Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7168451, 16 pages
https://doi.org/10.1155/2022/7168451

Research Article

Image Geolocation Method Based on Attention Mechanism Front Loading and Feature Fusion

Huayuan Lu[1,2], **Chunfang Yang**[1,2], **Baojun Qi**[1,2], **Ma Zhu**[1,2] **and Jingqian Xu**[1,2]

[1] *Henan Provincial Key Laboratory of Cyberspace Situational Awareness, Zhengzhou 450001, China*
[2] *Zhengzhou Science and Technology Institute, Zhengzhou 450001, China*

Correspondence should be addressed to Chunfang Yang; chunfangyang@126.com

Received 24 January 2022; Accepted 18 May 2022; Published 13 June 2022

Academic Editor: Ernestina Cianca

Copyright © 2022 Huayuan Lu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Image geolocation is an important technique for robotics and autonomous systems. The existing methods mainly extract local features from images directly and use global descriptors, which are aggregated by these local features, to retrieve candidate references from all references. Thus, the training efficiency is affected by the image noises and the accuracy is so limited that the further verification is extremely time consuming. To address these issues, this work proposes an image geolocation framework, which adds the noise filtering layer before local feature extraction. Based on this framework, an image geolocation method based on attention mechanism front loading and feature fusion is designed. In the noise filtering layer, the proposed method uses triplet attention to denoise images thus leading to higher training efficiency. In the feature aggregation layer, an improved SPP (spatial pyramid pooling) is designed to extract the local factors reflected by the position relationships among local features. Then, the local factors are incorporated with the global factors extracted by NetVLAD. The fused descriptors contain not only the statistic of the geometric elements but also the position relationships among them. The experimental results show that the proposed method outperforms NetVLAD in terms of the training convergence round and Recall@N ($N = 1, 5, 10, 20$); especially, the convergence round of Recall@5 reduces from 25 to 10, the convergence round of Recall@10 reduces from 25 to 7, Recall@1 increases from 79.45% to 84.01%, and Recall@5 increases from 90.10% to 92.81%.

1. Introduction

Image geolocation is a technique for geolocating a scene (in the image) or a camera based on the content of the image or its side information. This technique has been applied to many fields of IoT (Internet of things), such as autonomous driving [1–4], intelligent robotics [5], and virtual reality/augmented reality [6], and has received increasing attention in recent years [7–9].

Existing image geolocation methods can mainly be classified into image geolocation methods based on 3D point cloud [2, 10–17] and image geolocation methods based on image retrieval [18–21]. Image geolocation methods based on 3D point cloud use the SfM (structure from motion) [22] to construct a 3D point cloud from multiple images with different angles near the query (the image to be located) and then compare the constructed 3D point cloud with the local features of the query image point by point to obtain the exact shooting position and the estimation of the camera posture of the query image. Such methods have high positioning accuracy but require stronger conditions and a high computational cost. Image geolocation methods based on image retrieval extract the descriptors of the query and the reference (the image with known geographic location), then find the reference closest to the query according to the similarity between descriptors, and use its geographic location to estimate the location of the query. The geolocation precisions of these methods are greatly affected by the geographic labels of the references, but they are computationally efficient and robust and can be extended to global geolocation, which has received widespread attention [18–21]. Therefore, this paper focuses on image geolocation based on image retrieval.

The key to image geolocation based on image retrieval is the construction of image descriptors. According to different manners of image descriptor construction, the existing image geolocation methods based on image retrieval can be generally classified into two classes: (1) image geolocation methods based on the heuristic descriptors and (2) image geolocation methods based on deep learning.

Image geolocation methods based on the heuristic descriptor firstly detect key points using algorithms such as Harris [23] and MSER (maximally stable extremal regions) [24], then aggregate or directly use SIFT (scale invariant feature transform) [25], SURF (speeded up robust features) [26], or other local features as image descriptors [10, 13, 27–36], and finally, match the query descriptor with the reference descriptors to geolocate the query. Earlier studies on image geolocation mainly focus on such methods. For example, Johns and Yang [27] clustered the SIFT features of all references, then built a feature tree, scanned the feature tree to find the references closest to the query image, and finally used a geometric verification method to determine the final position. In 2012, Gálvez-López and Tardos [37] constructed the BoVW (bag of visual words) [38] histogram of the BRIEF [39] local features of FAST [40] key points as the descriptor of this image and then represented it in binary form for fast image geolocation. Cao and Snavely [10] constructed the BoVW [38] histogram of SIFT features of an image as its descriptor, then clustered references by descriptors, trained a model for each cluster to determine whether the query belongs to that cluster, and finally geometrically verified it against the references within the cluster to geolocate the query. Zemene et al. [30] proposed the DSC (dominant set clustering) algorithm to dynamically select the N nearest neighbors of the query's SIFT feature points from the feature point set of references and then determined the query's geolocation based on the similarity and adjacency relationships between feature points. The abovementioned methods consider the similarity and position relationships between feature points and can achieve high geolocation accuracy, strong interpretability, and high computational efficiency. The construction of their descriptors mainly considers edges and corners in images with drastic texture changes. But in foggy and weakly illuminated environments, the extraction of the edge and corner features in images is easily affected by noises, so these methods perform poorly in such environments [41].

Deep learning-based image geolocation methods use lots of images to train deep networks to extract features as descriptors and then match the descriptors to geolocate the query [20, 42–52]. These methods are easily affected by the training data but have strong robustness, such as high adaptability to the changes of brightness, color, and viewing angle [41]. Since Sharif et al. [53] first introduced CNN (convolutional neural networks) to image geolocation in 2014, this kind of method has become the hotspot in the field of image geolocation. For example, Sünderhauf et al. [45] trained a network to extract images' ROI (region of interest) features from images as region-level descriptors and matched them to geolocate the query. Anoosheh et al. [54] proposed ToDayGAN, which used GAN (generative adversarial network) to convert weakly illuminated images into normally illuminated images and then extracted its dense-VLAD [55] features as descriptors to geolocate the query. In 2016, Arandjelovic et al. [20] proposed NetVLAD, which extracted local features by VGG16 [56], then aggregated them as descriptors by a hot-plugging aggregation layer similar to VLAD (vector of locally aggregated descriptors) [57], and finally matched descriptors to geolocate the query. NetVLAD [20] is a milestone work in the field of image geolocation, since it first constructs a deep hot-plugging layer to aggregate local features as descriptors, making it possible to optimize the aggregation parameters automatically.

Based on NetVLAD [20], a number of improved methods have been proposed [2, 18, 19, 58]. For example, in 2020, Yu et al. [58] proposed SPE-VLAD, which divided an image into multiple nonoverlapping blocks, then concatenated the NetVLAD features of all blocks as descriptors, and matched them to geolocate the query. In 2021, Ge et al. [18] proposed an improved method of NetVLAD based on self-supervised learning, which used the descriptors obtained from NetVLAD to calculate the similarity between images and the similarity between images subregions, and then used the similarities to iteratively train the descriptor extraction network for image geolocation. The descriptors obtained by the abovementioned NetVLAD-based methods can effectively represent the whole image and achieve excellent geolocation performance. However, the global descriptors used by the abovementioned methods aggregate all local features and treat them equally, making task-irrelevant local features interfere with image geolocation. In addition, NetVLAD [20] represents the image with a sum of residuals, a global statistic, ignoring the positional relationships between local features.

To solve the problem that noisy local features tend to affect the accuracy of geolocation, researchers filter local features using attention mechanisms before aggregation and obtain more effective descriptors to improve geolocation accuracy [21, 59–65]. For example, in 2017, Kim et al. [59] proposed the CRN (contextual reweighting network), which upsampled filtered local features to get a weight matrix and reweighed local features, and then aggregated them as descriptors to geolocate the query. In 2018, Chen et al. [21] filtered the output of multiple layers in the feature extraction network and then constructed attention matrices to fuse the filtered results as descriptors for image geolocation. In 2021, Peng et al. [63] proposed SRALNet (semantic reinforced attention learning network), which clustered local features, and then aggregated the double-weighted residuals of local features and corresponding cluster centers as descriptors to geolocate the query. In 2021, Peng et al. [64] proposed APPSVR (attentional pyramid pooling of salient visual residuals for place recognition), which combined SRALNet with SPP (spatial pyramid pooling) [66] to fuse local features as descriptors for image geolocation. The abovementioned methods suppress task-irrelevant features and enhance task-relevant features to obtain more robust descriptors and can achieve better geolocation performance. To solve the problem of NetVLAD ignoring local factors, in 2021, Hausler et al. [19] proposed patch-NetVLAD, which retrieved candidate references using NetVLAD and reranked them based on the similarity between local features of two images to geolocate the query. However, on one hand, because the abovementioned methods directly extract local features from the images with noisy signals, the feature extraction networks are difficult to converge quickly. On the other hand, because the candidate references are retrieved from all references by the global descriptors, the accuracy is so limited that the further verification in patch-NetVLAD is extremely time consuming.

Therefore, this work is decided to improve the training efficiency and accuracy of image geolocation and mainly contains the following three contributions.

An image geolocation framework is proposed by adding the noise filtering layer before local feature extraction. The proposed framework consists of the noise filtering layer, the local feature extraction layer, the feature aggregation layer, and the descriptor matching layer.

According to the proposed framework, an image geolocation method based on attention mechanism front loading and feature fusion is designed. The noise filtering layer uses triplet attention [67] to denoise images thus improving the training efficiency. In the feature aggregation layer, the local factors extracted by an improved SPP are incorporated with the global factors extracted by NetVLAD. The fused descriptors contain not only the statistic of the geometric elements but also the position relationships among them.

SPP is improved by replacing the max grouping with GeM [68] when the number of SPP grids is 1 × 1. The improved SPP can extract the local factors reflected by the position relationships among local features.

The experimental results show that the proposed method can efficiently improve the accuracy of the model and the efficiency of training; especially, the convergence round of Recall@5 reduces from 25 to 10, convergence round of Recall@10 reduces from 25 to 7, Recall@1 increases from 79.45% to 84.01%, and Recall@5 increases from 90.10% to 92.81%.

The rest of the paper is organized as follows: Section 2 introduces the works strongly related to this paper, Section 3 details the proposed method, Section 4 shows the experimental results and analysis, and Section 5 concludes this paper.

2. Review of VLAD and NetVLAD

NetVLAD [20] is a milestone work in the field of image geolocation. It is derived from the classical VLAD [57] which extracts the descriptor as follows (see Figure 1):

(1) Extract the local features of the images

(2) Cluster the local features to obtain K clusters, each of which represents a type of feature (e.g., representing the corners of a window)

(3) Calculate the sum of residuals between the features in each cluster and their corresponding cluster center, as shown in equation (1) as follows

$$v_k = \sum_{i=1}^{N} a_k(x_i)(x_i - c_k), \tag{1}$$

where a local feature is denoted by a vector $x_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,D})$, D denotes the dimension of the local feature, $a_k(x_i)$ is 1 if feature x_i belongs to cluster k and 0 if otherwise, N is the number of local features, c_k is the center of the kth cluster, and v_k denotes the sum of residuals in the kth cluster

(4) Concatenate all $v_k, k = 1, 2, \cdots, K$ to obtain a single vector v as the descriptor

NetVLAD improves the traditional VLAD to a hot-plugging layer of deep networks that automatically learns better parameters and then extracts more robust descriptors to improve geolocation performance.

In NetVLAD, the piecewise function $a_k(x_i)$ is replaced by a derivable form a_k---(x_i) as shown by equation (2), to preserve the following property of $a_k(x_i)$ as much as possible, which is that when feature x_i is close to the kth cluster, the value of a_k---(x_i) is close to 1; otherwise, it is close to 0.

$$\overline{a_k}(x_i) = \frac{e^{-\alpha x_i - c_k{}^2}}{\sum_{k'} e^{-\alpha x_i - c_{k'}{}^2}}, \tag{2}$$

where α is a parameter (positive constant) that controls the decay of the response with the magnitude of the distance. $\|x_i - c_k\|^2$ denotes the square of the L2 norm of $x_i - c_k$, namely, the square of the Euclidean distance between feature x_i and the center of the kth cluster. Let $\omega_k = 2\alpha c_k$, $b_k = -\alpha c_k{}^2$, and then, equation (2) is transformed into a soft assignment of the following form:

$$\overline{a_k}(x_i) = \frac{e^{\omega_\kappa^T + b_k}}{\sum_{k'} e^{\omega_{k'}^T x_i + b_{k'}}}. \tag{3}$$

It can be seen that the expression of a_k---(x_i) is derivable. Essentially, the cluster operation in VLAD is transformed to find proper functions a_k---$(x_i), k = 1, 2, \cdots, K$, namely, to learn proper values of ω_k and b_k, which are the parameters in k convolution kernels with the size of 1 × 1. The final form of the NetVLAD layer is obtained by plugging the soft assignment (3) into the VLAD descriptor (1) resulting in

$$v_k = \sum_{i=1}^{N} \overline{a_k}(x_i)(x_i - c_k). \tag{4}$$

In general, NetVLAD extracts the descriptor as follows (see Figure 2):

(1) Extract local features $X = \{\cdots, x_{i-1}, x_i, x_{i+1}, \cdots\}$ using CNN

(2) Cluster local features to K cluster centers using K convolutions whose kernel size is 1 × 1

(3) Calculate the sum of residuals between the features in each cluster and their corresponding cluster center

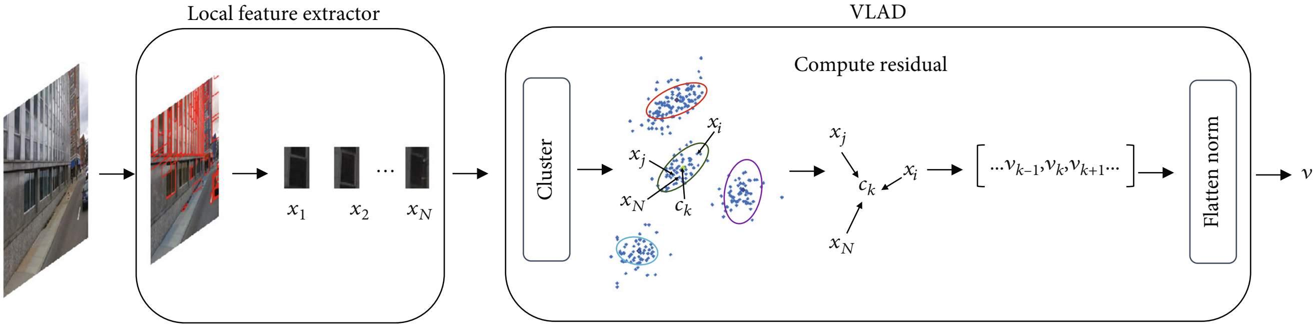


Figure 1: Traditional VLAD-based global descriptor extraction process.

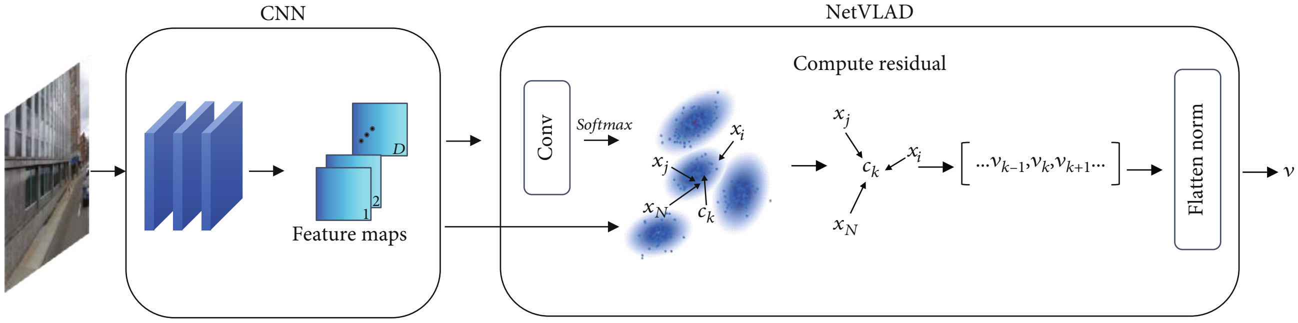


Figure 2: NetVLAD-based global descriptor extraction.

(4) Concatenate all residual sums into a single vector as the descriptor

3. Proposed Image Geolocation Framework

Image geolocation primarily utilizes the content information of images, but some task-irrelevant noises are inevitably introduced during image acquisition and processing. However, the existing deep learning-based image geolocation methods usually directly extract features from the query and references, which slows the learning speed of the model and affects the geolocation accuracy because of the interference of noises to the extracted features. Therefore, this section proposes an image geolocation framework by adding a noise filtering layer before local feature extraction. The proposed framework contains 4 superlayers as shown in Figure 3: noise filtering layer, local feature extraction layer, feature aggregation layer, and descriptor matching layer.

The noise filtering layer uses a filter to suppress the task-irrelevant signals and enhance the task-relevant signals in the image, which can accelerate the learning of the effective model and improve the geolocation accuracy.

The local feature extraction layer extracts local features from query and references. The local features can be traditional key point vector representations (such as SIFT [25] and SURF [26]) or feature maps extracted by the encoder of deep networks (such as VGG [56], ResNet [69], and EfficientNet [70]). In general, local features need to reflect the texture information, such as edges and corners, because this information can effectively distinguish the geographic location of the image.

The feature aggregation layer aggregates extracted local features into descriptors to geolocate the query. The aggregation methods can be traditional methods such as VLAD and BOVW or deep learning-based methods such as NetVLAD and GeM. The generated descriptors are used to retrieve images with similar content and should be robust to changes in a viewing angle and brightness.

The descriptor matching layer calculates the similarity between query and references to retrieve candidate references. Then, the geographic location of the candidate reference is regarded as the geographic location of the query. The mainstream similarity calculation methods include Euclidean distance, Manhattan distance, and cosine similarity.

4. Image Geolocation Based on Attention Mechanism Front Loading and Feature Fusion

In existing methods based on NetVLAD, the model convergence speed and accuracy are interfered with by the task-irrelevant noises in images and the positional relationships between local features are either ignored or used in an extremely time-consuming manner, such as the reranking of patch-NetVLAD. Therefore, under the guidance of the abovementioned framework, this section proposes an image geolocation method based on attention mechanism front loading and feature fusion, as shown in Figure 4.

In the proposed method, the encoder of VGG16 is used in the local feature extraction layer to extract local features and Euclidean distance is used in the descriptor matching

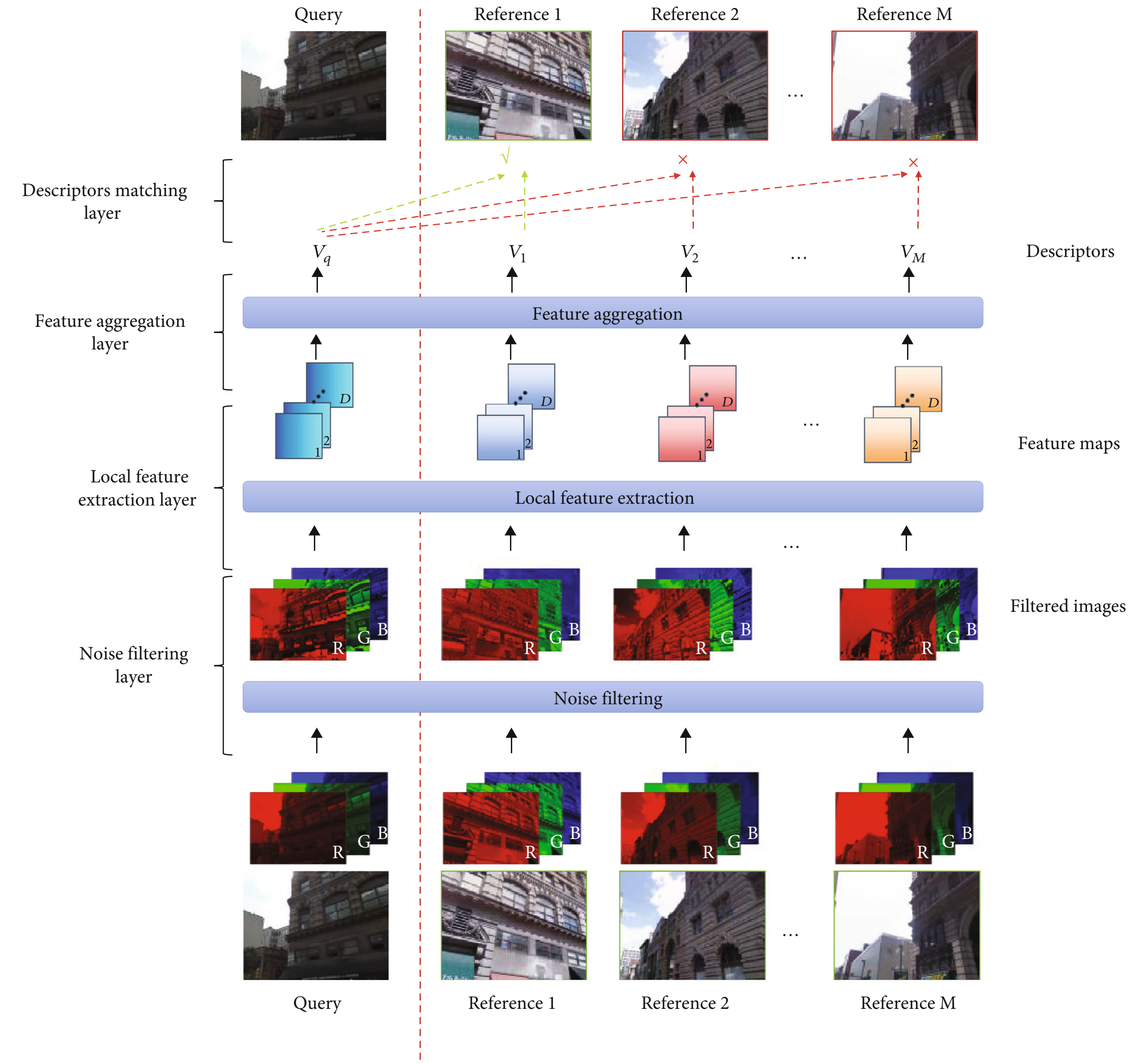


FIGURE 3: The proposed image geolocation framework.

layer to calculate the similarity between query and references, the same as NetVLAD and its improved versions such as patch-NetVLAD, CRN, and SARE [71]. Unlike existing image geolocation methods, triplet attention is plugged into as a noise filtering layer, to eliminate task-irrelevant noise while retaining task-relevant information of image contents. In the feature aggregation layer, the local features are aggregated by NetVLAD and an improved SPP; then, the aggregated results are concatenated as the descriptor.

In the following 2 subsections, the noise filtering layers and the feature aggregation layer will be described in detail.

4.1. Noise Filtering Layer Based on Triplet Attention. The noise filtering layer uses a filter to suppress the task-irrelevant signals in the image and enhance the task-relevant signals, which can accelerate the learning of the effective model and improve the geolocation accuracy.

The existing image noise filtering methods mainly use various correlations in the image to reduce the noises and keep the image contents. Among the existing methods for capturing various correlations in images, the attention mechanism models the correlation among information in the channel domain, spatial domain, or temporal domain, to effectively filter noises and achieve excellent performance on many tasks such as image classification, object detection, and semantic segmentation [72]. Scholars have proposed many attention mechanisms such as SKNet [73], SENet [74], residual attention network [75], CBAM [76], and triplet attention [67]. Among these methods, the triple attention [67] can model correlations in both the spatial and channel domain of images with almost no parameter increase and can achieve excellent performance. Therefore, the proposed noise filtering layer uses it to denoise images before extracting local features to eliminate the influence of noise information and improve the feature effectiveness.

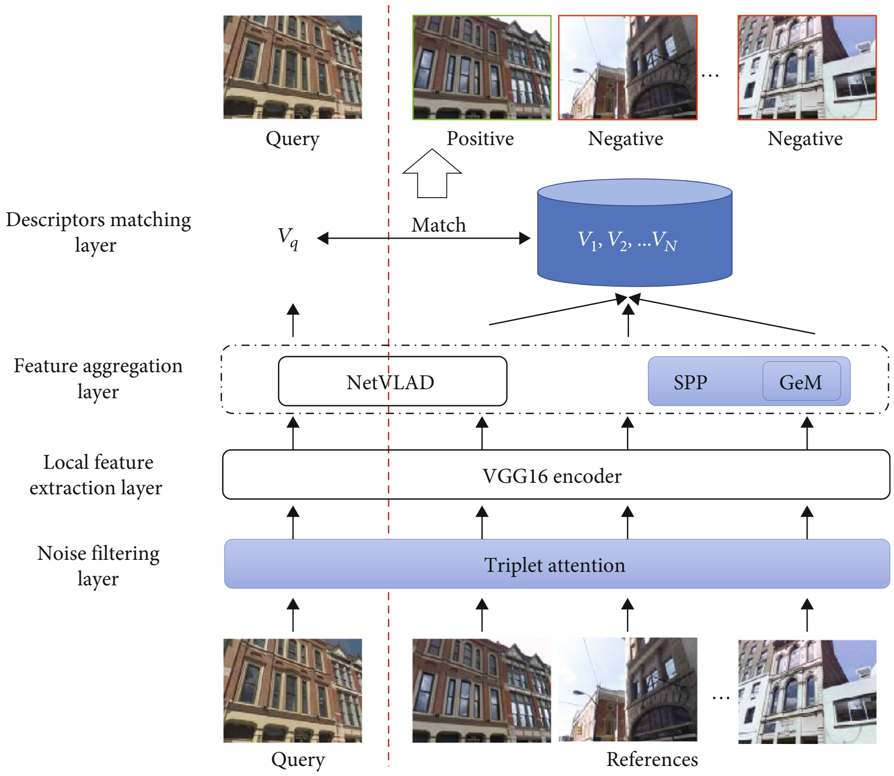


Figure 4: Image geolocation method based on attention mechanism front loading and feature fusion.

Let $x_{G\times H\times W}$ denote the input image, where G, H, W are the number of channels, height, and width of the input image, respectively. The specific steps of the adopted triplet attention are as follows (see Figure 5):

(1) Perform the dimensional rotation (permutation) operation on the input image, to obtain 2 tensors $x_{H\times G\times W}$ and $x_{W\times H\times G}$

(2) Perform the following operations on $x_{G\times H\times W}$ to obtain the weighted tensors $\hat{x}_{G\times H\times W}$

  (i) Perform max pooling and avg pooling (mean pooling) on $x_{G\times H\times W}$, which can obtain two tensors whose size is $1 \times H \times W$, and then, stack them and get $x^{*}_{2\times H\times W}$

  (ii) Set a convolution whose kernel size, step value, and padding value are 7×7, 1, and 3, respectively, and perform it on the $x^{*}_{2\times H\times W}$, which can obtain $x'^{*}_{1\times H\times W}$

  (iii) Perform batch normalization and sigmoid on $x'^{*}_{1\times H\times W}$, which can obtain the weight matrix $P_{1\times H\times W}$, and then, perform pointwise multiplication on it by the tensor $x_{G\times H\times W}$, which can obtain weighted tensor $\hat{x}_{G\times H\times W}$

(3) Perform the same steps in (2) on $x_{H\times G\times W}$ and $x_{W\times H\times G}$, to obtain the other 2 weighted tensors $\hat{x}_{H\times G\times W}$ and $\hat{x}_{W\times H\times G}$

(4) Inverse permute $\hat{x}_{H\times G\times W}$ and $\hat{x}_{W\times H\times G}$ to obtain 2 tensors $\hat{x}'_{G\times H\times W}$ and $\hat{x}'_{G\times H\times W}$, and then, perform element-wise addition and average operations on the 3 tensors $\hat{x}_{G\times H\times W}$, $\hat{x}'_{G\times H\times W}$, and $\hat{x}'_{G\times H\times W}$, to obtain the filtered tensor $\bar{x}_{G\times H\times W}$, whose size is the same as the input image

4.2. Feature Aggregation Layer Based on the Fusion of Global and Local Features. In the mainstream NetVLAD, a global statistical vector, containing the sums of residuals, is used as the descriptor, which weakens the role of the local factor. Recent excellent works, such as DOLG [77], DELG [78], and patch-NetVLAD, argue that the introduction of local factors can improve the retrieval and geolocation performance. However, directly using local features as descriptors will bring not only high computational complexity but also poor robustness. Therefore, the proposed feature aggregation layer incorporates the aggregation method in NetVLAD and an improved SPP method to maintain the local factor.

SPP not only can maintain position relationships among local features but can also pool the input feature maps with different sizes, while having high computational efficiency. Its effectiveness has been proven in many retrieval tasks [68, 77–79]. SPP grids a feature map into equally nonoverlapping parts and performs max pooling for every part. But in the original SPP, the importance difference among feature maps is ignored, which is proven useful by GeM for image retrieval and geolocation tasks [68, 77–79]. When the SPP grid number is 1×1, the pooing operation is similar

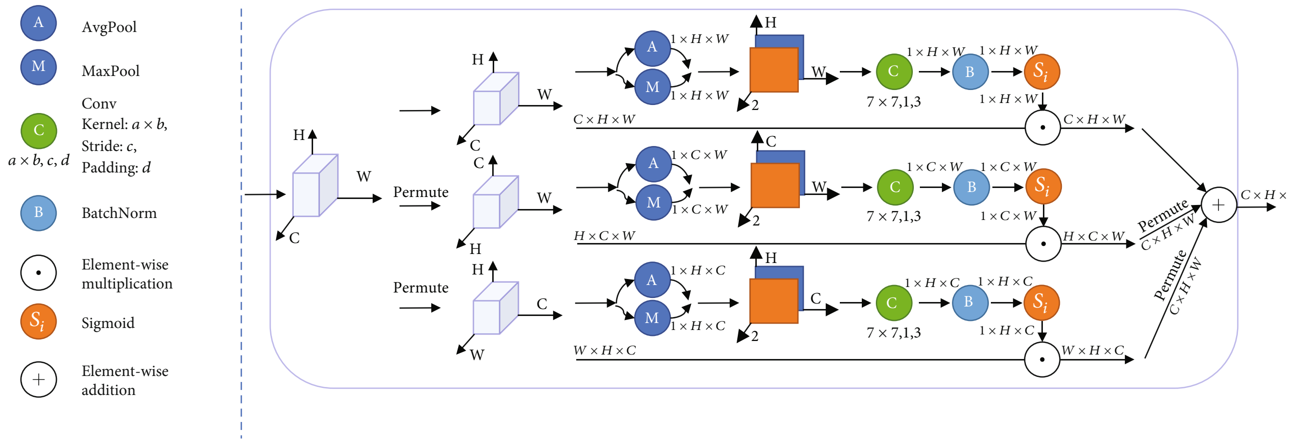


Figure 5: The triplet attention used in the proposed noise filtering layer.

to GeM, which can reflect the importance of different feature maps. Additionally, GeM has a simple and effective calculation method. Thus, in the proposed method, SPP is improved by replacing the max grouping with GeM when the number of SPP grids is 1×1.

Let $x_{g \times h \times w}$ denote the input feature maps, where g, h, w denotes the number of feature maps, height, and width of each feature map, respectively. The feature maps play the role of local features. As shown in Figure 6, the outputs of improved SPP and NetVLAD are concatenated as the final descriptor. The detailed steps are described as follows.

(1) Use NetVLAD to aggregate the input feature maps as output f^0

(2) Perform GeM pooling on all feature maps to obtain f^1 by

$$f^1 = \left[f_1^1, \cdots, f_k^1, \cdots, f_g^1\right]^T, f_k^1 = \left(\frac{1}{|\mathcal{X}_k|} \sum_{x \in \mathcal{X}_k} x^{p_k}\right)^{1/p_k}, \quad (5)$$

where p_k is the learnable parameter, which could also be adjusted manually, and f_k^1 is the pooling result of the kth feature map calculated by equation (4)

(3) Perform the following operations on $x_{g \times h \times w}$ to obtain the pooling result $f^L, L = 2, 3, 4$

(i) Divide each feature map in $x_{g \times h \times w}$ into $L \times L$ equally nonoverlapping feature submaps. When the height h (or width w) is not an integral multiple of L, the feature maps should be 0 padded until the height (or width) is the minimal integral multiple of L not less than h (or w)

(ii) Perform max pooling on each feature submap, arrange $L \times L$ max pooling results of $L \times L$ feature submaps as a column vector, and then, concatenate g column vectors obtained from g feature submaps as a vector $f^L$with the length of $(g \times L \times L)$

(4) Finally, concatenate f^0, f^1, f^2, f^3, f^4 and perform PCA on the concatenated feature to obtain the final descriptor

5. Experimental Results and Analysis

5.1. Experimental Setup. The proposed method was evaluated in the following setup as shown in Table 1. In the noise filtering layer, the learnable parameters of triplet attention were set as the corresponding line in Table 1. The feature extraction layer retained the NetVLAD setting, which used a partially pretrained VGG16 encoder to extract local features. That is, the last "ReLU & MaxPooling" of VGG16 pretrained on ImageNet was removed and other parts of it were used. Then, only the parameters of the last three convolutional layers were fine-tuned while other parameters were kept unchanged during training. In the feature aggregation layer, the parameters of SPP improved by GeM, $p_k, k = 1, 2, \cdots, 512$ were set as 3.0. In the descriptor matching layer, Faiss [80] was used to accelerate the feature matching process. The training procedure used query image q, positive reference p, and 10 negative references $\{n_1, n_2, \cdots, n_{10}\}$ to form 10 triplets $\{q, p, n_1\}, \{q, p, n_2\}, \cdots, \{q, p, n_{10}\}$, then used triplet loss to calculate loss, and used SGD to optimize the model.

The experimental dataset Pittsburgh 30K [81] contains 51840 Google Street View images captured at different times in the same year, which well simulates the real-application scenario. This dataset was roughly equally divided into 3 parts as train, validation, and test sets, and the number of queries and references contained in each part is shown in Table 2.

5.1.1. Evaluation Metrics. The performance of the proposed method was compared with the classical NetVLAD method in terms of Recall@N and the number of convergence rounds.

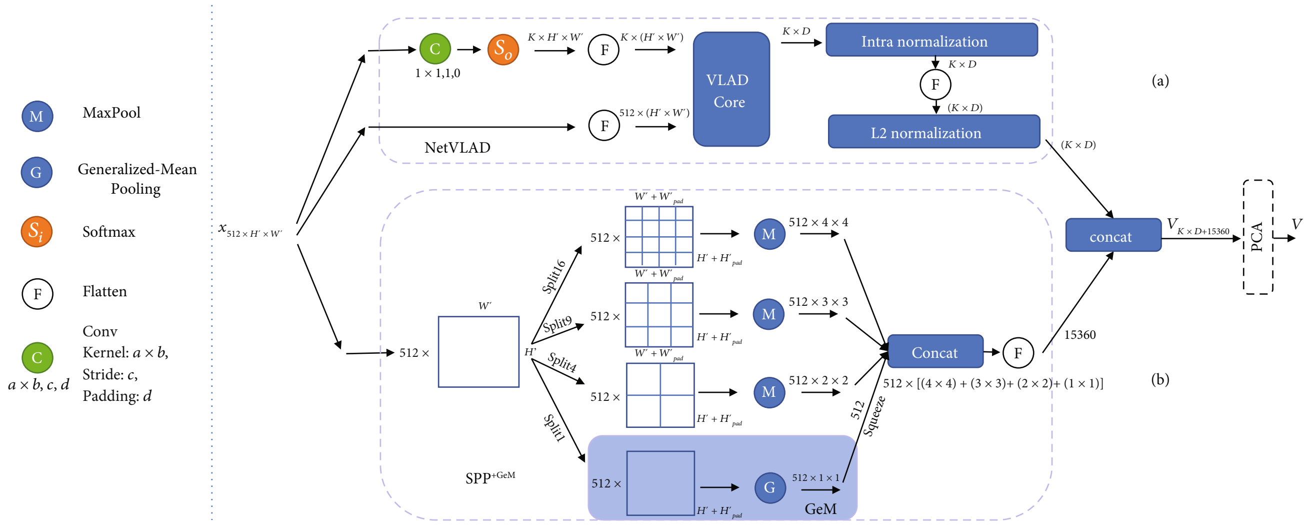


Figure 6: The proposed feature aggregation layer. The upper dashed box (a) represents the process of NetVLAD, and the lower dashed box (b) represents the process of SPP$^{+\text{GeM}}$.

Table 1: Experimental setup.

Layers/others	Setup
Noise filtering layer	Triplet attention, conv kernel size 7 × 7, stride 1, padding length 3
Feature extract layer	The same VGG16 as it is in NetVLAD, only fine-tune the last 3 conv layers
Feature aggregation layer	NetVLAD & SPP improved by GeM, $p_k = 3.0$, SPP's $L = 1, 2, 3, 4$
Descriptor matching layer	European distance implemented by Faiss
Training strategies	Loss function triplet loss, optimizer SGD with momentum = 0.9
Experiment platform	Ubuntu20.04, CUDA 11.0, PyTorch 1.7.1, Faiss 1.7.1.

Table 2: The experimental dataset splits.

Parts	Number of queries	Number of references
Train	7416	10000
Valid	7608	10000
Test	6816	10000

5.1.2. Recall@N. The Recall@N [81] of M queries is calculated in the manner described by the following formula.

$$\text{Recall@}N = \frac{\sum_{j=1}^{M} \min\left(\sum_{i=1}^{N} \alpha_j\left(r_{i,j}\right), 1\right)}{M}, \tag{6}$$

$$\alpha_j\left(r_{i,j}\right) = \begin{cases} 0, & \text{if } \operatorname{dis}\left(q_j, r_{i,j}\right) > 25\,\text{m}, \\ 1, & \text{if } \operatorname{dis}\left(q_j, r_{i,j}\right) \leq 25\,\text{m}. \end{cases} \tag{7}$$

If the distance between the jth query image q_j and the ith closet reference $r_{i,j}$ is less than 25 m, the output of $\alpha_j(r_{i,j})$ is 1 and otherwise 0. Taking query q_j as an example, $R = \{r_{i,j} | i = 1, \cdots, N$ denotes top N closet references of q_j, if the distances between the R and q_j are all greater than 25 m, then, the output of $\sum_{i=1}^{N} \alpha_j(r_{i,j})$ is 0, and the output of min $(\sum_{i=1}^{N} \alpha_j(r_{i,j}), 1)$ is also 0. Conversely, if the distance between any element of R and q_j is less than 25 m, the output of min $(\sum_{i=1}^{N} \alpha_j(r_{i,j}), 1)$ is 1.

Equations (5) and (7) mean that if one of the first N candidate locations of the query is correct, then, the geolocation result is considered correct. And the Recall@N value is the percentage of correctly geolocated queries.

5.1.3. Training Convergence Round. Let $A_E \in (0, 1)$ denote the Recall@N of the Eth round of training epoch E. If max $(A_{E+1}, A_{E+2}, \cdots, A_{E+10}) - A_E \leq 0.0005$, then, the convergence round is considered as E. The lower the value of E, the higher the training efficiency of the model.

5.2. Effectiveness Test of the Noise Filtering Layer. The 1st and 6th rows of Table 3 show the Recall@1, Recall@5, Recall@10, and Recall@20 of the original NetVLAD and the version improved by adding the noise filtering layer before local feature extraction. It can be seen that compared with the original NetVLAD, the improved version gets better accuracy. If performing PCA on the aggregated feature, the superiority of adding noise filtering layer is more significant,

Table 3: The performance comparisons between the proposed method and NetVLAD on the Pitts 30k [81].

Noise filtering layer	Feature aggregation layer	PCA	Dimension	R@1	R@5	R@10	R@20
×	NetVLAD	×	32768	79.45	90.10	92.77	95.19
×	NetVLAD	√	512	77.52	89.41	92.59	95.39
×	NetVLAD	√	1024	78.92	90.10	92.87	95.48
×	NetVLAD	√	2048	79.55	90.42	92.97	95.38
×	NetVLAD	√	4096	79.50	90.23	92.9	95.19
√	NetVLAD	×	32768	80.99	91.09	93.57	95.48
√	NetVLAD	√	512	80.34	91.37	93.97	95.76
√	NetVLAD	√	1024	81.24	91.61	94.07	95.77
√	NetVLAD	√	2048	81.41	91.62	93.85	95.85
√	NetVLAD	√	4096	81.29	91.37	93.63	95.6
√	NetVLAD + SPP^{+GeM}	×	48128	83.67	92.36	94.16	95.77
√	NetVLAD + SPP^{+GeM}	√	512	82.88	92.39	94.91	96.24
√	NetVLAD + SPP^{+GeM}	√	1024	83.73	92.81	94.91	96.20
√	NetVLAD + SPP^{+GeM}	√	2048	84.01	92.78	94.81	96.11
√	NetVLAD + SPP^{+GeM}	√	4096	83.92	92.62	94.5	95.95

"×" means that the operation corresponding to the column name is not applied to the model, and "√" means the opposite. "Dimension" denotes the dimension of the final descriptor; "R@N" denotes Recall@N.

as shown in the 2nd, 3rd, 4th, 5th, 7th, 8th, 9th, and 10th rows of Table 3. The reason should be that the noise filtering layer effectively filters the noises irrelevant to the image geolocation task.

Grad-CAM (gradient-weighted class activation mapping) [82] was used to visualize the comparison of the images before and after the noise filter layer on the R, G, B components of them, as shown in Figure 7. It can be seen that the edges and corners attract more attention after filtering and the focused effective areas are more abundant. The reason should be that in the original NetVLAD, due to the noises, the model may not be able to distinguish the local features in these areas from them in the noisy areas, resulting in the loss of local features important to the image geolocation. But adding the noise filtering layer suppresses the inference of noises in advances, which makes the effective areas attract more attention and the model easier to learn high-quality local features.

Figure 8 shows the training procedure of origin NetVLAD and the version improved by adding the noise filtering layer before local feature extraction. It can be seen that the improved version outperforms the origin NetVLAD in terms of the training convergence round, especially on Recall@10 and Recall@20. The reason may be that image denoising by triplet attention can eliminate the inference of noises, make the model focus on important areas, and then accelerate the learning of features.

The experimental results show that the noise filtering layer is effective in improving the image geolocation performance.

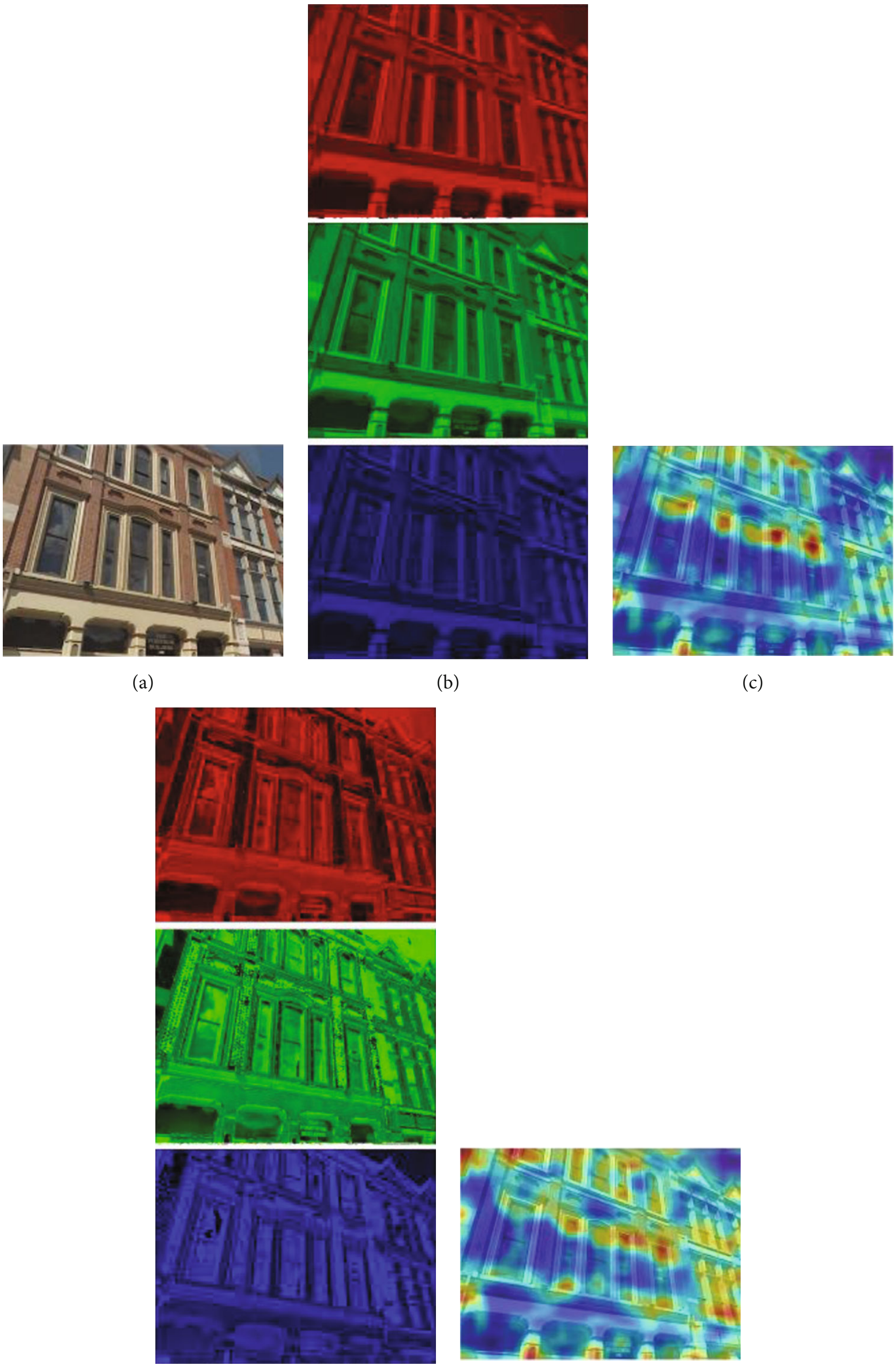
5.3. Effectiveness Test of the Proposed Method. The last 5 rows of Table 3 show the Recall@1, Recall@5, Recall@10, and Recall@20 of the proposed method, which improved by adding the noise filtering layer before local feature extraction and incorporating local factor into the aggregation layer. Experimental results show that the proposed method has a significant improvement in geolocation accuracy, especially, when PCA is used to reduce the dimensionality of the final descriptor. The performance improvement maybe because the proposed feature aggregation layer reduces the task-irrelevant components in the features by combining NetVLAD with SPP (improved by GeM) and performing PCA. The NetVLAD is used to compute the global factors, and the SPP improved by GeM is used to compute the local factors, which reflect the position relationships of local features.

Moreover, the computational efficiency also has been improved by PCA. That is, the computational efficiency and geolocation performance have both been improved by the proposed method.

Figure 9 shows an image geolocation example of the original NetVLAD method and the proposed method. It can be seen that the geometric elements in the references retrieved by the original NetVLAD have stronger similarities to those of the query and the geometric elements in the references retrieved by the proposed method not only have stronger similarity to them of the query but also own similar position relationships to them of the query. And the proposed method gets a more accurate image geolocation result. This is because the original NetVLAD only extracts the global factors, but the proposed method extracts both the global factors and the local factors, which reflect the position relationships among the geometric elements.

Furthermore, it can be seen in Figure 8 that the proposed method still remains the advantage of the version improved by adding a noise filtering layer in terms of training convergence rounds.

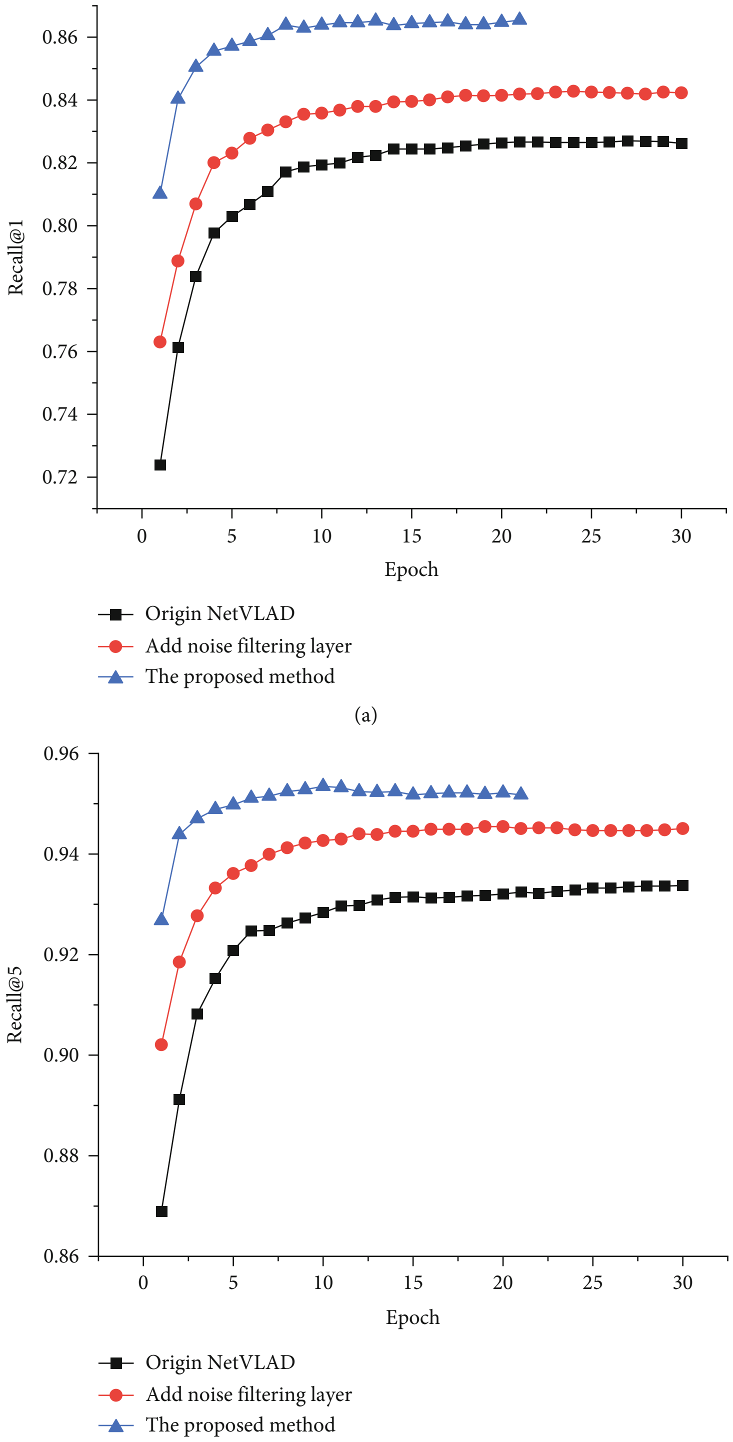
In a word, the proposed method outperforms classical NetVLAD in both geolocation accuracy and training speed.



(a) (b) (c)

(d) (e)

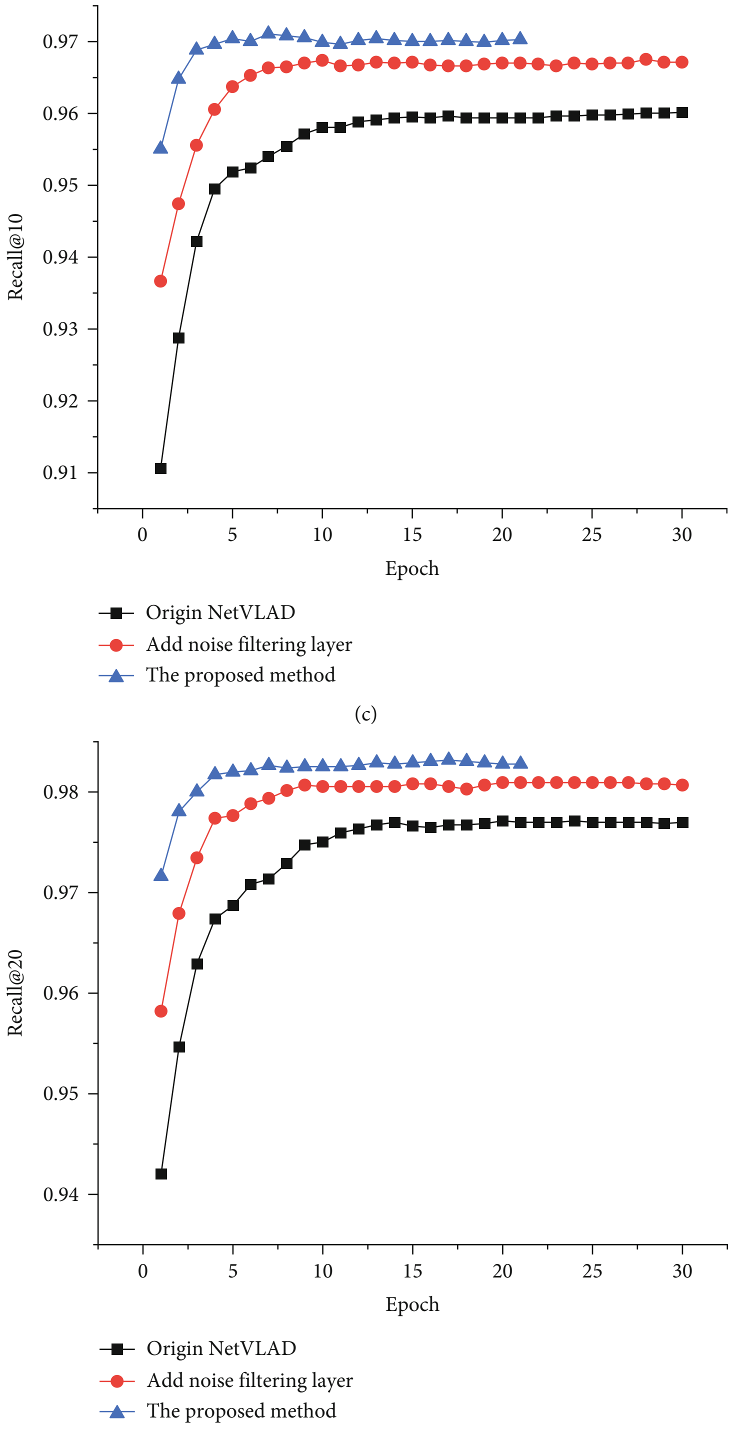
Figure 7: The Grad-CAM visualizations for the original NetVLAD and the version adding noise filtering layer. (a) The original image, (b) the R, G, B components of the original image, (c) the heat map of the original NetVLAD, (d) the R, G, B components of the image after the noise filtering layer, and (e) the heat map of the version improved by adding the noise filtering layer.



(a)

(b)

Figure 8: Continued.



(c)

(d)

Figure 8: The training performance comparison of classical NetVLAD and the proposed method.

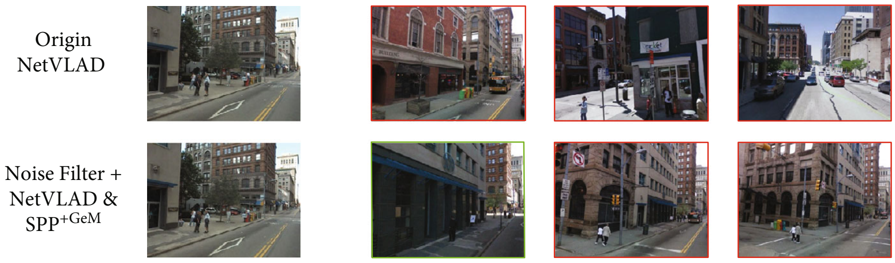


Figure 9: An example of geolocation results.

6. Conclusions

In this work, we have proposed a novel image geolocation framework by adding the noise filtering layer before feature extraction. Based on this framework, an image geolocation method based on attention mechanism front loading and feature fusion has been designed. Unlike original NetVLAD, our method uses triplet attention to denoise images and gets more effective descriptors by considering not only global factors but also local factors reflected by the relationships of local features extracted by an improved SPP. Experimental results show that our proposed method outperforms the original NetVLAD in terms of Recall@N($N = 1, 5, 10, 20$) and training convergence round.

Research works such as DELG and PatchNetVLAD show that the accuracy can be further improved by geometric verification. However, the verification procedure is extremely time consuming and its time complexity is closely related to the number of references under the same recall rate, viz., the value of N in Recall@N. Therefore, in future works, we will combine the proposed method with geometric verification, to reduce the time complexity and improve the accuracy of methods, such as DELG and patch-NetVLAD. Furthermore, we will try to extend the proposed method to other fields related to image retrieval.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

We thank Hao Li for helpful discussions. This work is supported by the National Natural Science Foundation of China (Grant nos. 61872448, 61772549, and U1804263) and Science and Technology Research Project of Henan Province (no. 222102210075), China.

References

[1] C. Lassance, Y. Latif, R. Garg, V. Gripon, and I. Reid, “Improved visual localization via graph filtering,” *Journal of Imaging*, vol. 7, no. 2, p. 20, 2021.

[2] P. E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, “From coarse to fine: robust hierarchical localization at large scale,” in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12708–12717, Long Beach, CA, USA, 2019.

[3] L. Heng, B. Choi, Z. Cui et al., “Project autovision: localization and 3d scene perception for an autonomous vehicle with a multi-camera system,” in *In International Conference on Robotics and Automation (ICRA)*, pp. 4695–4702, Montreal, QC, Canada, 2019.

[4] S. Zahedian, K. F. Sadabadi, and A. Nohekhan, “Localization of autonomous vehicles: proof of concept for a computer vision approach,” in *In ITS America Annual Meeting*, Washington, D.C, 2019.

[5] N. Pion, M. Humenberger, G. Csurka, Y. Cabon, and T. Sattler, “Benchmarking image retrieval for visual localization,” in *In International Conference on 3D Vision (3DV)*, pp. 483–494, London, UK, 2020.

[6] R. Castle, G. Klein, and D. W. Murray, “Video-rate localization in multiple maps for wearable augmented reality,” in *In 12th IEEE International Symposium on Wearable Computers*, pp. 15–22, Pittsburgh, PA, USA, 2008.

[7] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet, “A survey on visual-based localization: on the benefit of heterogeneous data,” *Pattern Recognition*, vol. 74, pp. 90–109, 2018.

[8] X. Xin, J. Jiang, and Y. Zou, “A review of visual-based localization,” in *In Proceedings of the International Conference on Robotics, Intelligent Control and Artificial Intelligence*, pp. 94–105, Shanghai, China, 2019.

[9] S. Lowry, N. Sunderhauf, P. Newman et al., “Visual place recognition: a survey,” *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[10] S. Cao and N. Snavely, “Graph-based discriminative learning for location recognition,” in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 700–707, Portland, OR, USA, 2013.

[11] H. Taira, M. Okutomi, T. Sattler et al., “InLoc: indoor visual localization with dense matching and view synthesis,” in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 7199–7209, Salt Lake City, UT, USA, 2018.

[12] H. Germain, G. Bourmaud, and V. Lepetit, “Sparse-to-dense hypercolumn mtching for long-term visual localization,” in

In 2019 International Conference on 3D Vision (3DV), pp. 513–523, Québec, Canada, 2019.

[13] W. Zhang and J. Kosecka, "Image based localization in urban environments," in *In Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*, pp. 33–40, Chapel Hill, USA, 2006.

[14] Z. Laskar, I. Melekhov, S. Kalia, and J. Kannala, "Camera relocalization by computing pairwise relative poses using convolutional neural network," in *In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV)*, pp. 920–929, Venice, Italy, 2017.

[15] V. Balntas, S. Li, and V. Prisacariu, "Relocnet: continuous metric learning relocalisation using neural nets," in *In Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 782–799, Munich, Germany, 2018.

[16] M. DIng, Z. Wang, J. Sun, J. Shi, and P. Luo, "CamNet: coarse-to-fine retrieval for camera re-localization," in *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 2871–2880, Seoul, Korea (South), 2019.

[17] T. Sattler, A. Torii, J. Sivic et al., "Are large-ccale 3D models really necessary for accurate visual localization?," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 6175–6184, Honolulu, HI, USA, 2017.

[18] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li, "Self-supervising fine-grained region similarities for large-scale image localization," in *In European Conference on Computer Vision (ECCV)*, pp. 369–386, Glasgow, UK, 2020.

[19] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-NetVLAD: multi-scale fusion of locally-global descriptors for place recognition," in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14141–14152, Nashville, TN, USA, June 2021.

[20] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307, Las Vegas, NV, USA, June 2016.

[21] Z. Chen, L. Liu, I. Sa, Z. Ge, and M. Chli, "Learning context flexible attention model for long-term visual place recognition," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4015–4022, 2018.

[22] J. L. Schonberger and J. M. Frahm, "Structure-from-motion revisited," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, Las Vegas, NV, USA, June 2016.

[23] C. Harris and M. Stephens, "A combined corner and edge detector," *In Alvey Vision Conference*, vol. 15, no. 50, 1988.

[24] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[26] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: speeded up robust features," in *In European Conference on Computer Vision (ECCV)*, pp. 404–417, Graz, Austria, 2006.

[27] E. Johns and G.-Z. Yang, "From images to scenes: compressing an image cluster into a single scene model for place recognition," in *In International Conference on Computer Vision (ICCV)*, pp. 874–881, Barcelona, Spain, Jan. 2011.

[28] D. Mishkin, M. Perdoch, and J. Matas, "Place recognition with WxBS retrieval," in *CVPR Workshop on Visual Place Recognition in Changing Environments*, Czech Technical University in Prague, 2015.

[29] J. Hays and A. A. Efros, "Large-scale image geolocalization," in *Multimodal Location Estimation of Videos and Images Estim Videos Images*, J. Choi and G. Friedland, Eds., pp. 41–62, Springer, Cham, 2015.

[30] E. Zemene, Y. T. Tesfaye, H. Idrees, A. Prati, M. Pelillo, and M. Shah, "Large-scale image geo-localization using dominant sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, vol. 41, pp. 148–161, 2019.

[31] S. Ardeshir, A. R. Zamir, A. Torroella, and M. Shah, "GIS-assisted object detection and geospatial localization," in *In European Conference on Computer Vision (ECCV)*, pp. 602–617, Zurich, Switzerland, 2014.

[32] D. Robertson and R. Cipolla, "An image-based system for urban navigation," in *In Proceedings of the 15th British Machine Vision Conference (BMVC)*, pp. 819–828, London, U.K, Sep. 2004.

[33] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–7, Minneapolis, MN, USA, 2007.

[34] J. Hays and A. A. Efros, "IM2GPS: Estimating geographic information from a single image," in *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, Anchorage, AK, USA, June 2008.

[35] M. Cummins and P. Newman, "FAB-MAP: probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research (IJRA)*, vol. 27, no. 6, pp. 647–665, 2008.

[36] A. R. Zamir and M. Shah, "Accurate image localization based on Google Maps Street View," in *In European Conference on Computer Vision (ECCV)*, pp. 255–268, Crete, Greece, 2010.

[37] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[38] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *In Proceedings of IEEE International Conference on Computer Vision(ICCV)*, pp. 1470–1477, Nice, France, 2003.

[39] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: binary robust independent elementary features," in *In European Conference on Computer Vision (ECCV)*, pp. 778–792, Heraklion, Crete, Greece, 2010.

[40] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *In European Conference on Computer Vision (ECCV)*, pp. 430–443, Graz, Austria, 2006.

[41] C. Masone and B. Caputo, "A survey on deep visual place recognition," *IEEE Access*, vol. 9, pp. 19516–19547, 2021.

[42] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *International Journal of Computer Vision(IJCV)*, vol. 124, no. 2, pp. 237–254, 2017.

[43] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *In European Conference on Computer Vision (ECCV)*, pp. 392–407, Zurich, Switzerland, Sep. 2014.

[44] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," in *In*

International Conference on Learning Representations(ICLR), pp. 1–12, San Juan, Puerto Rico, May 2016.

[45] N. Sünderhauf, S. Shirazi, A. Jacobson et al., "Place recognition with convnet landmarks: viewpoint-robust, condition-robust, training-free," *Robotics: Science and Systems XI*, vol. 11, pp. 1–10, 2015.

[46] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," in *In Proceedings of the 16th Australasian Conference on Robotics and Automation (ARAA)*, pp. 1–8, Melbourne, Australia, 2014.

[47] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from ConvNet for visual place recognition," in *In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9–16, Vancouver, BC, Canada, Sep. 2017.

[48] P. Neubert and P. Protzel, "Local region detector + CNN based landmarks for practical place recognition in changing environments," in *In 2015 European conference on Mobile robots (ECMR)*, pp. 1–6, Lincoln, UK, Sep. 2015.

[49] P. Panphattarasap and A. Calway, "Visual place recognition using landmark distribution descriptors," in *In Asian Conference on Computer Vision (ACCV)*, pp. 487–502, Taipei, Taiwan, 2017.

[50] N. Sunderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of ConvNet features for place recognition," in *In IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 4297–4304, Hamburg, Germany, 2015.

[51] Y. Hou, H. Zhang, and S. Zhou, "Convolutional neural network-based image representation for visual loop closure detection," in *In IEEE International Conference on Information and Automation*, pp. 2238–2245, Lijiang, China, Aug. 2015.

[52] H. Jégou and O. Chum, "Negative evidences and co-occurences in image retrieval: the benefit of PCA and whitening," in *In European Conference on Computer Vision (ECCV)*, pp. 774–787, Florence, Italy, Oct. 2012.

[53] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *In Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, Columbus, OH, USA, June 2014.

[54] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool, "Night-to-day image translation for retrieval-based localization," in *In International Conference on Robotics and Automation (ICRA)*, pp. 5958–5964, Montreal, QC, Canada, Aug. 2019.

[55] A. Torii, R. Arandjelovi, S. Masatoshi, and O. Tomas, "24/7 place recognition by view synthesis," in *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1808–1817, Boston, MA, USA, June. 2015.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *In Proceedings of International Conference on Learning Representations*, pp. 1–14, San Diego, CA, USA, May. 2015.

[57] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3304–3311, San Francisco, CA, USA, June. 2010.

[58] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition," *IEEE Transactions On Neural Networks And Learning Systems*, vol. 31, no. 2, pp. 661–674, 2020.

[59] H. J. Kim, E. Dunn, and J.-M. Frahm, "Learned contextual feature reweighting for image geo-localization," in *In IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 3251–3260, Honolulu, HI, USA, July 2017.

[60] G. Tolias, Y. Avrithis, and H. Jégou, "Image search with selective match kernels: aggregation across single and multiple images," *International Journal of Computer Vision*, vol. 116, pp. 247–261, 2016.

[61] 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, 2018.

[62] A. Khaliq, S. Ehsan, M. Milford, and K. McDonald-Maier, "Camal: context-aware multi-scale attention framework for lightweight visual place recognition," 2019, http://arxiv.org/abs/1909.08153.

[63] G. Peng, Y. Yue, J. Zhang, Z. Wu, X. Tang, and D. Wang, "Semantic reinforced attention learning for visual place recognition," in *In International Conference on Robotics and Automation (ICRA)*, pp. 13415–13422, Xi'an, China, June 2021.

[64] G. Peng, J. Zhang, H. Li, and D. Wang, "Attentional pyramid pooling of salient visual residuals for place recognition," in *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 885–894, Montreal, QC, Canada, Oct. 2021.

[65] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *In Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 3476–3485, Venice, Italy, Oct. 2017.

[66] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 37, no. 9, pp. 1904–1916, 2015.

[67] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: convolutional triplet attention module," in *In IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3139–3148, Waikoloa, HI, USA, Jan. 2021.

[68] V. Christlein, L. Spranger, M. Seuret, A. Nicolaou, P. Král, and A. Maier, "Deep generalized max pooling," in *In International conference on document analysis and recognition (ICDAR)*, pp. 1090–1096, Sydney, NSW, Australia, Sep. 2019.

[69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[70] M. Tan and Q. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," *In International Conference on Machine Learning (ICML)*, vol. 97, pp. 6105–6114, 2019.

[71] L. Liu, H. Li, and Y. Dai, "Stochastic attraction-repulsion embedding for large scale image localization," in *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2570–2579, Seoul, Korea (South), Nov. 2018.

[72] M. H. Guo, T. X. Xu, J. J. Liu et al., "Attention mechanisms in computer vision: a survey," http://arxiv.org/abs/2111.07624.

[73] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 510–519, Long Beach, CA, USA, June. 2019.

[74] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, Salt Lake City, UT, USA, Dec. 2018.

[75] F. Wang, M. Jiang, C. Qian et al., "Residual attention network for image classification," in *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164, Honolulu, HI, USA, July 2017.

[76] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *In Proceedings of European Conference on Computer Vision (ECCV)*, pp. 3–19, Munich, Germany, Sep. 2018.

[77] M. Yang, D. He, M. Fan et al., "DOLG: single-stage image retrieval with Deep orthogonal fusion of local and global features," in *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11772–11781, Montreal, QC, Canada, Oct. 2021.

[78] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *In European Conference on Computer Vision (ECCV)*, pp. 726–743, Glasgow, UK, Aug. 2020.

[79] Q. Zhou, T. Sattler, M. Pollefeys, and L. Leal-Taixe, "To learn or not to learn: visual localization from essential matrices," in *In International Conference on Robotics and Automation (ICRA)*, pp. 3319–3326, Paris, France, Aug. 2020.

[80] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," 2017, http://arxiv.org/abs/1702.08734.

[81] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 883–890, Portland, OR, USA, June 2013.

[82] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," in *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 618–626, Venice, Italy, Oct. 2017.