IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

UltraVPR: Unsupervised Lightweight Rotation-Invariant Aerial Visual Place Recognition

Chao Chen[1*], Chunyu Li[2*], Mengfan He[2], Jun Wang[1], Fei Xing[2] and Ziyang Meng[2]

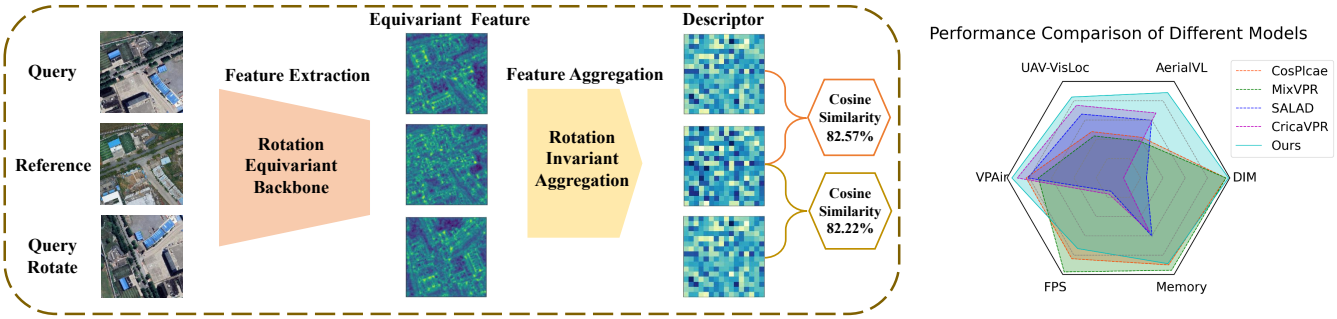

Fig. 1. **The left diagram** outlines the workflow of the proposed UltraVPR model, highlighting its ability to maintain high descriptor consistency under varying rotational viewpoints in aerial VPR. **The right radar chart** provides a comparative analysis of UltraVPR against several state-of-the-art VPR methods. Specifically, the chart displays the Recall@1 performances of UltraVPR and other algorithms on the *VP-Air*, *UAV-VisLoc*, and *AerialVL* datasets. It includes metrics such as frame rate, encoding dimensionality, and memory usage (lower dimensionality and memory consumption yield higher scores).

***Abstract*—Aerial Visual Place Recognition (VPR) is critical for Unmanned Aerial Vehicles (UAVs) localization, especially in environments with unstable or unavailable GPS signals. While neural network-based VPR methods have become mainstream, they face significant challenges on UAV platforms. Traditional CNN-based VPR models are highly sensitive to image rotation, degrading their performance in aerial-domain environments. Meanwhile, Transformer-based models have high computational complexity, making them less suitable for resource-constrained UAVs. In this letter, we propose a lightweight, rotation-invariant aerial VPR method. Our approach combines a rotation-equivariant backbone network with a rotation-invariant aggregation layer to ensure descriptor consistency across different orientations. Additionally, we propose an unsupervised training strategy that constructs higher-dimensional descriptors to optimize the model, while maintaining the lower descriptor dimensionality during application. Experimental results show that our method outperforms state-of-the-art methods across multiple aerial VPR datasets. The code will be released at https://github.com/cbbhuxx/UltraVPR.**

***Index Terms*—Feature Extraction, Image Retrieval, Visual Place Recognition, Deep Learning, Rotation Invariant.**

I. Introduction

VISUAL Place Recognition (VPR) [1] is a critical technology that employs environmental databases, such as image collections, for location identification, showcasing its significant potential in vision-based localization systems [2]. Recently, increasing attention has been directed towards VPR applications on unmanned aerial vehicle (UAV) platforms [3]. Advances in remote sensing technology have enabled the effective matching of images captured by UAVs with high-resolution satellite imagery, thereby achieving accurate positioning even in the absence of reliable GPS signals [4].

With the rapid development of deep learning technology, the introduction of neural networks has greatly advanced the VPR field [5]. Current VPR algorithms perform effectively on ground vehicles and portable devices, addressing challenges such as illumination changes and weather variations [6]. However, when applied to the UAV platform, VPR algorithms must also overcome unique viewpoint variations, particularly the image rotation problem in yaw [7]. The image rotation of UAV platforms differs from the horizontal rotation (left-to-right) observed in ground vehicle VPR. It involves in-plane image rotation, which significantly complicates the matching between query images and database images.

Existing VPR algorithms commonly fail to adequately address these UAV-specific challenges. For CNN-based VPR algorithms, convolutional layers extract features through local receptive fields. When input images undergo in-plane rotation, changes in the position and orientation of the receptive fields lead to inconsistent features. This renders the models insufficiently robust to rotations [8]. While Transformer-based VPR algorithms exhibit superior rotational equivariance than CNN-based algorithms after rotation-adaptive training, their high computational demands pose a significant challenge for resource-constrained UAV platforms.

Manuscript received: March, 6, 2025; accepted July, 1, 2025.

This letter was recommended for publication by Editor G. Loianno upon evaluation of the Associate Editor and Reviewers' comments. This work was supported in part by National Natural Science Foundation of China under Grant 62303040, and in part by Beijing Natural Science Foundation L233029. (*Corresponding author: Jun Wang.*)

[1]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).

[2]Chunyu Li, Mengfan He, Fei Xing 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).

*Chao Chen and Chunyu Li contribute equally to the article.

This letter has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

In this letter, we propose a rotation-invariant VPR model specifically designed for UAV platforms. To the best of our knowledge, this is the first work to address the rotation invariance of the aerial-VPR problem using a CNN model. A rotation-equivariant backbone network (E2ResNet) is integrated into the proposed model to extract features from the rotated input images. Subsequently, the features are aggregated into rotation-invariant descriptors through aggregation layers that do not rely on the spatial relationships of feature points. This novel design guarantees feature consistency across different rotation angles, thus enhancing the robustness and accuracy of VPR systems in UAV applications. Additionally, to further improve the representation capability of the descriptors, we introduce a model enhancement training strategy that uses the cluster centers of similar features as references, enabling more fine-grained descriptor representation for backpropagation optimization.

The contributions of our work are summarized as follows:

1) We propose a lightweight rotation-invariant VPR model that leverages rotation-equivariant networks for feature extraction and generates rotation-invariant descriptors, ensuring robust performance on UAV platforms.
2) An unsupervised training approach is designed that optimizes the model through feature clustering without increasing additional training parameters and descriptor dimensionality.
3) Extensive evaluations on multiple aerial VPR datasets demonstrate that the proposed method outperforms state-of-the-art (SOTA) algorithms. The strengths of the proposed method are shown in Fig. 1.

II. Related work

VPR research has primarily focused on general scenes, where significant progress has been made [5, 9–13]. For instance, MixVPR [14] enhances performance by integrating deep features with multi-layer perceptrons (MLPs) [15], while CricaVPR [16] improves the robustness of global image representations through cross-image self-attention mechanisms and multi-scale local feature fusion.

However, when VPR techniques are applied to aerial platforms, unique challenges arise due to the top-down perspective, severe illumination variations, and complex background interference [17]. To address these challenges, researchers have proposed several targeted solutions. MuSe-Net [18] introduces a style-adaptive mechanism to enhance robustness under varying imaging conditions. FSRA [19] leverages the global modeling capability of Transformers along with dynamic region alignment to handle positional shifts and scale uncertainties. Additionally, CV-Cities [20] and Sample4Geo [21] demonstrate strong advantages in cross-view matching tasks.

While the aforementioned methods have demonstrated significant advancements in aerial VPR tasks, UAV platforms are typically resource-constrained, necessitating models with low computational cost and memory usage. Given the need for large-scale databases in aerial VPR, high-dimensional descriptors can lead to excessive memory consumption on onboard computers, thus highlighting the importance of low-dimensional descriptors. To address this issue, researchers have explored various strategies for model compression and efficiency improvement. For instance, some studies have employed metric learning paradigms to design lightweight networks that compress high-dimensional image descriptors into low-dimensional binary hash codes, thereby reducing computational complexity and storage demands [22]. On the Other hand, collaborative strategies that employ low-bit quantization, lightweight architectures, and mixed-precision optimization have been proposed to achieve efficient VPR deployment on embedded devices [23]. Although these strategies effectively reduce deployment costs, they typically impose fixed low-dimensionality or compression constraints throughout training and inference. This approach may limit the model's potential to learn discriminative features during optimization.

In addition, image rotation consistency is also a core challenge in aerial VPR. The ultimate goal is to ensure that the final place descriptors remain invariant to rotational transformations of the input images. To address this problem, pioneering studies from autonomous driving scenarios provide valuable insights. RINet [24] introduces a ring-wise descriptor structure, and BEVPlace [25] achieves feature rotation equivariance by combining bird's-eye-view (BEV) representation with group convolution operations. In the domain of image matching, SE2-LoFTR [26] replaces LoFTR's CNN backbone with an E2CNN architecture, and Steerers [27] encodes image rotations into the descriptor space via linear transformations based on group representation theory. Both approaches significantly enhance keypoint matching stability under image rotation. For aerial VPR, the study in [28] proposes an online rotation adjustment strategy for aerial imagery, which partially mitigates the impact of rotational discrepancies on feature matching. However, image rotation consistency has not been well addressed in aerial VPR tasks.

In summary, although existing studies have achieved significant progress in both general and aerial VPR scenarios, there remains a gap in simultaneously addressing rotation invariance and resource constraints. To address this dual challenge, this letter proposes a lightweight, rotation-invariant aerial VPR method specifically tailored for UAV platforms. The proposed method introduces a global descriptor architecture that includes a rotation-equivariant backbone network and a rotation-invariant aggregation layer, endowing the model with robustness against image rotation. Additionally, it incorporates an unsupervised low-dimensional descriptor enhancement training module, which significantly boosts representational capacity without increasing model parameters. This makes the proposed method well-suited for deployment on resource-constrained UAV platforms.

III. Problem Description and Methodology

A. Problem Description

The VPR task can be formulated as retrieving frames from a pre-constructed map database that match the currently captured image, thereby determining the exact location of the image. In aerial VPR tasks, this map database is typically composed of high-resolution satellite maps of the flight area, segmented into several remote-sensing satellite

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

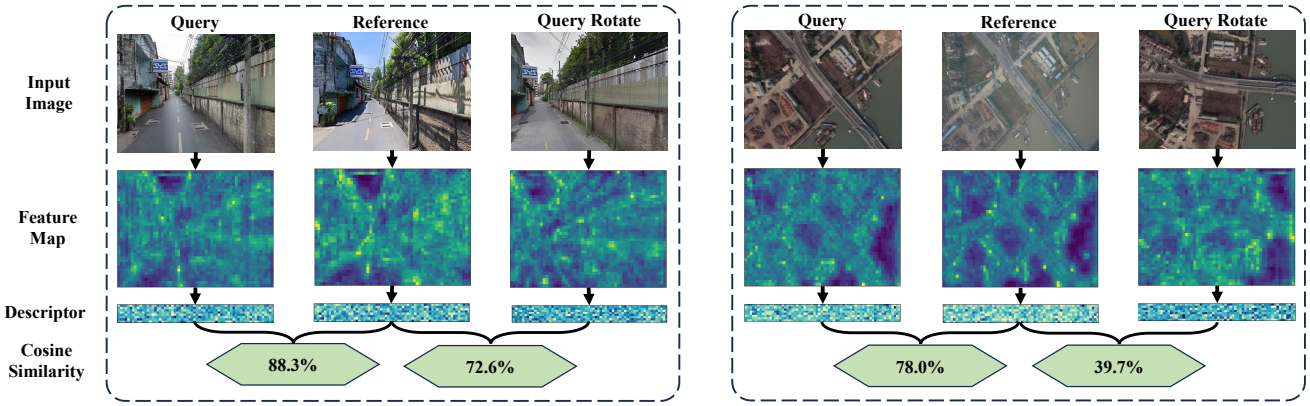

Fig. 2. **The impact of yaw rotation on descriptors varies significantly between ground and UAV platforms.** The left side illustrates the performance of the ground platform, while the right side illustrates the performance of the UAV platform. All results are generated using CNN-based network models (MixVPR). For ground-based VPR, descriptors are minimally affected by horizontal rotations caused by yaw. In contrast, for UAV-based VPR, in-plane rotations resulting from yaw have a substantial impact. These rotations cause significant distortions in the feature maps, leading to reduced descriptor similarity.

tiles. The location of the UAV is estimated by comparing the descriptors of the captured image with those of the remote-sensing satellite tiles. Current methods for generating image descriptors generally involve two main steps: feature extraction and feature encoding. Ensuring rotational equivariance during the feature extraction stage is crucial, as it facilitates the generation of rotation-invariant descriptors. This is particularly important for UAV platforms, which frequently experience viewpoint rotation.

As shown in Fig. 2, the rotational viewpoint changes on UAV platforms differ significantly from horizontal rotations on ground platforms. On ground platforms, descriptors are minimally affected by yaw changes. In contrast, on UAV platforms, these changes involve in-plane image rotation, which can significantly impact feature extraction. When the input and reference images are aligned in orientation, the extracted feature maps perform well, resulting in high descriptor similarity. However, a rotational viewpoint difference between the input and reference images causes severe distortions in the feature maps, leading to a significant reduction in descriptor similarity. This discrepancy can negatively affect image retrieval and localization, potentially causing failures under extreme conditions. Therefore, addressing the impact of rotational viewpoint changes on image descriptors is crucial for enhancing the accuracy and robustness of aerial VPR systems.

B. Methodology

In this study, we propose an unsupervised rotation-invariant aerial VPR model. The model consists of two key components: first, a rotation-equivariant backbone network is employed to extract rotation-equivariant features, which are then aggregated into a rotation-invariant descriptor. Second, an unsupervised enhancement training method is introduced, leveraging a higher-dimensional descriptor module for backpropagation, thereby further improving the discriminative capability of the original descriptor.

1) Rotation-Invariant VPR Model: The rotation-invariant VPR model primarily consists of a rotation-equivariant backbone network and a rotation-invariant aggregation module. The specific process for generating rotation-invariant descriptors is shown in Fig. 3.

Rotation-equivariant backbone. In this study, a rotation-equivariant backbone network, E2ResNet, is proposed. Following the same idea as previous work [26, 29], E2ResNet is also designed using E2CNN [30], replacing the traditional ResNet [31] architecture in CNNs. The convolutional filters in E2ResNet can process multiple orientations of the input image, enabling the extraction of multi-directional feature information. Specifically, the intermediate feature tensor generated by E2ResNet includes an additional orientation dimension, allowing the network to capture features across different orientations. The dimensions of these intermediate feature tensors are typically represented as $(N * D, H, W)$, where D denotes feature depth (the number of features captured at each spatial location), N represents the number of orientation channels. The value of $N * D$ is a fixed parameter determined by the basic ResNet structure, meaning that as N increases, D correspondingly decreases. However, since the feature depth D largely determines the representational capacity of feature points, it is necessary to balance the value between the number of discrete rotation angles N and the feature depth D. Based on empirical results, setting N to 8 achieves the best performance in the proposed E2ResNet. When using the ResNet50 structure, D should be set to 256. H and W denote the height and width of the feature maps, respectively. However, when the input image is rotated, the orientation channels within the intermediate feature tensor change accordingly, leading to a loss of rotation-equivariant in the intermediate feature representation.

To address this issue and maintain rotation-equivariance, unlike previous work [26, 29], we further incorporate a GroupPooling layer [30]. Given an intermediate feature tensor of dimensions $(N * D, H, W)$, the GroupPooling layer divides the tensor into $D * H * W$ groups, with each group containing N orientation channels. Within each group, max pooling is applied to aggregate the data, therefore preserving feature stability even when the order of orientation channels changes. This process effectively reduces the feature dimensionality, resizing the tensor from $(N * D, H, W)$ to (D, H, W).

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

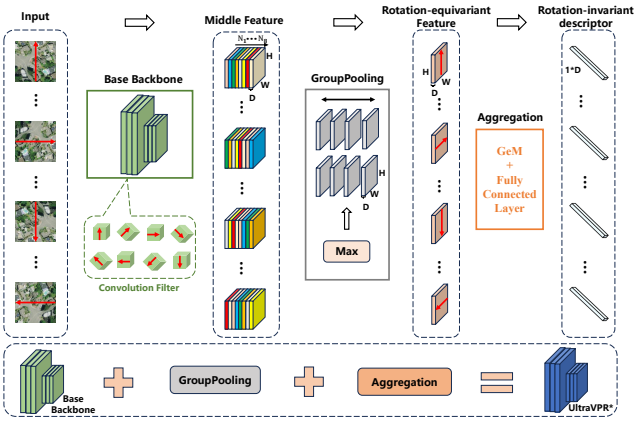

Fig. 3. **The schematic diagram of the proposed rotation-invariant VPR model.** Regardless of the input image's orientation, it is first processed by the E2ResNet backbone to generate cyclic features labeled from N_1 to N_8. Subsequently, the GroupPooling layer is applied to extract rotation-equivariant features. Finally, the rotation-invariant aggregation layer produces a globally rotation-invariant descriptor.

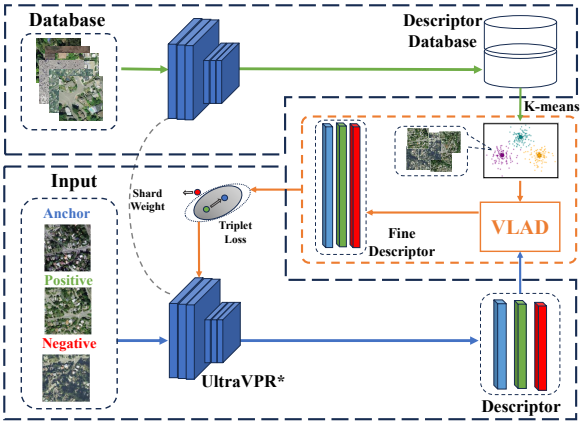

Fig. 4. **The flowchart of the proposed model enhancement strategy.** During the training process, high-dimensional descriptors are generated by using cluster centers formed from the database as reference points. These descriptors are then backpropagated to optimize the weights of the network model.

Building on this, E2ResNet achieves rotation equivariance at N specific rotation angles. Furthermore, by incorporating rotational augmentation as an auxiliary technique during the training phase, E2ResNet can approximate equivariance across arbitrary rotation angles.

Rotation-invariant aggregation. Subsequently, a low-parameter aggregation method to generate rotation-invariant global descriptors is introduced.

The positions of feature points in the rotation-equivariant feature maps change with the rotation of the input image. To extract a rotation-invariant global descriptor from the rotation-equivariant features, we design an aggregation strategy based on the statistical measures. First, GeM is applied over the entire spatial dimension of the feature map. GeM adaptively pools the features by learning the parameter p to select the optimal pooling method. This adaptive pooling effectively captures salient regions and maintains robustness against rotational changes. Then, a fully connected layer is incorporated to refine the feature vector processed by GeM, generating a rotation-invariant descriptor.

2) Model Enhancement: Inspired by VLAD [32] and NetVLAD [5], we propose an unsupervised training method to enhance the representation capability of the VPR model. This enhancement process is applied only during the training phase, where high-dimensional descriptors generated by VLAD are used to optimize the model weights, rather than being used as the final descriptors. On the other hand, during the application phase, only the lower-dimensional descriptors generated by the VPR model are used for retrieval and matching.

The proposed model enhancement method is illustrated in Fig. 4. Firstly, we use the VPR model to generate a D-dimensional feature descriptor for each image in the database.

$$\mathcal{D} = VPR(I), \tag{1}$$

where I represents the input image, and VPR refers to the encoding process of the VPR model. Next, we apply the K-means clustering algorithm to all generated descriptors. To encompass various levels of clustering granularity, we set the number of clusters as a power series of 2: $k = 2^0, 2^1, 2^2, \ldots, 2^m$, where m is an integer that can be set according to specific needs. This results in a total of $K = \sum_{n=0}^{m} 2^n = 2^{m+1} - 1$ cluster centers, denoted as $\mathbf{c}_j \in \mathbb{R}^D$ (j=1,...,K). The clustering centers are updated with each new epoch.

Subsequently, the training images are fed into the VPR model to generate D-dimensional descriptors $\mathbf{d}_t$ (as shown in the bottom part of Fig. 4), and the Euclidean distance to the cluster centers $\mathbf{c}_j$ is computed:

$$\text{dist}(\mathbf{d}_t, \mathbf{c}_j) = \|\mathbf{d}_t - \mathbf{c}_j\|. \tag{2}$$

Then, the inverse of these distances (adding a small constant ϵ to prevent division by zero) is computed and applied to the softmax function for obtaining normalized weights:

$$\text{Inv}(\mathbf{d}_t, \mathbf{c}_j) = \frac{1}{\text{dist}(\mathbf{d}_t, \mathbf{c}_j) + \epsilon}, \tag{3}$$

$$s(\mathbf{d}_t, \mathbf{c}_j) = \frac{\exp(\text{Inv}(\mathbf{d}_t, \mathbf{c}_j))}{\sum_{l=1}^{K} \exp(\text{Inv}(\mathbf{d}_t, \mathbf{c}_l))}. \tag{4}$$

Based on these normalized weights, we construct a higher-dimensional new feature representation $\mathbf{f} \in \mathbb{R}^{K \times D}$ using VLAD [32], which consists of the residuals between the original descriptor and each cluster center, weighted by the softmax-normalized inverse distances:

$$\mathbf{f}^{(j-1)D+i} = s(\mathbf{d}_t, \mathbf{c}_j) \cdot (\mathbf{d}_t^i - \mathbf{c}_j^i), \tag{5}$$

where, $i = 1, \ldots, D$ denotes the dimension index, and $\mathbf{d}_t^i$ and $\mathbf{c}_j^i$ are the i-th elements of the descriptor $\mathbf{d}_t$ and the cluster center $\mathbf{c}_j$, respectively. This weighted residual representation effectively captures the slight differences between similar features, facilitating the optimization of model weights through backpropagation.

To optimize the entire model, we adopt the Triplet Loss function [33], which aims to minimize the distance between positive samples while maximizing the distance between negative samples. Given a triplet $(\mathbf{f}_a, \mathbf{f}_p, \mathbf{f}_n)$, where $\mathbf{f}_a$ is the anchor feature vector, $\mathbf{f}_p$ is the positive sample feature vector, and $\mathbf{f}_n$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

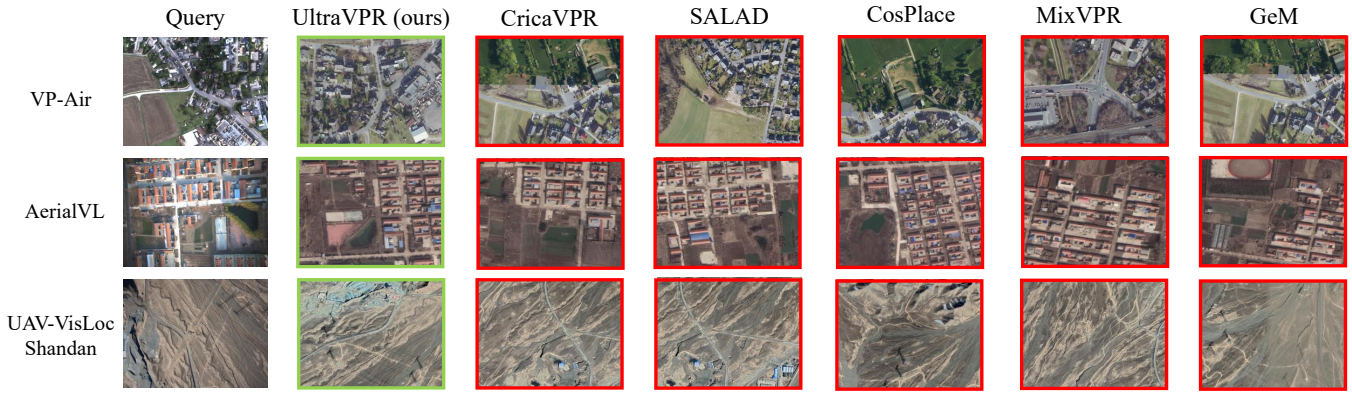

Fig. 5. **Retrieval examples across benchmark datasets for different VPR models.** The first row presents results from the *VP-Air* dataset, the second row from the *AerialVL* dataset, and the third row from Shandan in *UAV-VisLoc* dataset. In these examples, UltraVPR consistently retrieves the correct matches, whereas other methods return incorrect results. Despite other methods retrieve visually similar images from different locations, these results are all false positives.

is the negative sample feature vector, the Triplet Loss function is defined as follows:

$$L(\mathbf{f}_a, \mathbf{f}_p, \mathbf{f}_n) = \max(0, \mathbf{m} + d(\mathbf{f}_a, \mathbf{f}_p) - d(\mathbf{f}_a, \mathbf{f}_n)), \quad (6)$$

where $\mathbf{m}$ is the margin, used to control the required distance between positive and negative samples. In addition, d is a distance metric computed by the Euclidean distance:

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2. \quad (7)$$

IV. Experiments

Extensive experiments are conducted to validate the effectiveness of the proposed UltraVPR model. In particular, the proposed UltraVPR model is compared with SOTA techniques using multiple challenging benchmarks. The proposed backbone network, E2ResNet, is implemented based on ResNet50 and is referred to as E2ResNet50 here, with the number of discrete rotation angles N set to 8. The clustering granularity level m of the enhancement strategy is set to 5. The following content provides detailed information on the datasets, evaluation metrics, performance comparisons, and ablation studies.

Evaluation datasets: To comprehensively evaluate the proposed aerial VPR model, three aerial datasets: *UAV-VisLoc* [34], *VP-Air* [35], and *AerialVL* [36] are utilized in this letter. *UAV-VisLoc* is a large-scale image dataset covering various regions across China. It includes diverse terrain types, flight altitudes, seasonal variations, and detailed metadata. *VP-Air* focuses on high-altitude visual localization and encompasses diverse terrain features and significant lighting variations. *AerialVL* spans an area of approximately 20 square kilometers, featuring a range of terrain types, different flight altitudes and paths, and varying lighting conditions. The database satellite tiles are generated by cutting high-resolution images of the flight area, with a ground coverage similar to that of aerial views. Adjacent tiles are offset by one-third of the tile width and height in horizontal and vertical directions, respectively. Notably, the aerial images and their corresponding satellite tiles exhibit a relative rotation angle in these datasets.

Evaluation metrics: In the experiments, we follow common evaluation procedures and use the Recall@N (R@N) metric to evaluate the model's recognition performance [38–40]. This metric measures the proportion of relevant images among the top N retrieved database images that are within a specific threshold distance from the query image's true location. For the *UAV-VisLoc* and *AerialVL* datasets, we use geographic distance thresholds of 200 meters and 100 meters, respectively. For the *VP-Air* dataset, which consists of continuous frames, a consecutive frame threshold of ±1 frame is set.

Training: We first train the backbone of our model on the ImageNet-1000 classification task [41] to obtain pre-trained weights with feature extraction capabilities. Then, the model is fine-tuned using seasonal remote sensing satellite image pairs (a total of 11k pairs) provided by the *AerialVL* dataset. The fine-tuning is conducted on an NVIDIA GeForce RTX 4090 GPU, with the input image resolution set to 320×320. The Triplet Loss function is used in training, with hard negative mining performed using the method of Radenovic et al. [42]. Each training batch contained 16 image sets, each consisting of an anchor, a positive sample, and 10 negative samples. Anchor are input aerial images. Positive samples are the nearest satellite tiles geographically, and negative samples are those close in feature space but far in location. The model is optimized using the SGD optimizer, with an initial learning rate of 0.0032, which is halved every 20 epochs.

A. Comparisons with SOTA Models

The performance of the proposed UltraVPR model is compared with those of the SOTA VPR models. Including CNN-based generalist models (GeM [9], CosPlace [10], MixVPR [14]) and aerial models (Sample4Geo [21]), as well as Transformer-based generalist models (AnyLoc [12], SALAD [13], CricaVPR [16]) and aerial models (CV-Cities [20] and Game4Loc [37]). Notably, we retrain all models without "†" in Table I using the *AerialVL* dataset and closely reproduce the original implementations. Fig. 5 shows partial qualitative results of the evaluated methods. It indicates that the proposed UltraVPR achieves superior performance compared to SOTA methods.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I
COMPARISON OF MODEL RECALL WITH STATE-OF-THE-ART METHODS ON THE *UAV-VisLoc* DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND THE SECOND-BEST RESULTS ARE UNDERLINED. † THE MODEL AND WEIGHTS ARE PROVIDED BY THE ORIGINAL AUTHORS.

Method	Changjiang-20	Changjiang-23	Taizhou-1	Taizhou-6	Yunnan	Zhuxi	Huzhou-3	Huzhou-6	Huailai	Shandan
	R@1 / R@5	R@1 / R@5	R@1 / R@5	R@1 / R@5	R@1 / R@5	R@1 / R@5	R@1 / R@5	R@1 / R@5	R@1 / R@5	R@1 / R@5
CosPlace† [10]	22.22 / 49.49	24.65 / 45.75	4.17 / 8.59	16.40 / 29.54	54.55 / 82.66	39.71 / 72.94	5.04 / 12.79	4.44 / 13.32	27.08 / 52.08	31.02 / 59.15
MixVPR† [14]	32.03 / 52.67	36.51 / 54.81	0.91 / 4.56	49.46 / 60.30	42.49 / 76.74	49.12 / 75.29	3.00 / 8.91	2.22 / 4.57	31.25 / 56.94	11.19 / 26.10
SALAD† [13]	35.50 / 59.88	35.76 / 55.46	30.99 / 52.08	53.25 / 69.24	50.95 / 84.99	53.53 / 74.71	11.34 / 23.06 /	12.53 / 27.02	50.00 / 77.08	28.64 / 58.64
CricaVPR† [16]	29.00 / 50.07	28.57 / 50.51	7.68 / 18.75	30.35 / 42.68	34.25 / 70.82	27.06 / 63.53	5.04 / 13.57	4.83 / 14.49	31.94 / 72.92	19.66 / 46.44
Game4Loc† [37]	37.09 / 59.88	34.73 / 52.66	36.85 / 55.86	42.14 / 61.52	57.51 / 85.20	59.41 / 85.29	16.18 / 30.62	20.10 / 38.25	40.28 / 78.47	49.83 / 75.76
GeM [9]	27.56 / 49.64	27.26 / 47.25	20.31 / 43.62	25.47 / 48.92	53.28 / 83.30	46.76 / 79.71	14.73 / 29.94	13.97 / 34.33	29.86 / 70.83	47.97 / 68.64
CosPlace [10]	41.70 / 62.05	39.68 / 56.02	35.16 / 57.16	42.01 / 63.28	61.10 / 84.57	64.41 / 80.88	21.71 / 37.60	26.76 / 45.30	43.75 / 78.41	59.32 / 76.27
MixVPR [14]	37.95 / 61.18	36.23 / 51.82	35.41 / 58.33	47.97 / 68.02	61.95 / 86.26	57.94 / 82.06	18.90 / 33.82	25.72 / 47.65	33.33 / 67.36	49.15 / 69.32
Sample4Geo [21]	50.07 / 68.11	51.45 / 65.55	61.07 / 82.03	78.86 / 91.46	68.92 / 90.06	75.88 / 91.76	29.94 / 44.96	42.43 / 61.23	45.83 / 73.61	68.47 / 82.37
AnyLoc [12]	56.57 / **77.20**	48.37 / 68.16	68.49 / 82.94	78.59 / 91.33	69.98 / 93.23	79.41 / **97.06**	28.29 / 47.48	24.93 / 41.38	**66.67 / 88.89**	64.24 / 87.80
SALAD [13]	48.63 / 70.42	45.10 / 62.56	54.69 / 79.43	76.15 / 92.41	69.98 / 91.33	63.53 / 87.06	30.04 / 47.09	42.95 / 61.36	60.42 / 83.33	71.36 / 88.31
CricaVPR [16]	57.00 / **77.20**	51.17 / 66.48	70.05 / 88.02	83.06 / 94.04	71.67 / 92.39	72.65 / 89.71	36.72 / 50.78	52.74 / 71.93	54.86 / 81.25	77.12 / 89.15
CV-Cities [20]	53.82 / 76.62	52.66 / **69.47**	68.36 / 91.28	81.84 / 94.04	75.05 / **93.66**	70.29 / 92.76	34.79 / 49.32	47.91 / 67.49	43.06 / 82.64	78.81 / **90.51**
UltraVPR	**59.31** / 73.02	**53.59** / 64.71	**77.99 / 92.45**	**84.96 / 94.72**	**78.01** / 93.45	**83.24** / 95.29	**44.77 / 56.69**	**60.31 / 74.54**	61.81 / 84.03	**82.71 / 90.51**

TABLE II
COMPREHENSIVE PERFORMANCE COMPARISON WITH CNN-BASED AND TRANSFORMER-BASED VPR MODELS. FPS REPRESENTS THE FRAME RATE FOR PROCESSING A SINGLE IMAGE, DB(M) INDICATES THE DATABASE SIZE, AND MEMORY (MIB) REFERS TO THE MEMORY USAGE.

Method	DIM	FPS	DB	Memory	Recall(@1)
		3060 / NX		3060 / NX	*VP-Air / AerialVL*
GeM [9]	2048	160 / 34.1	61.3	468 / 392	45.57 / 30.71
CosPlace [10]	512	159 / 43.3	15.6	476 / 384	58.02 / 35.22
MixVPR [14]	512	185 / 42.6	15.6	424 / 360	60.38 / 32.92
Sample4Geo [21]	1024	37 / -	30.5	798 / -	64.30 / 35.99
AnyLoc [12]	49152	2 / -	1462.7	4768 / -	48.85 / 40.21
SALAD [13]	8448	26 / 7.8	251.8	760 / 1100	66.96 / 45.87
CricaVPR [16]	10752	32 / 6.8	320.3	802 / 1365	73.58 / 50.38
CV-Cities [20]	4096	24 / 8.7	122.3	808 / 1064	71.18 / 55.85
UltraVPR	256	139 / 14	8.0	488 / 1310	**76.98** / **63.05**

1) Comparison of Model Recall: We conduct a comparative evaluation of model recall across the 10 challenging VPR benchmarks provided by *UAV-VisLoc* dataset.

As listed in Table I, the proposed model significantly outperforms CNN-based generalist and aerial VPR methods. In the Taizhou-6 urban and suburban mixed scene, UltraVPR achieves a 42.95% improvement in R@1 over CosPlace and a 6.1% improvement over Sample4Geo. In the Huailai scene, primarily featuring farmland, UltraVPR's R@1 surpasses CosPlace by 18.06% and Sample4Geo by 15.98%. In the Shandan desert scene, our method outperforms CosPlace by 23.39% and Sample4Geo by 14.24% in R@1.

UltraVPR outperforms both Transformer-based general VPR models and aerial VPR models on multiple datasets. For example, on the Yunnan hilly dataset, UltraVPR achieves a 6.34% improvement in R@1 over CricaVPR and a 2.96% improvement over CV-Cities. On the terraced field scenario in Zhuxi, UltraVPR outperforms CricaVPR by 10.59% and CV-Cities by 12.95% in R@1. On the Huzhou-6 dataset, our method improves R@1 performance by 7.57% compared to CricaVPR and by 12.40% compared to CV-Cities.

The proposed UltraVPR algorithm demonstrates outstanding recall performance across various complex terrains and scenarios, showing significant advantages over CNN-based and Transformer-based generalist and aerial VPR models.

2) Comparison of Model Overall Performance: To further investigate the overall performance of the model, as shown in Table II, we test multiple metrics of the model on both the NVIDIA RTX 3060 and the Jetson Orin NX. The Recall results are obtained from the NVIDIA RTX 3060.

On the NVIDIA RTX 3060, the results show that CNN-based VPR models have advantages in terms of high frame rates and low resource consumption, but they fall short in recall accuracy. In contrast, Transformer-based VPR models exhibit lower frame rates and higher resource usage, yet they achieve better recall accuracy. Our UltraVPR model combines the strengths of both architectures. It not only achieves the lowest feature dimension (256) but also delivers a high frame rate of up to 139 FPS with relatively low resource consumption (488). Moreover, it achieves R@1 scores of 76.98% on VP-Air and 63.05% on AerialVL, demonstrating strong retrieval performance.

On the Jetson Orin NX platform, we conduct inference tests using TensorRT. During testing, the inference models for Sample4Geo and AnyLoc failed to export. Although TensorRT optimization for UltraVPR is limited, UltraVPR still outperforms Transformer-based VPR models in terms of speed, achieving a 60% improvement over CV-Cities. Notably, in terms of database storage, after encoding with the UltraVPR model, only 8 MB of space is required for 7,440 map tiles. It is significantly lower than that of other VPR models.

In summary, based on all test results, the proposed UltraVPR model offers clear advantages for deployment on resource-constrained UAV platforms.

B. Ablation Study

We perform a series of ablation experiments to validate the effectiveness of the proposed components in our method.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE III
ABLATION ON THE BACKBONE NETWORK. E2RESNET50, THE BACKBONE NETWORK DEVELOPED IN THIS LETTER, IS COMBINED WITH DIFFERENT AGGREGATION LAYERS.

Method	DIM	*VP-Air*		*AerialVL*	
		R@1	R@5	R@1	R@5
ResNet50+GeM	2048	45.57	62.27	30.71	50.67
Dinov2-s+GeM	384	62.12	74.98	35.51	55.85
E2ResNet50+GeM	**256**	**71.06**	**80.71**	**47.89**	**65.93**
ResNet50+CosPlace	512	58.02	71.88	35.22	54.32
Dinov2-s+CosPlace	512	65.45	77.86	38.29	56.53
E2ResNet50+CosPlace	512	73.58	82.82	52.59	**65.93**
E2ResNet50+CosPlace	**256**	**73.84**	**83.89**	**52.78**	65.26
ResNet50+MixVPR	512	60.38	74.69	32.92	51.73
Dinov2-s+MixVPR	512	63.97	**77.75**	42.42	60.75
E2ResNet50+MixVPR	512	**66.22**	75.54	**44.15**	57.49
E2ResNet50+MixVPR	**256**	65.93	76.02	43.19	**61.90**

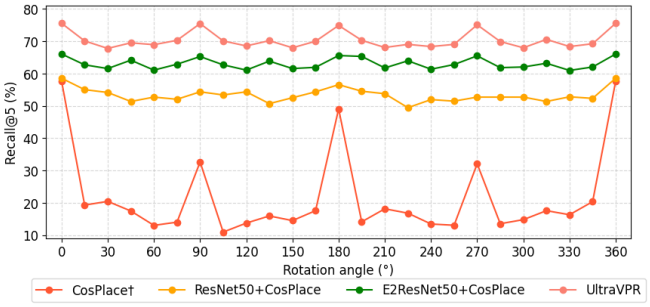

Fig. 6. **Ablation on rotation angles.** The figure shows the impact of different rotation angles on model recall. The † symbol indicates results obtained using the weights provided by the original authors.

1) Ablation on the Backbone Network: Table III lists the ablation study results for the proposed E2ResNet as the backbone network, in combination with different aggregation layers (GeM, CosPlace, and MixVPR).

The VPR model using E2ResNet50 as the backbone achieves high recall while maintaining a low embedding dimension. Specifically, E2ResNet50+GeM achieves R@1 scores of 71.06% and 47.89% on the *VP-Air* and *AerialVL* datasets, respectively, representing improvements of 25.49% and 17.18% over ResNet50+GeM, along with reduced feature dimensions. E2ResNet50+CosPlace improves R@1 performance by 8.39% and 14.49% on the corresponding datasets compared to Dinov2-s+CosPlace. E2ResNet50+MixVPR achieves R@1 scores of 65.93% and 43.19% on the above datasets, outperforming both ResNet50+MixVPR and Dinov2-s+MixVPR.

2) Ablation on Rotation Angles: To thoroughly investigate the sensitivity of the proposed model to rotation angles, we conduct tests on the *AerialVL* dataset by rotating input images to various angles. As shown in Fig. 6, CosPlace†, which lacks rotation-adaptive training, exhibits significant sensitivity to image rotation. Its recall fluctuates widely and drops below 15% at certain angles. In contrast, ResNet50+CosPlace, after undergoing rotation-adaptive training, maintains relatively stable recall despite rotational changes. However, its overall accuracy remains suboptimal. On the other hand, the combination of E2ResNet50 with CosPlace effectively resists the impact of rotation and achieves high recall accuracy while maintaining recall stability.

TABLE IV
ABLATION ON MODEL ENHANCEMENT STRATEGY. C INDICATES THE UTILIZATION OF THE ENHANCEMENT STRATEGY.

Method	DIM	*UAV-VisLoc*(ave)		*AerialVL*	
		R@1	R@5	R@1	R@5
UltraVPR*	256	66.35	80.74	60.08	73.32
UltraVPR*+C(m=4)	256	66.35	81.38	**63.63**	74.86
UltraVPR*+C(m=5)	256	**68.67**	**81.94**	63.05	**75.43**
UltraVPR*+C(m=6)	256	67.03	81.38	62.57	75.05

Additionally, we evaluate the UltraVPR algorithm incorporating all proposed components here. It demonstrates excellent performance in handling image rotation and achieves even higher recall levels. Notably, when images are rotated to angles other than 0°, 90°, 180°, 270°, and 360°, edge information loss occurs, which is one reason for the slight recall fluctuations observed in UltraVPR at these angles.

3) Ablation on Model Enhancement: Table IV presents the ablation results of the proposed model enhancement strategy. The recall on the *UAV-VisLoc* dataset represents the average result across its 10 sub-datasets. UltraVPR* represents the baseline model without any enhancement strategy, while +C denotes the model with the enhancement strategy. The numbers in parentheses (e.g., m=4 or m=5) indicate different granularity levels of the clustering method used. The total number of cluster centers K is given by $K = 2^{m+1} - 1$. When $m = 5$, $K = 63$. The descriptor dimensionality of UltraVPR* is $D = 256$, thus during training, the descriptor dimensionality for UltraVPR*+C (m=5) in backpropagation is $63 \times 256 = 16128$. Notably, this approach introduces no additional training parameters nor does it expand the descriptor dimensionality during the model application.

The results demonstrate that applying the enhancement strategy improves the recall on both the *UAV-VisLoc* and *AerialVL* datasets, with UltraVPR* achieving better performance when using C(m=5) clustering. Specifically, on the *UAV-VisLoc* dataset, the model with C(m=5) clustering achieves a 2.32% improvement in the R@1, while on the *AerialVL* dataset, the R@1 score increases by 2.97%.

V. CONCLUSION

In this letter, we address the challenges of image rotation sensitivity and resource limitations in aerial VPR tasks by proposing UltraVPR, a lightweight and rotation-invariant VPR method. The proposed method integrates a rotation-equivariant backbone network with a rotation-invariant aggregation layer to ensure consistent feature representation across various orientations. Additionally, an unsupervised training strategy is employed, utilizing feature clustering to enhance the representation capability of the descriptors and improve recognition accuracy. Experimental results demonstrate that the proposed method achieves superior performance across multiple aerial VPR datasets, significantly improving resource efficiency and making it suitable for deployment on resource-constrained UAV platforms.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

REFERENCES

[1] 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, 2016.

[2] 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.

[3] Y. Lu, Z. Xue, G.-S. Xia, and L. Zhang, "A Survey on Vision-Based UAV Navigation," *Geo-spatial Information Science*, vol. 21, no. 1, pp. 21–32, 2018.

[4] L. Wu and Y. Hu, "Vision-Aided Navigation for Aircrafts Based on Road Junction Detection," in *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems*, vol. 4, 2009, pp. 164–169.

[5] 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 (CVPR)*, 2016, pp. 5297–5307.

[6] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "EigenPlaces: Training Viewpoint Robust Models for Visual Place Recognition," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11 046–11 056.

[7] P. Yin, I. Cisneros, S. Zhao, J. Zhang, H. Choset, and S. Scherer, "iSimLoc: Visual Global Localization for Previously Unseen Environments With Simulated Images," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 1893–1909, 2023.

[8] T. S. Cohen and M. Welling, "Group Equivariant Convolutional Networks," in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016, pp. 2990–2999.

[9] 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, 2019.

[10] G. Berton, C. Masone, and B. Caputo, "Rethinking Visual Geo-localization for Large-Scale Applications," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4868–4878.

[11] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "TransVPR: Transformer-Based Place Recognition with Multi-Level Attention Aggregation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 638–13 647.

[12] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "AnyLoc: Towards Universal Visual Place Recognition," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1286–1293, 2024.

[13] S. Izquierdo and J. Civera, "Optimal Transport Aggregation for Visual Place Recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17 658–17 668.

[14] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "MixVPR: Feature Mixing for Visual Place Recognition," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 2998–3007.

[15] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, and M. Lucic, "MLP-Mixer: An All-MLP Architecture for Vision," in *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021, pp. 24 261–24 272.

[16] F. Lu, X. Lan, L. Zhang, D. Jiang, Y. Wang, and C. Yuan, "CricaVPR: Cross-Image Correlation-Aware Representation Learning for Visual Place Recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16 772–16 782.

[17] I. Moskalenko, A. Kornilova, and G. Ferrer, "Visual Place Recognition for Aerial Imagery: A Survey," *Robotics and Autonomous Systems*, vol. 183, p. 104837, 2025.

[18] T. Wang, Z. Zheng, Y. Sun, C. Yan, Y. Yang, and T.-S. Chua, "Multiple-environment Self-adaptive Network for Aerial-view Geo-localization[J]. Pattern Recognition," *Pattern Recognition*, vol. 152, p. 110363, 2024.

[19] M. Dai, J. Hu, J. Zhuang, and E. Zheng, "A Transformer-Based Feature Segmentation and Region Alignment Method For UAV-View Geo-Localization ," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4376–4389, 2021.

[20] G. Huang, Y. Zhou, L. Zhao, and W. Gan, "CV-Cities: Advancing Cross-View Geo-Localization in Global Cities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 1592–1606, 2025.

[21] F. Deuser, K. Habel, and N. Oswald, "Sample4Geo: Hard Negative Sampling For Cross-View Geo-Localisation," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 16 801–16 810.

[22] S. Roy, E. Sangineto, B. Demir, and N. Sebe, "Metric-Learning-Based Deep Hashing Network for Content-Based Retrieval of Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 2, pp. 226–230, 2020.

[23] O. Grainge, M. Milford, I. Bodala, S. D. Ramchurn, and S. Ehsan, "Design Space Exploration of Low-Bit Quantized Neural Networks for Visual Place Recognition," *IEEE Robotics and Automation Letters*, 2024.

[24] L. Li, X. Kong, X. Zhao, T. Huang, W. Li, F. Wen, H. Zhang, and Y. Liu, "RINet: Efficient 3D Lidar-Based Place Recognition Using Rotation Invariant Neural Network," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4321–4328, 2022.

[25] L. Luo, S. Zheng, Y. Li, Y. Fan, B. Yu, S.-Y. Cao, J. Li, and H.-L. Shen, "BEVPlace: Learning LiDAR-based Place Recognition using Bird's Eye View Images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8700–8709.

[26] G. Bökman and F. Kahl, "A Case for Using Rotation Invariant Features in State of the Art Feature Matchers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5110–5119.

[27] G. Bökman, J. Edstedt, M. Felsberg, and F. Kahl, "Steerers: A Framework for Rotation Equivariant Keypoint Descriptors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4885–4895.

[28] M. Bianchi and T. D. Barfoot, "UAV Localization Using Autoencoded Satellite Images," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1761–1768, 2021.

[29] J. Han, J. Ding, N. Xue, and G.-S. Xia, "ReDet: A Rotation-equivariant Detector for Aerial Object Detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2785–2794.

[30] M. Weiler and G. Cesa, "General E(2)-Equivariant Steerable CNNs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 14 334–14 345.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[32] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating Local Descriptors into A Compact Image Representation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.

[33] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[34] W. Xu, Y. Yao, J. Cao, Z. Wei, C. Liu, J. Wang, and M. Peng, "UAV-VisLoc: A Large-scale Dataset for UAV Visual Localization," *arXiv preprint arXiv:2405.11936*, 2024.

[35] M. Schleiss, F. Rouatbi, and D. Cremers, "VPAIR–Aerial Visual Place Recognition and Localization in Large-Scale Outdoor Environments," *ICRA 2022 Aerial Robotics Workshop arXiv:2205.11567*, 2022.

[36] M. He, C. Chen, J. Liu, C. Li, X. Lyu, G. Huang, and Z. Meng, "AerialVL: A Dataset, Baseline and Algorithm Framework for Aerial-Based Visual Localization With Reference Map," *IEEE Robotics and Automation Letters*, vol. 9, no. 10, pp. 8210–8217, 2024.

[37] Y. Ji, B. He, Z. Tan, and L. Wu, "Game4Loc: A UAV Geo-Localization Benchmark from Game Data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 3913–3921.

[38] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla, "Visual Place Recognition with Repetitive Structures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2346–2359, 2015.

[39] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 Place Recognition by View Synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 257–271, 2018.

[40] F. Warburg, S. Hauberg, M. López-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary Street-Level Sequences: A Dataset for Lifelong Place Recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2623–2632.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.

[42] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "CNN Image Retrieval Learns From BoW: Unsupervised Fine-Tuning With Hard Examples," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 3–20.