IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 5, 2019, accepted January 21, 2020, date of publication January 24, 2020, date of current version February 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969287

A Method of Hierarchical Image Retrieval for Real-Time Photogrammetry Based on Multiple Features

ZONGQIAN ZHAN, GAOFENG ZHOU, AND XUE YANG
School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China

Corresponding author: Gaofeng Zhou (gfzhou9608@163.com)

This work was supported in part by the National Key Research and Development Project of China under Grant 2016YFB0501403, and in part by the National Natural Science Foundation of China under Grant 61871295.

ABSTRACT In near real-time photogrammetry, the first step in processing each new added image is determining the most relevant image in pre-sequence unordered images quickly and exactly, which is pivotal for accurate image matching and 3D reconstruction. This paper presents a hierarchical image retrieval algorithm based on multiple features and details the choice for representation of multiple features which is critical to the improvement of accuracy of this algorithm. First, we represent global features using AlexNet-FC7(fully connected layers) or ResNet101-Pool5(pooling layers) and local features using SIFT (scale-invariant feature transform) in two parallel threads with support of GPU (Graphics Processing Unit). Next, we obtain candidates based on cosine similarities between global features of each pre-sequence image and new added image. Finally, we determine the most relevant image from those candidates according to feature matching results for each candidate and new added image. The experimental results confirm that the second step is rather fast and the third step is necessary to tackle the problem that global features cannot distinguish objects from the same class. The total time our algorithm takes is about 83.6ms for determining the most relevant image in 5063 pre-sequence unordered images of size 1024×768, which outperforms exhaustive pairwise matching, Bag of Words and multi-vocabulary trees. Accuracy of our algorithm also perform better than the state-of-the-art methods on three benchmark datasets. SIFT matching results obtained in the third step after eliminating mismatches with RANSAC (Random Sample Consensus) can also be used for high-precision incremental SFM (Structure from Motion) reconstruction.

INDEX TERMS Hierarchical image retrieval, multiple features, photogrammetry, real-time.

I. INTRODUCTION

For large scale real-time photogrammetry, we need to process each image transmitted to PC synchronously in real time. The first step is to quickly and accurately determine the most relevant images from large amounts of pre-sequence images for each image transmitted to PC after rectification without geographic location information. The quality of this has a direct impact on the subsequent matching and stereo model reconstruction. As photogrammetry and computer vision technology gain steam, content-based image retrieval technology can be adopted in this step. CBIR (content-based image retrieval) is the process of searching for the images with the same object or the same category of object in a database according to the object contained in the query image, which has been a longstanding research topic in the computer vision field. Since the 1990s, there are two kinds of retrieval methods—those that are based on local features, such as SIFT (scale-invariant feature transform), and based on global features, such as Convolutional Neural Networks (CNN)—in order to achieve accurate and quick correlated image retrieval [1] at a large scale.

Local features such as SIFT [2] are robust to rotation, scaling, view changes, affine transformation and noise. A common retrieval idea based on this is that SIFT features are extracted from all images first, and then they are matched according to the feature descriptors. This means that two images are set to be related when the number of matched

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca.

points is larger than a certain threshold value [3]. When the number of images is large, methods based on exhaustive pairwise matching are too inefficient to meet the requirement of real-time photogrammetry. Bow (Bag-of-Words) [3] can improve the retrieval efficiency compared with methods based on exhaustive pairwise matching. After extracting SIFT features for all images offline, the visual vocabulary is generated by clustering algorithm like K-means on all SIFT descriptors where each visual word represents one cluster center and is allocated a weight using TF-IDF (term frequency–inverse document frequency) algorithm. Next, each image can be expressed by some visual words to calculate the word frequency vector. Finally, according to a similar distance, such as the Euclidean distance, the most relevant image is determined. Nevertheless, these visual words are independent of each other, meaning that the accuracy of retrieval precision will be affected seriously due to the deletion of spatial position information between visual words. K-d tree and Random Forest [4]–[10] first traverse one or more k-d trees that are constructed by SIFT features to get the specified number of SIFT features adjacent to each SIFT feature of the query image. Each nearby feature belongs to one image in the dataset; thus, we can get corresponding images for these neighboring features as candidate images and count the number of neighboring features between the query image and each candidate image to determine the most relevant image. The bottleneck of this approach is that each traversal needs to backtrack, which means that, when dealing with high-dimensional data, the retrieval efficiency is greatly reduced [4].

In 2012, CNN was successfully applied to classification on ImageNet [11]. Convolution kernels of different sizes are employed on the entire input image to obtain global features that are different from local features such as SIFT. CNN models are mainly categorized into pre-trained CNN models—i.e., a network which weight parameters are obtained by training the network on the benchmark datasets in advance—and fine-tuned CNN models—i.e., a network in which weight parameters are fine-tuned on specific training sets. Several CNN models serve as good choices for extracting features, including AlexNet, VGGNet, GoogleNet and ResNet. Babenko Artem^{12} was the first to fine-tune a CNN model for generic image retrieval and compare the retrieval result of a fine-tuned CNN model with that of a pre-trained CNN model. The performance of adapted features of different layers was evaluated on benchmark datasets. The paper also investigated the performance of compressed neural codes, where plain PCA or a combination of PCA with discriminative dimensionality reduction result in very short codes with very good performance compared with VLADs, Fisher Vectors, or triangulation embedding [12]. Lin *et al.* [13] added a latent layer between F7 and F8 of the AlexNet and had neurons in this layer learn hash-like representations while fine-tuning it on the target domain dataset when retrieving similar images using a coarse-to-fine strategy that utilizes the learned hash-like binary codes and F7 features. Ng *et al.* [14] and Tolias *et al.* [15] also used the pre-training models to extract global features. Different convolution kernels have different receptive fields, and the activation map after convolution can be considered as the aggregation of multiple column features of size $1 \times 1 \times n$ where each column feature can be viewed as a description of a certain patch in the original image. Hariharan *et al.* [16] assembles column features of different convolutional layers to form the hyper-column feature. We find that the single-pass CNN methods tend to combine the individual steps in the SIFT-based model, and a forward propagation takes less time than SIFT feature extraction and description. Experiments also show that SIFT has difficulty extracting a certain number of feature matching points for images of the same target with different illumination, which does not exist in CNN. However, CNN also has the same problem as Hash. The distance between images is related to the image content itself, and it is difficult to determine an absolute threshold. Moreover, CNN-based retrieval methods all fine-tune the parameters of the existing model on the target data sets to improve the retrieval accuracy. When the amount of the target data set is limited, the reliability of the fine-tuning model needs to be further determined. Sarlin *et al.* [17] etc. exploited the coarse-to-fine localization paradigm: first, they performed a global retrieval to obtain location hypotheses and only later matched local features within those candidate places. A hierarchical localization approach based on a monolithic CNN was proposed, which simultaneously predicted local features and global descriptors for accurate 6-DoF localization. The coarse-to-fine localization paradigm improves the time efficiency and also ensures the orientation accuracy, which is also applicable to the determination of image overlaps. Two similar works were presented by Yan *et al.* [18] and Huang and Hang [19], who also fused CNN and SIFT from multiple levels. Yan et al. first obtained scene-level and object-level representations using a deeper network such as GoogLeNet and then chose the SIFT to represent point-level information in order to preserve geometric invariance in image representation. Finally, they directly concatenated these three-level features to generate the integrated representation for image retrieval. However, to obtain the object-level representation, they needed to extract deep features for hundreds of proposal regions produced by the object proposal approach and pool features to fixed-length feature vectors, which sacrifices time for accuracy and cannot meet the requirements of real-time photogrammetry. Huang et al. first filtered the images using global features extracted by AlexNet-FC6 and FC7 and then used the local features to re-rank them based on Bow. A codebook generated by the Oxford Building Dataset using the TF-maxIDF metric was needed during re-ranking, which brought up a problem that we need to determine the number of visual words in the codebook to achieve a compromise between accuracy and efficiency. Nowadays, CNN-based image retrieval has been successfully used in robust visual localization [20], [21]. We try to apply CNN-based image retrieval to close-range

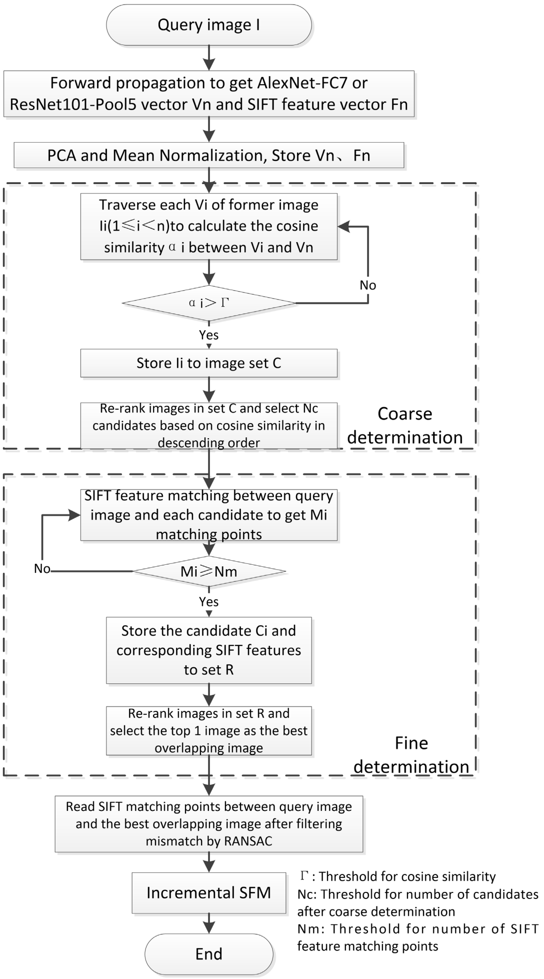

Query image I
Forward propagation to get AlexNet-FC7 or ResNet101-Pool5 vector Vn and SIFT feature vector Fn
PCA and Mean Normalization, Store Vn、Fn
Traverse each Vi of former image Ii(1≤i<n)to calculate the cosine similarity αi between Vi and Vn
αi>Γ
No
Yes
Store Ii to image set C
Re-rank images in set C and select Nc candidates based on cosine similarity in descending order
Coarse determination
SIFT feature matching between query image and each candidate to get Mi matching points
Mi≥Nm
No
Yes
Store the candidate Ci and corresponding SIFT features to set R
Re-rank images in set R and select the top 1 image as the best overlapping image
Fine determination
Read SIFT matching points between query image and the best overlapping image after filtering mismatch by RANSAC
Incremental SFM
End
Γ: Threshold for cosine similarity
Nc: Threshold for number of candidates after coarse determination
Nm: Threshold for number of SIFT feature matching points

FIGURE 1. Flowchart of the most relevant image hierarchy determination for query image at large scale in real time. For each query image, Nc candidates are selected firstly based on the cosine similarities between the AlexNet-FC7 feature vectors or ResNet101-Pool5 in descending order. Then, SIFT feature matching re-ranks the order of these Nc candidates and choose the top 1 retrieval image as the most relevant image for the following ISFM.

photogrammetry. Our experimental results have shown that exhaustive pairwise matching with an acceleration of the GPU performs better than Bow during re-ranking.

Different from feature fusion methods in Feng *et al.* [22], Yan *et al.* [18], our method first utilizes the global features obtained by AlexNet-FC7 or ResNet101-Pool5 to determine candidate overlapping images. Our experiments show that the retrieval speed based on CNN features is rather fast and can meet the precision requirements of roughly determining the image that may overlap with the query image. We then adopt exhaustive pairwise SIFT matching within those candidate images. Only a small amount of matching time is spent to accurately determine the final most relevant image. Meanwhile, SIFT features extracted in this step can be used for subsequent 3D reconstruction. Our method also performs excellently for UAV (Unmanned Aerial Vehicle) images. The flowchart is shown in Fig. 1.

Overall, our contributions are as follows:

1) Firstly, we verify that the extracted AlexNet-FC7 and ResNet101-Pool5 feature vectors are more robust to illumination change than SIFT features and very efficient, while SIFT features can solve the problem that AlexNet-FC7 and ResNet101-Pool5 feature vectors cannot distinguish objects from the same class, which explains why we fuse these two features hierarchically in our method;
2) Secondly, the distribution visualization of different AlexNet-FC7 and ResNet101-Pool5 feature vectors on three benchmark datasets proves that global features extracted by CNN can be used for coarse determination and experiments show that global feature represented by ResNet101-Pool5 achieves the better query results compared with global feature represented by AlexNet-FC7 for most datasets after PCA and Mean Normalization;
3) Finally, we test our method on the UAV images and also achieve satisfactory results.

The paper is organized as follows. Section II presents the basic idea of our method and reasons for choosing AlexNet or ResNet101 for global feature extraction and SIFT for local feature extraction. Section III first visualize the distribution of AlexNet-FC7 feature vectors and ResNet101-Pool5 feature vectors after Mean Normalization and PCA and then provides an evaluation, comparison and discussion of our method (AlexNet-FC7+SIFT and ResNet101-Pool5+SIFT), the CNN-based method (AlexNet-FC7 and ResNet101-Pool5) and other traditional methods, finally validates the effectiveness of our method for close-range photogrammetry application.

II. METHOD TO HIERARCHICALLY DETERMINATE THE MOST RELEVANT IMAGE FOR A QUERY IMAGE

A. COARSE DETERMINATION OF A CANDIDATE OVERLAPPING IMAGE BASED ON AlexNet-FC7 AND ResNet101-Pool5 FEATURE VECTORS

In recent years, CNN has been widely used in complex tasks such as scene categorization and object detection with excellent performance, which indicates that the CNN network has the ability to extract robust visual features from images. In 2012, Krizhevsky et al. achieved state-of-the-art recognition accuracy with the AlexNet in ILSRVC 2012, exceeding previous best results by a large margin. Since then, the focus of research has begun to transfer to deep learning-based methods, especially the convolutional neural network (CNN). We can use the outputs of deep hidden layers as a representation of each three-channel image after the forward propagation. For example, ResNet101-Pool5 outputs is a feature vector of size 2048 × 1. The distance between two feature vectors indicates the degree of correlation between the corresponding images. This means that two images are set to be related when the value of distance is larger than a certain threshold value [9]. In this paper, considering that the UAV images contain abundant ground objects, we use

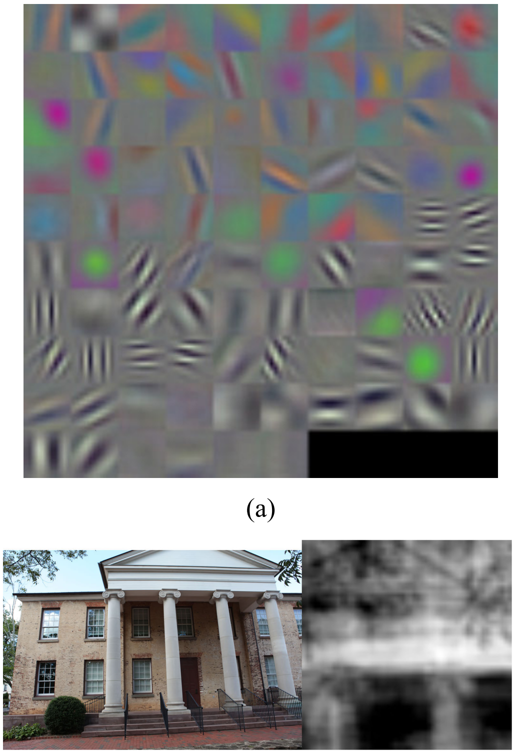


FIGURE 2. Visualization of kernels and feature map: (a) 96 kernels for the first layer of AlexNet; (b) Input image and corresponding response map.

AlexNet and ResNet101 to extract a global feature for each image. Fig. 2 gives the visualization of 96 kernels for the first layer of AlexNet and the feature map for an input image after the convolution operation using the 40th kernel.

To demonstrate that traditional exhaustive pairwise SIFT matching may fail when illumination changes significantly, we choose two image pairs with the same object, all collected during the day and night, respectively. Feature matching results are illustrated in Fig. 3; only seven corresponding points and three corresponding points were matched, which means that not each image pair is overlapping; worse, there were some mismatches. As a result, the accuracy of the 3D reconstruction will be affected and even fail if the most relevant image cannot be found. In contrast, global features extracted by CNN can still ensure reliable accuracy and cosine similarities for these two image pairs, reaching 0.957795, 0.894257 for AlexNet-FC7 and 0.96489, 0.876536 for ResNet101-Pool5. We find that the CNN-based method is more robust to the illumination change and ensures the reliability of the coarse candidate images. In practice, we collect images when the change of illumination is not significant and enough SIFT matching points can be obtained during SIFT matching consequently. However, the CNN-based method cannot distinguish objects from the same class accurately, whereas SIFT can tackle this problem well, which will be illustrated in section III. Hence, we need to re-rank the candidate images based on SIFT matching results.

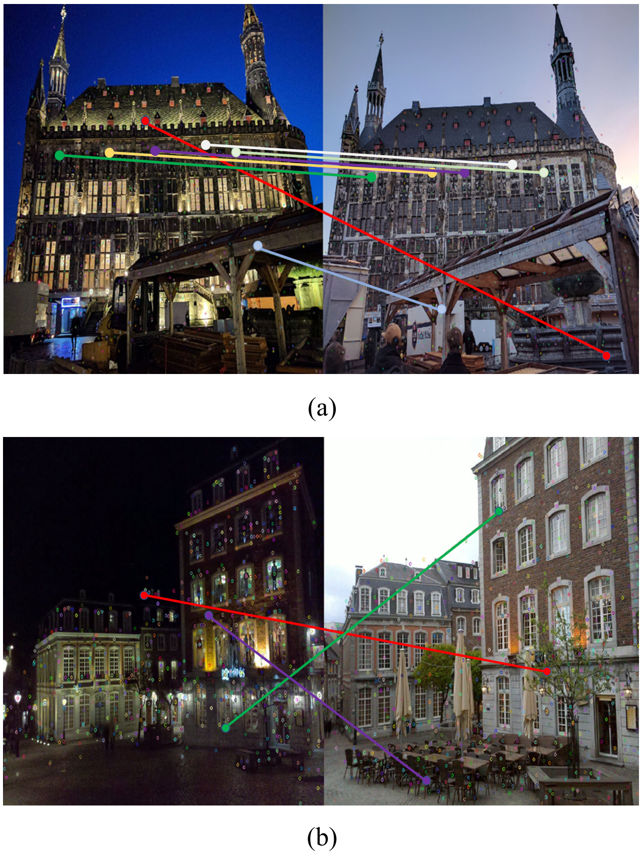


FIGURE 3. SIFT matching result for two image pairs under different illumination. (a) Image pair with seven corresponding points; (b) Image pair with three corresponding points.

B. DETERMINATION OF THE MOST RELEVANT IMAGE BASED ON SIFT FEATURE

After determining a certain number of candidate related images, the most relevant image is found accurately from those images based on SIFT feature matching in short time. The steps are as follows:

1) Read the SIFT features of the query image and its candidate related images from memory;
2) Count the number of matched SIFT feature points after filtering false matching points using RANSAC.

According to the number of matched SIFT feature points, the most relevant image is determined, and those feature points are stored for subsequent 3D reconstruction.

III. EXPERIMENTAL RESULTS AND ANALYSIS

Some instance-level datasets that have been commonly used in the field of image retrieval were adopted in the experiment, including the Holidays Dataset (1491 images composed of 500 groups of similar images) [23], the Oxford Buildings Dataset (5063 images collected by crawling images from Flickr using the names of 11 different landmarks in Oxford) [24] and the Oxford Paris (featured by 6,412 images crawled from 11 queries on specific Paris architecture) [25].

We also use a set of unmanned aerial vehicle (UAV) images (some images shown in fig.4), while most previous research aims at remote sensing images [26], [27]. Each forward

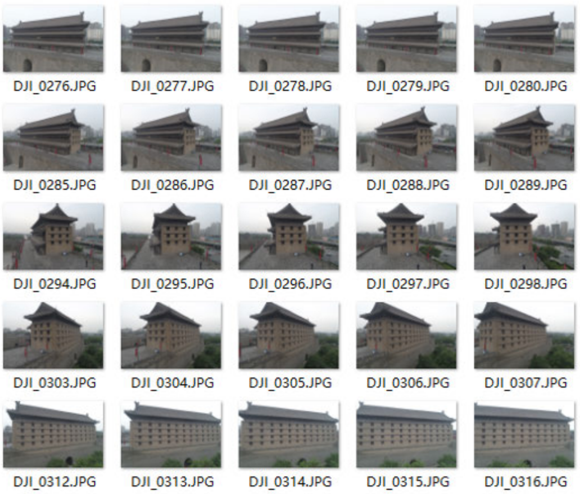


FIGURE 4. Unmanned aerial vehicle (UAV) images collected by DJI Phantom 4 PRO.

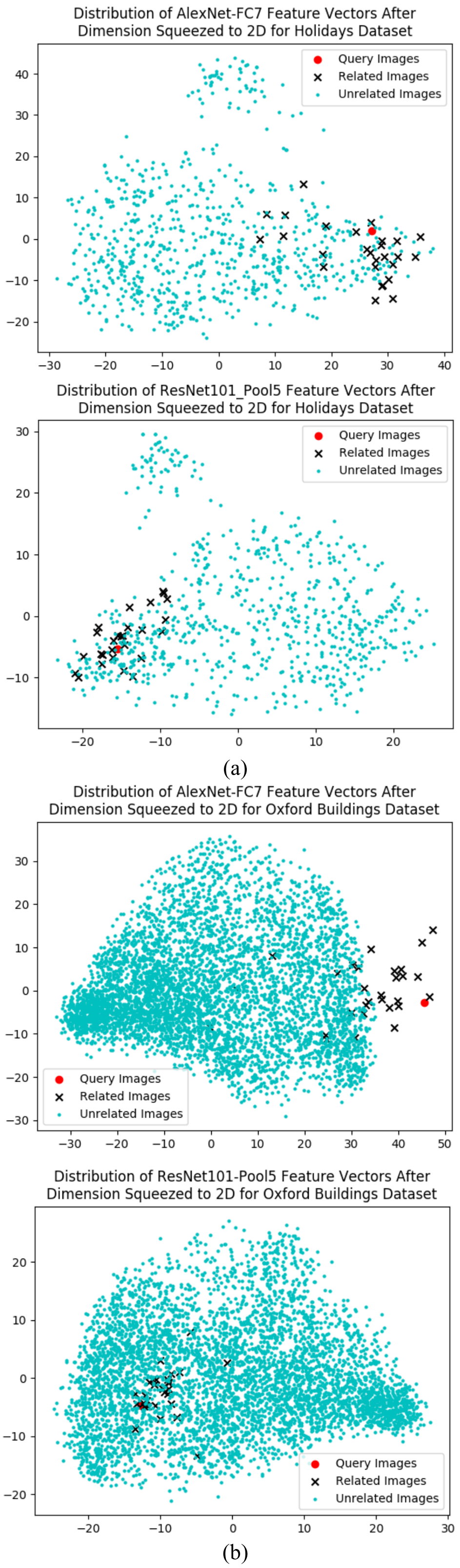
propagation of AlexNet is run on Windows with the support of Caffe. SiftGPU (provided by Changchang Wu from the University of North Carolina at Chapel Hill) is used to extract SIFT features for all candidate related images. For evaluation metrics, we chose the mean average precision (mAP). Typically, a larger mAP means a better retrieval performance.

A. DISTRIBUTION OF FEATURE VECTORS FOR AlexNet-FC7 AND ResNet101-Pool5

By visualizing the distribution of feature vectors for AlexNet-FC7 and ResNet101-Pool5 on three benchmark datasets, we can validate whether those feature vectors of related candidates are close enough to feature vector of query image so as to demonstrate the reliability of candidate related images returned by global features and choose the optimal CNN model feature outputs as global features. Take ResNet101-Pool5 feature vectors as example, the steps are as follows:

The benchmark datasets without labels need to be preprocessed first. We set labels of related candidates, which the overlap degree between them and query image is over 80%, as 1 and 0 otherwise by visual interpretation.Mean Normalization is adopted for all the ResNet101-Pool5 feature vectors after feature extraction. Then, we reduce dimension of those feature vectors from 2048D to 2D for distribution visualization using PCA.

On each benchmark dataset, we run the above steps for AlexNet-FC7 and ResNet101-Pool5 feature vectors distribution visualization, as depicted in fig.5. It can be seen from distribution visualization that most related candidates (represented by black ‘x’) basically concentrate around query image (represented by red point) and also some unrelated images (represented by cyan points). From the query results, the overlap degrees between those unrelated images and query image are not large enough to be used for subsequent



(a)

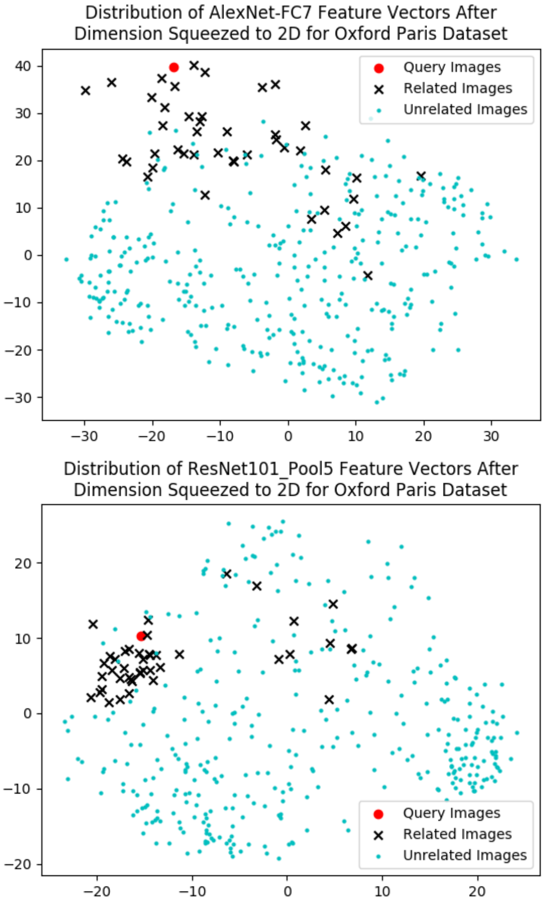
(b)

FIGURE 5. (a)~(c): Distribution of feature vectors for AlexNet-FC7 and ResNet101-Pool5 on Holidays Dataset, Oxford Buildings Dataset, and Oxford Paris Dataset. Red point represents the 2D feature vector of query image, cyan points represent the 2D feature vectors of unrelated images, black ‘x’ represent the 2D feature vectors of related images.

TABLE 1. Comparisons of time in seconds for a single image query with seven methods.

	Image Number	CNN (AlexNet-FC7) (s)	CNN (ResNe101t-Pool5) (s)	Exhaustive pairwise matching (SIFTGPU)(s)	Bag of Words (s)	Multi-Vocabulary Trees (s)	Our Method (AlexNet-FC7+SIFT) (s)	Our Method (ResNet101-Pool5+SIFT) (s)
Holidays Dataset	812	0.00212	0.003279	0.57	0.67	0.138	0.02142	0.024287
Oxford Buildings Dataset	5063	0.038762	0.039508	3.564	4.251	1.28	0.076759	0.080205
Oxford Paris Dataset	444	0.002776	0.001575	0.337	0.442	0.083	0.022498	0.015875

Time for all the methods in the table only include single image query time and does not include feature extraction time. For CNN-based method, we extract AlexNet-FC7 and ResNet101-Pool5 features; for exhaustive pairwise matching、Bag of Words and Multi-Vocabulary Trees, we extract SIFT features; for our method, we extract AlexNet-FC7, ResNet101-Pool5 features and SIFT features in two parallel threads. CNN features extraction (supported by GPU) takes about 0.007s for an image of size 2500×2000, while SIFT features extraction (supported by SIFTGPU) takes about 0.15s. So, the total time (feature extraction time and single image query time) of our method for each query image is less than exhaustive pairwise matching、Bag of Words and Multi-Vocabulary Trees.



(c)

FIGURE 5. ***(Continued.)*** **(a)~(c): Distribution of feature vectors for AlexNet-FC7 and ResNet101-Pool5 on Holidays Dataset, Oxford Buildings Dataset, and Oxford Paris Dataset. Red point represents the 2D feature vector of query image, cyan points represent the 2D feature vectors of unrelated images, black 'x' represent the 2D feature vectors of related images.**

SIFT feature matching and accurate 3D reconstruction or not relevant at all, so we set labels of those images as 0. That is why we will filter those unrelated images from candidates based on fine determination to find the most relevant image.

B. EXPERIMENTAL RESULTS BASED ON AlexNet-FC7 FEATURE VECTORS AND ResNet101-Pool5

The number of experimental images for Holidays Dataset, Oxford Buildings Dataset, and Oxford Paris Dataset is 812, 5063 and 444, respectively and each image size for these datasets is 2500 × 2000, 1024 × 768 and 1024 × 768, respectively. The retrieval experiment of single query image is carried out on these benchmark datasets based on different popular image retrieval methods, such as based on CNN (AlexNet-FC7 and ResNet101-Pool5), exhaustive pairwise matching (SIFTGPU), Bag of Words (we choose a codebook with 1 million SIFT visual words and Distributed Bag of Words library), multi-vocabulary trees (we use the method adopted in Wang *et al.* [9]) and our method (AlexNet-FC7+SIFT, ResNet101-Pool5+SIFT). Table 1 shows comparisons of the query time for a single image while table 2 lists the mean average precision for each method. For those methods, global features and local features are extracted with support of GTX 1080Ti. We use the default parameter settings of SiftGPU library and extract 800 feature points for each image of three benchmark datasets. Fig. 6 (a) and Fig. 6 (c) show the single image query results only based on cosine similarities between AlexNet-FC7 feature vectors and ResNet101-Pool5 feature vectors respectively for Holidays Dataset. Fig. 7(a) and Fig. 7(c) show the single image query results only based on cosine similarities between AlexNet-FC7 feature vectors and ResNet101-Pool5 feature vectors respectively for Oxford Buildings Dataset. Fig. 8(a) and Fig. 8(c) show the single image query results only based on cosine similarities between AlexNet-FC7 feature vectors and ResNet101-Pool5 feature vectors respectively for Oxford Paris Dataset. We sort cosine similarities in descending order and return the first ten images corresponding to these feature vectors as query results. We find that the CNN-based (AlexNet-FC7 and ResNet101-Pool5) methods take the shortest query time. Compared with the method based on exhaustive pairwise matching (SIFTGPU), Bag of Words, and multi-vocabulary trees, the CNN-based

TABLE 2. Comparisons of the mean average precision for a single image query by different methods.

	Image Number	CNN (AlexNet-FC7)	CNN (ResNet101-Pool5)	Exhaustive pairwise matching (SIFTGPU)	Bag of Words	Multi-Vocabulary Trees	Our Method (AlexNet-FC7+SIFT)	Our Method (ResNet101-Pool5+SIFT)
Holidays Dataset	812	0.85	0.957	0.943	0.786	0.761	0.968	0.947
Oxford Buildings Dataset	5063	0.7116	0.831	0.948	0.827	0.840	0.9415	0.988
Oxford Paris Dataset	444	0.982	0.9704	1.0	0.875	0.900	1.0	1

Table II shows mean average precision for a custom query image for three benchmark datasets. For CNN-based method, we extract AlexNet-FC7 and ResNet101-Pool5 features; for exhaustive pairwise matching、Bag of Words and Multi-Vocabulary Trees, we extract SIFT features; for our method, we extract AlexNet-FC7, ResNet101-Pool5 features and SIFT features in two parallel threads. For each benchmark dataset, we selected a custom image from the same dataset as the query image.

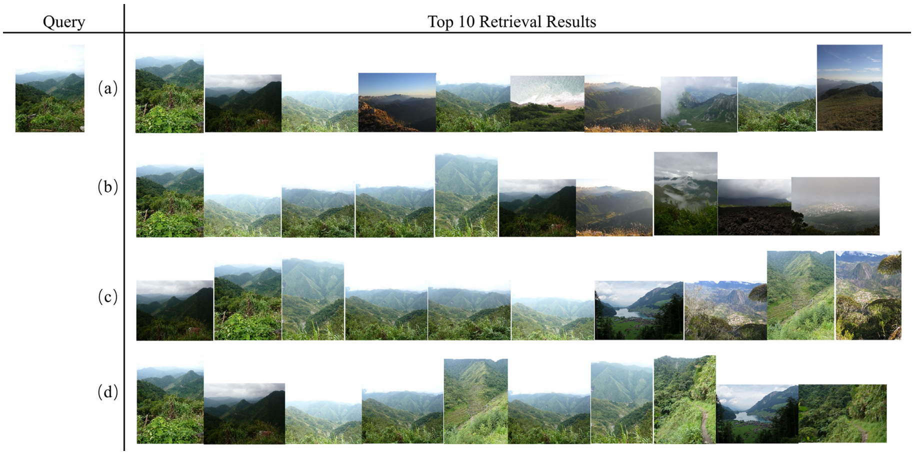


FIGURE 6. Query results based on (a) AlexNet-FC7 (b) AlexNet-FC7+SIFT (c) ResNet101-Pool5 (d) ResNet101-Pool5+SIFT for Holidays Dataset. The first image in line one is the query image. After the CNN-based image retrieval method and our method, we return the first ten images in descending order of cosine similarity.

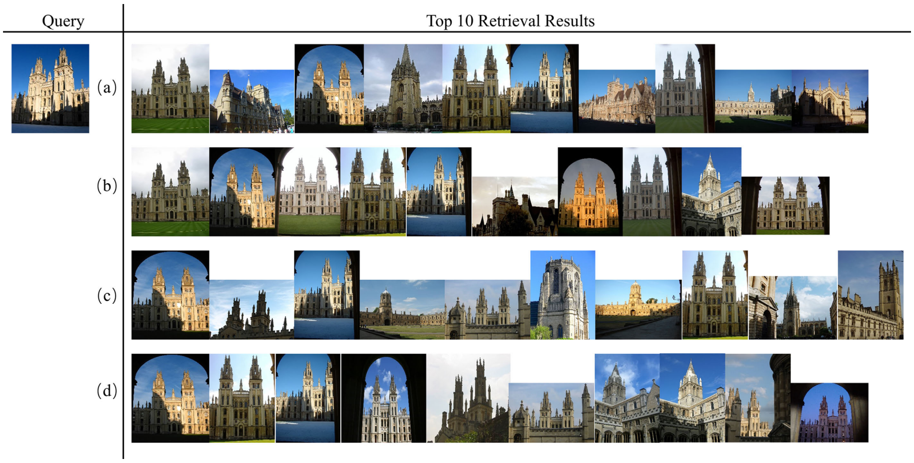


FIGURE 7. Query results based on (a) AlexNet-FC7 (b) AlexNet-FC7+SIFT (c) ResNet101-Pool5 (d) ResNet101-Pool5+SIFT for Oxford Buildings Dataset. The first image in line one is the query image. After the CNN-based image retrieval method and our method, we return the first ten images in descending order of cosine similarity.

method is faster by a factor of 35∼230 (see Table 1) depending on the size of the dataset even when the cosine similarities are calculated in pairs. Meanwhile, for Oxford Buildings Dataset, the mAP of CNN-based (AlexNet-FC7 and ResNet101-Pool5) methods is worse than method based on exhaustive pairwise matching (SIFTGPU), while that

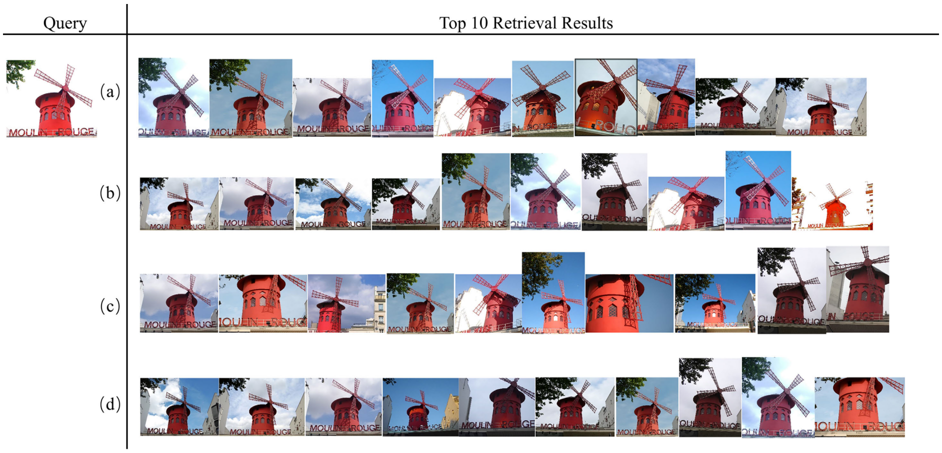


FIGURE 8. Query results based on (a) AlexNet-FC7 (b) AlexNet-FC7+SIFT (c) ResNet101-Pool5 (d) ResNet101-Pool5+SIFT for Oxford Paris Dataset. The first image in line one is the query image. After the CNN-based image retrieval method and our method, we return the first ten images in descending order of cosine similarity.

based on ResNet101-Pool5 is close to the corresponding result of Bag of Words and multi-vocabulary trees. For Holidays Dataset and Oxford Paris Dataset, the mAP of CNN-based (AlexNet-FC7 and ResNet101-Pool5) method is also a little worse than method based on exhaustive pairwise matching (SIFTGPU) but much higher than those based on Bag of Words and multi-vocabulary trees. We also test the accuracy and query time for pHash with support of OpenCV. For the same datasets, single image query takes only 0.000559s, 0.019724s and 0.000212s, respectively. However, the accuracy of image retrieval is 0.613, 0.4762 and 0.885, respectively. We find that pHash performs well on the datasets which images differ slightly on appearance like Oxford Paris Dataset.

For the CNN-based (AlexNet-FC7 and ResNet101-Pool5) method, the first 10 images returned are basically related images with a high degree of overlap with the query image. However, there are still some false images, only with similar but different objects; in other words, the CNN-based method cannot distinguish objects from the same class, which will be fatal to the 3D reconstruction of a single object. In Fig. 6 (a) and Fig. 6 (c), there are some irrelevant error results; in Fig. 7(a) and Fig. 7(c), although some of the returned images are also about churches, they are not the same church as in the query image. Therefore, these returned images cannot be used for the church reconstruction. In Fig. 8(a) and Fig. 8(c), the overlap degree between some returned images and the query image are not high. Generally, regarding query time, the CNN-based method improves on the other methods. Nevertheless, there will be some false results, as mentioned above. Considering that the most relevant image has been included in the returned candidate related images based on AlexNet-FC7 or ResNet101-Pool5 global feature vectors in a very short time, we will spend extra acceptable time determining the most relevant image accurately based on exhaustive pairwise matching (SIFT). This ‘‘from coarse to fine’’ retrieval strategy reduces the total number of time-consuming SIFT feature matching. It performs better in terms of accuracy than the other methods as well as guarantees the query efficiency, ensuring the feasibility of near real-time and high precision 3D reconstruction.

C. EXPERIMENTAL RESULTS BASED ON THE ‘‘FROM COARSE TO FINE’’ RETRIEVAL STRATEGY

Follow the steps in section II, re-ranking is carried out on the candidate related images. According to the number of images in the Holidays Dataset, Oxford Buildings Dataset and Oxford Paris Dataset, we set the number of returned candidate related images, rounding to 25, 50, and 15, respectively. Within the candidate overlapping images, it is quick and accurate to determine the most relevant image for the following reconstruction according to the result of SIFT feature matching with support of GPU. Table 1 provides a detailed quantitative comparison of the query time of our method and other popular methods. In our method, time for coarse determination and time for fine determination are summed as query time. Table 2 provides a detailed quantitative comparison of mean average precision of our method and other popular methods.

Our method takes a little more time than the CNN-based method due to re-ranking, but less than methods such as Bow, multi-vocabulary trees, and exhaustive pairwise matching. From the returned image perspective, false images only with similar objects are eliminated which cannot be filtered based on global features. Our method achieves the best accuracy compared with other traditional methods and the state-of-the-art methods [28]–[30]. Fig. 6 (b) and Fig. 6 (d) show the single image query results only based on our method for Holidays Dataset. Fig. 7(b) and Fig. 7(d) show the single image query results only based on our method for Oxford Buildings Dataset. Fig. 8(b) and Fig. 8(d) show the single

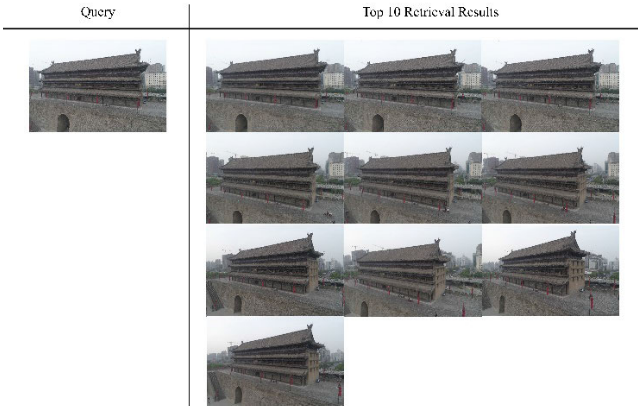


FIGURE 9. Query results for close-range photogrammetry. The first image in line one is the query image. We return the first ten related images through our method. All returned images are suitable for the ancient city wall 3D reconstruction.

image query results only based on our method for Oxford Paris Dataset.

D. EXPERIMENTAL RESULTS FOR CLOSE-RANGE PHOTOGRAMMETRY

We also test the 587 UAV images for an ancient city wall in Xi'an of size 5472×3648 collected by the DJI Phantom 4 PRO (see Fig. 4). For each query image, our method (ResNet101-Pool5+SIFT) takes only 0.417 seconds on average from feature extraction to acquisition of the most relevant image while methods based on Multi-Vocabulary Trees, exhaustive pairwise matching and Bag of Words take 0.733, 1.074 and 1.211 seconds respectively, the final first ten images returned for one UAV image in descending order are shown in Fig.9. Our method still performs excellently on UAV images irrespective of the query speed and mean average precision.

After determining the most relevant image, feature matching points between the query image and the corresponding most relevant image after filtering outliers are applied to the subsequent incremental SFM reconstruction.

IV. CONCLUSION

In this paper, a "from coarse to fine" image retrieval strategy is proposed which determines some candidate related images first based on global features obtained through AlexNet or ResNet101 forward propagation rapidly and then spends extra acceptable time determining the most relevant image from the candidate related images for each query image based on a local feature matching method such as SIFT. A single image query was taken from several benchmark datasets and UAV images to verify the accuracy and efficiency of our method. Experimental results show that our method performs excellently both in terms of accuracy and efficiency compared with former popular methods.

REFERENCES

[1] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May 2018.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[3] Sivic and Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.* Piscataway, NJ, USA: IEEE Press, 2003, p. 1470.

[4] S. Arya, "An optimal algorithm for approximate nearest neighbor searching in fixed dimensions," in *Proc. 5th ACM-SIAM Symp. Discrete Algorithms, Soc. Ind. Appl. Math.*, 1994, pp. 573–582.

[5] C. Silpa-Anan and R. Hartley, "Optimised KD-trees for fast image descriptor matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, Art. no. 4587638.

[6] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, 2009, pp. 331–340.

[7] M. Muja and D. G. Lowe, "Fast matching of binary features," in *Proc. 9th Conf. Comput. Robot Vis.*, May 2012, pp. 404–410.

[8] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2227–2240, Nov. 2014.

[9] X. Wang, Z. Zhan, and C. Heipke, "An efficient method to detect mutual overlap of a large set of unordered images for structure-from-motion," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 4, pp. 191–198, Jun. 2017.

[10] Z. Zhan, C. Wang, X. Wang, and Y. Liu, "Optimization of incremental structure from motion combining a random k-d forest and pHash for unordered images in a complex scene," *J. Electron. Imag.*, vol. 27, no. 01, p. 1, Feb. 2018.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1–9.

[12] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. ECCV*, 2014, pp. 584–599.

[13] K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen, "Deep learning of binary hash codes for fast image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 27–35.

[14] J. Y.-H. Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 53–61.

[15] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," in *Proc. ICLR*, 2016, pp. 1–12.

[16] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015.

[17] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12716–12725.

[18] K. Yan, Y. Wang, D. Liang, T. Huang, and Y. Tian, "CNN vs. SIFT for image retrieval: Alternative or complementary?" in *Proc. ACM Multimedia Conf. (MM)*. New York, NY, USA: ACM, 2016, pp. 407–411.

[19] S. Huang and H.-M. Hang, "Multi-query image retrieval using CNN and SIFT features," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 1026–1034.

[20] S. Xu, W. Chou, and H. Dong, "A robust indoor localization system integrating visual localization aided by CNN-based image retrieval with Monte Carlo localization," *Sensors*, vol. 19, no. 2, p. 249, Jan. 2019.

[21] Y. Chen, R. Chen, M. Liu, A. Xiao, D. Wu, and S. Zhao, "Indoor visual positioning aided by CNN-based image retrieval: Training-free, 3D modeling-free," *Sensors*, vol. 18, no. 8, p. 2692, Aug. 2018.

[22] Q. Feng, Q. Hao, Y. Chen, Y. Yi, Y. Wei, and J. Dai, "Hybrid histogram descriptor: A fusion feature representation for image retrieval," *Sensors*, vol. 18, no. 6, p. 1943, Jun. 2018.

[23] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. ECCV*, 2008, pp. 304–317.

[24] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[25] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

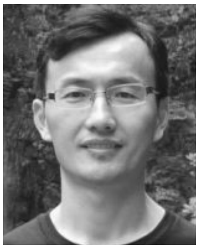
[26] M. Yang, W. Song, and H. Mei, "Efficient retrieval of massive ocean remote sensing images via a cloud-based mean-shift algorithm," *Sensors*, vol. 17, no. 7, p. 1693, Jul. 2017.

[27] X. Tang, X. Zhang, F. Liu, and L. Jiao, "Unsupervised deep feature learning for remote sensing image retrieval," *Remote Sens.*, vol. 10, no. 8, p. 1243, Aug. 2018.

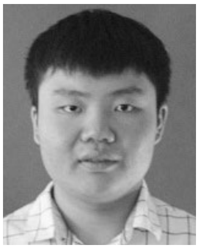
[28] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *Proc. ECCV*, 2016, pp. 241–257.

[29] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," Oct. 2016, *arXiv:1610.07940*. [Online]. Available: https://arxiv.org/abs/1610.07940

[30] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual instance retrieval with deep convolutional networks," in *Proc. ICLR workshops*, 2015, pp. 1–9.



ZONGQIAN ZHAN received the M.A. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2003 and 2007, respectively. He is currently an Associate Professor with the School of Geodesy and Geomatics, Wuhan University. His research interests include camera calibration, close-range photogrammetry, unmanned aerial vehicle (UAV) photogrammetry, oblique photogrammetry, deep learning, and remote sensing.



GAOFENG ZHOU received the Engineering degree from the School of Environment of Science and Spatial Informatics, China University of Mining and Technology, in 2018. He is currently pursuing the master's degree in photogrammetry and remote sensing with Wuhan University. His research interests include multiview stereo matching and structure from motion.

XUE YANG received the B.S. degree in geomatics engineering from Wuhan University, Wuhan, China. He is currently pursuing the M.S. degree in photogrammetry and remote sensing with the School of Geodesy and Geomatics, Wuhan University. His current research interests include video restoration, SLAM, and object detection.