

Received 17 October 2023, accepted 2 November 2023, date of publication 8 November 2023,
date of current version 17 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3331371

APPLIED RESEARCH

Fast Dual-Feature Extraction Based on Tightly Coupled Lightweight Network for Visual Place Recognition

XIAOFEI HU, YANG ZHOU^{id}, LIANG LYU^{id}, CHAOZHEN LAN^{id},
QUNSHAN SHI, AND MINGBO HOU

Institute of Geospatial Information, PLA Strategic Support Force Information Engineering University, Zhengzhou, Henan 450001, China

Corresponding author: Yang Zhou (zhouyang3d@163.com)

This work was supported by the National Natural Science Foundation of China under Grant 42001338.

ABSTRACT Visual place recognition (VPR) is a task that aims to predict the location of an image based on the existing images. Because image data can often be massive, extracting features efficiently is critical. To solve the problems of model redundancy and poor time efficiency in feature extraction, this study proposes a fast dual-feature extraction method based on a tightly coupled lightweight network. The tightly coupled network extracts local and global features in a unified model which has a lightweight backbone. Learned step size quantization is then performed to reduce the computational overhead in the inference stage. Additionally, an efficient channel attention module ensures feature representation ability. Efficiency and performance experiments on different hardware platforms showed that the proposed algorithm incurred significant runtime savings for feature extraction, and the inference was 2.9–4.0 times faster than that in the general model. The experimental results confirmed that the proposed method can significantly improve VPR efficiency while ensuring accuracy.

INDEX TERMS Visual place recognition, dual-feature extraction, tightly coupled, learned step size quantization.

I. INTRODUCTION

Visual place recognition (VPR) aims to estimate the location of a given image based on a group of existing images, and it is widely used in robotics, social media, and computer vision [1]. Central to this problem is the representations used to describe images. There are two types of image representations: global and local features. Global features are better at recall, whereas local features are used for precision processing [2]. Owing to the complementarity of the two feature types, the dual-feature combination method has become a mainstream method in VPR [3], [4], [5]. For example, in coarse localization, global features are first used for recall, after which local features are used for geometric verification [6]. Similarly, in fine localization, global features are required for fast retrieval, after which local features are used

for precise 6-degree-of-freedom (6-DOF) localization [7]. Because the image data are often massive, the efficiency of feature extraction is critical along with limited computational resources. In this work, we aim to seek a fast dual-feature extraction technique that is also robust for VPR tasks.

In the last decade, researchers have attempted to apply several dual-feature combination methods for VPR [8], [9], [10]. However, most require the extraction of local and global features separately with different models. This is undesirable as both require similar feature extraction computations, resulting in redundant processing and unnecessary complexity, thereby increasing memory usage and latency. Recently, researchers attempted to extract global and local features jointly in a unified model. However, the focus was on improving accuracy, while execution efficiency was ignored. In practical applications, the terminals for visual localization may be devices with low computational performance, such as mobile phones and robots. It is difficult to apply these feature

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo^{id}.

extraction methods to low-performance devices. Therefore, improving the efficiency of feature extraction while ensuring accuracy is important.

This study proposes a fast extraction method based on a tightly coupled lightweight model to solve the problem of poor time efficiency in dual-feature extraction. First, a tightly coupled network is designed to extract local and global features in a unified model to reduce redundancy, and the ordinary convolution operation is replaced by a lightweight backbone to improve the network efficiency. We leverage the learned step size quantization (LSQ) aware training method, which reduces the computational overhead in the inference stage, to further enhance the efficiency of this approach. In addition, an efficient channel attention (ECA) module ensures feature representation ability. The main contributions of this study are as follows:

- 1) A tightly coupled dual-feature extraction lightweight network model is proposed. Local and global features share model parameters that can reduce redundancy in feature extraction. A lightweight backbone network is developed, and the ECA module is added to improve the network performance. Furthermore, the LSQ aware training method is adopted to decrease the computational time. The experiment proves that the proposed method can significantly improve the efficiency while ensuring accuracy for VPR.
- 2) The proposed method shows excellent performance in VPR-related tasks. In terms of feature extraction, our method is 2.9–4.0 times faster than traditional methods. The method improves execution efficiency and exhibits good performance in multiple tasks. Compared to traditional methods, the mean average precision (mAP) and matching score decreased by only 0.2% and 0.4% in image retrieval and local feature point matching, respectively.

We organize the rest of this paper as follows. In Section II, we introduce related works on dual-feature extraction. Section III introduces the proposed framework in detail. The details and results of the extensive experiments are provided in Section IV. Finally, the limitations and conclusions are presented in Section V.

II. RELATED WORK

VPR refers to the process of identifying and obtaining the geographic location of a given query image in a pre-constructed image database. The core of the VPR task lies in the manner to effectively describe the images with global and local features. Taken into account practicality, the resource consumption of features generation efficiency is also an important indicator.

In earlier VPR systems, hand-crafted techniques such as SIFT [11] were used to extract local features, followed by methods such as VLAD [12], [13], FV [14] to encode local features and obtain the global feature vector image representation. Recently, owing to the advantages of deep learning techniques, deep learning-based feature extraction

methods [15], [16] have shown superior performance in VPR. The dual-feature extraction methods can be classified into tightly coupled and loosely coupled, depending on the coupling degree of dual features. Loose coupling refers to using independent models or methods to extract local and global features. Fang et al. [17] used NetVLAD [18] global features to search for images at the map level and Geodesc [8] local features to match the images with the retrieved 3D points for visual localization. The loosely coupled method uses mutually independent feature extraction modules, inevitably increasing the complexity of training and execution. Our work leverages a tightly coupled feature extraction method that generates global and local features from the same model to reduce feature redundancy.

Existing dual-feature extraction methods use complex models to improve accuracy, but the execution efficiency remains poor. Yang et al. [19] and Noh et al. [20] used a complex ResNet [21] network to perform dual-feature extraction and achieved good performance in large-scale image retrieval and VPR tasks. While these methods achieve better performance, they rely on a large backbone, and they are difficult to apply to low-performance devices. In practical applications, the end terminals for visual localization may be devices with low computing performance, such as mobile phones and robots. Therefore, it is crucial to improve the efficiency of feature extraction. In recent years, many deep neural network architectures have been designed for an optimal trade-off between accuracy and efficiency [22], [23], [24], [25], [26]. Some works apply lightweight network to VPR systems [27], [28], [29], [30], [31]. LSDNet [27] utilizes a dual-distillation method to efficiently generate global feature for describing images. Zaffar et al. [28] proposed a local features extraction method based on oriented gradient histograms and achieved a balance between memory usage and execution efficiency. However, these methods generate either global or local features. In dual-feature applications, an additional module is required. Sarlin et al. [29] designed the HF-NET for dual-feature extraction based on a lightweight network which supervised by the VGG16 network. However, it did not adequately learn the intrinsic patterns of the teacher model, which resulted in a decrease in accuracy, and the extraction efficiency also requires further improvement. Therefore, the proposed method uses ECA module to improve performance and uses the quantization method to considerably improve the efficiency of network models. Quantization methods such as DFQ [32], LSQ [33], and LSQ+ [34] not only reduce the memory size but also accelerate the inference speed of the model, and hence, have been applied in other fields [35]. To date, little literature has explored VPR from this perspective.

III. PROPOSED METHOD

Our goal is to improve the efficiency of feature extraction and reduce time consumption while maintaining good feature representation. To achieve this, we first design a tightly

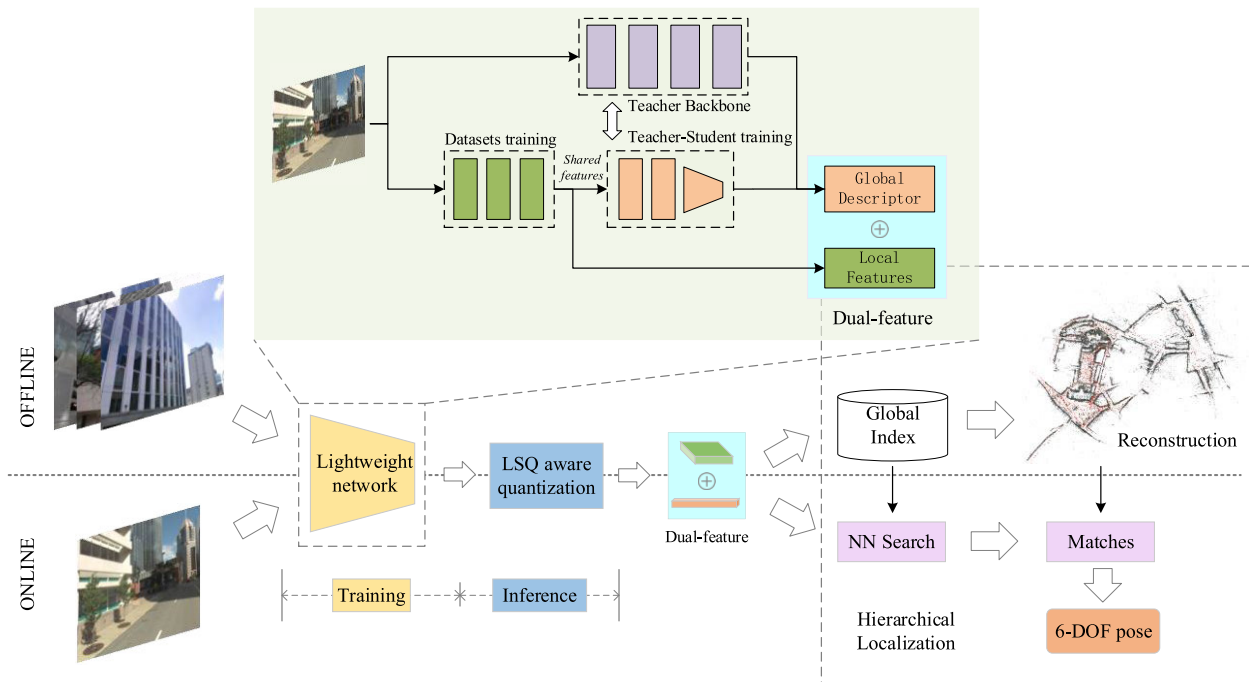


FIGURE 1. Algorithm flow of the proposed method. A combination of datasets and Teacher–Student training is used to train the lightweight network. We implement effective lightweight operations in both the training and inference stages.

coupled network to extract the dual feature comprising global and local features of the input image. These two features will be used together in the subsequent visual localization process. Then, the LSQ method performs quantization-aware training on the network in the inference stage. A combination of datasets and Teacher–Student training was used for network training. Our VPR system is loosely based on the hierarchical localization framework. Fig. 1 shows the algorithm flow of the tightly coupled feature extraction method.

A. TIGHTLY COUPLED LIGHTWEIGHT BACKBONE

To improve the efficiency of feature extraction, we designed a lightweight backbone network. Among existing lightweight backbone networks, MobileNets [25] proposed a simple model structure that showed good performance in target detection and image classification tasks. Inverted residuals and linear bottlenecks are the main methods used by MobileNetV2 to achieve lightweight models. MobileNetV3 optimizes the activation function and adds the channel attention model, which improves the feature extraction performance and convolutional operational efficiency. The lightweight design is mainly referred to MobileNetV3. To achieve lightweight local and global feature extraction, we replaced ordinary convolutions with bottleneck convolutions to reduce the amount of computation and model parameters. A 5-stage backbone network was set up, where the first layer used an ordinary convolution, while the remaining layers used the bottleneck convolution. We used the *h-swish* activation function in the first convolution layer by

TABLE 1. Backbone architecture.

Input	Operator	EF	C	NL	L	S
$1 * H * W$	Conv2d	-	16	HS	1	2
$16 * H/2 * W/2$	bottleneck	1	32	RE	1	1
$32 * H/2 * W/2$	bottleneck	3	64	RE	2	2
$64 * H/4 * W/4$	bottleneck	2	128	RE	2	2
$128 * H/8 * W/8$	bottleneck	2	256	RE	1	1

Each line describes a sequence of one or more identical layers, repeated L times. C denotes the output channels, EF denotes the expansion factor, NL denotes the type of nonlinearity used, and S denotes the stride.

referring to MobileNetV3. We applied the channel attention mechanism to global feature extraction and used a lighter ECA module. Table 1 shows the specific structure of the backbone.

The lightweight backbone ultimately produces a middle-level image representation. To generate the dual features, we design a feature extraction architecture composed of three prediction heads: i) key points, ii) local feature descriptors, and iii) a global image-wide descriptor. The global descriptor is computed from aggregating local feature maps, which might be useful for predicting local features. Fig. 2 shows the proposed dual feature extraction architecture.

B. LEARNED STEP SIZE QUANTIZATION

Based on a lightweight backbone, the LSQ method is further used to improve the inference speed of the network. Quantization is one of the most impactful ways to decrease

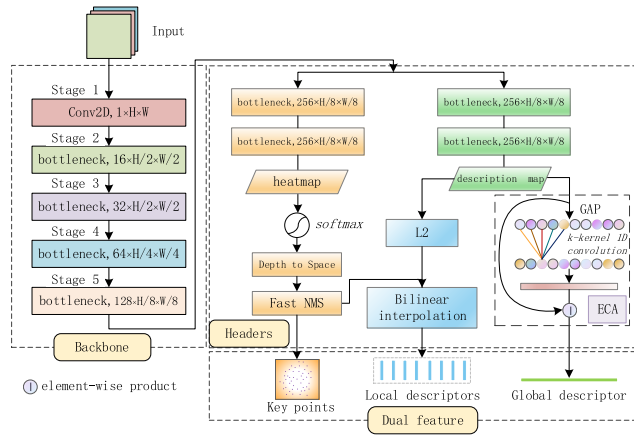


FIGURE 2. Tightly coupled local-global features extraction network. The global descriptor, key points, and local descriptors share the same encoder, which can effectively reduce model redundancy.

the computational time and energy consumption of neural networks [36]. The key to quantization is to close the gap to full-precision accuracy since low bit-width quantization introduces noise to the network, which can lead to an accuracy drop. The LSQ algorithm, a quantization-aware training method, learns the step size parameters through backpropagation and obtains the scaling factors for the weight and activation for each layer through training, which further improves the efficiency while reducing the quantization loss. In particular, we first convert the float values to integers with low bit width, in which case quantization can be expressed as

$$\hat{x} = q(x; s, z, n, p) = s \times \left(\text{clamp} \left(\left\lfloor \frac{x}{s} \right\rfloor + z; n, p \right) - z \right) \quad (1)$$

where \hat{x} is the output after the quantization operation, $q(\cdot)$ is the quantization operation, x is the input floating data, s is the quantization scale factor, z is the zero point, and (n, p) is the range of quantization truncation.

In general, assuming b to be the quantization bit width for signed quantization, the value range of (n, p) is $(-2^{b-1}, 2^{b-1} - 1)$, where $\lfloor \cdot \rfloor$ is the round-to-nearest operator and clamping is defined as

$$\text{clamp}(x; n, p) = \begin{cases} x, & n \leq x \leq p \\ n, & x < n \\ p, & x > p \end{cases} \quad (2)$$

where s and z , known as quantization parameters, are the key to quantization, considering that the noise impacts the quantization result. Pseudo-quantization operations, where the quantized weight and activation are used in forward propagation whereas gradients update floating-point weights during backpropagation, should be used during the training process. The key to pseudo-quantization is computing the pseudo-gradients. Backward propagation in neural networks is based on the chain rule for derivatives. However, the quantization operation is a step function whose derivative does not

exist. Therefore, training the quantization parameters can be challenging. We leverage the STE method [37] to calculate the pseudo-gradients. The quantization parameters gradient is given as follows

$$\frac{\partial \hat{x}_i}{\partial s} = \frac{\partial}{\partial s} \left(s \times \text{clamp} \left(\left\lfloor \frac{x_i}{s} \right\rfloor; n, p \right) \right) = \begin{cases} \frac{-x_i}{s} + \left\lfloor \frac{x_i}{s} \right\rfloor, & Q_{\min} \leq x_i \leq Q_{\max} \\ n, & x_i < Q_{\min} \\ p, & x_i > Q_{\max} \end{cases} \quad (3)$$

$$\frac{\partial \hat{x}_i}{\partial z} = \begin{cases} 0, & Q_{\min} \leq x \leq Q_{\max} \\ -s, & \text{otherwise} \end{cases} \quad (4)$$

where (Q_{\min}, Q_{\max}) represents the dequantization value range obtained from the current quantization parameters.

C. FEATURES GENERATION WITH ECA ATTENTION

Relevant work has proved that attention can enhance performance; hence, we used ECA [38], a lightweight channel attention module based on SE-NET, to ensure performance while reducing the feature extraction time. Avoiding dimensionality reduction and cross-channel interaction is the key to improving the performance of channel attention. ECA proposes a local cross-channel interaction strategy without dimensionality reduction, which can be efficiently implemented via one-dimensional (1D) convolution, and adaptively selects the kernel size of convolution to increase the interaction ability of local cross-channels. Compared to the fully connected network module, 1D convolution maintains attention performance and reduces the complexity of the model.

The input feature map $\chi \in \mathbb{R}^{W \times H \times C}$, where $\{W, H, C\}$ represents the dimensions of the feature map, first undergoes a global average pooling operation according to the channel dimension. Then, it integrates the spatial information to obtain a feature representation. Furthermore, adaptive k -core 1D convolution is performed on this representation tensor. k is calculated as

$$k = \psi(C) = \left\lfloor \frac{\log_2(c)}{\gamma} + \frac{b}{\gamma} \right\rfloor_{\text{odd}} \quad (5)$$

where $\lfloor \cdot \rfloor_{\text{odd}}$ represents the nearest odd number and $\gamma = 2, b = 1$ are the hyperparameters. Additionally, the method adapts convolution kernel size for different channel dimensions.

Normalized attention weights are obtained using the activation function. The original feature map is then multiplied by the weight vector to enhance attention. This module is inserted into the descriptor extraction process to obtain more representative global features. The performance is guaranteed considering there is no dimensionality reduction and correlation between channels. Additionally, 1D convolution with fewer parameters guarantees efficiency.

For local features, two decoders are used for feature point detection and descriptor generation. The feature point detection decoder uses convolution operations to generate a feature

map. The *Softmax* function was used to normalize the feature map to compute the distribution probability. Because the feature map size was 1/8 of the original, it was necessary to up-sample the feature map to the full image size. We leveraged sub-pixel convolution to efficiently reconstruct the feature map from low- to high-resolution images. The output heatmap had a size of $\{W, H, 1\}$. Finally, the fast non-maximum suppression (NMS) operation was performed on the feature map to predict feature points.

The feature description decoder are used to generate the descriptors. In particular, the feature map undergoes convolution to obtain a semi-dense description map with dimensions $\{W_c, H_c, 256\}$. Depending on the coordinates of the feature points, the bicubic interpolation sampling method was used to obtain a 256-dimensional representation for each feature point.

D. LOSS FUNCTION

The loss function consists of two parts: local feature loss (L_1) and global feature loss (L_2). The specific local feature loss function is given as follows:

$$L_1 = L_p + L_{p'} + \lambda L_d \quad (6)$$

where P, P' are the inputs of the loss function calculation, which is an image pair generated using the random homography transformation. $L_p, L_{p'}$ are the feature point losses of the images P, P' , and L_d is the description loss between the pair of images.

To calculate L_d , we used the sparse loss of feature descriptors in the training process to improve the training efficiency. Unlike dense loss, sparse loss selects M matching and N incorrect matching point pairs in an image pair as positive and negative samples, respectively. The computation of the loss function decreased from $(H_c \times W_c)^2$ to $M \times N$, which in turn improved the training speed. Hinge loss was used to calculate the error between feature descriptors as

$$L_d = \lambda_1 \frac{1}{M} \sum_{i=1}^M \max(0, m_p - d_i^T d_i') + \frac{1}{MN} \sum_{j=1}^{MN} \max(0, d_j^T d_j' - m_n) \quad (7)$$

where M and N are the numbers of positive and corresponding negative samples selected, respectively. m_p, m_n are the boundary parameters of Hinge loss, d_i, d_i' and d_j, d_j' are the positive and negative samples selected from the feature descriptor of the image pair, respectively.

Assuming x_i, x_i' are the feature point coordinate vectors corresponding to d_i, d_i' and τ is the threshold of the coordinate difference between two points after homography transformation, the method to determine whether the two descriptors are positive samples will be

$$G(H, x_i, x_i') = \begin{cases} \text{True,} & \|Hx_i - x_i'\|^2 < \tau \\ \text{False,} & \text{otherwise} \end{cases} \quad (8)$$

The feature point loss, L_p or $L_{p'}$, can be calculated using the binary cross-entropy loss as follows:

$$L_p = \frac{1}{W_c H_c} \text{bce_loss}(\text{soft_max}(\chi_s), \chi_t) \quad (9)$$

where χ_s is the feature map generated by the feature point detection decoder and χ_t is the corresponding real label.

This study adopted the Teacher–Student knowledge distillation method to obtain global feature representations. It is a method that transfers the learned feature knowledge from the teacher to the student with a simpler model. DELG [19] trains the complex ResNet network on a large-scale data set and obtains a high-performance global feature extraction model, thereby demonstrating stable performance in many fields, such as image retrieval and location recognition. Therefore, this study adopted the DELG as a teacher model to train global features. Additionally, the loss function of the global feature is represented using cosine distance as follow:

$$L_2 = 1 - \cos(D_g, D_g') = 1 - \frac{D_g^T D_g'}{\|D_g\| \|D_g'\|} \quad (10)$$

where D_g is the image global feature descriptor generated by our method and D_g' is the global feature descriptor generated by the teacher model.

IV. EXPERIMENT AND ANALYSIS

The following experiments were designed to verify that the proposed method can significantly improve the efficiency of feature extraction while maintaining good feature performance for VPR.

A. IMPLEMENTATION

We trained the network before conducting the experiments. The training process was divided into three steps: local feature module training, global feature module training, and quantization training. First, training was conducted on the local feature module to generate the simulated data and initialize the feature point detector. The random homography transformation matrix was calculated to transform the original data, and the corresponding point coordinates were recorded as labels. The Adam loss optimization method was used during training, and the learning step size was set to 0.0001. Then, we trained the model on the COCO dataset with 170 K iterations. The DELG method was used to generate the global feature descriptors for each image for use as supervised data for the global feature module training. During training, the weight parameters of the backbone were frozen, and the training iterations were 50 K. Finally, the quantization-aware training was performed with 400 iterations. The hyperparameters used were $\tau = 4$, $m_p = 1.0$, $m_n = 0.5$, $\lambda = 1.2$, and $\lambda_1 = 200$. We implemented the network in Python/Pytorch framework and the experiment was performed in the following hardware specifications: Windows 10 64-bit with Intel(R) Core (TM) i9-9880H 2.30 GHz, 32 GB memory, and a 3090Ti GPU.

B. VISUAL PLACE RECOGNITION

6-DOF VPR: To validate the practicability of the proposed method, we designed a 6-DOF visual localization experiment on Aachen dataset [39]. The global feature was used for image retrieval, and we extracted the TOP-20 images as a candidate set. The candidate set was then matched using local features, and the results were analyzed using multi-view clustering. Finally, the EPnP method was used to achieve precise 6-DoF localization.

Multiple sets of experiments were conducted to prove the image place recognition accuracy of the proposed method. Among them, NV+SP adopted a combination of NetVLAD and SuperPoint to extract the global and local point image features, respectively. NV+SIFT adopted the SIFT local point features extraction method. HF-Net is a feature extractor using a lightweight network combined with NetVLAD, and CSL is another commonly used 2D–3D image-based place recognition method.

TABLE 2. Experimental results of 6-DOF VPR performance.

Methods	$L_{2+0.25}$	$L_{5+0.5}$	L_{10+5}
NV+SP	79.7	88.0	93.6
HF-NET	75.7	84.3	90.9
NV+SIFT	82.8	88.1	91.1
CSL	52.3	80.0	94.3
Ours	77.3	85.1	91.0

We report the recall [%] at different distance and orientation thresholds and the bold values represent the results of our method.

TABLE 3. Experimental results of visual location recognition time efficiency (ms).

Methods	Global feature extraction	Local feature extraction	Global feature retrieval	Local feature matching	Position	Total
NV+SIFT	661	677	13	182	12	1,545
NV+SP	661	518	12	146	9	1,346
HF-NET	411		12	146	9	578
Ours	256		13	145	9	423

The bold values represent the results of our method

Table 2 shows the accuracy of the experimental results for the different place recognition methods. The accuracy was divided into three levels depending on the position and orientation: $L_{2+0.25}$, $L_{5+0.5}$, and L_{10+5} , which indicate position errors of less than 0.25, 0.5, and 5 m and orientation errors of less than 2° , 5° , and 10° , respectively. The performance of the proposed method is superior to the HF-NET lightweight feature extraction method. The best performance is achieved by using complex network models, and our method differs from the optimal results by 5.5%, 3.0%, and 2.6% in the three positioning levels; however, the difference was not

numerically significant. Take into account the time consumption shown in Table 3, it can be said that our proposed method has the comprehensive advantages.

To compare the execution efficiency of different methods, Table 3 summarizes the specific times for global feature extraction, local feature extraction, global feature retrieval, local feature matching, and positioning. Compared with the NV+SIFT and NV+SP methods, the proposed method exhibits $3.6\times$ and $3.2\times$ faster execution, respectively. Compared to the HF-NET method, the proposed method reduces the time consumption by 155 ms. The previous experiments prove that the designed unified network can significantly reduce the feature extraction time, and the proposed method can significantly improve the efficiency of visual localization while ensuring accuracy.

IR-VPR: We further evaluated the performance and efficiency of our method in Image Retrieval VPR (IR-VRR), the global descriptor of query images used to retrieve similar images from the database. Two benchmark datasets, Pitts30k and Pitts250k [18], were employed for the experiment. Following the standard evaluation protocol for the employed datasets, model performance is evaluated by Recall@N. Table 4 and Fig. 3 present the performance and efficiency of different methods, respectively. The results show that our method performs similarly to LSDNet in performance and is significantly better than other methods. While the latency time of our method outperforms all other methods, including LSDNet. In terms of the competitiveness, our method has achieved a satisfying balance between performance and efficiency.

TABLE 4. Experimental results of IR-VPR performance.

Methods	Pitts30k				Pitts250k			
	@1	@5	@10	@20	@1	@5	@10	@20
NetVLAD	75.6	89.3	92.9	95.6	73.4	87.4	91.0	93.8
HF-NET	70.3	86.5	91.1	94.6	65.3	81.6	86.4	90.0
LSDNet	83.2	92.2	94.7	96.4	82.8	91.8	94.4	96.2
Ours	82.3	92.2	94.5	96.1	82.5	92.5	94.3	96.3

The bold values represent the results of our method.

C. MORE RESULTS AND DISCUSSION

1) EFFICIENCY OF FEATURE EXTRACTION

Comparative analyses were performed on various platforms to verify the efficiency of the proposed method. The experimental equipment comprised a personal computer, Huawei MetaPad Pro tablet, and Nvidia AGX AXIVAR embedded development kit.

Table 5 summarizes the experimental results of feature extraction time efficiency acquired on a personal computer. Feature extraction operations were performed on images with different sizes, and the execution time, which included the encoder and total extraction times, was recorded. A random

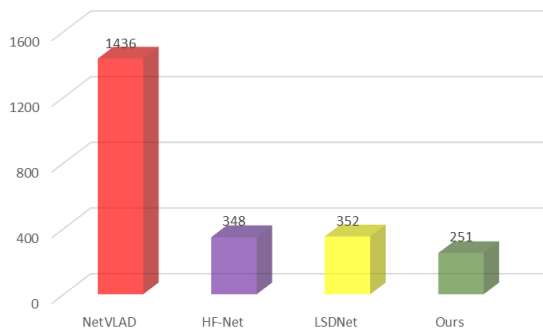


FIGURE 3. Latency times (ms) of different methods. Combined with Table 4, the figure reveals clearly advantages of our method.

TABLE 5. Experimental results of feature extraction time (personal computer).

Method	Image Size				Average Speedup
	240×320	480×640	960×1280	1920×2560	
VGG-F	83ms	228	871	3207	1×
VGG-Q	62	132	499	1860	1.6×
M-F	32	76	377	1671	2.5×
M-Q	28	32	90	329	7.3×
Post-processing	16	41	190	231	
Speedup (Encode)	3.0×	7.1×	9.6×	9.8×	7.3×
Speedup (Overall)	2.2×	3.6×	3.8×	6.1×	3.9×

The feature extraction time (ms) of different methods and the bold values represent the results of our method.

group of 20 images was selected for this experiment, and no more than 1,000 feature points were extracted from each image. The M-Q model refers to the model processed using the method proposed in this study, and the VGG-F model is the commonly used VGG16 model. VGG-Q is the quantized model of VGG-F, whereas M-F is the full-precision model of M-Q. The time performance was affected by the hardware environment. Considering the quantization model is generally applied in a low-performance environment, this experiment was performed in a CPU environment.

The following experimental results were obtained: i) The time efficiency of feature extraction in the M-Q model was significantly higher. Compared to the VGG-F model, the proposed method exhibited 7.3× faster encoding and 3.9× faster overall processing. The average encoding time on images of different sizes was only 14.5% of VGG-F. Owing to the smaller number of parameters and the faster fixed-point operation speed of the M-Q model, the encoding speed improved significantly compared to that of the general VGG model. ii) The larger the image size, the better was the time optimization effect. When the image size was 240 × 320, the speedup ratio of encoding was 3.0×, and when the image size was increased to 1920 × 2560, the encoding speedup ratio was 9.8×, considering the size of the model is not linearly related

TABLE 6. Experimental results of feature extraction time (AGX AXIVAR).

Method	Image Size			Average Speedup
	240×320	480×640	960×1280	
VGG-F	335	1075	3564	1×
VGG-Q	253	530	1612	1.8×
M-F	186	469	1243	2.3×
M-Q	89	216	512	5.2×
Post-processing	17	72	321	-
Speedup (Encode)	3.7×	4.9×	7.0×	5.2×
Speedup (Overall)	3.3×	3.9×	4.7×	4.0×
VGG-F	70	235	973	1×
VGG-Q	67	217	739	1.1×
M-F	36	68	155	3.8×
M-Q	37	53	134	4.5×
Post-processing	16	69	302	-
Speedup (Encode)	1.9×	4.4×	7.2×	4.5×
Speedup (Overall)	1.6×	2.5×	2.9×	2.3×

The feature extraction time (ms) of different methods on Nvidia AGX AXIVAR platform. The bold values represent the results of our method.

to the time consumed by the computer. As the amount of data increased, the computer required more overhead to perform calculations, thereby demonstrating the actual performance of the proposed algorithm in a practical environment.

Furthermore, the feature extraction time efficiency experiment was carried out on the embedded development terminal AGX AXIVAR device. Table 6 summarizes the experimental results. M-Q was run separately in the CPU and GPU environments, and the execution times on images of different sizes were recorded. The results showed that the proposed method completed the feature extraction efficiently. Compared to the VGG-F method, the speedup ratio of the M-Q method reached 3.3×, 3.9×, 4.7× on the CPU platform and 1.6×, 2.5×, 2.9× on the GPU platform under the three different image sizes. Additionally, the quantization results showed that quantization significantly improved the efficiency of feature extraction. Compared to that of the float models, the encoding speed of the quantization models on the CPU platform was 1.8× and 2.3× higher. Compared to that on CPU platforms, the performance of quantization on GPU platforms was slightly poor. Analysis of the causes of the phenomenon, although quantization decreased the storage space, it was affected by the hardware computing unit. Additionally, it increased the complexity of the calculation process and generated additional CPU and GPU communication overheads. Therefore, the optimization effect on the inference speed was relatively poor.

Table 7 shows the time efficiency of feature extraction on the MetaPad Pro tablet. The encoding time performance

TABLE 7. Experimental results of feature extraction time (MetaPad Pro).

Image size	VGG-F (ms)	M-Q (ms)	Post-processing (ms)	Speedup (Overall)
240×320	352	125	15	2.6×
480×640	1076	307	83	2.9×
960×1280	3,203	768	296	3.3×

The bold values represent the results of our method.

of the M-Q and VGG-F models was analyzed under different image sizes. The proposed method improved the feature extraction time by 2.6×, 2.9×, and 3.3× for the 480 × 640, 960 × 1280, and 1920 × 2560 image sizes, respectively.

2) IMPACT OF LSQ

To demonstrate the effectiveness of LSQ, we used the signal-to-quantization noise ratio (SQNR), which reflects the relationship between the signal strength and quantization error. The higher the SQNR, the smaller the quantization error.

Assuming x is the original signal and \hat{x} is the quantized signal, the specific calculation formula for L_{SQNR} is given as

$$L_{SQNR} = 20\log_{10} \frac{\|x\|}{\|x - \hat{x}\|} \quad (11)$$

Fig. 4 shows the quantization error of the feature points, feature descriptors, and global descriptor extracted by the quantization model with different iterations. The performance of direct post-training quantization (PTQ) was poor, and the SQNR values of the three features were less than 0. As the training iterations increased, the SQNR of the signal gradually increased, which indicates that the quantization features could better retain the original feature information, thereby ensuring their application effect.

3) PERFORMANCE OF GLOBAL FEATURE WITH ECA ATTENTION

Instance retrieval was used to evaluate the image representation performance of the global feature, and the experimental results are summarized in Table 8. The experiment was based on two public datasets: Oxford 5K [40], which contained 5,063 images of 11 different buildings in Oxford, and Paris 6K [41], which contained 6,412 images of 11 landmark buildings in Paris. Mean average precision (mAP), which is the average of retrieval average precision (AP) of all N query images, was used as the evaluation index.

$$mAP = \frac{1}{N} \sum_{k=1}^N AP(k) \quad (12)$$

The comparison methods included NetVLAD, M-F, M-Q, and M-Q-E. M-Q-E is the M-Q method without the ECA module, and NetVLAD is a mainstream method in image retrieval and visual localization. Compared to NetVLAD and M-F, the proposed method showed a slightly lower mAP

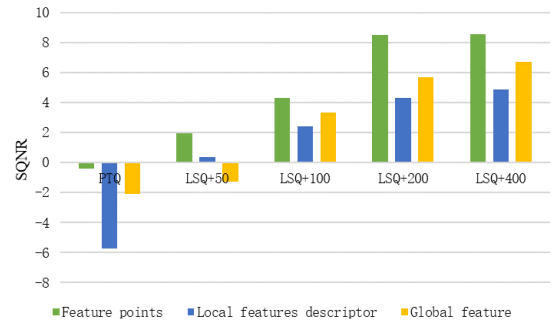


FIGURE 4. Experimental results of the feature extraction quantization error. The SQNR gradually increased, indicating that the quantized vectors could better retain the information of the original vectors.

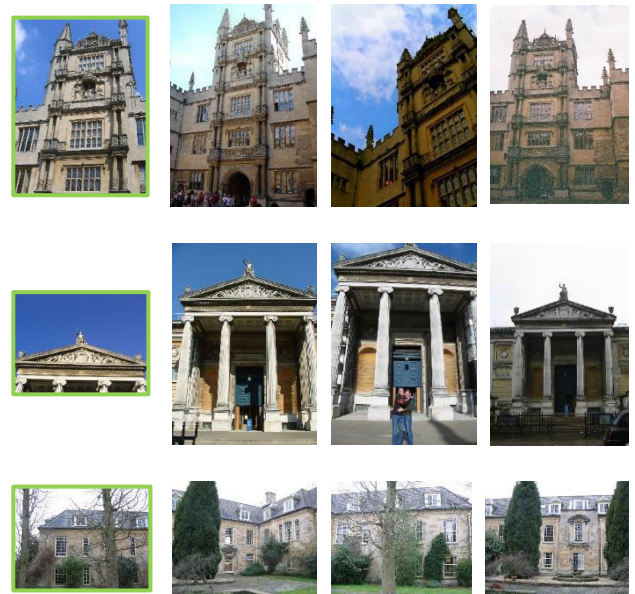


FIGURE 5. Examples of partial search results. The first column is the query images.

TABLE 8. Experimental results of global feature retrieval.

Method	Oxford 5K	Paris 6K
NetVLAD	0.698	0.73
M-F	0.710	0.723
M-Q	0.686	0.706
M-Q-E	0.614	0.65

The bold values represent the results of our method.

value on the two datasets. However, the difference was not numerically significant. Additionally, the mAP values of the M-Q-E method decreased by 0.072 and 0.056 compared to those of M-Q, which indicates that the representation ability of global features was enhanced after using the ECA method. Fig. 5 shows some image retrieval results.

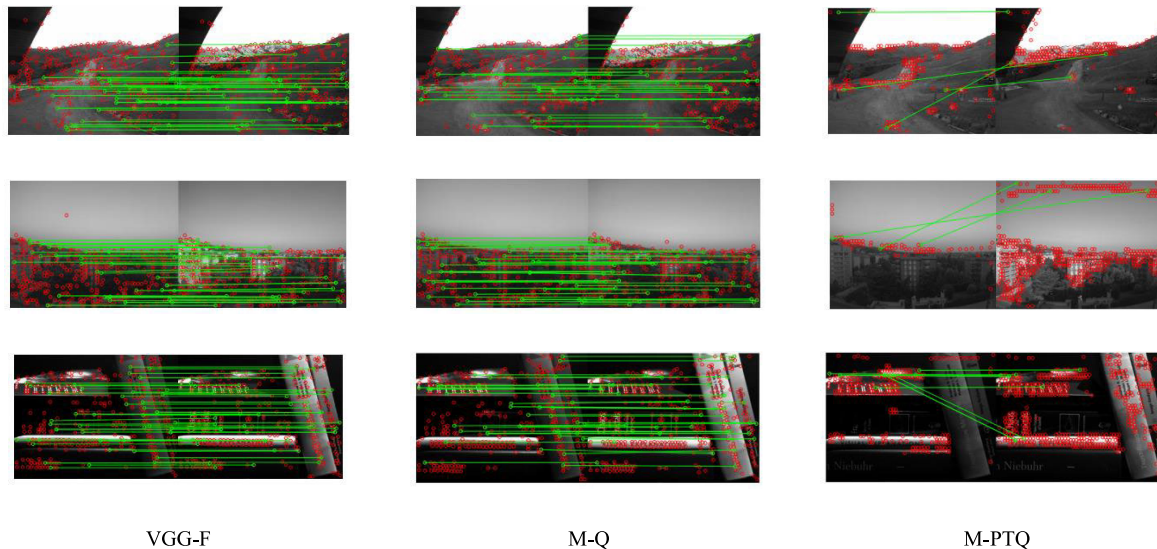


FIGURE 6. Comparison of local feature matching results. The matching results were not significantly different compared to those of VGG-F.

4) PERFORMANCE OF LOCAL FEATURES

We designed a local feature matching experiment to analyze the performance of local features. The experimental results are shown in Table 9. The compared methods included VGG-F, M-PTQ, SIFT, and M-Q. M-PTQ was PTQ static quantification for the network. Four experimental indicators were used: repetition rate (Rep), localization error (Le), matching score (Ms), and mAP. Rep measures the probability that a point is detected in the second image.

$$\text{Rep} = \frac{1}{N_1 + N_2} (\sum_i f_{\text{cr}}(x_i) + \sum_j f_{\text{cr}}(x_j)) \quad (13)$$

where $N_1(N_2)$ is the points count in the first (second) image, $f_{\text{cr}}(x)$ represents point x is detected in the other image.

Le is used to describe the error between the feature point coordinates and the true value. It is expressed as

$$\text{Le} = \frac{1}{N} \sum_{i=1}^N \|x_i - \check{x}_i\| < \varepsilon \quad (14)$$

Ms is the ratio of the number of matched points after cross-validation (N_{inliers}) to the total number of feature points ($N_1 + N_2$). The larger the value, the better the matching performance.

$$\text{Ms} = \frac{N_{\text{inliers}} \times 2}{N_1 + N_2} \quad (15)$$

mAP counts the matching recall rate under different thresholds, which range between 0 and 1. The larger the value, the better the matching performance.

The experiments showed that there was no significant drop in the matching performance of the M-Q algorithm. Compared to the original and M-Q models, the M-PTQ algorithm showed significantly poor matching performance, with a matching score of only 0.039. While there was no significant difference between the M-Q algorithm and the original model

TABLE 9. Experimental results of local feature matching.

	Rep	Le	Ms	mAP
VGG-F	0.696	0.898	0.506	0.854
M-PTQ	0.281	1.936	0.039	0.358
M-Q	0.697	0.885	0.468	0.859
SIFT	0.576	0.623	0.401	0.621

The bold values represent the results of our method.

for all indicators, except for a difference of 0.04 in the matching score. Therefore, the experimental results confirmed that the M-Q quantization algorithm could effectively reduce the quantization errors and maintain the feature extraction ability of the original model. The local features obtained met the requirements of the local feature matching task.

Fig. 6 shows the matching results after feature extraction by different models (columns 1, 2, and 3, represent the matching results of the VGG-F, M-Q, and M-PTQ methods, respectively). The feature points extracted by the M-PTQ algorithm were unevenly distributed, with a poor matching effect; hence, none of the three images were correctly matched. In contrast, the feature points extracted by the M-Q method were not significantly different with VGG-F, and the overall matching effect was better than that of M-PTQ.

V. CONCLUSION

This study proposed a fast dual-feature extraction method for visual place recognition to address the issue of feature extraction efficiency. We designed a tightly coupled feature extraction network that extracts local and global features in a unified model based on a lightweight backbone. The LSQ method was used in the inference stage to further optimize the

efficiency. The experimental results showed that the proposed method significantly improved the efficiency while ensuring accuracy for VPR. Additionally, the proposed method was able to achieve comparable performance to the optimal method at a significantly lower cost, and it was $2.9\times$ to $4.0\times$ faster than the mainstream feature extraction model. This shows our method can significantly improve feature extraction efficiency while maintaining good representation performance.

This study has some limitations, and the proposed method can be further improved. Firstly, in the training stage, we only considered optimization for the encoder, and improving the post-processing efficiency requires in-depth research. Then, the quantization effect varies across different hardware platforms, designing more suitable quantization methods for different platforms requires extensive experiments and analysis. These issues will be addressed in the future.

DATA AVAILABILITY STATEMENT

The Aachen dataset could be found at <https://www.visuallocalization.net/datasets/>

The COCO dataset could be found at <http://cocodataset.org/>

The HPatches dataset could be found at <https://github.com/hpatches/hpatches-dataset/>

The Oxford 5K dataset could be found at <https://www.robots.ox.ac.uk/~vgg/data/oxbuildings/>

The Paris 6K dataset could be found at <http://www.robots.ox.ac.uk/~vgg/data/parisbuildings/>

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] R. Ji, Y. Gao, W. Liu, X. Xie, Q. Tian, and X. Li, "When location meets social multimedia: A survey on vision-based recognition and mining for geo-social multimedia analytics," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 1, pp. 1–18, Mar. 2015.
- [2] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *Proc. ECCV*, Glasgow, U.K., 2020, pp. 726–743.
- [3] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "InLoc: Indoor visual localization with dense matching and view synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7199–7209.
- [4] S. Dian, Y. Yin, C. Wu, Y. Zhong, H. Zhang, and H. Yuan, "Loop closure detection based on local-global similarity measurement strategies," *J. Electron. Imag.*, vol. 31, no. 2, Mar. 2022, Art. no. 023004.
- [5] F. Xue, I. Budvytis, D. O. Reino, and R. Cipolla, "Efficient large-scale localization by global instance recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 17327–17336.
- [6] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May 2018.
- [7] M. Yu, L. Zhang, W. Wang, and H. Huang, "Loop closure detection by using global and local features with photometric and viewpoint invariance," *IEEE Trans. Image Process.*, vol. 30, pp. 8873–8885, 2021.
- [8] Z. Luo et al., "GeoDesc: Learning local descriptors by integrating geometry constraints," in *Proc. ECCV*, Munich, Germany, 2018, pp. 168–183.
- [9] Y. Fang, K. Yang, R. Cheng, L. Sun, and K. Wang, "A panoramic localizer based on Coarse-to-Fine descriptors for navigation assistance," *Sensors*, vol. 20, no. 15, p. 4177, Jul. 2020.
- [10] S. J. Lee, D. Kim, S. S. Hwang, and D. Lee, "Local to global: Efficient visual localization for a monocular camera," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2230–2239.
- [11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [12] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.
- [13] R. Arandjelovic and A. Zisserman, "All about VLAD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1578–1585.
- [14] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3384–3391.
- [15] L. G. Camara and L. Přeucil, "Visual place recognition by spatial matching of high-level CNN features," *Robot. Auton. Syst.*, vol. 133, Nov. 2020, Art. no. 103625.
- [16] J. Mao, X. Hu, X. He, L. Zhang, L. Wu, and M. J. Milford, "Learning to fuse multiscale features for visual place recognition," *IEEE Access*, vol. 7, pp. 5723–5735, 2019.
- [17] Y. Fang, K. Wang, R. Cheng, and K. Yang, "CFVL: A coarse-to-fine vehicle localizer with omnidirectional perception across severe appearance variations," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Las Vegas, NV, USA, Oct. 2020, pp. 1885–1891.
- [18] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5297–5307.
- [19] M. Yang, D. He, M. Fan, B. Shi, X. Xue, F. Li, E. Ding, and J. Huang, "DOLG: Single-stage image retrieval with deep orthogonal fusion of local and global features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 11752–11761.
- [20] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," 2016, *arXiv:1612.06321*.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. ECCV*, Amsterdam, The Netherlands, 2016, pp. 630–645.
- [22] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1 MB model size," 2016, *arXiv:1602.07360*.
- [23] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 1577–1586.
- [24] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "ShuffleNet v2: Practical guidelines for efficient CNN architecture design," in *Proc. ECCV*, 2018, pp. 116–131.
- [25] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
- [26] Y. Chen et al., "Mobile-former: Bridging mobilenet and transformer," in *Proc. IEEE/CVF CVPR*, New Orleans, LA, USA, Jun. 2022, pp. 5270–5279.
- [27] G. Peng, Y. Huang, H. Li, Z. Wu, and D. Wang, "LSDNet: A lightweight self-attentional distillation network for visual place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Kyoto, Japan, Oct. 2022, pp. 6608–6613.
- [28] M. Zaffar, S. Ehsan, M. Milford, and K. McDonald-Maier, "CoHOG: A light-weight, compute-efficient, and training-free visual place recognition technique for changing environments," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1835–1842, Apr. 2020.
- [29] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 12708–12717.
- [30] C. Shi, J. Li, J. Gong, B. Yang, and G. Zhang, "An improved lightweight deep neural network with knowledge distillation for local feature extraction and visual localization using images and LiDAR point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 177–188, Feb. 2022.
- [31] P. E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, and C. Cadena, "Leveraging deep visual descriptors for hierarchical efficient localization," in *Proc. Conf. Robot Learn.*, 2018, pp. 456–465.
- [32] M. Nagel, M. van Baalen, T. Blankevoort, and M. Welling, "Data-free quantization through weight equalization and bias correction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1325–1334.
- [33] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha, "Learned step size quantization," 2019, *arXiv:1902.08153*.

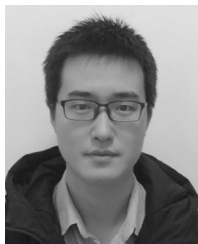
- [34] Y. Bhalgat, J. Lee, M. Nagel, T. Blankevoort, and N. Kwak, "LSQ+: Improving low-bit quantization through learnable offsets and better initialization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2978–2985.
- [35] Z. Tan, K. Li, and Y. Wang, "Differential evolution with adaptive mutation strategy based on fitness landscape analysis," *Inf. Sci.*, vol. 549, pp. 142–163, Mar. 2021.
- [36] M. Nagel, M. Fournarakis, R. Ali Amjad, Y. Bondarenko, M. van Baalen, and T. Blankevoort, "A white paper on neural network quantization," 2021, *arXiv:2106.08295*.
- [37] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," 2013, *arXiv:1308.3432*.
- [38] Q. Wang, B. Wu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," 2019, *arXiv:1910.03151*.
- [39] T. Sattler et al., "Benchmarking 6DOF outdoor visual localization in changing conditions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8601–8610.
- [40] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.
- [41] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 1–8.



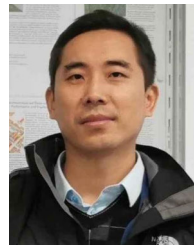
LIANG LYU received the B.S. degree in measurement and control engineering and the M.S. degree in photogrammetry and remote sensing from the Zhengzhou Institute of Surveying and Mapping, China, in 2011 and 2014, respectively, and the Ph.D. degree in surveying and mapping from PLA Strategic Support Force Information Engineering University, China, in 2019. He is currently a Lecturer with PLA Strategic Support Force Information Engineering University. His research interests include photogrammetry, remote sensing, digital earth information resources, space situational awareness, and visualization.



CHAOZHEN LAN received the B.S. and M.S. degrees in photogrammetry and remote sensing and the Ph.D. degree in surveying and mapping from the Zhengzhou Institute of Surveying and Mapping, China, in 2002, 2005, and 2009, respectively. He is currently an Associate Professor and a Master Supervisor with PLA Strategic Support Force Information Engineering University. His research interests include photogrammetry and UAV remote sensing.



XIAOFEI HU received the B.S. degree in geographic information system from the North China University of Water Resources and Electric Power, in 2013, and the M.S. degree in surveying and mapping from PLA Strategic Support Force Information Engineering University, China, in 2017, where he is currently pursuing the Ph.D. degree in surveying and mapping. His research interests include photogrammetry, cyberspace visualization, and virtual reality.



QUNSHAN SHI received the B.S. and M.S. degrees in photogrammetry and remote sensing from the Zhengzhou Institute of Surveying and Mapping, China, in 2008 and 2011, respectively, and the Ph.D. degree in surveying and mapping from PLA Strategic Support Force Information Engineering University, China, in 2015. He is currently a Lecturer with PLA Strategic Support Force Information Engineering University. His research interests include photogrammetry, remote sensing, and virtual reality.



YANG ZHOU received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from the Zhengzhou Institute of Surveying and Mapping, China, in 1996, 2002, and 2009, respectively. He is currently a Professor and a Doctoral Supervisor with PLA Strategic Support Force Information Engineering University. His research interests include remote sensing, digital photogrammetry, and cyberspace mapping.



MINGBO HOU received the B.S. degree in surveying and mapping engineering from PLA Strategic Support Force Information Engineering University, China, in 2022, where he is currently pursuing the master's degree. His research interests include photogrammetry, remote sensing, and cyberspace surveying and mapping.

...