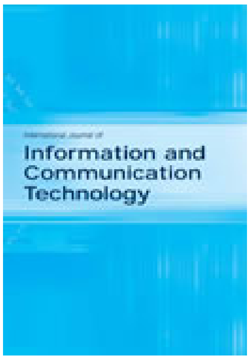


International Journal of
Information and
Communication
Technology

**International Journal of Information and Communication Technology**

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

# Single-stage landmark retrieval with texture feature fusion

Kun Tong, GuoXin Tan

**Article History:**

| | |
|---|---|
| Received: | 30 July 2024 |
| Last revised: | 18 September 2024 |
| Accepted: | 20 September 2024 |
| Published online: | 13 December 2024 |

# Single-stage landmark retrieval with texture feature fusion

Kun Tong* and GuoXin Tan

National Research Center of Cultural Industries,
Central China Normal University,
WuHan, 430079, China
Email: ncrilgxs@outlook.com
Email: gxtan@mail.ccnu.edu.cn
*Corresponding author

**Abstract:** Existing landmark retrieval models typically fuse global and local feature descriptors of target images to generate feature vectors for landmark retrieval. However, these models often exhibit poor resilience to complex viewpoints, occlusions, and lighting conditions. Moreover, the fused feature descriptors still contain substantial redundant information, leading to decreased retrieval accuracy. To address these issues, this paper proposes a novel single-stage image retrieval model enhanced by texture augmentation. The model incorporates a texture enhancement module that leverages texture feature encoding to reconstruct the original feature maps, amplifying the influence of texture features in deep feature vectors across different scales. This approach ensures robust feature representation under extreme angles, occlusions, or varying lighting conditions. To mitigate the problem of redundant features, the model introduces an innovative feature fusion module. This module optimises local features from multi-scale feature descriptors using a mapping fusion technique, eliminating redundant information and generating more compact and discriminative feature descriptors. Extensive experiments demonstrate that the proposed model achieves significant improvements in retrieval performance compared to state-of-the-art image retrieval models, while maintaining acceptable retrieval times.

**Keywords:** landmark retrieval; feature fusion; multi-scale fusion; feature enhancement.

**Reference** to this paper should be made as follows: Tong, K. and Tan, G. (2024) ‘Single-stage landmark retrieval with texture feature fusion’, *Int. J. Information and Communication Technology*, Vol. 25, No. 9, pp.43–59.

**Biographical notes:** Kun Tong is a Research Fellow of the National Cultural Industry Research Center of the Central China Normal University, member of the Chinese Computer Federation, Member of the IEEE Computer Society, member of the Association for Computing Machinery, and member of the Association for the Advancement of Artificial Intelligence. His research fields: recommender system, representation learning and deep learning.

GuoXin Tan is the Deputy Director, Professor, and Doctoral Supervisor of the National Cultural Industry Research Center of Central China Normal University, ‘3551’ talent introduction program of Wuhan East Lake National Innovation Demonstration Zone, Post-doctoral Fellow at the University of Tokyo, member of the Multimedia Professional Committee of the Chinese

Computer Society, and member of the Wuhan Cultural and Technological Integration Expert Group, Secretary-General of the Wuhan Animation Association. His research fields: geographic information system (GIS) and deep learning.

## 1 Introduction

Landmark retrieval is an important part of the computer vision field. Given a landmark image, landmark retrieval aims to retrieve landmark images that contain the same subject target as the queried image in the image database through an algorithmic model. The landmark retrieval model has a wide range of real-world applications, such as travel recommendation. Due to the diversity of shooting conditions, the quality of the image to be queried will be affected to a certain extent. For example, lighting, shadows, viewing angle, distance and proximity, occlusion or jitter, etc. (Teichmann et al., 2019). These uncertainties may affect the quality of the landmark image. These uncertainties will bring great challenges to the landmark retrieval model.

Given an image, existing landmark retrieval models usually obtain the most similar image by extracting the representation vectors that can describe the features of the image and measuring the similarity between the vectors. There are two types of image representation vectors: global and local (Giorgos et al., 2020). Global representations are feature descriptors obtained through deep, fine-grained image feature extraction. These typically express abstract semantic relationships within images, such as the connection between targets and their surrounding backgrounds. However, they often lose shallow image features, including the local spatial structure and texture characteristics of the main subject (Wang et al., 2022). Conversely, local representations focus on capturing feature descriptors from multiple regions of interest within an image. These descriptors usually contain shallow representations of local areas, such as local spatial structures and texture patterns, but may overlook the intrinsic connections between multiple regions of interest (Chen et al., 2022).

Generally, global representations contribute to improving recall rates, while local representations enhance precision. Global representations demonstrate robust retrieval capabilities for targets in various poses, whereas local representations are more advantageous for localising the main subject and excluding interference from other objects in the image. Consequently, most state-of-the-art landmark retrieval models employ both global and local representations (Song et al., 2022a, 2022b; Mohtashami and Jaggi, 2023; Liu et al., 2024; Do and Sinha, 2024). To address the semantic fusion challenge between these multi-scale feature vectors, various feature fusion algorithms have been proposed to semantically complement the advantages of both, thereby enhancing the model's practical application and robustness. However, these models typically face the following issues:

1 Existing landmark retrieval models often use output feature maps from intermediate and deep network layers for local and global feature branches, respectively. This process, involving multiple convolution and pooling operations, results in significant loss of shallow semantic information. However, in complex retrieval conditions such as distorted viewpoints, varying lighting, and occlusions, shallow features,

particularly texture features, often offer superior discriminative power (Do et al., 2022). Texture features provide rich semantic information about spatial structures and local details of landmarks, remaining distinctive even under challenging conditions. As surface-level features with geometric invariance, textures maintain their spatial structure under distorted viewpoints, offering robustness in complex environments. The loss of these crucial features leads to inadequate image representation and suboptimal retrieval performance.

2 While existing multi-scale feature fusion algorithms can connect global and local representations in the form of feature point pairs and merge them into a highly reliable feature vector, the fused feature vector often contains numerous redundant features (Wu et al., 2023). These redundant features not only slow down the model's retrieval speed in subsequent processes but may also introduce irrelevant feature vectors that interfere with the landmark retrieval process, potentially decreasing the model's retrieval accuracy.

To address the aforementioned issues, we propose a single-stage retrieval model with texture feature enhancement. The model incorporates a novel texture feature enhancement module designed to extract texture information embedded in shallow feature maps, generating corresponding texture feature maps that are propagated to subsequent global and local feature branches. This approach amplifies the influence of texture features in deep feature maps, resulting in more discriminative final feature representations. Within the local feature module, a self-attention mechanism is employed to select more distinctive local features, thereby enhancing the representational capacity of local feature vectors. The model culminates with a feature fusion module that integrates local and global features, effectively eliminating redundant features and mitigating the impact of irrelevant features on subsequent retrieval processes. This fusion process generates the final feature descriptor for landmark retrieval and matching. Experimental results demonstrate that the retrieval model augmented with the texture enhancement module significantly improves both image retrieval accuracy and feature extraction efficiency. The main contributions are summarised as follows:

- We propose a feature enhancement module to improve the expressiveness of multi-scale feature descriptors. By propagating relevant texture features from shallow layers to deep feature descriptors and reconstructing them, we enrich the representation of feature descriptors, enhancing their robustness in complex environments.
- We introduce a novel multi-scale feature fusion module that optimises local features from multi-scale feature descriptors generated by the backbone network. This module eliminates redundant features and performs subsequent feature fusion, resulting in more compact and discriminative fused feature descriptors.
- Through extensive experiments on the Revisited Oxford and Paris datasets, we demonstrate that our proposed model achieves superior retrieval accuracy and performance compared to other state-of-the-art models.

## 2 Related work

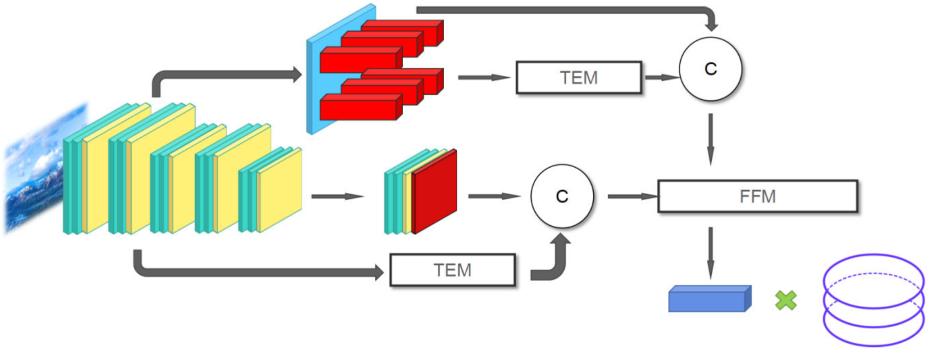
Contemporary landmark retrieval models predominantly utilise cross-scale fused features as guiding features for subsequent retrieval processes. This approach involves employing feature fusion methods to merge multi-scale feature vectors into a compact feature representation. A primary objective of feature fusion is to complement the strengths of global and local features, thereby obtaining more informative semantic features and consequently enhancing retrieval accuracy (Jerome et al., 2019; Weinzaepfel et al., 2022; Weyand et al., 2020; Xiao et al., 2020). In landmark retrieval, local features play a crucial role by capturing the most distinctive partial characteristics of the object being searched, thereby enabling precise target localisation. With the advancement of deep learning technologies, an increasing number of researchers have proposed local feature extraction methods based on various neural network architectures (Gordo et al., 2016; Philbin et al., 2007; Noh et al., 2016). Global features, on the other hand, compensate for the inherent discreteness of local features and their lack of consideration for overall geometric structures, offering a more comprehensive representation. In scenarios requiring high-precision object localisation within images, global features often outperform local features (Babenko and Lempitsky, 2015; Sain et al., 2021; Tolias et al., 2015). As the depth of feature extraction networks in existing models continues to increase, the semantic representations contained in the extracted global and local feature maps become increasingly abstract. This shift moves away from focusing on the extraction of feature points from the original image towards exploring the relationships between feature point pairs or subsets (Zhang et al., 2023). Consequently, in scenarios with strong interfering factors such as distorted viewpoints, occlusions, and unnatural lighting conditions, retrieval models may experience performance degradation due to their inability to effectively mine these inter-feature relationships (Salih and Abdulla, 2023). In response to this challenge, some researchers have begun to explore methods of feature enhancement for deep feature maps (Baldrati et al., 2022; Suo et al., 2024). One promising direction involves introducing shallow semantic features into deep feature spaces, such as colour (Asadi Amiri et al., 2022), geometric contours (Liu et al., 2022), and texture features (Liu and Yang, 2023; Varish, 2022). Among these, texture feature models have gained particular favour among researchers due to their geometric invariance, demonstrating significant improvements in various domains (Öztürk et al., 2023; Kelishadrokhi et al., 2023; Ahmad, 2022).

Beyond feature enhancement, the methods used for feature fusion are also a focal point of research. Local and global features differ in their attention to image content due to the varying scales of their feature extraction receptive fields. Global features focus on macro-scale and highly abstract semantic content, while local features concentrate on specific regions within the image (Xie et al., 2023). The efficient and accurate fusion of these different types of feature descriptors remains a critical research topic. Existing feature fusion methods often employ pyramid structures (Pipanmekaporn et al., 2023; Ao and Wu, 2023) to achieve feature fusion, leading to the development of various network architectures such as serial branch structures (Xu et al., 2023; Lu et al., 2024), parallel branch structures (Weng et al., 2023; Chen et al., 2024), and multi-scale feature output networks (Bhatti et al., 2023). While these model structures construct feature descriptors with strong representational capabilities, most of these fusion methods focus on semantic matching between feature points across different scale feature maps. They often neglect the optimisation of the feature descriptors themselves, resulting in fused feature vectors that contain significant redundant features, potentially impacting the accuracy and performance of landmark retrieval models.

## 3 Method

In this chapter, a single-stage landmark retrieval model with texture feature fusion (TFFLR) is presented. The model extracts shallow texture semantic information and fuses it with global depth semantic information and local depth semantics, respectively, to obtain texture-enhanced feature descriptors. In particular, this paper uses a cross-scale semantic feature descriptor fusion module to fuse the global semantic feature descriptor with the local semantic feature descriptor, while retaining the two kinds of descriptive information to form a feature descriptor that can better describe the complete features of the image. The overall model is shown in Figure 1.

**Figure 1** Landmark retrieval model with texture feature fusion (see online version for colours)



In Figure 1, 'TEM' and 'FFM' denote texture enhancement module and feature fusion module, respectively. 'C' denotes concatenation operator. In this model, the input image is processed through a multi-layer backbone network into a feature descriptor that contains the overall abstract semantic information of the image. The feature maps in the middle layer of the network are used as input to the local feature extraction network. After the local feature extraction network, one of the features is processed by the texture enhancement module to get the enhanced texture feature descriptor, and the other one is connected by skip connection to get the final local semantic feature descriptor. The feature descriptors containing the overall abstract semantics of the image are also connected using the same structure as the local feature extraction branch network to obtain the final global semantic feature descriptors. In the texture enhancement module, the quantisation coding method is used to quantise and code the features of the feature map obtained from the backbone network, and then the quantisation coding map and the statistical feature map of the feature are obtained respectively. The two feature maps are then used as inputs to the texture enhancement algorithm to obtain the final texture-enhanced feature descriptors. In the feature fusion module, in order to better fuse the global semantic features with the local semantic features, this paper uses the mapping fusion module to compute the orthogonal difference mapping of the local semantic features with respect to the global semantic features, and obtain the orthogonal mapping feature map. Finally, the mapping map is merged with the original global semantic features and fed into a single-layer fully connected network to obtain the final fused feature map. The feature map is used as the retrieval and matching benchmark, which is fed into the image database for retrieval, and the most similar images are retrieved, and the labels corresponding to these images are output as the results to obtain the final retrieval results.

### *3.1 Texture enhancement module*

Existing landmark retrieval methods typically employ high-level abstract semantic features to retrieve similar information, which can indeed achieve favourable retrieval results in certain specific scenarios. While utilising high-level abstract semantic features alone allows retrieval systems to comprehend the characteristics of the retrieval target from a holistic perspective, in the domain of landmark retrieval, fine-grained features often play a crucial role. For instance, the edges and shapes of buildings are particularly significant. However, directly applying existing image retrieval methods often results in the loss of this vital information. These detailed features are encapsulated within shallow texture features. Therefore, for applications such as landmark retrieval, it is imperative to integrate texture features into multi-scale feature vectors to achieve superior retrieval results.

To address this, we introduce a texture enhancement module based on a specialised texture feature quantisation operator. This module extracts comprehensive texture feature encodings from shallow feature maps. Furthermore, we propose a feature enhancement algorithm that utilises texture feature encodings to reconstruct the original feature maps. This process amplifies the influence of texture features within feature vectors at different scales, enabling feature vectors across all scales to better capture the structural details of landmark images. Consequently, this approach enhances the model's robustness under complex viewing angles and lighting conditions, thereby improving the accuracy and accelerating the retrieval speed of landmark retrieval models.

#### *3.1.1 Texture feature quantisation operator*

Most deep neural networks for image processing have a fully convolutional layer, but the fully convolutional layer is sensitive to small local variations in the image, which can lead to the inability to accurately represent the statistical features of texture. In this paper, we propose to use the texture feature quantisation operator to characterise the statistical features of texture. In particular, the purpose of this operator is to quantise the input texture features into multiple layers, and quantise and encode the features in each layer to obtain the texture quantisation encoding matrix, and finally perform regular counting on this quantisation matrix to obtain the texture feature encoding. The specific structure of the operator is shown in Figure 2 as follows:

1 Quantise

   Suppose the size of the original feature map $A$ is $R^{C \times H \times W}$, and the feature vector after global average pooling is $G$, whose size is $R^{C \times 1 \times 1}$. First, calculate the cosine similarity $S$ between the feature map $A$ and the feature map $G \in R^{1 \times H \times W}$:

$$S_{i,j} = \frac{G \cdot A_{i,j}}{\| G \|_2 \cdot \| A_{i,j} \|_2} \tag{1}$$

where $S_{i,j}$ denotes the value of $S$ at $(i, j)$ spatial locations, and when $S_{i,j}$ is computed for all locations, the final cosine similarity matrix $S$ is obtained. Subsequently, the matrix $S$ is deformed into a long vector $S_{reshape} \in R^{HQ}$, and $S_{reshape}$ is divided equally into $N$ intervals, and the $L_n$ for each interval is obtained by the following computation.

$$L_n = \frac{\max\left(S_{reshape}\right) - \min\left(S_{reshape}\right)}{N} \cdot n + \min\left(S_{reshape}\right) \tag{2}$$

Finally, the quantisation coding matrix $E \in R^{N \times HW}$ is calculated from $S_{reshape}$ obtained from equation (1) and $L_n$ obtained from equation (2):

$$E_{n,i} = \begin{cases} 1 - \left| L_n - S_i \right| & \text{if } -\frac{0.5}{N} \leqslant L_n - S_i < \frac{0.5}{N} \\ 0 & \text{else} \end{cases} \tag{3}$$

where $i \in [1, HW]$, $n \in [1, N]$, $S_i$ at each spatial position is computed with $L$ at all intervals to get the quantisation coding vector at that position, when the quantisation coding vectors at all positions have completed the computation of equation (3), the final result, the quantisation coding matrix $E$ is obtained.

2 Regular count

Given a quantisation coding matrix $E$, the quantisation counting map associated with it can be obtained $Q \in R^{N \times 2}$, where the first dimension represents the position n of the interval in which the counting is to be performed, while the second dimension represents the corresponding regular counts under that interval. The quantisation count map is derived from the following equation:

$$Q = Cat\left( L, \frac{\sum_{i=1}^{HW} E_{i,n}}{\sum_{n=1}^{N} \sum_{i=1}^{HW} E_{i,n}} \right) \tag{4}$$

where the *Cat* operator denotes the concatenation operation, the positions corresponding to the two vectors are directly connected to obtain the merged regular count matrix $Q$.

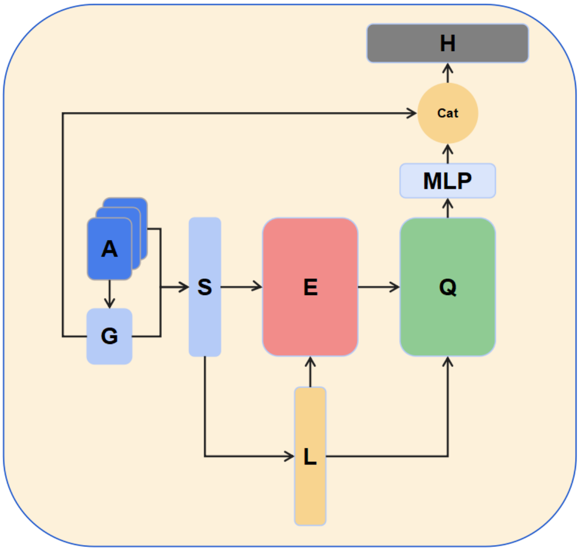
3 Average feature coding

The regular count matrix obtained from equation (4) reflects the feature distribution in the image feature map, which belongs to the statistically relevant information. In order to further obtain more effective feature information, it is necessary to merge and recode the average pooled feature map $G$ with the regular count matrix $Q$ to obtain the final average feature encoding:

$$H = Cat\left(MLP(Q), G\right) \tag{5}$$

where $H$ denotes the average feature encoding, $MLP$ denotes multilayer perceptron and $G$ denotes the feature vector after average pooling in the first step, the purpose of the MLP operator is to resize the $Q$ matrix to match $G$. This is usually achieved using a two-layer perceptron, where the first layer uses the leaky ReLU function as the activation function.

**Figure 2** Structure of the texture feature quantisation operator (see online version for colours)



### 3.1.2 Texture enhancement operator

The role of the texture feature quantisation operator is to extract semantic information describing image textures from shallow feature maps of images and form corresponding texture feature encodings. However, in landmark retrieval models, the multi-scale deep feature maps extracted through deep neural networks often lose a significant amount of texture information. To recover texture feature information in deep feature maps, this paper proposes a texture feature enhancement algorithm. This algorithm utilises texture feature encodings to amplify the influence of texture features within multi-scale deep feature maps. By doing so, it enhances the ability of deep feature maps to discern structural details in images, thereby improving the retrieval accuracy of the model.

Firstly, the shallow feature map is processed by the texture feature quantisation operator to obtain the quantisation coding matrix $E$ [equation (3)], and the average feature coding $H$ [equation (5)]. Subsequently, the reconstructed quantisation matrix $J$ is obtained by the following equation:

$$L = soft\max\left(\phi_1(H)^T \cdot \phi_2(H)\right) \tag{6}$$

$$J = \phi_3(H) \cdot L \tag{7}$$

where $\phi_1$, $\phi_2$ and $\phi_3$ denote different 1 × 1 convolutions respectively and softmax denotes the softmax function. Finally, the final reconstructed coding matrix $U$ is calculated by the following equation:

$$U = J \cdot E \tag{8}$$

The reconstruction coding matrix can be viewed as the contribution weight of each pixel point on the image in the shallow texture features, which is used to determine which regions in the shallow features have a greater impact on subsequent retrieval. Finally, to mitigate the potential for vanishing gradients, we employ a skip connection architecture to concatenate the reconstructed encoding matrix with the original input feature map *A*. This results in the texture-augmented feature vector *V*:

$$V = Cat(U, A) \tag{9}$$

where the *Cat* operator denotes the concatenation operation, i.e., the positions corresponding to two vectors are directly connected.

### *3.2 Multi-scale feature fusion module*

In any processing task involving images, the coding quality of the image's own representations directly affects the efficiency and accuracy of the task. The feature descriptors of an image are usually divided into two categories: global features and local features. Global features are more informative and contain the most distinguishing features of the things described in the image, which will directly affect the accuracy of the task, but global features are greatly affected by the view angle and illumination of the image, and small changes in the view angle or illumination will cause large fluctuations in the performance of the model, especially in the field of landmark retrieval, where the impact is particularly significant. Local features contain detailed information about the image description, and due to the principle of local geometric invariance, they are less affected by the view angle or other noises, but too many local features will produce local feature redundancy, which leads to a decrease in the operating efficiency of the model. In order to compensate each other for their respective shortcomings, this paper chooses feature fusion as a means of combining global features with local features. Although the final feature map obtained by the general feature fusion method can get a better feature representation, the global features and local features are usually extracted from multiple independent branches of the backbone network, and there is overlapping information between their feature representations, which will lead to redundancy of the representations when they are fused. Based on this, this paper proposes a new feature fusion method, the mapping fusion module, which reconstructs the local features, removes the redundant feature information that already exists in the global features, and finally connects the local features after removing the redundant representations with the global features, which makes the feature representations more robust and improves the efficiency of the subsequent models as well as their robustness.

The structure of the fusion module used in this paper is shown in Figure 3, where *Cat* denotes the concatenation operator and FC denotes the fully connected layer. Let the global feature map be $F_g$, the size be $C \times 1$, and the local feature map be $F_l$, $C \times H \times W$. Firstly, for each location point on the local feature map, compute its mapping vector to the global feature map $F_{pro}$:

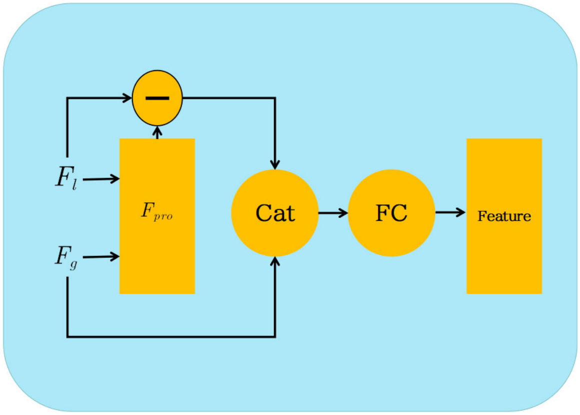
$$F_{pro}^{i,j} = \frac{F_l^{i,j} \cdot F_g}{\left|F_g\right|^2} F_g \tag{10}$$

where $i, j$ denote the coordinates where the computed points are located, $F_l^{i,j} \cdot F_g$ denotes the dot product between the local and global feature maps at the $i, j$ coordinate points, and $|F_g|^2$ denotes the N2 paradigm of the global feature map. The corresponding difference vector $F_{diff}^{i,j}$ is then computed for each coordinate point:

$$F_{diff}^{i,j} = F_l^{i,j} - F_{pro}^{i,j} \tag{11}$$

When the difference vectors corresponding to all points $F_{diff}^{i,j}$ are calculated, the overall differential feature map $F_{diff}$ can be obtained, which can be regarded as the feature representation complement of the local feature map relative to the global feature map mathematically, and can be regarded as an information supplement to the global feature map representation in specific applications, and from this point of view, it is only necessary to merge the differential feature map with the original global feature map. In this paper, the difference vectors $F_{diff}^{i,j}$ corresponding to all points are used to subsequently append the global feature map $F_g$ to get the final feature map $F_{out}$. In order to better couple this feature map with subsequent tasks, in practice, we use a global pooling layer as well as a fully-connected layer at the end to transform the size of the feature map into vectors of the size 512 × 1, which makes the feature characterisation more compact.

**Figure 3** Structure of the feature fusion module (see online version for colours)



### 3.3 Optimisation

Based on the approach proposed in Wang et al. (2022), the final feature descriptor obtained after feature fusion is followed by placing a prediction layer with a weight size of $\omega \in R^{512 \times N}$, where $N$ denotes the number of labelled categories in the data. The final loss function required is as follows:

$$L = -\log\left(\frac{\exp\left(\gamma \times AF\left(\hat{\omega}_k^T \hat{g}, 1\right)\right)}{\sum_n \exp\left(\gamma \times AF\left(\hat{\omega}_n^T \hat{g}, y_n\right)\right)}\right) \tag{12}$$

where $\gamma$ is the impact factor, $\hat{\omega}_n$ denotes the weight after regularisation by $L_2$, $\omega$ the magnitude of the weight for the $n^{\text{th}}$ class, which mathematically denotes the $k^{\text{th}}$ row of $\omega$ after regularisation. $y_n$ denotes the output unique heat coding vector, and k denotes the index of the category where the ground-truth is located. The AF function denotes the ArcFace function, which has the form shown below:

$$AF(x, y) = \begin{cases} \cos\left(a\cos(x) + m\right) & \text{if } y = 1 \\ x & \text{if } y = 0 \end{cases} \tag{13}$$

where $x$ denotes the cosine similarity, $m$ denotes the ArcFace edge value, and $y$ indicates whether the class is ground-truth or not.

## 4 Experiments

### *4.1 Dataset and parameter details*

The training dataset used for the experiments in this paper is Google Landmark V2 dataset (Giorgos et al., 2020), and the network is trained and then tested using Oxford and Paris datasets. Google Landmark V2 dataset is a dataset released by Google in 2020 for landmark detection and retrieval, which is suitable for large-scale, fine-grained landmark retrieval, and contains 200K labels of different landmark instances, totalling 5M images. The dataset is suitable for large-scale, fine-grained landmark retrieval and contains labels for 200K different landmark instances, totalling 5M images. All the images in the dataset have real-world lighting conditions and viewpoints, and contain a large number of unknown attractions in addition to well-known landmarks. The source of the data is the Wikimedia Common media repository and some of the images uploaded by users of the travel app. The Oxford dataset and Paris dataset are another kind of dataset for landmark retrieval model, with data sizes of 4,993 and 6,322 images, respectively, which are smaller in size, faster in detection and more in line with the real situation, and are therefore commonly used for landmark retrieval model testing. In the design of experimental test indexes, according to the idea of Cao et al. (2020) and Zhu et al. (2021), the mean average precision (mAP), the average precision of the top ten results (mp@10) and the average response time (time) are used as the reference indexes, which are compared with the mainstream landmark retrieval models to get the final results.

In terms of parameter and model settings, the model proposed in this paper uses Res50 as the backbone network and the initial weight values of this backbone network use ImageNet pre-training weights. All the models involved in the experiment use Google LandMark V2 as the training dataset, where randomly 80% of this dataset is used for training and the remaining 20% is used for validation. The experimental platform was 8 V100 GPUs for training, with the parameters set as follows: weight decay was fixed at 0.0001, learning rate was 0.05, edge value m in the ArcFace function was set to 0.15, the impact factor of the loss function was set to 30, the batch size of training was 128, and the number of training iterations was 100.

### *4.2 Experiment results*

The experimental results of mAP and mp@10 are shown in Table 1.

**Table 1** Comparison of mAP and mp@10 results

| | *DataSets* | | | |
|---|---|---|---|---|
| *Method* | *ROxford* | | *RParis* | |
| | *mAP* | *mp@10* | *mAP* | *mp@10* |
| AlexNet+GEM (Philbin et al., 2007) | 43.3 | 62.1 | 58.1 | 91.6 |
| VGG16-GEM (Philbin et al., 2007) | 61.9 | 82.7 | 69.3 | 97.9 |
| ResNet101-GEM (Philbin et al., 2007) | 67.3 | 84.7 | 77.2 | 98.1 |
| ResNet101-GEM+SOLAR (Filip et al., 2017) | 69.9 | 86.7 | 81.6 | 97.1 |
| ResNet101-AdvBCT (Pan et al., 2023) | 76.9 | 93.8 | 83.7 | 98.3 |
| ResNet101-DToP (Song et al., 2023) | 77.5 | 94.2 | 84.9 | 99.1 |
| ResNet101-CiDeR (Song et al., 2024) | 78.2 | 94.7 | 86.2 | 99.4 |
| DELF-ASMK+SP (Xiao et al., 2020) | 67.8 | 87.9 | 76.9 | 99.3 |
| DELF-R-ASMK+SP (Xiao et al., 2020) | 76.0 | 93.4 | 80.2 | 99.1 |
| ResNet50-DF-SBIR (Chaudhuri et al., 2023) | 69.3 | 88.7 | 79.9 | 99.2 |
| ResNet50-DELG (Weyand et al., 2020) | 69.7 | 89.1 | 81.6 | *99.5* |
| ResNet50-CBIR-SNN (Hu et al., 2022) | 75.5 | 94.1 | 84.4 | 99.3 |
| ResNet50-TFFLR | *79.6* | *95.5* | *87.1* | *99.5* |

In Table 1, we compare the retrieval metrics of mainstream retrieval models with the model proposed in this paper on the two datasets ROxford and RParis. In terms of the backbone network, the retrieval models utilising ResNet as the backbone network demonstrated superior performance compared to those using AlexNet and VGG16 as the backbone network. Contrast, among the models utilising ResNet as the backbone, the model following the incorporation of the TFFLR module demonstrated the most optimal performance across all metrics on both ROxford and RParis. The mAP metric on the ROxford dataset reaches 79.6, which represents an improvement of 4.1 compared to the CBIR-SNN model that uses the same backbone network. The mAP metric on the RParis dataset reaches 87.1, which represents an improvement of 2.7 over the CBIR-SNN model. This indicates that the TFFLR model performs better overall compared to the CBIR-SNN model with the use of the same backbone network, and that the retrieval accuracy has been improved to some extent. With regard to the model comprising ResNet101 as the backbone network, the CiDeR model demonstrated the most optimal experimental performance, achieving 78.2 on ROxford-mAP and 86.2 on RParis-mAP. In contrast, TFFLR demonstrates superior performance when ResNet50 is employed as the backbone network. It outperforms the CiDeR model by 1.4 and 0.9 points, respectively, in the two aforementioned metrics. Furthermore, on the ROxford-mp@10 and RParis-mp@10, the improvement was 0.8 and 0.1, respectively.

The experimental results demonstrate that the TFFLR model performs exceptionally well on both test datasets, whether compared to the baseline model with ResNet101 as the backbone or models using ResNet50 as the backbone network. The improvement is particularly significant on the R0xford dataset. A possible explanation is that while deeper backbone networks can generate more compact deep feature vectors, they lose a considerable amount of shallow features. For landmark images, shallow features, especially texture features, contain semantic information about the spatial structure and local details of landmarks, and their impact on retrieval performance cannot be ignored.

In contrast, the TFFLR model incorporates a texture feature enhancement module that amplifies the shallow texture features in multi-scale feature maps. This ensures that the deep feature maps processed by the backbone network still retain rich shallow texture semantic information. Consequently, the extracted feature vectors maintain good expressiveness even under complex viewing angles and lighting conditions. Additionally, the model's feature fusion module eliminates redundant features from the original feature maps, reducing the influence of irrelevant features on the model's retrieval process, thereby enhancing the performance of the landmark retrieval model. These findings indicate that the TFFLR model can significantly improve retrieval accuracy while reducing the depth of the backbone network. This achievement results in a more streamlined overall model with fewer parameters, potentially increasing the training speed of the entire retrieval model.

In the field of landmark retrieval, in addition to the mAP and mp@10 indicators, the response time is another indicator that cannot be ignored, the response time is negatively correlated with the retrieval speed, the longer the response time, the slower the retrieval speed, and the speed of the retrieval speed directly indicates that the retrieval model's operational efficiency as well as immediacy, which is a major criterion for judging whether the retrieval model is practical or not. The experimental results of response time are shown in Table 2.

**Table 2** Comparison of average response time

| *Method* | *Time (ms)* |
|---|---|
| AlexNet+GEM | 120 |
| VGG16-GEM | 230 |
| ResNet101-GEM | 134 |
| ResNet101-GEM+SOLAR | 150 |
| ResNet101-AdvBCT | 192 |
| ResNet101-DToP | 247 |
| ResNet101-CiDeR | 205 |
| DELF-ASMK+SP | 510 |
| DELF-R-ASMK+SP | 2,260 |
| ResNet50-DF-SBIR | 207 |
| ResNet50-DELG | *118* |
| ResNet50-CBIR-SNN | 187 |
| R50-TFFLR | 213 |

As illustrated in Table 2, the TFFLR model exhibits inferior performance in terms of average response time relative to other mainstream retrieval models, in which for the CiDeR model, the average response time exceeds 8ms, and the retrieval speed is slower than that of the CiDeR model. Using the same backbone network, the TEIR model has the largest gap with the DELG model, with an average response time exceeding 95 ms However, combining the mAP and mp@10 metrics, TFFLR outperforms the DELG model in both the ROxford-mAP and RParis-mAP metrics, with an improvement of 9.9 and 5.5. In the ROxford-mp@10 metrics, TFFLR improves 6.4 compared to the DELG model, indicating that the model sacrifices some of the retrieval speed but achieves a large accuracy improvement. The mean value of the average response time of the models compared in the experiment is 380 ms, while the average response time of the TFFLR model is 213ms. Although the retrieval efficiency decreases when comparing the minimum value of the response time of 118ms, it is much smaller than the mean value of the average response time, which indicates that the method still has a strong immediacy in the practical application scenario. And from the perspective of dataset, TFFLR improves significantly under the ROxford dataset, compared with ResNet101-CiDeR, both metrics are improved by 1.4 and 0.8, respectively, while on the RParis dataset, it is improved by 0.9 and 0.1. The main reason is that the images captured by RParis are of high quality and the number of noisy images is small, so that even the ordinary retrieval models can also obtain better results, and the room for improvement is smaller than ROxford.

In conclusion, the proposed TFFLR retrieval model, although not outperforming some existing retrieval models in terms of response time, maintains an acceptable latency for practical implementations. Notably, it exhibits substantial enhancement in mean average precision when compared to contemporary models. This improvement in accuracy, coupled with its reasonable computational efficiency, renders the TFFLR model highly valuable for real-world retrieval scenarios. The model's performance characteristics suggest that it is well-suited to address the requirements of a wide range of landmark retrieval applications, striking an effective balance between precision and operational feasibility.

## 5 Conclusions

We present a texture-enhanced single-stage landmark retrieval model designed to address the sensitivity of existing models to factors such as illumination, viewpoint, and occlusion by leveraging texture features. The model initially employs a texture enhancement module to reconstruct original feature maps, amplifying the influence of texture features in deep feature descriptors and thereby improving the expressiveness and robustness of multi-scale feature maps under complex conditions. To obtain more comprehensive feature descriptors, the model introduces a novel feature fusion module that integrates global and local feature maps, utilising a mapping fusion approach to optimise local features within feature descriptors at various scales and eliminate redundant features, resulting in more complete and discriminative fused feature descriptors. Experimental results demonstrate that this model significantly outperforms existing methods in retrieval accuracy while maintaining acceptable retrieval speeds. As the modalities used to describe landmarks continue to expand, future research directions will explore the potential for multi-modal landmark retrieval, such as implementing landmark retrieval for video data.

## Acknowledgements

Kun Tong: writing, investigation, supervision and GuoXin Tan: re-writing and data check.

## Data availability statement

All data, models, and code generated or used during the study appear in the submitted article.

## References

Ahmad, F. (2022) 'Deep image retrieval using artificial neural network interpolation and indexing based on similarity measurement', *CAAI Transactions on Intelligence Technology*, Vol. 7, No. 2, pp.200–218.

Ao, Y. and Wu, H. (2023) 'Feature aggregation and refinement network for 2D anatomical landmark detection', *Journal of Digital Imaging*, Vol. 36, No. 2, pp.547–561.

Asadi Amiri, S., Mohammadpoory, Z. and Nasrolahzadeh, M. (2022) 'A novel content-based image retrieval system using fusing color and texture features', *Journal of AI and Data Mining*, Vol. 10, No. 4, pp.559–568.

Babenko, A. and Lempitsky, V. (2015) 'Aggregating deep convolutional features for image retrieval', in *ICCV*, DOI: 10.48550/arXiv.1510.07493.

Baldrati, A., Bertini, M., Uricchio, T. et al. (2022) 'Effective conditioned and composed image retrieval combining clip-based features', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.21466–21474.

Bhatti, U.A., Huang, M., Neira-Molina, H. et al. (2023) 'MFFCG-multi feature fusion for hyperspectral image classification using graph attention network', *Expert Systems with Applications*, Vol. 229, p.120496, DOI: 10.1016/j.eswa.2023.120496.

Cao, B., Araujo, A. and Sim, J. (2020) 'Unifying deep local and global features for image search', in *Eur. Conf. Comput. Vis*., DOI: 10.1007/978-3-030-58565-5_43.

Chaudhuri, A., Bhunia, A.K., Song, Y-Z. and Dutta, A. (2023) 'Data-free sketch-based image retrieval', *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp.12084–12093, DOI: 10.1109/CVPR52729.2023.01163.

Chen, Y., Xia, R., Yang, K. et al. (2024) 'MFFN: image super-resolution via multi-level features fusion network', *The Visual Computer*, Vol. 40, No. 2, pp.489–504.

Chen, Y., Zhang, S., Liu, F. et al. (2022) 'Transhash: transformer-based hamming hashing for efficient image retrieval', *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pp.127–136.

Do, T. and Sinha, S.N. (2024) 'Improved scene landmark detection for camera localization', *2024 International Conference on 3D Vision (3DV)*, IEEE, pp.975–984.

Do, T., Miksik, O., DeGol, J. et al. (2022) 'Learning to detect scene landmarks for camera localization', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.11132–11142.

Filip, R., Giorgos, T. and Ondrej, C. (2017) 'Fine-tuning CNN image retrieval with no human annotation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, pp.1655–1668, DOI: 10.1109/TPAMI.2018.2846566.

Giorgos, T., Tomas, J. and Ondřej, C. (2020) 'Learning and aggregating deep local descriptors for instance-level recognition', in *European Conference on Computer Vision*, Springer, pp.460–477.

Gordo, A., Almazan, J., Revaud, J. et al. (2016) *Deep Image Retrieval: Learning global Representations for Image Search*, Springer, Cham, DOI: 10.1007/978- 3-319-46466-4_15.

Hu, T., Kwiatkowski, M., Matern, S. and Hellwich, O. (2022) 'Content-based landmark retrieval combining global and local features using siamese neural networks', arxiv preprint arxiv:2208.04201.

Jerome, R., Jon, A., Rafael, S.R. and Cesar, R.d.S. (2019) 'Learning with average precision: training image retrieval with a listwise loss', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Vol. 2, No. 6, pp.5107–5116.

Kelishadrokhi, M.K., Ghattaei, M. and Fekri-Ershad, S. (2023) 'Innovative local texture descriptor in joint of human-based color features for content-based image retrieval', *Signal, Image and Video Processing*, Vol. 17, No. 8, pp.4009–4017.

Liu, C., Huang, H., Ma, Z. et al. (2024) 'AIR-HLoc: adaptive image retrieval for efficient visual localisation', arXiv preprint arXiv:2403.18281.

Liu, G.H. and Yang, J.Y. (2023) 'Exploiting deep textures for image retrieval', *International Journal of Machine Learning and Cybernetics*, Vol. 14, No. 2, pp.483–494.

Liu, J., Singh, A.K. and Lin, C.T. (2022) 'Predicting the quality of spatial learning via virtual global landmarks', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 30, pp.2418–2425, DOI: 10.1109/TNSRE.2022.3199713.

Lu, F., Liu, G.H. and Gao, X.Z. (2024) 'Image retrieval using multilayer feature aggregation histogram', *Cognitive Computation*, pp.1–14, DOI: 10.1007/s12559-024-10334-9.

Mohtashami, A. and Jaggi, M. (2023) 'Landmark attention: Random-access infinite context length for transformers', arXiv preprint arXiv:2305.16300.

Noh, H., Araujo, A., Sim, J. et al. (2016) *Large-Scale Image Retrieval with Attentive Deep Local Features*, DOI: 10.48550/arXiv.1612.06321.

Öztürk, Ş., Çelik, E. and Çukur, T. (2023) 'Content-based medical image retrieval with opponent class adaptive margin loss', *Information Sciences*, Vol. 637, p.118938, DOI: 10.1016 /j.ins.2023.118938.

Pan, T., Xu, F., Yang, X., He, S., Jiang, C., Guo, Q. and Chu, W. (2023) 'Boundary-aware backward-compatible representation via adversarial learning in image retrieval', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.15201–15210.

Philbin, J., Chum, O., Isard, M., Sivic, J. and Zisserman, A. (2007) 'Object retrieval with large vocabularies and fast spatial matching', in *IEEE Conf. Comput. Vis. Pattern Recog.*, Vol. 1, No. 2, pp.1–8.

Pipanmekaporn, L., Kamonsantiroj, S., Ratanavilisagul, C. et al. (2023) 'Spatial pyramid attention enhanced visual descriptors for landmark retrieval', *Image*, Vol. 11, pp.359–366, DOI: 10.18178/joig.11.4.359-366.

Sain, A., Bhunia, A.K., Yang, Y. et al. (2021) 'StyleMeUp: towards style-agnostic sketch-based image retrieval', DOI: 10.1109/CVPR46437.2021.00840.

Salih, S.F. and Abdulla, A.A. (2023) 'An effective bi-layer content-based image retrieval technique', *The Journal of Supercomputing*, Vol. 79, No. 2, pp.2308–2331.

Song, C.H., Han, H.J. and Avrithis, Y. (2022a) 'All the attention you need: global-local, spatial-channel attention for image retrieval', *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.2754–2763.

Song, Y., Zhu, R., Yang, M. et al. (2022b) 'Dalg: deep attentive local and global modeling for image retrieval', arXiv preprint arXiv:2207.00287.

Song, C.H., Yoon, J., Choi, S. and Avrithis, Y. (2023) 'Boosting vision transformers for image retrieval', in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.107–117.

Song, C.H., Yoon, J., Hwang, T., Choi, S., Gu, Y.H. and Avrithis, Y. (2024) 'On train-test class overlap and detection for image retrieval', arxiv preprint arxiv:2404.01524.

Suo, Y., Ma, F., Zhu, L. et al. (2024) 'Knowledge-enhanced dual-stream zero-shot composed image retrieval', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.26951–26962.

Teichmann, M., Araujo, A., Zhu, M. and Sim, J. (2019) 'Detect-to-retrieve: efficient regional aggregation for image search', in *CVPR*, pp.5104–5113, DOI: 10.1109/CVPR.2019.00525.

Tolias, G., Sicre, R. and Jégou, H. (2015) 'Particular object retrieval with integral max-pooling of CNN activations', *Computer Science*, DOI: 10.48550/ arXiv.1511.05879.

Varish, N. (2022) 'A modified similarity measurement for image retrieval scheme using fusion of color, texture and shape moments', *Multimedia Tools and Applications*, Vol. 81, No. 15, pp.20373–20405.

Wang, M., Zhou, W., Tian, Q. et al. (2022) 'Deep graph convolutional quantization networks for image retrieval', *IEEE Transactions on Multimedia*, Vol. 25, pp.2164–2175, DOI: 10.1109/TMM.2022.3143694.

Weinzaepfel, P., Lucas, T., Larlus, D. et al. (2022) 'Learning super-features for image retrieval', arXiv e-prints, DOI:10.48550/arXiv.2201.13182.

Weng, L., Pang, K., Xia, M. et al. (2023) Sgformer: a local and global features coupling network for semantic segmentation of land cover', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 16, pp.6812–6824, DOI: 10.1109/JSTARS.2023 .3295729.

Weyand, T., Araujo, A., Cao, B. et al. (2020) 'Google landmarks dataset v2-A large-scale benchmark for instance-level recognition and retrieval', IEEE, DOI: 10.1109/CVPR42600. 2020.00265.

Wu, H., Wang, M., Zhou, W. et al. (2023) 'Asymmetric feature fusion for image retrieval', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.11082–11092.

Xiao, Y., Wang, C. and Gao, X. (2020) 'Evade deep image retrieval by stashing private images in the hash space', *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, DOI: 10.1109/CVPR42600.2020.00967.

Xie, T., Wang, L., Wang, K. et al. (2023) 'FARP-Net: local-global feature aggregation and relation-aware proposals for 3D object detection', *IEEE Transactions on Multimedia*, Vol. 26, pp.1027–1040, doi: 10.1109/TMM.2023.3275366.

Xu, Y., Bin, Y., Wei, J. et al. (2023) 'Multi-modal transformer with global-local alignment for composed query image retrieval', *IEEE Transactions on Multimedia*, Vol. 25, pp.8346–8357, DOI: 10.1109/TMM.2023.3235495.

Zhang, X., Bai, C. and Kpalma, K. (2023) 'OMCBIR: offline mobile content-based image retrieval with lightweight CNN optimization', *Displays*, Vol. 76, p.102355, DOI: 10.1016 /j.displa.2022.102355.

Zhu, L., Ji, D., Zhu, S. et al. (2021) 'Learning statistical texture for semantic segmentation', DOI: 10.48550/arXiv.2103.04133.