IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

GeoCluster: Enhancing Visual Place Recognition in Spatial Domain on Aerial Vehicle Platforms

Chao Chen, Mengfan He, Jun Wang, *Member, IEEE*, and Ziyang Meng, *Senior Member, IEEE*

***Abstract*— Visual Place Recognition (VPR) is a critical technology for achieving robust long-term visual geo-localization. During the past few years, VPR research mainly focused on ground-based platforms in the street-level captured scenes with deep learning methods (e.g. NetVLAD, GeM), but little attention was paid to the VPR task on aerial vehicles. The algorithms and models designed for ground-based platforms are always directly applied to the aerial VPR problem. However, the viewpoint variance on Unmanned Aerial Vehicles (UAV) is much larger than the ground-based platforms. Due to the sparse distribution of aerial image features, when the viewpoint of the camera changes, the features of the query image are largely inconsistent with the descriptors in the database, which results in the failures of image retrieval and visual geo-localization. In this letter, we propose an aerial VPR enhancement module called *GeoCluster*, which presents a feature aggregation method using spatial clustering information to improve the robustness and consistency of the global descriptors for UAV-captured frames. Moreover, it can be applied to any NetVLAD-based VPR method and boost the pre-trained model without any further training process. By integrating GeoCluster into an existing state-of-the-art localization method, we can achieve about 10% improvement for aerial image retrieval tasks and have more accurate and robust geo-localization results.**

I. INTRODUCTION

Visual Place Recognition (VPR) is a crucial technology [15] applied in the vision-based localization system. In the typical VPR problem, a place retrieval task is performed based on the environment database, such as images and 3D point clouds. The traditional VPR technology has extensive applications ranging from autonomous vehicles [24] to mobile devices [5]. In the visual Simultaneous Localization and Mapping (SLAM) system [17], VPR is closely related to the loop closure detection module to eliminate the accumulated drift of the pure odometry method. In addition, based on the pre-built map database of the environment, VPR can also be applied to retrieve the co-visible images and perform the global localization later.

Currently, most of the VPR methods focus on ground-based platforms with street-view observation. On the other hand, less attention has been paid to VPR tasks on aerial platforms, where the satellite map is usually used as the database to retrieve the query images captured by aerial vehicles. In particular, the majority of public datasets are collected by vehicles or hand-held devices such as Pitts250k [26], Tokyo24/7 [25], and MSLS [27], which hinders the evaluation of the VPR method on the aerial platforms. Although the Shopping Street dataset [16] and EuRoc dataset [4] use the Unmanned Aerial Vehicle (UAV) as the data collection platform, the appearances of these datasets are very similar to the street view images since the flight area and altitude are limited. Considering the aerial vehicle platforms which are used at much higher altitudes and in larger-scale areas, the Global Navigation Satellite System (GNSS) signals are often unavailable in certain sceneries, which highlights the necessity to optimize the existing VPR methods for the aerial-based visual geo-localization task. The aerial-based VPR problem has quite different scenarios from the traditional techniques designed for the ground-based VPR task. The main challenges of aerial-based VPR include:

1) *Descriptor Unreliability:* Visual information collected by aerial platforms varies greatly from ground images. Slight differences in the viewpoints (i.e., rotation, scaling) can cause dramatic changes in visual appearance, such that the primary features-based global descriptor is inconsistent with the descriptors of the database.
2) *Data Scarcity:* Due to the significant differences between the satellite imagery and aerial frames, a learning-based module and an offline training process are also needed for the aerial-based VPR problem. However, the ability of the learning-based methods is limited due to the lack of a large-scale aerial-based VPR dataset.

In this letter, we proposed an aerial VPR enhancement method called *GeoCluster*, which can improve the domain adaptation of pre-trained VPR models in a non-training manner. The images captured from the aerial platform contain more landscapes. The different classes of features are strongly spatially correlated. Motivated by these characteristics, we design a spatial clustering method for the learning-based vector of locally aggregated descriptors (VLAD). We re-weight the deep feature map extracted from the base model of VLAD by using the spatial relationship between

Manuscript received: September, 12, 2023; accepted January, 25, 2024. Date of publication 7 February 2024; date of current version 19 February 2024.This letter was recommended for publication by Editor S. Behnke upon evaluation of the Associate Editor and Reviewers' comments. This work was supported in part by National Natural Science Foundation of China under Grant 62273195, 62303040, and in part by Beijing Natural Science Foundation L233029.(*Chao Chen and Mengfan He contributed equally to this work.*) (*Corresponding author: Jun Wang.*)

Chao Chen and Jun Wang are with the College of Information Science & Technology, Beijing University of Chemical Technology, Beijing, 100029, China (e-mail: chenchao@buct.edu.cn; wangjunrob@buct.edu.cn).

Mengfan He and Ziyang Meng are with the Department of Precision Instrument, Tsinghua University, Beijing, 100084, China (e-mail: hmf21@mails.tsinghua.edu.cn; ziyangmeng@mail.tsinghua.edu.cn).

To foster future research, we make the code and datasets in this work publicly available for any researcher at https://github.com/cbbhuxx/GeoCluster.

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2024.3363536, provided by the authors.

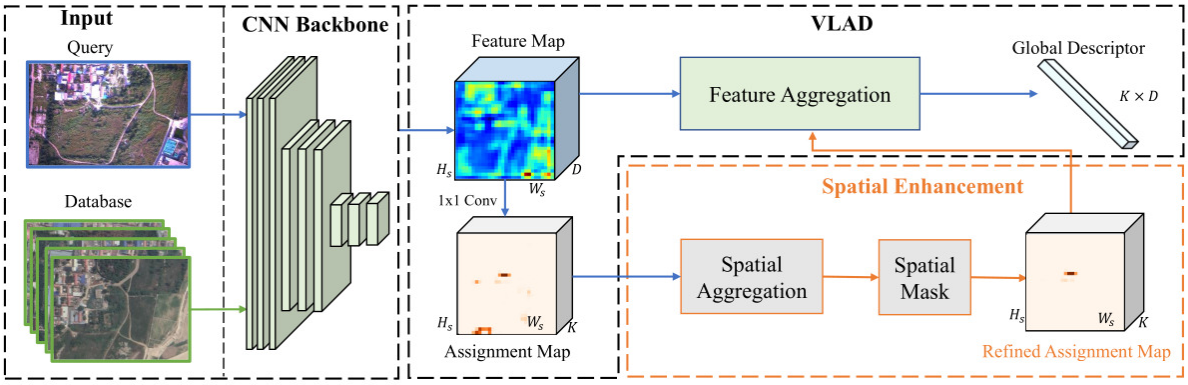

Fig. 1. The architecture of proposed method. We give a versatile enhancing approach for the traditional NetVLAD-based descriptors which focuses on the assignment map rather than the extracted multi-dimensional features. The feature map extracted by the backbone is structured with sampled shaped $H_s \times W_s$ and the feature channel D. The assignment map with K clusters has the same tensor shape as the feature map and is refined by the proposed spatial enhancement module. Finally, the extracted feature is aggregated by the refined assignment map to get the global descriptor with $K \times D$ dimension.

different clustering categories, as shown in Fig. 1. Thus, a more reliable global descriptor is achieved for the aerial VPR tasks. Instead of adding new training parameters in terms of extracting features as well as soft assignments, the aggregation module which combines the spatial relationships of the feature points is designed for reweighting. The contributions of our work are summarized as follows:

1) An enhancement method of the global descriptor for aerial VPR algorithm is proposed in this work. In particular, the spatial information of clusters is utilized in the encoding process to improve the global descriptor consistency when the viewpoint changes.
2) We show that the proposed method can be easily integrated into any existing VLAD-based neural network as an individual module. Moreover, this module achieves significant enhancement without any extra learnable parameters for the trained model.
3) We conduct several dataset evaluations and real-world experiments for aerial visual geo-localization. Compared with the state-of-the-art methods, the proposed method achieves 9.8% and 9.5% improvements on average retrieval recall of three evaluated datasets and has faster convergence speed in localization experiments.

The rest of this letter is organized as follows. In Sec. II, we thoroughly review the VPR techniques related to this letter. Then, we present the derivation of the basic method and the enhancement design in Sec. III. Then, we give the implementation and evaluation details of the proposed method in Sec. IV. Finally, in Sec. V, we summarize this letter.

II. Related Work

Researchers have proposed various global image feature descriptors to extend VPR technology into more applications. These designs aim to improve the performance and robustness of global feature extraction in large-scale environments and under different conditions, such as changes in illumination, weather, and season. Here, we will introduce the current research on global image descriptors in VPR and highlight the advantages and limitations of different global feature extraction methods. We further discuss the potential of global descriptors designed to improve the performance and stability of aerial visual place recognition.

1) Hand-Crafted Features in Global Descriptors: Global image descriptor is the essential part of VPR and Loop Closure Detection (LCD), such as Bag of Words (BoW) [7], Fisher Vectors (FV) [2], and Vector of Locally Aggregated Descriptors (VLAD) [9]. These methods are mainly based on the hand-crafted image local features including LBP [18], SIFT [14], SURF [3], and ORB [22] to obtain the representation of the global descriptor. Utilizing the aggregation of local region features makes the global descriptor collects the statistic feature at the image level and more distinctive. However, the global descriptor based on hand-crafted features might not be able to adequately capture complex and subtle visual patterns and fail to generalize well to various conditions.

2) Learning-based Global Descriptors: In recent years, learning-based global descriptors have received rapid development for VPR. Several works attempted to replace the hand-crafted features with learning-based ones. Different from VLAD, NetVLAD [1] computes the image-wise global descriptor from the learned features and makes significant progress in the image retrieval problem. Subsequently, more researchers [11], [28], [32], [20] focus on constructing novel features based on NetVLAD. Some researchers [8][13] took both local descriptors and global descriptors into consideration to correctly identify the embedded global features. The other works [21], [29], [12] designed end-to-end networks to evaluate the similarity between different images relying on more complex convolution operations. These features are more capable of capturing complex and stable visual patterns than hand-crafted features since they are learned from a large number of training images.

Although learning-based global descriptor extraction methods showed promising results for the ground-based VPR task, few researchers focused on visual place recognition solutions for aerial platforms. Recent work [31] evaluates existing ground-based VPR methods on a self-built aerial-based dataset to test the general ability of ground-based VPR techniques to the aerial platforms. However, the evaluated dataset has a similar appearance to the street-level images

and does not cover the high-altitude situation.

The state-of-the-art ground-VPR methods mainly construct global descriptors by capturing the primary features in the image. In high-altitude situations, target images are captured with a downward-facing camera from the UAV, which can cause dramatic changes in the image features even if the viewpoint has slight differences. Therefore, the primary feature-based ground-VPR methods make it difficult to provide global descriptors that are consistent with the database when the viewpoint of the aerial platform changes. Furthermore, datasets collected from the aerial platforms are insufficient compared to the ground-based datasets, which makes it also hard to train the corresponding learning model according to the specific mission. Therefore, we propose a novel learning-free algorithm that is both distinctive and robust without any extra training process. A spatial clustering method is designed to re-weight the deep feature map from the base NetVLAD model. In addition, more reliable global descriptors can be achieved by using the spatial relationship between different clusters.

III. Problem Description and Methodology

A. Problem Description

The visual geo-localization problem is defined as a task to find the corresponding location of the captured frame by retrieving it from a pre-built map database. The key to correctly retrieving the query image is the discriminative description of the image. The classical image description method is constructed from two parts: feature extraction and feature encoding. Fig. 2 shows these two stages and the effects of viewpoint variation on the description of the images that respectively token from the ground platform and aerial platform. As shown in Fig. 2, alteration in the viewpoint variation could directly lead to an impact on the difference of the feature map. Since the aerial platform has a broader field of view than the ground platform, the change of viewpoints generates a more significant difference in the extraction of feature maps than those of the ground platform. Consequently, the image global descriptor of aerial VPR is affected more seriously by the viewpoint change as the cosine similarity shows. The comparison results indicate the limitation of the traditional ground VPR method on the aerial VPR task.

B. Methodology

The proposed method mainly focuses on the global descriptors of queried images for the aerial VPR methods. The encoding strategy for the global descriptor is outlined in three parts as follows. Firstly, an initial global descriptor is calculated by using the CNN model to extract a feature map and soft assignment map. Then, based on the spatial domain of the soft assignment map, a more reliable soft assignment map is obtained. Lastly, a spatial discriminative mask is added to re-weight the assignment map on the feature graph to achieve a more robust global descriptor. An illustration of the refinement is shown in Fig. 1.

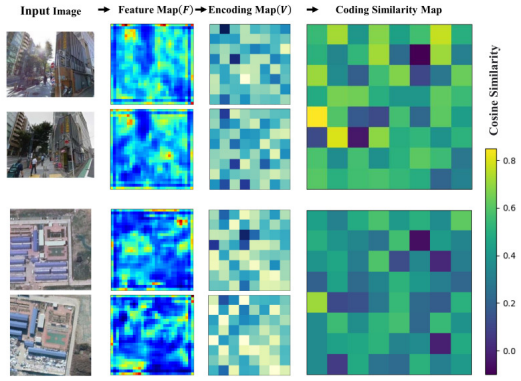

Fig. 2. The effects of point-of-view variations on the ground and aerial platforms. The feature maps of the aerial platform have larger differences than those of the ground platform with viewpoint variations. Thus, a lower cosine similarity of the global descriptor is obtained by the aerial platform.

1) VLAD Representation with Learning-based Feature Map: Note that the initial global descriptor can be achieved by using various VLAD-based encoding methods. We briefly introduce the NetVLAD approach as an example for calculating the initial global descriptor here. NetVLAD model applies multi-layer convolutional blocks as the autoencoder to extract a feature map $\mathcal{F}$ ($\mathcal{F} \in \mathbb{R}^{H_s \times W_s \times D}$), where the parameters H_s and W_s represent the size of the extracted feature map. The parameter D denotes the number of channels (or the length of the global descriptor). This multi-dimensional tensor gives the dense feature of every single pixel position in this down-sampled image map.

Then, in the encoding stage, a convolution block with 1×1 convolution kernel and K convolution depth is applied to obtain the assignment situation of each cluster center. K denotes the number of cluster centers in the database and it is a manually set parameter. We can refer this assignment map as $\mathcal{A}$ ($\mathcal{A} \in \mathbb{R}^{H_s \times W_s \times K}$). To make the sum of the assignment equal to 1, a softmax function is applied as the following activation layer. Finally, the output of initial global description $\mathcal{V}$ ($\mathcal{V} \in \mathbb{R}^{K \times D}$) is achieved after the feature aggregation and L_2 normalization. A more detailed explanation of the NetVLAD framework can be found in [1].

2) Reliable description based on cluster in spatial domain: The assignment map $\mathcal{A}$ obtained in the VLAD-based encoding process is also called a "soft" assignment tensor. The soft assignment results are highly dependent on the training dataset and sensitive to the change in visual appearance. To improve the robustness of the VLAD-based algorithms to the viewpoint variation of aerial VPR, we design an enhancing method by utilizing the assignment information in the image spatial domain, as shown in Fig. 3.

Each pixel in the down-sampling of an image is considered as a feature point. The set of feature points are defined as $\mathbb{P} = \{(i,j) | 0 \leq i < H_s, 0 \leq j < W_s\}$. We define the pixel belongs to a certain cluster as follows:

$$p_{ij} \in C_k, k = argmax(\mathcal{A}(p_{ij})), \tag{1}$$

where p_{ij} denotes the pixel at the coordinate (i,j) in the feature map, and C_k is one of the aggregated cluster center. $\mathcal{A}(p_{ij})$ represents a $1 \times 1 \times C$ vector of probability distribution

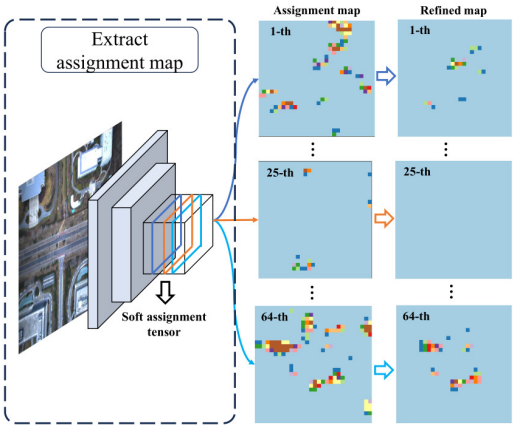

Fig. 3. The illustration of the enhancement to the soft assignment map. NetVLAD extracts an assignment tensor with 64 channels (left), and each channel represents the probability distribution of the feature points belonging to a certain cluster. The probability value is marked using different colors in the figure, and the background with light blue indicates zero probability (middle). To eliminate the influence of unstable information with a low probability, a selecting approach is proposed by taking spatial information into account (right).

at p_{ij}. In the original NetVLAD module, the aggregate description at the C_k cluster center is defined as:

$$\mathcal{V}_k = \sum_{(i,j)\in\mathbb{P}} \mathcal{A}_k(p_{ij}) \cdot (\mathcal{F}(p_{ij}) - \mathcal{C}_k)\,, \tag{2}$$

where $\mathcal{C}_k$ denotes the precomputed feature vector of the C_k cluster center, and $\mathcal{F}(p_{ij})$ is the encoded feature. $\mathcal{A}_k(p_{ij})$ represents the probability of feature point p_{ij} belonging to cluster k. And the representation feature is aggregated from every pixel and weighted by the soft assignment tensor.

Eq. (2) demonstrates that the expression ability of global descriptors mainly depends on the nearness relation between the encoded features and their corresponding cluster centers. More specifically, the cluster center that the features belong to has a greater contribution to obtaining a stable global descriptor. Conversely, the cluster centers in other categories even have a negative effect on the global description $\mathcal{V}$ as the noise. This discovery can be further extended to the proposition that a better performance of the assignment map can be achieved by considering the distinctive region rather than the overall area of $\mathbb{P}$. Motivated by this observation, a selection approach is proposed to choose the target region in this letter. The target region is composed of several patches around the pixel of the assignment map which belongs to a certain cluster. We define $\hat{\mathbb{P}}_k$ as the target region of the C_k cluster center and the selection approach of $\hat{\mathbb{P}}_k$ is given below:

$$\hat{\mathbb{P}}_k = \{p_{ij} \mid \|(i,j) - (i',j')\|_2 \le m \text{ and } p_{i'j'} \in C_k\}\,, \tag{3}$$

where (i', j') denotes the coordinate of the feature point that belongs to the cluster C_k, and m represents the threshold distance to limit the size of selected patches around the feature points (i', j'). In Eq. (3), the original region $\mathbb{P}$ of aggregation region is refined to $\hat{\mathbb{P}}_k$. In addition, a new probability of feature point $\bar{\mathcal{A}}_k(p_{ij})$ is achieved, subsequently. The noise from the irrelevant features is removed and only the principal component is taken into consideration. The aggregate description at the C_k cluster center is derived as:

$$\mathcal{V}_k = \sum_{(i,j)\in\hat{\mathbb{P}}_k} \bar{\mathcal{A}}_k(p_{ij}) \cdot (\mathcal{F}(p_{ij}) - \mathcal{C}_k)\,, \tag{4}$$

3) Spatial discriminative mask: In the aerial VPR issue, more paired features with similar coordinates in the image plane denote a higher probability that these two images represent the same place. Thus, we further improve the reliability of the assignment map $\bar{\mathcal{A}}$ by designing a spatial discriminative mask. The re-weighting mask formula is defined below:

$$\hat{\mathcal{A}}_k = \mathcal{M} \times \bar{\mathcal{A}}_k, \quad \mathcal{M} = \mathcal{M}_c \times \mathcal{M}_d, \tag{5}$$

$$\mathcal{M}_{c(ij)} = \lambda_1 - (H_s/2 - i)^2 - (W_s/2 - j)^2, \tag{6}$$

$$\mathcal{M}_{d(ij)} = 1/d(C_l, C_k), l \in (1, K = 64)), \tag{7}$$

$$\begin{aligned} d(C_l, C_k) = &\sum_{p_{ij}\in C_k} (d(C_l, p_{ij}))/num(C_k) \\ &+ \sum_{p_{i'j'}\in C_l} (d(C_k, p_{i'j'}))/num(C_l) + \lambda_2, \end{aligned} \tag{8}$$

$$d(C_l, p_{ij}) = \min[(i^{'} - i)^2 + (j^{'} - j)^2],\ p_{i'j'} \in C_l, \tag{9}$$

where $\mathcal{M}_c$ and $\mathcal{M}_d$ are the re-weight masks with the same size as the assignment map $\bar{\mathcal{A}}$. $\mathcal{M}_c$ determines the effectiveness of feature points at different coordinates of the assignment map $\bar{\mathcal{A}}$. If the feature point is near the center of the feature map, it implies that this feature would be less affected by changes in viewpoint, and a higher weight is set to the feature at this coordinate. In addition, we set a parameter λ_1 to limit the highest weight for the feature point in the center in Eq. (6).

$\mathcal{M}_d$ is another discriminative mask that represents the spatial relationship between the clusters. We mention that the nearer cluster neighbors will have more effect on the feature aggregation. Therefore, we set the matrix element at (i, j) of $\mathcal{M}_d$ to the reciprocal of the distance between cluster C_l and C_k in Eq. (7). The distance between two clusters is composed of the average point-to-cluster distance as given in Eq. (8) and we give the definition of point-to-cluster distance as the minimum distance among the target feature points belonging to the cluster. As indicated in Eq. (9), if the feature point p_{ij} belongs to the target cluster C_l, the distance $d(C_l, p_{ij})$ equals to 0. Therefore, a parameter λ_2 is set as the minimum distance margin. Finally, the above definition of the spatial discriminative mask aggregates more features of the closer clusters and leaves out other unrelated features.

IV. Evaluation

A. Evaluation Datasets

We evaluate the proposed enhancement module using three different datasets including the self-built *City* dataset, *ALTO* [6] dataset, and *VP-Air* [23] dataset. *City* dataset is a self-built dataset that employs multiple terrain-class satellite images including urban, farmland, and desert. We

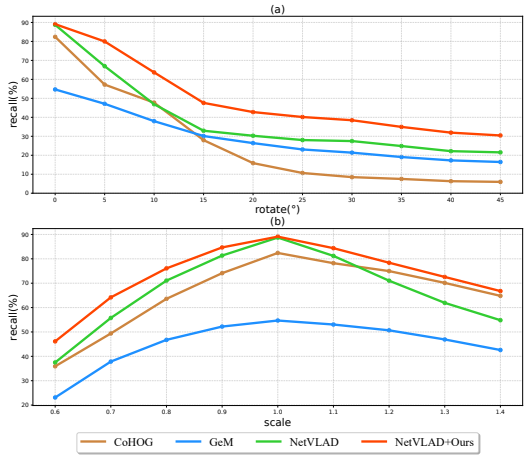

Fig. 4. Image retrieval evaluation on the *City* dataset: (a) retrieval performance with changing the query image's rotation, (b) retrieval performance with changing the query image's scale.

select maps from different years to simulate the database and UAV-captured frames. The well-aligned satellite maps are introduced with several image transformations such as rotation, scaling, and translation to simulate viewpoint variance in real-flight scenarios. *ALTO* dataset is captured from a helicopter with a long flight trajectory. There are several terrains including urban, suburban, forested, and rural. This dataset also provides the corresponding satellite imageries as well as the aerial captured frames. Moreover, satellite imageries with several offsets are also provided, which enable us to test the retrieval under some translation. *VP-Air* dataset is another aerial-based VPR dataset consisting of several database-query image pairs, which also encompasses various challenging regions, such as urban, forest and farmland. All the experiment results are obtained on a Windows computer with NVIDIA GeForce RTX3060 GPU, Intel Core i7-12700H 2.30 GHz, and 16GB running memory.

B. Image Retrieval Evaluation

1) Model Enhancement: In this section, we evaluate the image retrieval performance on the *City* dataset to prove the improvement of the proposed algorithm. Besides the original NetVLAD model, other classical VPR methods such as CoHOG [30] and GeM [21] are also evaluated as the baseline. To explore the ability of our algorithm to handle image viewpoint changes, we use different image augmentations on the query image. The evaluation results with different rotations and scaling are shown in Fig. 4(a) and Fig. 4(b), respectively. We evaluate the image encoding performance using the evaluation metric of the top-k% recall, which denotes the percentage of correctly retrieved images among the top-k% ranked database images. Both of the two curves show the top-2% recall.

In Fig. 4(a), the results without any rotation are regarded as the comparison baseline, where the query images are well aligned with the database image. Although the evaluated algorithms perform well in the baseline, the proposed algorithm can still make progress based on the original NetVLAD algorithm. In addition, it also shows that when the rotation angle increases from 0° to 45°, the performances of CoHOG, GeM and NetVLAD fall significantly. Even so, the proposed method can still make a large improvement up to about 10%$\sim$15% against the original NetVLAD model.

The performances of four methods with scaling factors from 0.6 to 1.4 are shown in Fig. 4(b). According to the results, the proposed method can significantly improve the performance of NetVLAD with different scales. Though the recall rate decreases with the increase of scaling changes, the proposed algorithm has a relatively lower decline. And, the proposed algorithm performs much better than the other three methods, because it can improve the consistency of the global descriptors by considering the spatial information to select more informative features and exclude useless features.

Additionally, we set different image translations to simulate the viewpoint change. We randomly select the translation within 20% of the image resolution and give the top-k% recall (denoted as R@k%) in Table I. The results show that the translation change also leads to substantial decreases in the retrieval recall of the baseline algorithms. That is because the primary features change in the image with the viewpoint variance, therefore the descriptors of query image and database are largely inconsistent. However, the proposed method achieves better performance by containing more robust spatial information.

2) Method Generalization: As mentioned in the introduction section, the proposed method can be integrated into any VLAD-based model as an individual module. To prove the generalization of the proposed method, besides the original NetVLAD method, we implement it into other state-of-the-art VLAD-based algorithms, including MultiRes(MR)-NetVLAD [11], Patch-NetVLAD [8] and AnyLoc [10]. We use the *ALTO* dataset and *VP-Air* dataset as the evaluation datasets.

ALTO dataset offers five kinds of reference images with different offsets. For example, db_N_20 presents the paired image with 20 meters offset in the north direction and db_S_20 means the offset in the south direction. We separately test the performance of MR-NetVLAD using its own trained weight (marked as I) and the initial weight trained by NetVLAD (marked as II). In addition, we use the RANSAC algorithm as the verification step in Patch-NetVLAD and employ the DINOv2 [19] vision transformer as the feature extractor in AnyLoc with soft assignment map. The evaluation results are presented in Table II. Applying the proposed module to the original NetVLAD model is capable of achieving a 9%$\sim$10% improvement on *ALTO* dataset. The MR-NetVLAD achieves a 3%$\sim$8% improvement against the NetVLAD model due to the extra consideration of multi-

TABLE I

RECALL ON THE *City* DATASET WITH DIFFERENT TRANSLATIONS.

Method	R@0.5%	R@1.0%	R@2.0%
CoHOG	45.32	53.83	64.23
GeM	32.14	40.31	49.03
NetVLAD	52.01	62.15	73.83
Ours	**65.86**	**74.72**	**83.56**

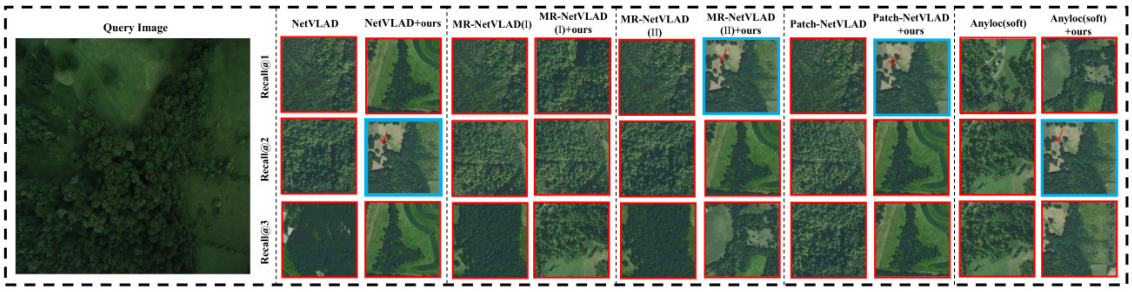

Fig. 5. Evaluations of NetVLAD, MR-NetVLAD, Patch-NetVLAD, Anyloc, and the corresponding improved algorithms on the *ALTO* dataset. We provide the top three ranked retrieved images. The correct and incorrect retrieved images are marked with the blue border and the red border, respectively. Compared to the original models, the proposed method achieves an obvious improvement in retrieving the correct candidate images.

TABLE II

EVALUATION OF THE PROPOSED GENERALIZATION METHOD. THE RESULTS ARE TESTED ON THE *ALTO* DATASET

Method	Well Aligned			db_N_20			db_S_20			db_N_40			db_S_40		
	R@0.5%	R@1.0%	R@2.0%	R@0.5%	R@1.0%	R@2.0%	R@0.5%	R@1.0%	R@2.0%	R@0.5%	R@1.0%	R@2.0%	R@0.5%	R@1.0%	R@2.0%
NetVLAD	25.48	39.08	52.37	23.90	36.52	48.51	26.50	40.20	52.30	21.14	33.75	46.13	25.59	38.80	50.37
NetVLAD+Ours	**36.17**	**50.12**	**62.60**	**33.61**	**46.93**	**60.92**	**36.45**	**49.74**	**62.50**	**30.25**	**43.95**	**56.43**	**35.30**	**47.74**	**59.83**
MR-NetVLAD(I)	33.02	44.79	55.13	30.60	42.31	53.31	33.72	45.67	55.45	29.09	40.97	50.86	32.21	43.32	54.47
MR-NetVLAD+Ours(I)	**37.36**	**50.05**	**60.67**	**35.37**	**47.56**	**58.43**	**38.38**	**49.88**	**60.60**	**31.51**	**42.80**	**54.92**	**35.75**	**47.53**	**57.87**
MR-NetVLAD(II)	27.41	41.01	53.24	24.04	36.94	50.58	27.23	41.22	53.24	22.19	35.26	46.97	26.85	39.40	50.37
MR-NetVLAD+Ours(II)	**41.92**	**56.19**	**68.24**	**39.26**	**54.29**	**66.18**	**42.06**	**55.77**	**68.28**	**36.03**	**50.12**	**62.50**	**39.71**	**54.05**	**66.88**
Patch-NetVLAD	39.36	44.86	46.82	35.71	41.25	43.32	39.92	45.07	47.31	32.35	37.22	39.53	38.27	42.72	45.07
Patch-NetVLAD+Ours	**41.60**	**47.31**	**50.47**	**40.09**	**46.23**	**49.49**	**43.07**	**48.75**	**51.98**	**35.47**	**41.88**	**44.37**	**39.95**	**45.56**	**48.93**
AnyLoc(soft)	51.17	66.53	76.62	49.21	64.42	74.55	50.51	64.42	75.67	44.79	60.67	73.26	48.26	61.37	73.50
AnyLoc(soft)+Ours	**69.12**	**79.64**	**87.10**	**67.09**	**77.74**	**85.38**	**69.09**	**79.39**	**86.86**	**63.72**	**76.10**	**83.56**	**66.28**	**78.62**	**86.93**

TABLE III

RESULTS PERFORMANCE ON THE *VPAIR* DATASET

Method	R@0.5%	R@1.0%	R@2.0%	Time(ms)
NetVLAD	33.00	46.00	62.50	24.3
NetVLAD+Ours	**46.00**	**57.50**	**70.50**	26.5
MR-NetVLAD(I)	40.00	53.00	64.00	58.1
MR-NetVLAD+Ours(I)	**42.50**	**54.50**	**69.50**	90.8
MR-NetVLAD(II)	33.00	46.00	62.00	55.6
MR-NetVLAD+Ours(II)	**38.00**	**58.50**	**72.00**	87.5
Patch-NetVLAD	62.50	65.00	66.50	1338
Patch-NetVLAD+Ours	**66.00**	**70.50**	**71.00**	1345
AnyLoc(soft)	56.00	76.00	87.50	1726
AnyLoc(soft)+Ours	**71.50**	**92.00**	**97.00**	1754

resolution inputs. With the enhancement of the proposed algorithm, the recall rate can further achieve about a 5% improvement compared to the original MR-NetVLAD method. And for the MR-NetVLAD without extra training, the improvements are up to over 10% by applying the proposed method, which is even better than the MR-NetVLAD with extra training. Although Patch-NetVLAD contains an additional verification step, the proposed GeoCluster module can still achieve a 2%$\sim$5% improvement based on the original encoding method. The AnyLoc with soft feature assignment has the best performance against other retrieval methods, due to its robust visual feature. After adding the GeoCluster module, there is also a significant enhancement with an increase of more than 10%. This proves that our method can enhance the global descriptors as long as the method is based on VLAD and the improvement is independent of the training dataset and network architecture. Fig. 5 shows the results after being enhanced by the proposed algorithms on the *ALTO* dataset. According to the visualization results, the original VPR models fail to retrieve the correct image since the database images have similar appearance. However, the proposed method is able to help the models to retrieve the correct reference image by utilizing more robust global descriptions.

Additionally, we test the same models on *VP-Air* dataset and implement the GeoCluster module to evaluate the generalizing ability. As shown in Table III, GeoCluster also makes comprehensive improvements over all kinds of the evaluated methods. Even for the well-performed AnyLoc, the proposed model can also achieve a greater improvement with the more discriminative visual feature. In particular, the time consumption of the proposed GeoCluster module is also listed in Table III. According to the results, although the proposed module adds an extra spatial enhancement process, the overall time consumption does not exhibit a significant increase. Since the time consumption of this module is relatively fixed, compared to time-consuming methods such as Patch-NetVLAD and AnyLoc, the proportion of GeoCluster is quite small.

C. Localization Experiment

We conduct a field test in Beijing, China, with a quadrotor equipped with a downward-facing camera and high-precision GNSS receiver as the position ground truth. The entire flight lasts about 10 minutes with a 3.8 km flight route and about 200 meters flight height. We also download the geo-

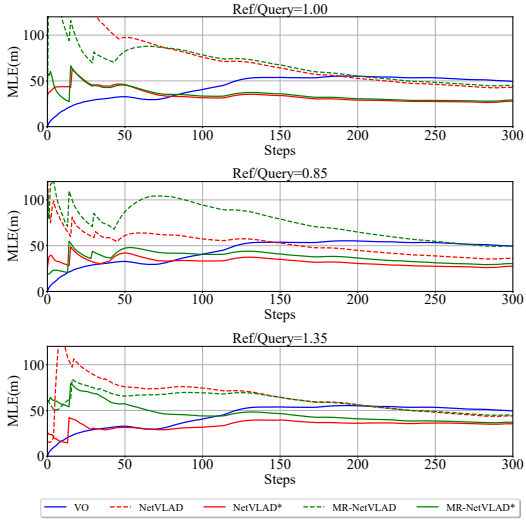

Fig. 6. The average localization errors of every step in visual odometry (VO), particle filter using NetVLAD and MR-NetVLAD. We also evaluate the models refined by GeoCluster (denoted as NetVLAD* and MR-NetVLAD*). The average position of particles with the higher weight is used as the localization result.

referenced satellite images of the flight area from Google Earth as the map database. We extract the global descriptors of the reference images using the proposed methods to compose the localization database in advance. To evaluate the improvement of the proposed algorithm for localization performance, we use particle filtering as the localization framework to integrate the proposed method. The weights of particles are updated based on the Euclidean distance in the image descriptors between the query image and the database image. We randomly initialize 5000 particles in the test area and the number of particles is gradually reduced in each resampling step until there are only 500 particles left. We test the localization performance of the NetVLAD model and MR-NetVLAD model with particle filter framework, as well as the models incorporating GeoCluster. The results are shown in Table.IV, where the results of the pure visual odometer (VO) are also listed as the baseline.

As for the localization result, we set different sizes of the database images to simulate the flight altitude variance. Scale factor 1 denotes the prepared database images have the same size as the query image. And the bigger scale factor represents the larger database image. As illustrated in the localization result, the accuracy of VPR methods is better than the pure VO method which is independent to the map database. Besides that, the proposed algorithm achieves better performances on both localization metrics than the original NetVLAD and MR-NetVLAD methods. In different scale factors, the average errors of NetVLAD are decreased by 6.5 m to 13.7 m using the proposed method. There is also an obvious decrease in localization error for MR-NetVLAD method. As to the convergence speed, when the localization error is less than 50 m, we consider the initialization process as converged. Similarly, the convergence speed of the particle filter is listed in the last row of Table IV. It shows that the proposed method converges 30% to 66% faster than the original NetVLAD and MR-NetVLAD methods. Furthermore, the mean localization errors for each step in the particle filter are shown in Fig. 6. For the overall localization performance, the enhancement module shows significant improvement in particle convergence speed and localization accuracy against both of the original methods. To show the details of each iteration of the particle filter pipeline, we present the distributions of the first four iterations in Fig. 7. It is shown that the particle distribution of the proposed algorithm is more concentrated compared with the original methods.

TABLE IV

LOCALIZATION PERFORMANCE USING PARTICLE FILTER.

Metrics	**Methods**	Map Scaling Factor		
		1.00	0.85	1.35
Mean Localization Error (m)	VO	48.59	48.59	48.59
	PF with NetVLAD	40.61	34.10	41.34
	PF with NetVLAD*	**26.92**	**26.41**	**34.54**
	PF with MR-NetVLAD	42.66	46.32	42.75
	PF with MR-NetVLAD*	**28.59**	**29.01**	**36.02**
Best Match Error (m)	PF with NetVLAD	44.54	49.31	80.62
	PF with NetVLAD*	**36.46**	**31.23**	**48.77**
	PF with MR-NetVLAD	54.16	72.26	63.50
	PF with MR-NetVLAD*	**38.52**	**43.30**	**49.63**
Convergence Speed (step)	PF with NetVLAD	70	69	106
	PF with NetVLAD*	**49**	**51**	**47**
	PF with MR-NetVLAD	89	96	132
	PF with MR-NetVLAD*	**48**	**54**	**45**

V. CONCLUSION

In this letter, to address the aerial VPR problem, we propose the GeoCluster algorithm: an enhancement method for the VLAD-based global descriptor by utilizing spatial information of clusters. In particular, we give more attention to the assignment map rather than the extracted feature map for feature aggregating. We select the informative region which has a spatial relationship with the cluster features and exclude the irrelevant areas. Experiment results prove that the proposed method can make significant improvements in the aerial image retrieval problem. Moreover, in contrast to the currently existing approaches, our system can be utilized as a separate module that can be directly integrated into any VLAD-based descriptor without extra training parameters. We also present the results in a field localization test by implementing our algorithm into the framework of the particle filter. The results show the efficiency and robustness of the proposed method in the visual geo-localization task.

REFERENCES

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.

[2] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *IEEE International Conference on Computer Vision*, 2015, pp. 1269–1277.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Lecture Notes in Computer Science*, vol. 3951, pp. 404–417, 2006.

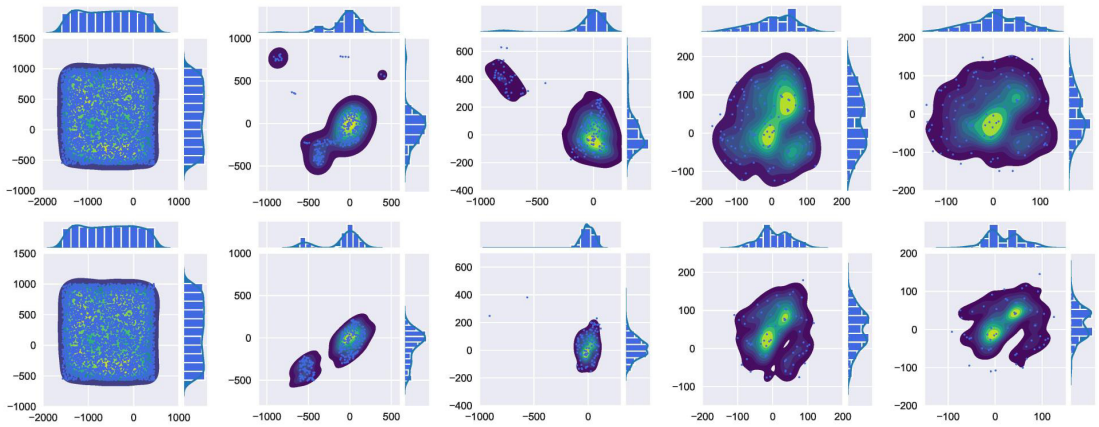

Fig. 7. Comparison of localization methods with different coding strategies on the distribution of major particles. Global localization performances using NetVLAD (upper row) and NetVLAD* (lower row) are shown here. In each subplot, we plot the distributions of particles with large weights and the particle density in the beginning five iterations. In these figures, X-Y positions are relative to the target position.

[4] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[5] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk, "City-scale landmark identification on mobile devices," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 737–744.

[6] I. Cisneros, P. Yin, J. Zhang, H. Choset, and S. Scherer, "Alto: A large-scale dataset for uav visual place recognition and localization," *ArXiv Preprint ArXiv:2207.12317*, 2022. [Online]. Available: https://github.com/MetaSLAM/ALTO

[7] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," in *IEEE International Conference on Robotics and Automation*, 2007, pp. 3921–3926.

[8] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 141–14 152.

[9] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.

[10] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," *arXiv preprint arXiv:2308.00688*, 2023.

[11] A. Khaliq, M. Milford, and S. Garg, "Multires-NetVLAD: Augmenting place recognition training with low-resolution imagery," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3882–3889, 2022.

[12] S. Lee, H. Seong, S. Lee, and E. Kim, "Correlation verification for image retrieval," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5374–5384.

[13] Z. Li, C. D. W. Lee, B. X. L. Tung, Z. Huang, D. Rus, and M. H. Ang, "Hot-NetVLAD: Learning discriminatory key points for visual place recognition," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 974–980, 2023.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[15] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[16] F. Maffra, Z. Chen, and M. Chli, "Viewpoint-tolerant place recognition combining 2d and 3d information for UAV navigation," in *IEEE International Conference on Robotics and Automation*, 2018, pp. 2542–2549.

[17] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source slam system for monocular, stereo, and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[18] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[19] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[20] G. Peng, Y. Huang, H. Li, Z. Wu, and D. Wang, "LSDNet: A lightweight self-attentional distillation network for visual place recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022, pp. 6608–6613.

[21] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[22] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *IEEE International Conference on Computer Vision*, 2011, pp. 2564–2571.

[23] M. Schleiss, F. Rouatbi, and D. Cremers, "VPAIR-Aerial Visual Place Recognition and Localization in Large-scale Outdoor Environments," *arXiv preprint arXiv:2205.11567*, 2022.

[24] C. Toft, W. Maddern, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, T. Pajdla, *et al.*, "Long-term visual localization revisited," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2074–2088, 2020.

[25] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1808–1817.

[26] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 883–890.

[27] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2626–2635.

[28] Y. Xu, J. Huang, J. Wang, Y. Wang, H. Qin, and K. Nan, "ESA-VLAD: A lightweight network based on second-order attention and NetVLAD for loop closure detection," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6545–6552, 2021.

[29] M. Yang, D. He, M. Fan, B. Shi, X. Xue, F. Li, E. Ding, and J. Huang, "Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 772–11 781.

[30] M. Zaffar, S. Ehsan, M. Milford, and K. McDonald-Maier, "CoHOG: A light-weight, compute-efficient, and training-free visual place recognition technique for changing environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1835–1842, 2020.

[31] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, K. Alexis, and K. McDonald-Maier, "Are state-of-the-art visual place recognition techniques any good for aerial robotics?" *ArXiv Preprint ArXiv:1904.07967*, 2019.

[32] J. Zhang, Y. Cao, and Q. Wu, "Vector of locally and adaptively aggregated descriptors for image feature representation," *Pattern Recognition*, vol. 116, p. 107952, 2021.