Forests, Natural, Nordland, and MSLS. The Forests dataset contains forest, rainforest, and bamboo forest scenes, including about 30000 images. A lot images of Forests were collected by ourselves. In forest scenes, it is hard to recognize a place due to the varied shape, color, size of leaves or grass, leading to a significant variation in appearance. The natural dataset includes natural scenes from rivers, lakes, oceans, etc. It contains identified locations with drastic lighting and viewpoint changes. It is challenging and includes 48 place categories of about 200,000 images. Nordland (Zaffar et al. 2021) presents an exceptionally challenging benchmark. It was collected across four seasons via a camera mounted on the front of a train, it includes scenes spanning from snowy winters to sunny summers, featuring drastic variations in appearance. MSLS (Warburg et al. 2020) is a notable dataset collected through the use of car dashcams. What makes it stand out is that it encompasses a broad spectrum of viewpoint variations and significant changes in illumination. This diverse range of visual conditions captured by car dashcams makes MSLS a valuable resource for computer vision and related fields.

5.1.2 Training

ForestsNet was trained using all the above four datasets. The categorical cross entropy that was used as loss function has achieved excellent performance in VPR. Adaptive Moment Estimation (Adam) was used as it performs better than Stochastic Gradient Descent (SGD) in terms of *BNNs*, with a momentum of 0.9 and a learning rate of 0.0001. For training, 224×224 images with up to 100 iterations were used.

5.2 Comparison with Prevailing Binary Neural Networks

This work constructed a binarization backbone network to reduce model size while resisting accuracy degradation for resource-constrained forestry robots. Recall and *BCR* were utilized to compare the recognition capability and binarization efficacy with other *BNNs*, i.e. BinaryNet, ShowllowNet, FloppyNet, BiRealNet, XNOR-Net, and RealToBiNet. Their baseline networks were ResNet-18/34/50 or AlexNet architectures, and their recognition performance and model size are given in Table 1 for convenient comparison.The average Recall@1 is obtained by averaging the Recall@1 of four datasets.

5.2.1 Evaluation of Recognition Capability

Table 2 gives the Recall of *BNNs* on Forests and Natural datasets. According to the comparison between

Table 1 Recall and model size of baseline networks

Method \ Dataset	Forests	Natural Environment	Nordland	MSLS	Average Recall@1	Model Size [MiB]
	Recall@1	Recall@1	Recall@1	Recall@1		
AlexNet	71.64%	47.50%	10.35%	57.93%	46.85%	223.18
ResNet-18	76.81%	51.58%	33.57%	62.58%	56.13%	42.76
ResNet-34	77.14%	51.58%	36.17%	66.17%	57.77%	81.35
ResNet-50	77.64%	52.00%	37.42%	74.93%	60.50%	94.28

Table 2 Recall of *BNNs* on Forests and Natural datasets

Method \ Dataset	Baseline Network	Forests	Natural		
		Recall@1	Recall@1	Recall@5	Recall@10
BiRealNet-34	ResNet-34	65.40%	38.69%	73.31%	86.86%
BinaryNet	AlexNet	68.75%	35.90%	71.83%	84.48%
FloppyNet	AlexNet	73.79%	44.40%	77.35%	89.21%
ShallowNet	AlexNet	–	40.35%	74.21%	88.04%
DoReFaNet	AlexNet	68.33%	44.68%	78.31%	90.56%
XNOR-Net	ResNet-34	–	–	–	–
RealToBinNet-34	ResNet-34	–	26.83%	62.98%	80.02%
ForestsNet-18	ResNet-18	77.01%	51.86%	82.51%	91.24%
ForestsNet-34	ResNet-34	77.55%	52.76%	83.47%	92.58%
ForestsNet-50	ResNet-50	78.08%	53.13%	94.94%	94.50%

»–« is no data and represents gradients vanishing

Table 1 and 2 for Forests dataset, ForestsNet-50 achieved the highest Recall@1 at 78.08%, while BiRealNet-34 showed the lowest at 65.40% (11.74% below ResNet-34). ForestsNet-18/34/50 outperformed their ResNet-18/34/50 baselines by 0.5%. XNOR-Net, ShallowNet, and RealToBinNet experienced gradient vanishing during training. For Natural dataset, XNOR-Net suffered gradient vanishing. ForestsNet-18/34/50 improved over ResNet-18/34/50 baselines by 0.67%, 1.34%, and 0.75%, respectively, unlike other *BNNs* that showed performance degradation. ForestsNet-18/34/50 outperformed the other methods on Forests and Natural datasets.

Table 3 gives the Recall of *BNNs* on Nordland and MSLS datasets. According to the comparison between Table 1 and 3 for Nordland dataset, RealToBinNet-34 suffered gradient vanishing. BinaryNet showed the largest Recall@1 drop of 23% (which is compared with its baseline network AlexNet), while FloppyNet declined by 11.64%. ForestsNet-18/34/50 showed minimal losses of 0.93–1.17% (which is compared with ResNet-18/34/50). On MSLS, RealToBinNet-34 had the steepest decline in Recall@1 of 22.68%, while ForestsNet-18/34/50 dropped by 0.75–5.41% compared to their baselines. Therefore, on Nordland and MSLS, ForestsNet-18/34/50 perform better than other methods.

Across Forests and Natural datasets, ForestsNet outperformed ResNet by 0.26–1.18% in Recall@1. On Nordland and MSLS datasets, its Recall@1 loss of

Table 3 Recall of *BNNs* on Nordland and MSLS datasets

Method \ Dataset	Baseline Network	Nordland	MSLS		
		Recall@1	Recall@1	Recall@5	Recall@10
BiRealNet-34	ResNet-34	13.48%	60.46%	82.89%	88.63%
BinaryNet	AlexNet	7.35%	55.37%	71.49%	86.68%
FloppyNet	AlexNet	18.71%	65.88%	66.97%	79.36%
ShallowNet	AlexNet	8.04%	42.10%	61.99%	87.49%
DoReFaNet	AlexNet	17.49%	43.79%	63.75%	88.83%
XNOR-Net	ResNet-34	12.54%	47.11%	70.25%	88.49%
RealToBinNet-34	ResNet-34	–	22.68%	38.33%	58.33%
ForestsNet-18	ResNet-18	28.16%	80.97%	80.47%	87.40%
ForestsNet-34	ResNet-34	32.45%	81.26%	82.11%	89.26%
ForestsNet-50	ResNet-50	35.00%	83.42%	85.53%	90.43%

»–« is no data and represents gradients vanishing

0.93–1.17% was minimal compared to other *BNNs*. This superiority stems from its Mixer Forests design. The linear projection of Mixer Forests in its top layer further enhanced environmental adaptability, enabling consistent performance.

5.2.2 Evaluation of *BCR*

BCR, shown in Eq. (13), was used to compare the model compression trade-offs between ForestsNet and other *BNNs*. Based on the results of Recall@1 in 5.2.1, the model size reduction and recall reduction could be computed, so *BCR* could be obtained.

Table 4 shows the *BCR* of *BNNs*, where higher *BCR* indicates more favorable binarization outcome. From Table 4, it can be seen that ForestsNet outperformed other *BNNs* in average Recall@1. RealToBinNet-34 had the largest accuracy loss of 48.80%, while FloppyNet lost 8.65%. ForestsNet-18/34/50 showed minimal loss of 0.48%, 0.21%, and 0.57%. XNOR-Net showed the lowest *BCR* of 1.19, while FloppyNet reached 6.87. For ForestsNet-18/34/50, ForestsNet-34 achieved the highest *BCR* of 281.68, and ForestsNet *BCR* score was 20–152 times higher than that of other *BNNs*.The binarization of neural network causes significant information loss, leading to low recall. Owing to the proposed Leaky Sign function of ForestsNet, image features are preserved, while Mixer Forests aggregates them effectively, enabling robust handling of viewpoint/lighting changes and optimizing memory-recall balance.

Unlike the evaluation of standalone memory or accuracy metrics, *BCR* offers a comprehensive evaluation of *BNNs*. For instance, when comparing XNOR-Net and RealToBinNet-34 using separate metrics, conflicting results are obtained: XNOR-Net achieves 90.33% model reduction with 46.30% accuracy loss, while RealToBinNet-34 attains 96.79% model reduction at the cost of 48.80% accuracy loss. It demonstrates that XNOR-Net excels in accuracy preservation, whereas RealToBinNet-34 outperforms in model compression. Such discrepancies highlight the inadequacy of single metric in assessing overall effectiveness. In contrast, *BCR* integrates these trade-offs, revealing that RealToBinNet-34 slightly outperforms XNOR-Net in comprehensive compression performance. By harmonizing memory reduction and precision loss, *BCR* provides a robust framework for evaluating binarization methods across diverse baselines, thus advancing model compression research.

5.3 Comparison with Prevailing Aggregation Methods

This work proposed Mixer Forests to incorporate global information into local features to improve the robustness of VPR resisting environment changes. This section compares ForestsNet with a few prevailing aggregation methods in terms of recognition robustness and model allocation efficiency. The prevailing aggregation methods include AVG, GeM, and CosPlace. It is noticed that these aggregation methods are full-precision networks. The metrics used are Recall and MAE.

5.3.1 Evaluation of Recognition Capability

Table 5 gives the Recall of all aggregation methods on Forests and Natural datasets. From Table 5, it can be seen that ForestsNet outperforms other aggregation models across 18, 34, and 50 layers. For instance, ForestsNet-18 achieved 77.01% Recall@1 on Forests dataset, surpassing AVG-18 with 70.60%, Gem-18 with 73.79%, and CosPlace-18 with 73.10%. Similarly, on Natural dataset, ForestsNet-18 achieved the highest

Table 4 *BCR* of *BNNs*

Method	Average Recall@1(↑)	Model Size [MiB](↓)	Baseline Network	Model Reduction(↓)	Accuracy Loss(↓)	*BCR*(↑)
BiRealNet-34	44.51%	2.55	ResNet-34	96.74%	16.20%	3.65
BinaryNet	41.84%	8.55	AlexNet	96.17%	16.01%	3.60
FloppyNet	52.20%	0.44	AlexNet	99.08%	8.65%	6.87
ShallowNet	22.62%	6.67	AlexNet	97.01%	34.23%	1.69
DoReFaNet	42.50%	8.60	AlexNet	96.15%	14.35%	4.02
XNOR-Net	14.91%	7.87	ResNet-34	90.33%	46.30%	1.19
RealToBinNet-34	12.41%	2.61	ResNet-34	96.79%	48.80%	1.21
ForestsNet-18	59.90%	1.53	ResNet-18	96.42%	0.48%	120.48
ForestsNet-34	61.00%	2.73	ResNet-34	96.64%	0.21%	281.68
ForestsNet-50	62.41%	3.99	ResNet-50	95.77%	0.57%	105.82

Table 5 Recall of aggregation methods on Forests and Natural datasets

Method \ Dataset	Forests	Natural		
	Recall@1	Recall@1	Recall@5	Recall@10
AVG-18	70.60%	46.69%	80.35%	91.37%
GeM-18	73.79%	44.77%	78.60%	88.44%
CosPlace-18	73.10%	46.63%	80.25%	90.94%
ForestsNet-18	77.01%	51.86%	82.51%	91.24%
AVG-34	71.17%	47.31%	83.04%	93.17%
GeM-34	75.00%	43.32%	76.96%	88.90%
CosPlace-34	74.06%	47.19%	81.19%	91.81%
ForestsNet-34	77.55%	52.76%	83.47%	92.58%
AVG-50	73.18%	48.13%	84.25%	93.94%
GeM-50	73.79%	45.35%	80.04%	91.50%
CosPlace-50	75.78%	49.85%	83.25%	93.42%
ForestsNet-50	78.08%	53.13%	94.94%	94.50%

Recall@1/5/10 (51.86%/82.51%/91.24%), outperforming AVG-18, Gem-18, and CosPlace-18.

Table 6 shows the results on Nordland and MSLS datasets. For metric Recall@1 on both datasets, CosPlace series achieved a higher Recall@1 than ForestsNet series, while both series significantly outperformed the AVG and GeM series. This result is expected, since ForestsNet is binary neural network and CosPlace is full-precision network, meanwhile, there are many structured environments in these two datasets that ForestsNet excels at handling unstructured natural environments. Mixer Forests, which is a dense-layer feature aggregation method, enable ForestsNet to handle environmental variations better in natural scenarios by integrating global relations into local feature maps.

Table 6 Recall of aggregation methods on Nordland and MSLS datasets

Method \ Dataset	Nordland	MSLS		
	Recall@1	Recall@1	Recall@5	Recall@10
AVG-18	8.63%	67.32%	82.56%	83.55%
GeM-18	8.99%	74.95%	80.01%	91.29%
CosPlace-18	28.97%	81.75%	82.36%	89.03%
ForestsNet-18	28.16%	80.97%	80.47%	87.40%
AVG-34	10.64%	68.66%	84.35%	88.77%
GeM-34	15.74%	73.18%	82.21%	78.88%
CosPlace-34	33.95%	83.85%	83.36%	91.20%
ForestsNet-34	32.45%	81.26%	82.11%	89.26%
AVG-50	15.30%	70.09%	86.02%	90.01%
GeM-50	20.83%	76.53%	84.01%	79.30%
CosPlace-50	35.81%	84.32%	86.19%	92.57%
ForestsNet-50	35.00%	83.42%	85.53%	90.43%

5.3.2 Comparison of Memory Allocation Efficiency

MAE, shown in Eq. (12), was compared for ForestsNet, AVG, GeM, and CosPlace across 18-, 34-, and 50-layer networks in 4 datasets. Results are shown in Table 7. Average Recall@1 was calculated by averaging the Recall@1 values in 4 datasets. For model size, ForestsNet-18 is 1.53 MiB, much smaller than AVG-18/GeM-18/CosPlace-18 (45.80/44.40/45.81 MiB). ForestsNet-34 and ForestsNet-50 are 2.73 MiB and 3.99 MiB, just 3.21% and 4.68% of the size of GeM-34 and GeM-50. For MAE, ForestsNet-18/34/50 achieves the highest η_{m} with 38.89, 22.34, and 15.64. Moreover, its value decreases with layers increasing. Thus, the results show that our method is superior in Memory Allocation Efficiency performance. This is due to the fact that Leaky Sign binarization shrinks the network size, suiting the storage needs of deep learning, and Mixer Forests combine binary network backbone features and integrate global features into each map, boosting environmental change robustness.

Table 7 MAE of aggregation methods

Method	Average Recall@1	Model Size [MiB]	Our proposed MAE $\eta_{\mathrm{m_{new}}}(\uparrow)$
AVG-18	48.31%	45.80	1.05
GeM-18	50.63%	44.40	1.14
CosPlace-18	57.61%	45.81	1.26
ForestsNet-18	59.50%	1.53	38.89
AVG-34	49.45%	86.04	0.57
GeM-34	51.81%	85.24	0.61
CosPlace-34	61.76%	86.24	0.76
ForestsNet-34	61.00%	2.73	22.34
AVG-50	51.68%	98.44	0.53
GeM-50	54.13%	94.28	0.57
CosPlace-50	63.19%	98.44	0.64
ForestsNet-50	62.41%	3.99	15.64

5.4 Ablation Experiments

In this section, we conduct multiple ablation experiments to further validate the design of ForestsNet. First, the combination of Leaky Sign and DoReFaNet's weight function was validated. Then, the number of aggregation layers and dimensions of descriptors in ForestsNet were determined through a series of experiments, thus optimizing the network structure.

5.4.1 Binarization Function

Phase-wise comparative experiments were conducted to evaluate the performance of Leaky Sign. First, on the Forests dataset, we assessed the Recall@1 performance of three activation functions – Approximate Sign, STE Sign, and DoReFaNet Activation – when paired with various binarization weighting functions, aiming to identify the best one. Subsequently, leveraging the identified best binarization weighting function, an extended comparative analysis was performed of Recall@1 between Leaky Sign and other mainstream binarization activation functions.

In experiments, BiResNet-18 was used as backbone network. Five binarized activation functions were compared, including Leaky Sign, Leaky tanh (Geiger et al. 2020), Approximate sign (Liu et al. 2018), STE sign, and DoReFaNet activation functions (Zhou et al. 2016). Three binarization weight functions were used, namely STE sign, Magnitude Aware Sign (Liu et al. 2018), and DoReFaNet's Weight function. The results are shown in Table 8. Experiments were divided into 4 groups. In the first to third group, STE Sign, Magnitude Aware Sign and DoReFaNet's Weight were used as binarization weight function, respectively, and STE Sign, Approx Sign, DoReFaNet Activation were used as binarization activation function in each group for comparison. It can be seen that DoReFaNet Activation obtains the highest Recall@1 in each group with 71.50%, 73.08%, and 75.25%. Considering the DoReFaNet Activation as activation function, DoReFaNet's Weight achieves the highest value with 75.25%. Thus, in the first three groups, the combination of DoReFaNet Activation and DoReFaNet's Weight function was the best one. Finally, in the 4th group, Leaky tanh and Leaky Sign were used as activation function, and combined with DoReFaNet's Weight function. Results in the 3rd and 4th group show that the highest Recall@1 was achieved with the combination of Leaky sign and DoReFaNet's Weight function, namely 76.00%. Thus, our proposed Leaky Sign function had higher accuracy in recognition performance than other binarization activation functions.

The proposed Leaky Sign significantly improves the model's recognition ability due to its unique design, which can retain more information. DoReFaNet's activation function, Approx Sign, and STE Sign methods compressed input activation values to –1 and +1, leading to significant information loss. As a result, their performance was inferior to that of Leaky Tanh and Leaky Sign. In particular, Leaky Sign employed more binarization factors than Leaky Tanh, enabling it to preserve relevant information while removing irrelevant details and enhancing the binarization performance.

Table 8 Comparison of different combinations of binarization activation and weight functions

Group	Binarization Activation Function	Binarization Weight Function	Recall@1
1	STE Sign	STE Sign	66.50%
	Approx Sign	STE Sign	68.37%
	DoReFaNet Activation	STE Sign	71.50%
2	STE Sign	Magnitude Aware Sign	67.40%
	Approx Sign	Magnitude Aware Sign	71.75%
	DoReFaNet Activation	Magnitude Aware Sign	73.08%
3	STE Sign	DoReFaNet's Weight	68.00%
	Approx Sign	DoReFaNet's Weight	73.00%
	DoReFaNet Activation	DoReFaNet's Weight	75.25%
4	Leaky tanh	DoReFaNet's Weight	75.55%
	Leaky Sign	DoReFaNet's Weight	76.00%

5.4.2 Hyperparameters

To investigate the effect of mixer features, experiments were conducted to explore the number of L (L is the number of stacked feature mixer blocks) based on multiple backbone networks, i.e. BiResNet-18/34/50. First, BiResNet-18/34/50 was trained with L in 1–8, and the descriptor dimension was set to 20 (MixVPR setting), with the aim to find the best L.

Table 9 shows experimental results for different L. In ForestsNet, BiResNet18/34/50 was used as backbone network and Mixer Forests as aggregation layer. In this study, BiResNet-18/34/50 used Leaky Sign and DoReFaNet's weight function to binarize the network backbone of ResNet-18/34/50, while the shortcut connection of ResNet-18/34/50 was left unbinarized. The BiResNet-18 was the backbone network; when L=5, the Recall@1 reached the maximum of 74.79%, and when

Table 9 Recall@1 of BiResNet-18/34/50 with different L

Baseline Network / xL	BiResNet-18 Recall@1	BiResNet-34 Recall@1	BiResNet-50 Recall@1
1	73.33%	73.54%	70.83%
2	73.33%	71.46%	70.42%
3	72.92%	73.33%	70.63%
4	74.17%	73.33%	70.63%
5	74.79%	71.58%	70.21%
6	73.96%	73.75%	68.75%
7	73.13%	72.71%	68.75%
8	73.48%	71.48%	69.72%

L=4, the Recall@1 was 74.17%. When ForestsNet used BiResNet-34 as the baseline network, the Recall@1 reached the maximum of 73.75% when *L*=6, and the Recall@1 was 73.54% when *L*=1, achieving the balance of performance. When ForestsNet used BiResNet-50 as the backbone, the Recall@1 reached the highest value of 70.83% when *L*=1. When *L*=4, BiResNet-18/54 achieved the second-highest Recall@1, while BiResNet-34 ranked third. Despite the limited depth, its performance was competitive with the top result, demonstrating a favorable trade-off. Therefore, *L*=4 was selected for ForestsNet.

5.4.3 Descriptor Dimension

The structure of ForestsNet allows the dimension of output descriptor to be configured by the dimension of fully connected layer. After determining the fully connected layer dimensions as 256, 512, 1024, 2048, and 4096, ablation experiments were performed to determine the optimal dimension.

Table 10 shows the Recall@1 obtained with different descriptor dimensions for ForestsNet aggregation layer. In this experiment, BiResNet-18 with the dimension of 512 reaches the highest value of Recall@1 - 75.38%, and with the dimension of 1024, the second highest value of Recall@1 is reached. BiResNet-34 with the dimension of 1024 gets the highest Recall@1 - 76.75%, BiResNet-50 with the dimension of 1024 gets the highest Recall@1 - 76.08%. Therefore, a dimension of 1024 was selected for the descriptor.

Table 10 Different description dimensions of ForestsNet aggregation layer

Baseline Network / Descriptor dimension	BiResNet-18 Recall@1	BiResNet-34 Recall@1	BiResNet-50 Recall@1
4096	73.15%	72.50%	73.21%
2048	73.50%	73.75%	72.67%
1024	74.54%	76.75%	76.08%
512	75.38%	75.92%	74.88%
256	74.33%	75.54%	72.79%

5.5 Qualitative Results

Fig. 4 gives some image examples that highlight significant challenges, including variations in viewpoint, season, and illumination. The 1st column gives the images captured to query, and columns 2–6 present the retrieval results from five different methods, respectively. It can be seen that ForestsNet successfully retrieved all correct images, while other methods either retrieved similar images from different locations or identified places that were geographically close but still below the threshold, thus failing to retrieve all the correct results. Thus, this experiment demonstrated the superior performance of ForestsNet in forest environments, and showed strong robustness against these challenges.

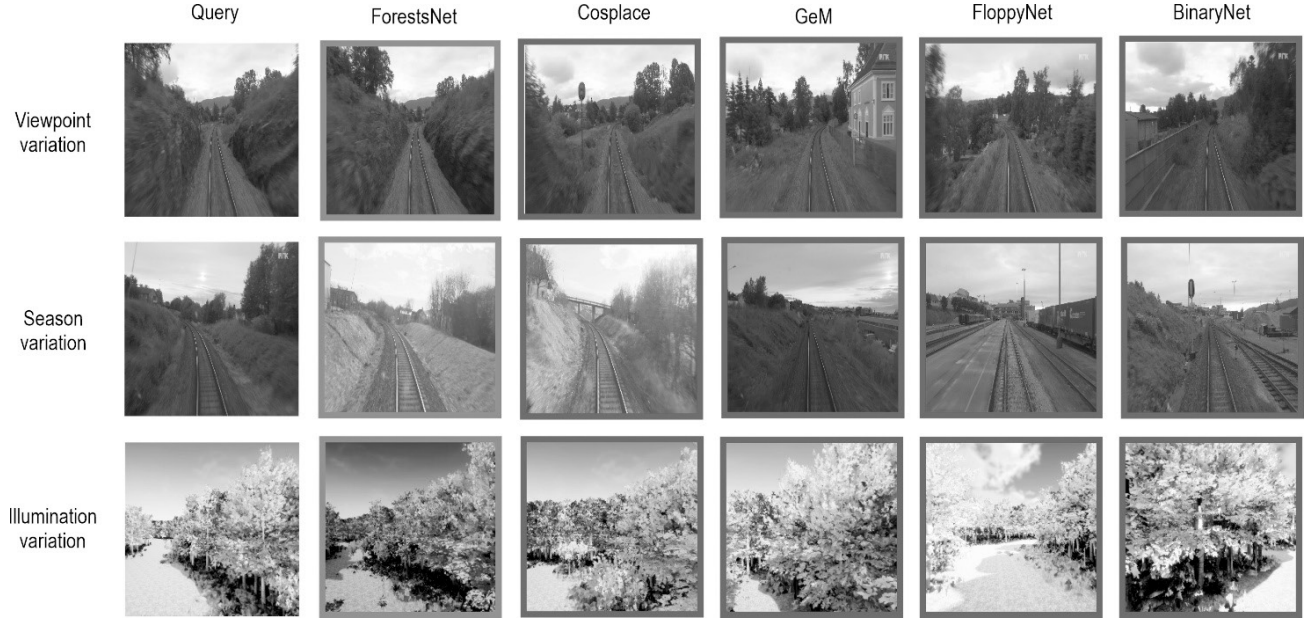

Fig. 4 Comparison of retrieval results on challenging forest environments

6. Discussion and Conclusion

This work proposed ForestsNet, a lightweight VPR model, to address problems of appearance variability and storage constraints that forestry robots confront. First, a Binary Neural Network (*BNN*) was constructed to reduce model size and resist accuracy degradation. Then, a novel multi-layer perceptron-based aggregation method, Mixer Forests, was introduced to resist changes of forest environments. Moreover, in order to better quantify the trade-off between memory efficiency and place recognition accuracy, two novel evaluation metrics, MAE and *BCR* were designed. A series of experiments were carried out to verify the performance of our proposed ForestsNet on 4 datasets (including public datasets and datasets captured by ourselves). Experimental results demonstrate that ForestsNet outperforms other prevailing *BNNs* in terms of recall and *BCR*. It stems from the proposed Leaky Sign function, which preserves more image features, and Mixer Forests aggregates them effectively, enabling robust handling of viewpoint/lighting changes and optimizing memory-recall balance. Moreover, ForestsNet was also compared with a few prevailing aggregation methods. Results show that ForestsNet exhibits notable competence in terms of recognition robustness and model allocation efficiency, and it validates the effectiveness of Mixer Forests in resisting challenging changing environments. Qualitative evaluation experiments demonstrate that ForestsNet is able to recognize the correct location in challenging forest environments.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62203059 and Grant 32071680.

7. References

Lowery, S., Sünderhauf, N., Newman, P., Leonard, J.J., Cox, D., Corke, P., Milford, M.J., 2016: Visual place recognition: A survey. IEEE Transactions on Robotics 32(1): 1–19. https://doi.org/10.1109/TRO.2015.2496823

Li, P., Wen, S., Xu, C., Qiu, T.Z., 2024: Visual Place Recognition for Opposite Viewpoints and Environment Changes. IEEE Transaction on Instrumentation Measurement 73: 1–9. https://doi.org/10.1109/TIM.2024.3350152

Chen, J., Xie, F., Huang, L., Yang, J., Liu, X., Shi, J.A., 2022: Robot Pose Estimation Optimized Visual SLAM Algorithm Based on CO-HDC Instance Segmentation Network for Dynamic Scenes. Remote Sensors 14(9): 2114. https://doi.org/10.3390/rs14092114

Zhao, J., Zhai, Q., Zhao, P., Huang, R., Cheng, H., 2023: Co-Visual Pattern-Augmented Generative Transformer Learning for Automobile Geo-Localization. Remote Sensors 15(9): 2221. https://doi.org/10.3390/rs15092221

Chen, C., Wang, B., Lu, C.X., Trigoni, N., Markham, A., 2023: Deep Learning for Visual Localization and Mapping: A Survey. IEEE Transaction on Neural Network and Learning Systems 35(12): 17000–17020. https://doi.org/10.1109/TNNLS.2023.3309809

Lee, J., Bahn, H., 2022: Analyzing Memory Access Traces of Deep Learning Workloads for Efficient Memory Management. 12th International Conference on Information Technology in Medicine and Education (ITME), Xiamen, China, Nov. 18, 389–393. https://doi.org/10.1109/ITME56794.2022.00090

Fan, C., Zhou, Z., He, X., Fan, Y., Zhang, L., Wu, X., Hu, X., 2022: Bio-Inspired Multisensor Navigation System Based on the Skylight Compass and Visual Place Recognition for Unmanned Aerial Vehicles. IEEE Sensors Journal 22(15): 15419–15428. https://doi.org/10.1109/JSEN.2022.3187052

Ferrarini, B., Waheed, M., Waheed, S., Ehsan, S., Milford, M., McDonald-Maier, K.D., 2019: Visual place recognition for aerial robotics: Exploring accuracy-computation trade-off for local image descriptors. 2019 NASA/ESA Conference on Adaptive Hardware and Systems,Colchester, United Kingdom, 22 July 103–108. https://doi.org/10.1109/AHS.2019.00011

Ferrarini, B., Milford, M.J., McDonald-Maier, K.D., Ehsan, S., 2022: Binary Neural Networks for Memory-Efficient and Effective Visual Place Recognition in Changing Environments. IEEE Trans. Robot 38(4): 2617–2631. https://doi.org/10.48550/arXiv.2010.00716

Courbariaux, M., Bengio, Y., 2016: BinaryNet: Training deep neural networks with weights and activations constrained to 1 or-1. Available online: https://arxiv.org/abs/1602.02830. (accessed on April 5, 2024)

Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J., 2016: NetVLAD:CNN Architecture for Weakly Supervised Place Recognition. 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 26, 5297–5307. https://doi.org/10.1109/TPAMI.2017.2711011

Radenović, F., Tolias, G., Chum, O., 2018: Fine tuning CNN image retrieval with no human annotation. IEEE Transaction on Pattern Analysis and Machine Intelligence 41(7): 1655–1668. https://doi.org/10.1109/TPAMI.2018.2846566

Berton, G., Masone, C., Caputo, B., 2022: Rethinking visual geo-localization for large-scale applications. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, June 21, 4868–4878. https://doi.org/10.1109/CVPR52688.2022.00483

Ali-Bey, A., Chaib-Draa, B., Giguére, P., 2023: MixVPR:Feature Mixing for Visal Place Recognition 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV),Waikoloa, HI, USA, January 3, 2997–3006. https://doi.org/10.1109/WACV56688.2023.00301

Daou, A., Pothin, J.-B., Honeine, P., Bensrhair, A., 2023: Indoor Scene Recognition Mechanism Based on Direction-Driven Convolutional Neural Networks. Sensors 23(12): 5672. https://doi.org/10.3390/s23125672

Heidari, M.J., Najafi, A., Alavi, S., 2018: Pavement Deterioration Modeling for Forest Roads Based on Logistic Regression and Artificial Neural Networks. Croatian Journal of Forest Engineering 39(2): 271–287.

Proto, A.R, Sperandio, G., Costa, C., Maesano, M., Antonucci, F., Macrì, G., Mugnozza, G.S., Zimbalatti, G., 2020: A Three-Step Neural Network Artificial Intelligence Modeling Approach for Time, Productivity and Costs Prediction. Croatian journal of forest engineering 41(1): 35–47. https://doi.org/10.5552/crojfe.2020.611

Petrakis, G., Partsinevelos, P., 2023: Keypoint Detection and Description through Deep Learning in Unstructured Environments. Robotics 12(5): 137. https://doi.org/10.3390/robotics12050137

Yan, F., Gong, Y., Feng, Z., 2015: Combination of Artificial Neural Network with Multispectral Remote Sensing Data as Applied in Site Quality Evaluation in Inner Mongolia. Croatian Journal of Forest Engineering 36(2): 307–319.

Zhang, J., Chen, L., Shi, R., Li, Y., 2025: Detection of bruised apples using structured light stripe combination image and stem/calyx feature enhancement strategy coupled with deep learning models. Agriculture Communications 3(1): 2949–7981. https://doi.org/10.1016/j.agrcom.2025.100074

Yu, H., Wang, Q., Yan, C., Feng, Y., Sun, Y., Li, L., 2024: DLD-SLAM: RGB-D Visual Simultaneous Localisation and Mapping in Indoor Dynamic Environments Based on Deep Learning. Remote Sensors 16(2): 246. https://doi.org/10.3390/rs16020246

Babenko, A., Lempitsky, V., 2015: Aggregating local deep features for image retrieval. 2015 IEEE/CVF International Conference on Computer Vision (CVPR), Boston, USA, June 7, 1269–1277. https://doi.org/ 10.1109/ICCV.2015.150

Tolias, G., Sicre, R., Jégou, H., 2015: Particular object retrieval with integral max-pooling of cnn activations. Available online https://arXiv:1511.05879 (accessed on 5 April 2024).

Häusler, S., Garg, S., Xu, M., Milford, M., Fischer,T., 2021: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 24 June, 14141–14152. https://doi.org/10.1109/CVPR46437.2021.01392

Yang, Y., Ma, B., Liu, X., Zhao, L., Huang, S., 2021: GSAP: A Global Structure Attention Pooling Method for Graph-Based Visual Place Recognition. Remote Sensors 13(8): 1467. https://doi.org/10.3390/rs13081467

Wang, R., Shen, Y., Zuo, W., Zhou, S., Zheng, N., 2022: Transvpr: Transformer-based place recognition with multi-level attention aggregation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, June 21, 13648–13657. https://doi.org/10.1109/CVPR52688.2022.01328

Wang, C., Chen, S., Song, Y., Xu, R., Zhang, Z., Zhang, J., Yang, H., Zhang, Y., Fu, K., Du, S., Xu, Z., Gao, L., Guo, L., Xu, S., 2025: Focus on Local: Finding Reliable Discriminative Regions for Visual Place Recognition. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) 39(7): 7536–7544. https://doi.org/10.1609/aaai.v39i7.32811

Han, S., Pool, J., Tran, J., Dally, W.J., 2015: Learning both weights and connections for efficient neural networks. Neural Information Processing Systems (2015), Montreal, Canada, 7 December, 1135–1143.

Courbariaux, M., Bengio, Y., David, J.-P., 2014: Training deep neural networks with low precision multiplications. Available online: https://arxiv.org/abs/1412.7024. https://doi.org/10.48550/arXiv.1412.7024

Simons, T., Lee, D.-J., 2019: A review of binarized neural networks. Electronics 8(6): 661. https://doi.org/10.3390/electronics8060661

Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A., 2016: Xnor-Net: ImageNet classification using binary convolutional neural networks. European Conference on Computer Vision 2016 (ECCV 2016), Amsterdam, the Netherlands, October 10, 525–542. https://doi.org/10.1007/978-3-319-46493-0_32

Liu, Z., Wu, B., Luo, W., Yang, X., Liu, W., Cheng, K.-T., 2018: Bi-real net:enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. Europeon Conference on Computer Vision 2018 (ECCV 2018) Munich, Germany, Sept 18, 722–737. https://doi.org/10.1007/978-3-030-01267-0_44

Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., Zou, Y., 2016: Dorefa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients. Available online:https://arxiv.org/abs/1606.06160. https://doi.org/10.48550/arXiv.1606.06160

Esser, S.K., McKinstry, J.L., Bablani, D., Appuswamy, R., Modha, D.S., 2020: Learned step size quantization. 2020 International Conference on Learning Representations (ICLR), Millennium Hall,Addis Ababa, Ethiopia, April 26.

Li, F., Liu, B., 2016: Ternary weight networks. Available online: https://arXiv:1605.04711 (accessed on April 5, 2024). https://doi.org/10.48550/arXiv.1605.04711

Zhu, C., Han, S., Mao, H., Dally, W.J., 2017: Trained ternary quantization. IEEE International Conference on Robotics and Automation (2017), Marina Bay Sands, Singapore, May 29. https://doi.org/10.48550/arXiv.1612.01064

Hubara, L., Courbariaux, M., Soudry, D., El-Yaniv, R., Bengio, Y., 2017: Quantized neural networks: Training neural networks with low precision weights and activations. Journal of Machine Learning Research 18(187): 6869–6898.

Lin, X., Zhao, C., Pan, W., 2017: Towards accurate binary convolutional neural network. Neural Information Processing

Systems (2017) Long Beach, CA, USA, Dec 4, 344–352. https://doi.org/arxiv-1711.11294

Phan, H., Huynh, D., He, Y., Savvides, M., Shen, Z., 2020: MoBiNet: A Mobile Binary Network for Image Classification.2020 IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, Aspen, CO, USA, March 1, 3442–3451. https://doi.org/10.1109/WACV45572.2020.9093444

Dong, P., Li, L., Zhong, Y.D., Du, D.Y., Fan, R., Chen, Y.H., Tang, Z.H., Wang, Q., Xue, W., Guo, Y.K., Chu, X.W., 2025: Breaking the 1-Bit Barrier with Structured Binary LLMs. 2025 International Conference on Learning Representations (ICLR), Online, April 26.

Ding, R., Chin, T.-W., Liu, Z., Marculescu, D., 2022: Regularizing activation distribution for training binarized deep networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Long Beach, CA, USA, June 16, 11400–11409. https://doi.org/10.1109/CVPR.2019.01167

He, K., Zhang, X., Ren, S., Sun, J., 2016: Deep residual learning for image recognition.2016 IEEE/CVF IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 26, 770–778.

Geiger, L., Team, P., 2020: Larq: An open-source library for training binarized neural networks. Journal of Open Source Software 5(45): 1746–1750. https://doi.org/10.21105/joss.01746

Bannink, T., Bakhtiari, A., Hillier, A., Geiger, L., de Bruin, T., Overweel, L., Neeven, J., Helwegen, K, 2020: Larq compute engine Design, benchmark,and deploy state-of-the-art binarized neural networks. Available online: https://arxiv.org/abs/2011.09398. https://doi.org/10.48550/arXiv.2011.09398

Zaffar, M., Garg, S., Milford, M., Kooij, J., Flynn, D., McDonald--Maier, K., 2021: Ehsan.S.Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. International Journal of Computer Vision 129(7): 2136–2174. https://doi.org/10.1007/s11263-021-01469-5

Warburg, F., Hauberg, S., López-Antequera, M., Gargallo, P., Kuang, Y., Civera, J., 2020: Mapillary street-level sequences:A dataset for lifelong place recognition. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Virtual, June 16, 2626–2635. https://doi.org/10.1109/CVPR42600.2020.00270

Authors' addresses:

Junshuai Wang
e-mail: junshuai1043@bjfu.edu.cn
Junyu Han, PhD
e-mail: hanjun0801@bjfu.edu.cn
Prof. Ruifang Dong, PhD *
e-mail: ruifang_dong@bjfu.edu.cn
Prof. Jiangming Kan, PhD
e-mail: kanjm@bjfu.edu.cn
Beijing Forestry University
School of Technology
No. 35 Qinghua East Road, Haidian District
100083, Beijing
CHINA

* Corresponding author

Received: April 07, 2024
Accepted: February 12, 2026