ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XI-2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

Evaluation of Visual Place Recognition Methods for Image Pair Retrieval in 3D Vision and Robotics

Dennis Haitz[1]*, Athradi Shritish Shetty[1], Michael Weinmann[2], Markus Ulrich[1]

[1]Karlsruhe Institute of Technology, Institute of Photogrammetry and Remote Sensing, Karlsruhe, Germany - (dennis.haitz, markus.ulrich)@kit.edu, athradi.shetty@student.kit.edu
[2] Delft University of Technology, Department of Intelligent Systems, The Netherlands - m.weinmann@tudelft.nl

Keywords: Visual Place Recognition, Image Retrieval, RGB-D Registration, Gaussian Splatting Registration, Loop Closure

Abstract

Visual Place Recognition (VPR) is a core component in computer vision, typically formulated as an image retrieval task for localization, mapping, and navigation. In this work, we instead study VPR as an *image pair retrieval* front-end for registration pipelines, where the goal is to find top-matching image pairs between two disjoint image sets for downstream tasks such as scene registration, SLAM, and Structure-from-Motion. We comparatively evaluate state-of-the-art VPR families - NetVLAD-style baselines, classification-based global descriptors (CosPlace, EigenPlaces), feature-mixing (MixVPR), and foundation-model-driven methods (AnyLoc, SALAD, MegaLoc) - on three challenging datasets: object-centric outdoor scenes (Tanks and Temples), indoor RGB-D scans (ScanNet-GS), and autonomous-driving sequences (KITTI). We show that modern global descriptor approaches are increasingly suitable as off-the-shelf image pair retrieval modules in challenging scenarios including perceptual aliasing and incomplete sequences, while exhibiting clear, domain-dependent strengths and weaknesses that are critical when choosing VPR components for robust mapping and registration.

1. Introduction

Visual place recognition (VPR) is a computer vision task that seeks to retrieve similar images of the same scene based on a query image. While image retrieval is a task that solely relies on information derived from the image itself, VPR methods often further utilize the acquisition position, i.e. images are assumed to be georeferenced or geotagged. From a set of search images, the top-k most similar images w.r.t. the query image are to be retrieved, where the k search images ideally show the same scene content as the query image. Challenges in VPR are different viewpoints and viewing angles, illumination and weather changes, or even object manipulations or displacements, e.g. through moving cars at different acquisition times. Additionally, perceptual aliasing plays a role, including recurring texture or patterns through the scene or sequence. Typical use-cases of VPR are large-scale outdoor scenes, e.g. in urban and rural environments.

While the performance of early CNN-based approaches is impacted by their limited receptive fields, the introduction of Vision transformers (ViT) with their attention mechanisms allows to better capture long-range dependencies. A wide range of the state-of-the-art (SOTA) methods are now built on top of DINOv2 (Oquab et al., 2023) as a ViT-based visual feature encoder.

In this contribution, we put our attention on investigating the suitability of different conceptual approaches for finding a fast and reliable association between two sets of RGB images with overlapping scene content, without taking further availability of geometry data into account to avoid further computational overhead. This is of great relevance for the registration of RGB-D images or 3D Gaussian Splatting (3DGS) models (Kerbl et al., 2023) with overlapping scene content, where the image sets or the 3DGS models may not initially be located in the same coordinate system and a georeferencing, e.g. a coarse image acquisition position with heading direction assumed through geographic or UTM coordinates, in the same coordinate system is not available. In order to tackle this task, geometric approaches can be applied, where depth maps can be projected into 3D space based on known camera parameters to get a point cloud. This allows applying known point cloud registration methods to receive an initial coarse registration, in the form of an estimated transformation matrix. While methods based on iterative-closest-point (ICP) (Besl and McKay, 1992) are often prone to local minima, they cannot establish the relative pose if point clouds do not overlap without further computation efforts.

We aim to obtain information on the suitability and performance of SOTA VPR methods to initially find a fast and reliable association between two sets of RGB images (Fig. 1) with overlapping scene content by comparing various widely used datasets, prepared for the task of image-driven scene registration. Regarding runtime efficiency, deep-learning-based methods can process hundreds of images within few seconds with GPU support. For image sequences, the search space can further be reduced if keyframes are taken into account. In the case of RGB-D registration, stereo-camera setups with known intrinsic and extrinsic camera parameters can be considered as potential application scenarios. Regarding 3DGS, cameras with intrinsics and extrinsics are already assumed to be known beforehand, in order to be able to train a 3DGS model. The RGB images for image-driven registration can then be rendered through given extrinsics from such models. Besides common indoor and outdoor datasets, we want to evaluate VPR methods on challenging scenes with ambiguous content, e.g. repeating textures or objects. Those scenes are often captured with object-centric trajectories. We prepare scenes of the Tanks-and-Temples (Knapitsch et al., 2017) dataset for this specific case. In addition, we also focus on indoor and outdoor setups, represented by ScanNet-GSReg (Chang et al., 2024) and KITTI-

* Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-647-2026 | © Author(s) 2026. CC BY 4.0 License.

Figure 1. Example images from sets A (left) and B (right), taken from the Caterpillar scene of the Tanks and Temples dataset (Knapitsch et al., 2017). Green framed views depict potential matches in the overlap area, implied by the rear of the wheel loader. A particular challenge for VPR models in such scenes is perceptual aliasing, e.g. indicated by the wheels, which are visible on both sides and in addition appear in similar relative positions.

odometry (Geiger et al., 2012) , respectively. Our evaluation is therefore set apart from the usual VPR training and test case, which includes large-scale data. Only recently, MegaLoc (Berton and Masone, 2025) took ScanNet as an indoor dataset into account, which is the superset of the ScanNet-GSReg dataset.

Our contributions are the following:

- An evaluation pipeline with scale-independent matching criteria for photogrammetric reconstructions, targeted at typical 3D vision and robotics applications
- Metrics that elevate standard-practice in VPR towards image pair matching
- Providing prepared subsets of well-known datasets (KITTI, Tanks and Temples) for the application of image-driven scene registration.

2. Related Work

In the following, we provide a brief survey of the developments regarding VPR.

Hand-crafted feature approaches. Early developments from the pre-deep-learning era focused on the representation of scene characteristics shown in images in terms of hand-crafted features. Representations based on *global features* such as GIST descriptors (Oliva and Torralba, 2001) and their respective variants (Murillo et al., 2013) aggregate the gradient characteristics of an image in a single vector representation to capture the holistic scene structure. In contrast, other representations leverage sets of distinctive *local features*, such as SIFT, SURF, and ORB descriptors that are invariant to scale and rotation. The conversion of distributions of local descriptors into compact global descriptors, can be achieved based on computing bag-of-words representations (Sivic and Zisserman, 2003; Csurka et al., 2004; Nistér and Stewénius, 2006) or vectors of locally aggregated descriptors (VLAD) (Jégou et al., 2010; Arandjelovic and Zisserman, 2013) that both rely on the initial generation of codebook descriptors (e.g., based on K-means or mean-shift clustering) and a subsequent assignment of the individual descriptors to their closest codebook entry (i.e., word) to get a histogram-like representation or the aggregation of the offset vectors of the descriptors assigned to a particular codebook entry, respectively. Furthermore, sequence-based approaches (Milford and Wyeth, 2012; Bampis et al., 2016) leverage temporal image matching for place recognition, i.e. the matching of sequences of images rather than individual frames, thereby improving robustness to appearance changes. While computationally efficient, these methods often fail under severe illumination or seasonal variations as well as significant changes of view conditions.

Deep-feature approaches. With the advent of deep learning, representations via CNN features (e.g., obtained from networks trained on large datasets like ImageNet) were demonstrated to outperform handcrafted descriptors for landmark retrieval (Razavian et al., 2014; Babenko et al., 2014). Inspired by VLAD-based image representations based on hand-crafted local features, NetVLAD (Arandjelovic et al., 2016) introduced a trainable VLAD layer to aggregate CNN features into a single robust image descriptor, whereas others introduced intermediate layer pooling (Chen et al., 2017) to exploit mid-level features to capture both semantic and structural cues. Focusing on better pooling functions, GeM pooling (Generalized-Mean pooling) was introduced to unify average and max pooling and improve compact global descriptors (Radenović et al., 2019). Whereas global descriptors are efficient but may miss fine geometric cues and local features are accurate but accompanied by higher computational demands, several works (e.g., (Cao et al., 2020; Yang et al., 2021)) focused on combining the advantages of both of these approaches within hybrid architectures that unify local and global representations.

Increased robustness to viewpoint variations and cross–domain conditions. To provide additional robustness for place recognition under varying viewpoints, recent developments particularly focused on multi-view training, patch-level descriptors, multi-scale fusion, and transformer backbones that leverage attention for region selection and robustness. Examples include the training of lightweight CNNs with varied viewpoints (Khaliq et al., 2020) and the training on multiple views of the same scene with a per-place clustering of images from diverse viewpoints into the same label to enforce the network to learn descriptors invariant to perspective changes (Berton et al., 2023). Others leveraged local feature matching at retrieval time, e.g., in terms of performing dense correspondence matching after initial retrieval to verify matches across wide baselines (Berton et al., 2021).Furthermore, Patch-NetVLAD (Hausler et al., 2021) computes patch descriptors from NetVLAD residuals and fuses local and global descriptors at multiple scales to gain robustness against viewpoint changes. Further work includes the use of context-flexible attention models to focus on stable landmarks for long-term place recogni-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-647-2026 | © Author(s) 2026. CC BY 4.0 License.

tion (Chen et al., 2018; Liu et al., 2019) as well as the use of multi-level attention to highlight task-relevant regions and combine global and key-patch descriptors (Wang et al., 2022). An alternative strategy to address viewpoint gaps consists in reformulating the problem as an overlap prediction task, where spatial verification techniques are used to predict the overlap or relative pose between images to determine if they depict the same place (Wei et al., 2025).

To address scalability issues of contrastive pairwise training and training based on triplet losses and hard negative mining for large training datasets, CosPlace (Berton et al., 2022) partitions the scene into many discrete ”place classes” and trains a classifier to predict the place label from an image, leading to state-of-the-art results with significantly more compact descriptors and greatly reduced memory usage. Building on that idea, MixVPR (Ali-Bey et al., 2023) used feature mixing as a form of data augmentation for VPR to further improve the generalization capabilities.

Another recent trend is the exploitation of foundation models for VPR, e.g. in terms of using a large pre-trained ViT (such as DINOv2) as a frozen feature extractor (Keetha et al., 2023; Izquierdo and Civera, 2024) or the fine-tuning of such foundation features for VPR (Izquierdo and Civera, 2024). To improve the aggregation, SALAD (Izquierdo and Civera, 2024) reformulates NetVLAD’s soft assignment as an optimal transport problem, introducing a Sinkhorn-based clustering (with a “dustbin” to drop non-informative features) to produce a more discriminative global descriptor.

Further improvements were achieved by cross-domain models that are applicable across diverse domains such as indoor and outdoor scenarios, day/night conditions, or seasonal changes. Examples include the training of networks on large datasets of time-varying outdoor webcam and street-view data, leading to improved robustness to illumination and seasonal changes (Chen et al., 2017; Hausler et al., 2021; Wang et al., 2022), as well as the use of unsupervised fine-tuning strategies and dataset-wise automatic supervision using 3D reconstructions or SfM to mine positives/negatives (Radenović et al., 2019) that improved retrieval performance without manual annotation. More recently, the training on a mix of datasets has also been followed with recent, matured models (Berton and Masone, 2025) that leverage previous insights on data augmentation, multi-domain training, and large-batch learning to handle different conditions with a single model, leading to robust performance and generalization for across autonomous driving scenes (KITTI), indoor scans (ScanNet), and object-centric datasets (e.g., Tanks and Temples).

3. Methodology

This section lays out the method for evaluating SOTA VPR methods as detectors for image pairs that show a common scene area. We first introduce common VPR methods, their evaluation procedure and lastly, present our evaluation setup for image pair retrieval using VPR. Additionally, we reduce the problem of image pair retrieval to two image sets without loss of generality, as it can certainly be the case to find image pairs in multiple image sets, showing the same scene. Therefore, our setup can be considered as a subtask of a multi-dataset problem.

3.1 Preliminaries for VPR

Neural-network-based VPR methods are usually formulated as image retrieval problems, often in conjunction with runtime efficiency, included in the broad VPR literature and usually denoted as *extraction time* (Berton et al., 2022, 2023; Ali-Bey et al., 2023). In real-time scenarios like (online) SLAM, where VPR methods potentially can be used for loop-closure or relocalization, this is of high importance. Thus, it is necessary to utilize an image representation that yields high computational efficiency, e.g. realized through vectors. Whereas the matching of vectors is a well-researched topic in the field of information retrieval, many VPR methods do not aim at fast vector matching but rather focus on obtaining robust representational vectors that preserve high matchability in challenging scenarios. These scenarios include perceptual aliasing, large viewpoint distances, and large view-direction angles. In the following, we discuss core aspects of VPR methods in the three categories of *backbone and aggregation*, denoting general architecture, *loss functions and training* and *reranking*, which is usually the second part of two-stage approaches.

Backbone and Aggregation. Convolutional Neural Networks (CNNs) parameterized by their weights θ provide the possibility to extract local feature characteristics and derive a vector representation of an image $f_\theta : I \mapsto v$. Early deep-learning-based methods like NetVLAD (Arandjelovic et al., 2016) approached this by training a CNN to produce image representations v that are directly targeted at VPR by minimizing the Euclidean distance between the representation of two images. The backbones of such methods were often pretrained ResNet models of different depth, usually depths of 18, 50, or 101. Beyond these options, VGG16 (Simonyan and Zisserman, 2015) can also be utilized in CosPlace (Berton et al., 2022) and EigenPlaces (Berton et al., 2023) as the underlying architecture. Instead, more recent methods like SALAD (Izquierdo and Civera, 2024), MegaLoc (Berton and Masone, 2025) or AnyLoc (Keetha et al., 2023) are built on ViTs. Especially DINOv2 (Oquab et al., 2023), which is pretrained on a large dataset of 142 million images and aims at providing meaningful domain-independent visual features, is often utilized as a feature extraction backbone.

Attached to the backbone is an aggregation strategy to obtain the global image representation or descriptor v from local features, obtained from, for example, feature maps in CNNs or patch embeddings in ViTs. Aggregations consist of an often small amount of additional network layers, like fully-connected, convolution, or activation layers. Fully-connected and convolution layers also imply that a training is necessary if they are included in the aggregation. The NetVLAD layer Arandjelovic et al. (2016) has been widely adopted for aggregation. It learns the position of a pre-defined number of cluster centers at training time. The descriptor is then calculated from the residuals of the output of the backbone and the cluster centers, weighted by a soft-assignment to further refine the feature-assignment to a cluster-induced visual context. Applying a Generalized Mean Pooling (GeM) (Radenović et al., 2019) in addition to fully-connected and L_2-normalization layers is a further strategy for aggregation. GeM pools its input depending on a learnable parameter, which determines an interpolation between average and maximum pooling.

Loss functions. Contrastive and triplet losses are common for a lot of VPR approaches. The Euclidean distance of the descriptors of positive and negative samples are usually maximized through these losses. For triplet losses, an additional anchor is added as a third loss input. Descriptors as representations of images are defined as positives, negatives, or anchors either through labeling, based on georeferencing with distance and angular thresholds, or images of the same place at different

times, inducing different visual cues. On the other hand, methods like CosPlace (Berton et al., 2022) and EigenPlaces (Berton et al., 2023) formulate their training procedure as a classification problem, utilizing a cross-entropy-based loss. Classes represent grid cells including places, where a place is represented through a point in a coordinate system. In case of CosPlace, this place stems from georeferencing and is further assigned to the grid cell, defined across all places at a certain resolution. To mitigate place ambiguities induced by similar places with vastly differing view directions per cell, each cell is further split into multiple classes, representing heading bins within a cell.

Reranking. Reranking is often added as a second stage within a VPR pipeline and has the objective of further refining the top results. PatchNetVLAD (Hausler et al., 2021) is a method that utilizes local cues from the descriptor building process to add a second stage. Because the feature vectors per position are downsampled and condensed representations of images, these vectors represent actual patches with additional information about their neighboring patches through convolution operations. This patch-based approach is also utilized within ViT pipelines, because a ViT creates patch embeddings, which consequently can be reused for reranking on a patch level (Zhang et al., 2023). However, methods like SALAD (Izquierdo and Civera, 2024) or MixVPR (Ali-Bey et al., 2023) show and explicitly state that modern global-descriptor-only approaches outperform reranking, which often also add a significant computational overhead that is not feasible for real-time applications. The use of only global descriptors is further underlined by Shao et al. (2023), who also argue that a lightweight refinement of global descriptors can be sufficient for superior results regarding local reranking.

3.2 Preliminaries for VPR Evaluation

The common evaluation procedure in the VPR literature involves comparing a retrieved subset (retrieval set) of top-k images from a search set, typically provided by an image database, to a query image. To classify an image in the retrieval set as positive w.r.t. the query image, different criteria emerged across different datasets and methods. Because images are often georeferenced in VPR scenarios, a true positive image within the retrieval set is defined by the metric viewpoint position distance from the query image to the retrieval image. This distance is set to 25 meters in evaluation scenarios across all methods for reasons of comparability, e.g. in (Arandjelovic et al., 2016; Ali-Bey et al., 2023; Izquierdo and Civera, 2024; Berton et al., 2023). With the introduction of the Mapillary dataset (Warburg et al., 2020), an additional view direction difference of smaller than 40 degrees was taken into account for the ground truth definition of a match. The Nordland dataset (Sünderhauf et al., 2013) was acquired along a railroad from a train in different seasons, with the camera facing in driving direction. Though in the evaluation in MixVPR, the ground truth for this dataset is for the viewpoints of the images to lie within a distance of 25 meters, Malone et al. (2025) set a number of frames two viewpoints may lie apart. Therefore, the sequential nature of such datasets can also be taken into account.
Retrieval performance is typically evaluated based on the *recall@k* metric (R@k). This metric is calculated as the average of all retrievals (which is the number of all utilized query images), that contain at least one true positive within the retrieval set of size k, according to the aforementioned definitions of ground truth.

3.3 Image Pair Retrieval Evaluation

We aim to perform evaluation on datasets Tanks and Temples (T&T) (Knapitsch et al., 2017) and ScanNet-GSReg (Dai et al., 2017; Chang et al., 2024) based on SfM-reconstructions (Schönberger and Frahm, 2016), which yield camera poses and intrinsics for further tasks. KITTI-odometry (KITTI) (Geiger et al., 2012) as our third dataset is used with its own ground truth camera poses and intrinsics. All three datasets contain either scenes (T&T, ScanNet-GSReg) or sequences (KITTI), which we also denote as scenes throughout this work. A scene is split into subsets A and B with both sets including subsets of images showing the same scene area. We evaluate SOTA methods regarding their ability to match image pairs that cover such areas. The top-k image pair matches are utilized as the retrieval set in order for tasks such as RGB-D registration to yield a sufficient baseline for matching images. Although in all three datasets images were acquired as sequences, we perform a brute-force matching without prior information to simulate image collection scenarios. Such scenarios are typically more challenging due to their inherent lack of image interconnection information and for the retrieval task constitute an upper bound for runtime performance.

Metrics. As described in Sec. 3.2, *Recall@k* is the standard metric in VPR, which we also include in our task of image pair matching, yielding information if there is at least one true positive match within the retrieval set. However, information about the *quantity* of true positives within top-k pairs is relevant especially for image-driven scene registration. This holds true if the number of true positives is too low, so that even robust methods like RANSAC-based transformation estimation might fail in a registration scenario. We therefore employ the *Precision@k* (P@k) metric to gather information about the ratio of true positives to false positives within the retrieval set. As a third metric, we utilize *mean-Average-Precision@k* (mAP@k), which exploits information about the ranking of true positives within a retrieval set. We therefore aim to gather information about how to set k.
All three metrics are averaged across all scenes per dataset. By doing so, we follow the usual practice in VPR evaluation, where R@k is averaged across all queries, i.e. all retrieval sets per dataset. Also, as common in VPR literature, the retrieval set sizes are $k \in \{1, 5, 10\}$.

Matching. For brute-force matching of images $I_i^A \in \mathcal{I}_A$ and $I_j^B \in \mathcal{I}_B$ we first compute a global descriptor of dimension D for every image using the employed VPR method $\psi : I \to \mathbb{R}^D$. We then define the descriptor sets as

$$\mathcal{D}_X = \{\mathbf{d}_k^X = \psi(I_k^X) \mid I_k^X \in \mathcal{I}_X\}, \quad X \in \{A, B\}. \tag{1}$$

In the second step, the top-k descriptor pairs are obtained by computing the cosine similarity across all possible pairs and ranking them in descending order:

$$\mathcal{P}_k = \underset{i,j}{\arg\,\text{top-k}} \frac{(\mathbf{d}_i^A)^\top \mathbf{d}_j^B}{\|\mathbf{d}_i^A\|_2 \, \|\mathbf{d}_j^B\|_2}, \tag{2}$$

where $\mathcal{P}_k$ contains the indices (i, j) of the top-k pairs, sorted by decreasing cosine similarity.

Ground truth definition. Using SfM (Schönberger and Frahm, 2016) camera poses for T&T and SN-GS and ground truth poses from KITTI, the first criterion to be met is the view

direction angle of two images. Based on a task-driven motivation, we set the angular threshold to $\tau_{view} = 75°$. This threshold is higher compared to $40°$ in VPR literature. We argue that specifically registration tasks can still perform sufficiently well with large camera orientation differences, as long as valid keypoint matches can be obtained from image pairs. Let $R_i^A, R_j^B \in SO(3)$ denote the world-to-camera rotation matrices of images A and B, respectively. Their viewing directions in world coordinates are defined as

$$\mathbf{d}_A = (R_i^A)^\top \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}, \qquad \mathbf{d}_B = (R_j^B)^\top \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}. \tag{3}$$

The angle between both viewing directions is then given by

$$\phi_{\text{view}} = \arccos\left(\frac{\mathbf{d}_A^\top \mathbf{d}_B}{\|\mathbf{d}_A\|_2 \, \|\mathbf{d}_B\|_2} \right), \tag{4}$$

and the first criterion is fulfilled if

$$\phi_{\text{view}} \leq \tau_{view}. \tag{5}$$

Opposed to the benchmark dataset used for VPR (Sünderhauf et al., 2013; Torii et al., 2013, 2015; Warburg et al., 2020), our camera poses from SfM reconstructions are only defined up to a scale w.r.t. their position. Therefore, we cannot utilize a reliable distance measure as common in VPR methods and pointed out in Sec. 3.2. However, in computer vision tasks such as the estimation of homography, fundamental or essential matrix, and 3D transformations, it is common to rely on matched keypoints. Therefore, we define a second criterion as follows: We extract and match SIFT keypoints (Lowe, 2004) from image pairs and, given intrinsic camera parameters, estimate the essential matrix E via RANSAC-based 5-point algorithm (Nistér, 2004). Utilizing SVD, E is decomposed into rotation matrix $R_E^{A \to B} \in SO3$ and unit translation vector $||t_i||_2$. For better readability, mapping $A \to B$ is left out in the following. We check if R_E approximates the rotation matrix R_{SfM} of the relative pose of both images using the geodesic distance of both rotation matrices:

$$d_R(R_E, R_{\text{SfM}}) = \arccos\left(\frac{tr\big(R_E R_{\text{SfM}}^\top\big) - 1}{2} \right), \tag{6}$$

and set a threshold of $\tau_{dev} = 10°$ maximum angular deviation, so that

$$\frac{d_R \cdot 180°}{\pi} < \tau_{dev}. \tag{7}$$

The threshold value is set rather conservatively and motivated to account for small inaccuracies caused by local feature matching, RANSAC-based estimation of E, or pose noise. To add further robustness, an inlier threshold of $\tau_{in} = 0.25$ is set for the estimation of E. For the purpose of comparability, this second criterion is also applied for the KITTI scenes, even though the view position is given in absolute units (meters) per image and therefore could be compared analog to VPR evaluation, i.e. through a distance threshold.
We also account for the runtime without applying typical reranking, as this step often requires several seconds and would exceed our goal of fast initial image pair retrieval (Ali-Bey et al., 2023; Izquierdo and Civera, 2024; Hausler et al., 2021).

4. Experiments

In this section, we describe the experimental setup with the datasets, methods, and the used hardware.

4.1 Datasets

In order to evaluate the described methods, we customized T&T (Knapitsch et al., 2017) and KITTI (Geiger et al., 2012) scenes by dividing them into two disjoint subsets A and B with overlapping scene content in both subsets as already pointed out in Sec. 3.3. Both A and B contain between 30 and 60 images. The subsets are divided subsequently or with a gap to create challenging scenarios with wider baselines. From T&T, we utilized the *Caterpillar Barn*, *Train*, *Truck*, *Palace*, *Playground* and *Lighthouse* scenes as the object-centric scenarios. With object centric trajectories, we want to focus on ambiguous texture and repeating objects in disjoint scene regions, which is usually refered to as perceptual aliasing. From the KITTI dataset, we utilized sequences *00*, *02*, *03*, *05*, *06*, *07*, *08*, *09*, and *10*. We left out sequences with a significant amount of moving objects, because we focus on scenes with mostly static content. KITTI represents an outdoor scenario with the focus on larger (sub-)urban environments, similar to large-scale benchmark datasets used for VPR training and evaluation, but with an emphasis on autonomous driving.
For indoor scenarios, we utilize the ScanNet-GSReg (SN-GS) dataset from Chang et al. (2024), which is an already prepared dataset for 3DGS registration, based on the ScanNet dataset (Dai et al., 2017). We utilized scenes *0000_01*, *0009_00*, *0018_00*, *0050_00*, *0100_02*, *0111_01*, *0170_02*, *0218_01*, *0222_01*, *0309_01*, *0328_00*, *0369_01*, *0420_00*, *0455_00*, *0541_00*, *0568_00*, *0588_00*, *0591_01*, *0629_00*, *0630_01*, *0666_00*, *0667_00*, *0682_00*, *0696_00*, *0701_00* and *0703_00* in our experiments. The scenes were chosen randomly, as all scenes are indoor with similar acquisition configurations. It is important to note that MegaLoc (Berton and Masone, 2025), which is used in our evaluation, was trained on the original ScanNet dataset and therefore is *heavily biased* regarding the aforementioned scenes. Nevertheless, we included the results of this method on this dataset for the sake of completeness. Another important aspect of this dataset is that some scenes contain a small number of identical images in A and B. We did not modify these scenes, in order for future research to be applied on the SN-GS dataset.
In total, we use 16 scenes from T&T, 26 scenes from SN-GS, and 26 scenes from KITTI.

4.2 Methods

We compare a wide range of SOTA methods based on the datasets from Sec. 4.1 and metrics from Sec. 3.3. Runtimes are averaged over all scenes per dataset and we additionally report standard deviations. To account for application scenarios, descriptor extraction, matching and ranking are included in runtime observation, with the evaluation pipeline excluded. The methods include PatchNetVLAD (Hausler et al., 2021), CosPlace (Berton et al., 2022), EigenPlaces (Berton et al., 2023), MixVPR (Ali-Bey et al., 2023), AnyLoc (Keetha et al., 2023), SALAD (Izquierdo and Civera, 2024), and MegaLoc (Berton and Masone, 2025). The methods CosPlace, EigenPlaces, SALAD, and MixVPR provide different backbone or descriptor configurations. For our evaluation, different configurations of those methods are used as listed in Table 1. PatchNetVLAD as an older method is included, because it is close to

Method	Backbone	Descriptor size
CosPlace512	ResNet18	512
CosPlace2048	ResNet101	2048
EigenPlaces512	ResNet18	512
EigenPlaces2048	ResNet101	2048
MixVPR512	ResNet50	512
MixVPR4096	ResNet50	4096
SALAD2048	DINOv2-b14	2048
SALAD8192	DINOv2-b14	8192

Table 1. Different backbone and descriptor size configurations used for the methods CosPlace (Berton et al., 2022), EigenPlaces (Berton et al., 2023), MixVPR (Ali-Bey et al., 2023), and SALAD (Izquierdo and Civera, 2024).

the original NetVLAD-based method and further includes local reranking. Regarding AnyLoc, we did not use an initialization with specific cluster centers, as such cluster centers are biased towards certain domains they have been optimized to. AnyLoc, SALAD and MegaLoc utilize ViT backbones (DINOv2), whereas MixVPR, CosPlace, EigenPlaces and PatchNetVLAD utilize CNN backbones (ResNet).

4.3 Hardware

For all methods, the experiments were conducted on a PC with an Nvidia RTX 3090 GPU and an Intel i9 10850 CPU with 32 Gb RAM.

5. Results

Qualitative Results We show qualitative results in Figure 2 for top-5 retrievals of six methods for the T&T Barn scene and top-5 retrievals for two selected methods in Figure 3 for a scene *07_02* of the KITTI dataset (Geiger et al., 2012). Both are selected for certain perceptual aliasing challenges and wide-baseline image matching to show the performance of current SOTA VPR methods regarding image pair retrieval.

Quantitative Results The quantitative results of the image pair retrievals across all datasets are shown in Table 2. Metrics P@k, R@k, and mAP@k are utilized as described in Sec. 3.3 with $k \in \{1, 5, 10\}$ for P@k and R@k. For mAP@k, only $k \in \{5, 10\}$ are used, as mAP@1 yields the same results as R@1.

6. Discussion

The results in Table 2 show different retrieval performances per dataset. Due to the matters regarding the SN-GS as mentioned in Sec. 4.1, the high performances are easily explainable. It is to be expected that identical images are not only found (R@k), but also ranked at the top (mAP@k). Also as mentioned already in Sec. 4.1, MegaLoc (Berton and Masone, 2025) as ViT-based method was trained on ScanNet among others, which probably leads to overly optimistic retrievals across all scenes in the dataset. Cross-domain generalization for this method is therefore better assessed in T&T and KITTI datasets, where it outperforms the other methods in most cases. Other ViT-based methods are SALAD (Izquierdo and Civera, 2024) and AnyLoc (Keetha et al., 2023), whereby AnyLoc performs poorest of this group. As stated in (Izquierdo and Civera, 2024), a fine-tuning of the general purpose ViT-model (DINOv2) can improve performance, which was done for the SALAD method as opposed to AnyLoc. The CNN-based methods mostly perform a bit worse than the ViT methods. EigenPlaces (Berton et al., 2023) as a CNN-based method in both configurations is close to the ViT-based method, and outperforms AnyLoc across all metrics.

Across all methods, there is no significant drop between mAP@5 and mAP@10. This induces, that a small k can be sufficient if computation times for following tasks (e.g., transformation estimation) are critical. For robustness and offline use, a higher k is to be preferred, if both P@k and mAP@k yield sufficient numbers. For P@k, usually 20 % are the lower bound for robust methods like RANSAC to obtain enough inliers. It is also worth noting that the descriptor size and backbone depth in CosPlace (Berton et al., 2022) and EigenPlaces (Berton et al., 2023) only have a small impact. In some metrics (e.g., in T&T), the smaller models with smaller descriptor sizes even perform better in P@k, meaning over all retrieved pairs, the number of true positives is higher. The parameter p of the GeM-pooling that interpolates between average- and maximum-pooling is between 2.6 (CosPlace2048) and 3.0 (Eigenplaces512), the other two being at around 2.9. This pooling parameter has direct impact on the descriptor. However, it is unclear if it correlates to model complexity and descriptor size. For SALAD (Izquierdo and Civera, 2024), small descriptor sizes seem to perform better across almost all metrics and datasets. For MixVPR (Ali-Bey et al., 2023), the opposite is the case. The size of the descriptor on its own is not a clear indicator of retrieval performance. We therefore opt for experimentation and choose the best working configuration for a given task.

To have a more detailed view of the retrieved image pairs, Fig. 2 shows the top-5 retrievals for six methods. This scene was chosen because it yields the challenge of perceptual aliasing. PatchNetVLAD (Hausler et al., 2021) obtains false positives only, even though the local reranking would induce a taming of perceptual aliasing in the form of similar content at different positions, e.g., as is the case for the brown door in the images. Even MegaLoc (Berton and Masone, 2025) includes a false positive in the top-5. SALAD (Izquierdo and Civera, 2024), and MixVPR (Ali-Bey et al., 2023) exceed in this case. MixVPR, however, also performs worst on T&T in general. In other cases of perceptual aliasing, where similar objects also appear at the similar position in the image, MixVPR fails. The domain gap between the training data and the test data might as well be too large for MixVPR to generalize well to such object-centric scenes. Fig.3 shows two methods for KITTI scene with the additional challenge of missing frames, besides the perceptual aliasing. We chose CosPlace (Berton et al., 2022), which completely fails for the top-5, and MegaLoc (Berton and Masone, 2025), which exceeds in this example. It is notable that MegaLoc appears to be exceptionally robust against perceptual aliasing, which is mostly represented by cars and vegetation, whereas for CosPlace (Berton et al., 2022) it is clearly visible that it fails for such aliasing.

Finally, the runtime performance indicates clearly that CNN-based methods outperform ViT-based methods by a large margin. With the exception of PatchNetVLAD (Hausler et al., 2021), all CNN-based methods perform below one second, whereas the ViT-based methods take more than one or multiple seconds on average. This is important for real-time tasks. Both configurations of EigenPlaces (Berton et al., 2023) seem to yield the best balance of runtime and retrieval performance.

Summarized, our experiments suggest the following practical insights:

- **Object-centric scenes (T&T):** Scenes with strong per-

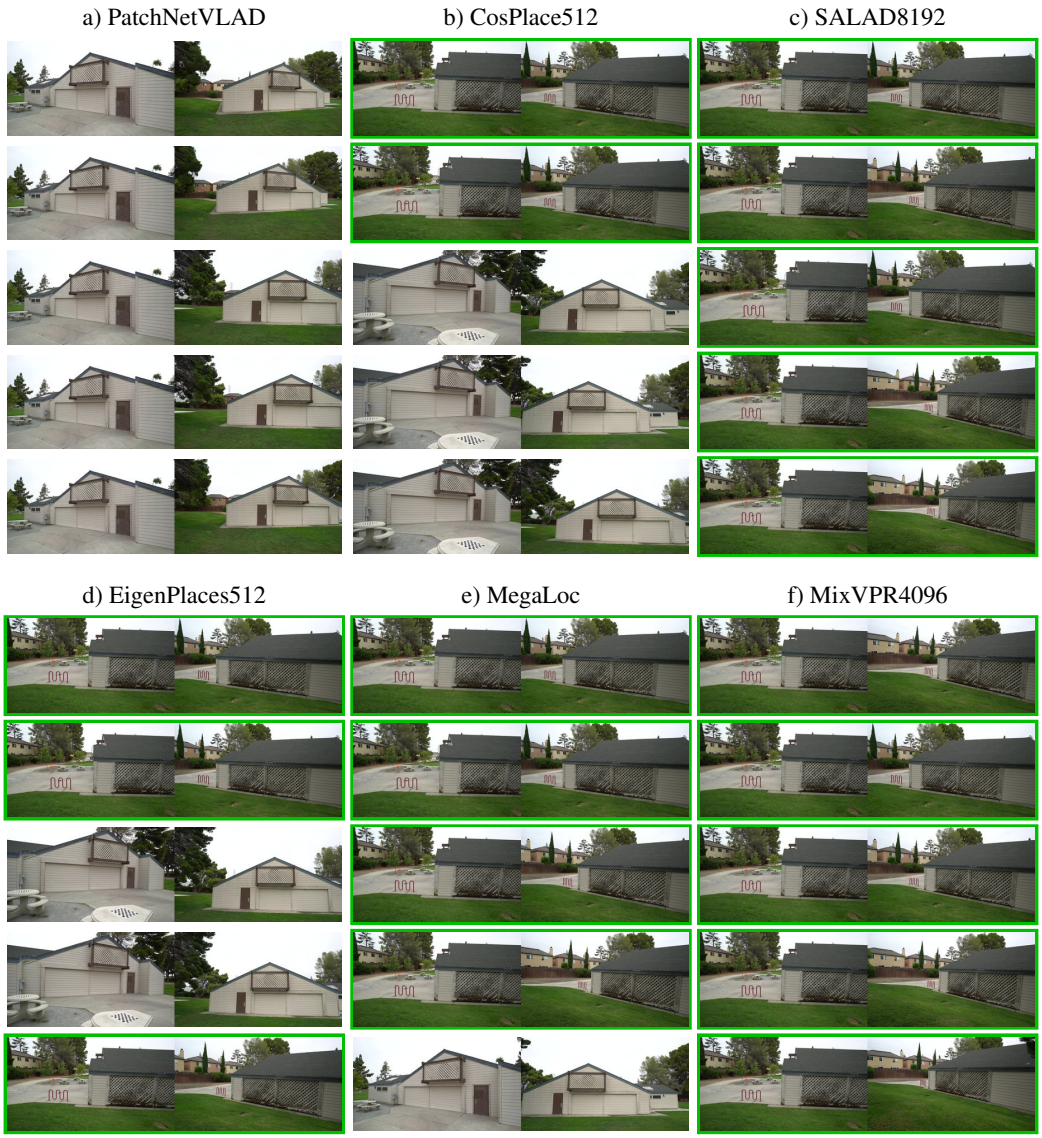

Figure 2. Qualitative top-5 retrieval results for different VPR methods on the T&T scene *Barn*. This scene is especially prone to perceptual aliasing, as indicated by the brown door, which is at different positions on both sides and the color of the floor in front of the house. Overlap areas are the back of the house, which is completely correctly retrieved by SALAD8192 and MixVPR4096. Columns correspond to methods (a–f) and, within each column, pairwise image matches are shown from rank $k = 1$ (top) to $k = 5$ (bottom) based on descending cosine similarity results. True positive matches are framed in green.

ceptual aliasing (e.g., repeated structures such as wheels or façades) remain challenging for all methods. ViT-based models such as SALAD and MegaLoc often produce more robust top-k lists in the most ambiguous cases (cf. Fig. 2), while strong CNN-based methods like EigenPlaces achieve comparable average P@k and R@k at substantially lower runtime.

- **Indoor scenes (SN-GS):** All methods perform extremely well, which we attribute to the high overlap between the disjoint sets A and B and, in some scenes, even a small number of identical images across both sets.

- **Outdoor autonomous navigation (KITTI):** For typical SLAM and odometry scenarios, CNN-based methods such as EigenPlaces offer a favorable trade-off between retrieval quality and runtime. The performance gap in P@k and R@k to the best-performing ViT-based methods is only minor, while the runtime advantage is substantial.

7. Conclusion

In this work, we presented an extensive evaluation of state-of-the-art VPR methods for the task of image pair retrieval. Our evaluation pipeline is specifically designed for photogrammetrically reconstructed scenes, and our metrics are tailored to typical 3D vision and robotics tasks. We showed that VPR methods that were especially trained for out-of-domain performance (e.g., MegaLoc) can be used as strong off-the-shelf front-ends for the initial matching of images from two disjoint datasets. This is particularly relevant for RGB-D registration, 3D Gaussian Splatting registration, relocalization, loop-closure detection, and boosting SfM correspondence search. In real-time scenarios, CNN-based methods appear preferable due to their favourable trade-off between retrieval quality and runtime, whereas ViT-based methods achieve higher retrieval performance in highly challenging scenes with strong perceptual aliasing and missing frames in image sequences.
By design, this work isolates and evaluates VPR purely at

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-647-2026 | © Author(s) 2026. CC BY 4.0 License.

a) CosPlace512

b) MegaLoc

Figure 3. Qualitative Top-5 retrieval results from CosPlace512 (a) and MegaLoc (b) for KITTI scene 07_02. The challenge here is that some images were left out between A and B. This is especially important for tasks such as SLAM relocalization caused from missing frames. CosPlace (Berton et al., 2022) completely fails here, likely because of perceptual aliasing, indicated by two different black cars in A and B and other similarities, whereas MegaLoc (Berton and Masone, 2025) yields only true positives. The cue for correct results is the houses in the center of the A images, which appear towards the left border of the B images. Pairwise image matches are shown from rank $k = 1$ (top) to $k = 5$ (bottom) based on descending cosine similarity results. True positive matches are framed in green.

the retrieval stage, independently of any particular registration pipeline. Building on these results, future work will integrate the most promising VPR configurations into complete scene registration and SLAM pipelines and quantify their impact on pose accuracy and robustness while preserving runtime efficiency. Furthermore, the demonstration of the general capability of VPR methods to match pairs of images showing the same scene content under often challenging conditions such as per-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-647-2026 | © Author(s) 2026. CC BY 4.0 License.

Dataset	Method	P@1	P@5	P@10	R@1	R@5	R@10	mAP@5	mAP@10	t_μ	t_σ
T&T	PatchNetVLAD	25.00	21.25	17.50	25.00	37.50	43.75	25.10	26.27	1.102	0.105
	CosPlace512	56.25	43.75	36.88	56.25	56.25	62.50	55.62	52.69	0.133	0.014
	CosPlace2048	56.25	40.00	32.50	56.25	56.25	62.50	53.33	51.54	0.645	0.070
	EigenPlaces512	56.25	47.50	36.88	56.25	56.25	62.50	54.37	54.35	0.134	0.016
	EigenPlaces2048	56.25	45.00	35.62	56.25	62.50	62.50	57.19	53.67	0.649	0.070
	MixVPR512	18.75	6.25	5.00	18.75	25.00	31.25	21.88	21.56	0.130	0.012
	MixVPR4096	25.00	13.75	8.12	25.00	25.00	31.25	23.44	23.28	0.133	0.012
	AnyLoc	31.25	21.25	16.25	31.25	43.75	50.00	32.84	31.04	6.116	3.771
	SALAD2048	56.25	46.25	34.38	56.25	56.25	56.25	55.03	53.70	2.740	0.305
	SALAD8192	50.00	38.75	29.38	50.00	50.00	56.25	47.25	46.78	2.755	0.307
	MegaLoc	56.25	47.50	35.62	56.25	56.25	56.25	53.46	51.72	2.799	0.316
SN-GS	PatchNetVLAD	92.31	93.08	93.46	92.31	96.15	96.15	93.21	93.47	2.541	0.187
	CosPlace512	100.00	98.46	99.23	100.00	100.00	100.00	98.81	98.92	0.375	0.028
	CosPlace2048	100.00	99.23	99.23	100.00	100.00	100.00	99.81	99.55	1.817	0.126
	EigenPlaces512	100.00	99.23	99.23	100.00	100.00	100.00	99.57	99.44	0.372	0.025
	EigenPlaces2048	100.00	98.46	99.23	100.00	100.00	100.00	98.85	98.95	1.825	0.124
	MixVPR512	96.15	94.62	93.85	96.15	96.15	100.00	95.40	96.06	0.301	0.022
	MixVPR4096	92.31	91.54	93.08	92.31	100.00	100.00	94.04	93.59	0.308	0.023
	AnyLoc	92.31	94.62	93.85	92.31	96.15	100.00	94.36	94.73	12.501	2.821
	SALAD2048	100.00	100.00	99.62	100.00	100.00	100.00	100.00	100.00	9.458	0.669
	SALAD8192	100.00	99.23	99.62	100.00	100.00	100.00	99.57	99.53	9.560	0.683
	MegaLoc	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	9.642	0.694
KITTI	PatchNetVLAD	34.62	24.62	21.54	34.62	38.46	38.46	34.36	31.47	1.683	0.293
	CosPlace512	34.62	35.38	33.46	34.62	38.46	38.46	36.54	35.82	0.120	0.045
	CosPlace2048	34.62	35.38	32.31	34.62	38.46	38.46	36.54	36.36	0.551	0.094
	EigenPlaces512	34.62	34.62	32.31	34.62	34.62	34.62	34.62	34.18	0.112	0.019
	EigenPlaces2048	34.62	34.62	32.31	34.62	34.62	38.46	34.62	34.91	0.546	0.092
	MixVPR512	19.23	19.23	16.92	19.23	26.92	30.77	23.25	23.18	0.198	0.035
	MixVPR4096	30.77	22.31	20.77	30.77	38.46	38.46	32.97	31.76	0.204	0.036
	AnyLoc	23.08	23.08	21.92	23.08	26.92	30.77	24.17	24.10	8.516	3.163
	SALAD2048	34.62	35.38	33.08	34.62	38.46	38.46	36.54	36.54	1.799	0.311
	SALAD8192	34.62	34.62	33.08	34.62	34.62	34.62	34.62	34.55	1.813	0.316
	MegaLoc	34.62	35.38	33.08	34.62	38.46	38.46	36.54	36.47	1.848	0.319

Table 2. Image pair retrieval across different datasets, consisting of multiple scenes (quantity in braces) for each dataset: T&T (16), SN-GS (26), and KITTI (26). Each dataset is evaluated on seven VPR methods, split into eleven configurations. Metrics are calculated for the k best ranked image pairs, leading to Precision@k (P@k), Recall@k (R@k), and mean-Average-Precision@k (mAP@k), all given in percent. Average time in seconds for all retrievals per scene is included as t_μ with t_σ as the standard deviation, respectively. The measured time includes descriptor extraction, matching and ranking. The evaluation procedure is excluded, as we focus on application performance.

ceptual aliasing provided in this work might be extended by a more detailed analysis. In this regard, cases with valid matches but rejection due to a false negative result (e.g. a low number of keypoints with wrong matches in largely homogeneous environments) might artificially lower the measured performance of the VPR methods. While our work provides per-dataset evaluations, a future study might also investigate failure cases per scene as opposed in more detail.

References

Ali-Bey, A., Chaib-Draa, B., Giguère, P., 2023. MixVPR: Feature mixing for visual place recognition. *WACV*.

Arandjelovic, R., Gronát, P., Torii, A., Pajdla, T., Sivic, J., 2016. NetVLAD: CNN architecture for weakly supervised place recognition. *CVPR*.

Arandjelovic, R., Zisserman, A., 2013. All about VLAD. *CVPR*, 1578–1585.

Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V. S., 2014. Neural codes for image retrieval. *ECCV*, 584–599.

Bampis, L., Amanatiadis, A., Gasteratos, A., 2016. Encoding the description of image sequences: A two-layered pipeline for loop closure detection. *IROS*.

Berton, G. M., Masone, C., Paolicelli, V., Caputo, B., 2021. Viewpoint invariant dense matching for visual geolocalization. *ICCV*.

Berton, G., Masone, C., 2025. Megaloc: One retrieval to place them all. *CVPR Workshops*.

Berton, G., Masone, C., Caputo, B., 2022. Rethinking visual geo-localization for large-scale applications. *CVPR*.

Berton, G., Trivigno, G., Caputo, B., Masone, C., 2023. Eigenplaces: Training viewpoint robust models for visual place recognition. *ICCV*.

Besl, P. J., McKay, N. D., 1992. A Method for Registration of 3-D Shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2), 239–256.

Cao, B., Araújo, A., Sim, J., 2020. Unifying deep local and global features for image search. *ECCV*.

Chang, J., Xu, Y., Li, Y., Chen, Y., Feng, W., Han, X., 2024. Gaussreg: Fast 3d registration with gaussian splatting. *ECCV*.

Chen, Z., Jacobson, A., Sünderhauf, N., Upcroft, B., Liu, L., Shen, C., Reid, I. D., Milford, M., 2017. Deep learning features at scale for visual place recognition. *ICRA*.

Chen, Z., Liu, L., Sa, I., Ge, Z., Chli, M., 2018. Learning Context Flexible Attention Model for Long-Term Visual Place Recognition. *IEEE Robot. Autom. Lett.*, 3(4), 4015–4022.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints. *ECCV Workshop on Statistical Learning in Computer Vision*, 1–22.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *CVPR*.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. *CVPR*.

Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T., 2021. Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition. *CVPR*.

Izquierdo, S., Civera, J., 2024. Optimal transport aggregation for visual place recognition. *CVPR*.

Jégou, H., Douze, M., Schmid, C., Pérez, P., 2010. Aggregating local descriptors into a compact image representation. *CVPR*.

Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K. M., Scherer, S., Krishna, M., Garg, S., 2023. AnyLoc: Towards Universal Visual Place Recognition. *IEEE Robot. Autom. Lett.*

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4), 139:1–139:14.

Khaliq, A., Ehsan, S., Chen, Z., Milford, M., McDonald-Maier, K. D., 2020. A Holistic Visual Place Recognition Approach Using Lightweight CNNs for Significant Viewpoint and Appearance Changes. *IEEE Trans. Robot.*, 36(2), 561–569.

Knapitsch, A., Park, J., Zhou, Q.-Y., Koltun, V., 2017. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Trans. Graph.*, 36(4).

Liu, K., Wang, H., Han, F., Zhang, H., 2019. Visual place recognition via robust ℓ_2-norm distance based holism and landmark integration. *AAAI*.

Lowe, D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.*, 60(2), 91–110.

Malone, C., Hussaini, S., Fischer, T., Milford, M., 2025. A hyperdimensional one place signature to represent them all: Stackable descriptors for visual place recognition. *ICCV*.

Milford, M., Wyeth, G. F., 2012. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. *ICRA*.

Murillo, A. C., Singh, G., Kosecká, J., Guerrero, J. J., 2013. Localization in Urban Environments Using a Panoramic Gist Descriptor. *IEEE Trans. Robot.*, 29(1), 146–160.

Nistér, D., 2004. An Efficient Solution to the Five-Point Relative Pose Problem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(6), 756–777.

Nistér, D., Stewénius, H., 2006. Scalable recognition with a vocabulary tree. *CVPR*, 2161–2168.

Oliva, A., Torralba, A., 2001. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *Int. J. Comput. Vis.*, 42(3), 145–175.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. et al., 2023. DINOv2: Learning Robust Visual Features without Supervision. *arXiv:2304.07193*.

Radenović, F., Tolias, G., Chum, O., 2019. Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(7), 1655–1668.

Razavian, A. S., Azizpour, H., Sullivan, J., Carlsson, S., 2014. CNN features off-the-shelf: An astounding baseline for recognition. *CVPR*, 512–519.

Schönberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *CVPR*.

Shao, S., Chen, K., Karpur, A., Cui, Q., Araujo, A., Cao, B., 2023. Global features are all you need for image retrieval and reranking. *ICCV*.

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556*.

Sivic, J., Zisserman, A., 2003. Video google: A text retrieval approach to object matching in videos. *ICCV*, 1470–1477.

Sünderhauf, N., Neubert, P., Protzel, P., 2013. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. *ICRA Workshop on Long-Term Autonomy*.

Torii, A., Arandjelović, R., Sivic, J., Okutomi, M., Pajdla, T., 2015. 24/7 place recognition by view synthesis. *CVPR*.

Torii, A., Sivic, J., Pajdla, T., Okutomi, M., 2013. Visual place recognition with repetitive structures. *CVPR*.

Wang, R., Shen, Y., Zuo, W., Zhou, S., Zheng, N., 2022. Transvpr: Transformer-based place recognition with multi-level attention aggregation. *CVPR*.

Warburg, F., Hauberg, S., López-Antequera, M., Gargallo, P., Kuang, Y., Civera, J., 2020. Mapillary street-level sequences: A dataset for lifelong place recognition. *CVPR*.

Wei, T., Lindenberger, P., Matas, J., Barath, D., 2025. Breaking the frame: Visual place recognition by overlap prediction. *WACV*.

Yang, M., He, D., Fan, M., Shi, B., Xue, X., Li, F., Ding, E., Huang, J., 2021. Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. *ICCV*.

Zhang, H., Chen, X., Jing, H., Zheng, Y., Wu, Y., Jin, C., 2023. Etr: An efficient transformer for re-ranking in visual place recognition. *WACV*.