Revisit Anything: Visual Place Recognition via Image Segment Retrieval

Kartik Garg*[1], Sai Shubodh Puligilla*[2], Shishir Kolathaya[1], Madhava Krishna[2], and Sourav Garg[3]

[1] Indian Institute of Science (IISc), Bengaluru, India
[2] International Institute of Information Technology, Hyderabad, India
[3] University of Adelaide, Australia
kartikgarg@iisc.ac.in, p.saishubodh@gmail.com, shishirk@iisc.ac.in, mkrishna@iiit.ac.in, sourav.garg@adelaide.edu.au

Abstract. Accurately recognizing a revisited place is crucial for embodied agents to localize and navigate. This requires visual representations to be distinct, despite strong variations in camera viewpoint and scene appearance. Existing visual place recognition pipelines encode the *whole* image and search for matches. This poses a fundamental challenge in matching two images of the same place captured from different camera viewpoints: *the similarity of what overlaps can be dominated by the dissimilarity of what does not overlap*. We address this by encoding and searching for *image segments* instead of the whole images. We propose to use open-set image segmentation to decompose an image into 'meaningful' entities (i.e., things and stuff). This enables us to create a novel image representation as a collection of multiple overlapping subgraphs connecting a segment with its neighboring segments, dubbed SuperSegment. Furthermore, to efficiently encode these SuperSegments into compact vector representations, we propose a novel factorized representation of feature aggregation. We show that retrieving these partial representations leads to significantly higher recognition recall than the typical whole image based retrieval. Our segments-based approach, dubbed SegVLAD, sets a new state-of-the-art in place recognition on a diverse selection of benchmark datasets, while being applicable to *both* generic and task-specialized image encoders. Finally, we demonstrate the potential of our method to "revisit anything" by evaluating our method on an object instance retrieval task, which bridges the two disparate areas of research: visual place recognition and object-goal navigation, through their common aim of recognizing goal objects specific to a place. Source code: https://github.com/AnyLoc/Revisit-Anything.

Keywords: Visual Place Recognition · Image Segmentation · Robotics

1 Introduction

Visual Place Recognition (VPR) is an important capability for embodied agents to localize and navigate autonomously. A predominant solution for VPR is to encode

* Equal contribution

an image into a global vector and retrieve similar vectors as coarse localization hypotheses [19, 44, 60, 67]. Thus, for almost a decade, researchers have focused on learning/finetuning image encoders so that global descriptors are induced with invariance to appearance [3,56,69], viewpoint [3,8], and clutter [27]. On the other hand, there is a vast literature on local descriptors (point/pixel-level), mainly relevant for geometric reranking in hierarchical VPR [11, 16, 25, 57, 68]. In the middle of local and global descriptors exists a variety of methods that use regions/patches [4, 25], lines/planes [17], objects (things/stuff) [21, 31, 45, 65], and segments [26, 33, 50] to represent images. However, these methods are still only aimed at either improving global descriptors based coarse retrieval or local feature matching based reranking. In this work, in contrast to conventional retrieval-based VPR, we explore an alternative: **retrieval via encoding segments instead of the whole image**. This is particularly enabled by recent advances in open-set image segmentation [37] which can meaningfully deconstruct a place into 'things' (and/or 'stuff') [10]. Thus, we reformulate the VPR problem of revisiting places as that to *revisiting things* by enabling recognition of these specific things within the context of their place. While such a segment-level place recognition approach provides a direct link to higher-level semantic tasks, such as object-goal navigation [12, 20, 24, 43], it also addresses a fundamental issue in matching partially-overlapping images from across significant viewpoint change. Segments-based partial image representation avoids the mismatches caused by the whole-image representation when *the similarity of what overlaps is dominated by the dissimilarity of what does not overlap*. Our novel segments-based VPR method, dubbed *SegVLAD* (Segment based Vector of Locally Aggregated Descriptors), is illustrated in Figure 1, which makes the following novel contributions:

1. an image representation as a collection of multiple overlapping subgraphs of segments, dubbed *SuperSegments*, which enables accurate recognition across partially-overlapping images;
2. a factorized representation of feature aggregation to effectively accommodate both segment-level information as well as segment neighborhood information; and
3. a similarity-weighted ranking method to convert segment-level retrieval into image-level retrieval.

Using a diverse set of data sources, we demonstrate that our proposed segments-based retrieval enables place recognition under wide viewpoint variations, where global descriptor based retrieval suffers. SegVLAD achieves a new state-of-the-art on multiple challenging datasets. We also introduce an evaluation of our method on an instance-level object retrieval task – a novel capability of our pipeline unlike conventional VPR methods. We conduct several ablations and parameter studies to justify the design choices and emphasize the effectiveness of our method as an open-set segments-based coarse retriever.

2 Related Works

Image retrieval-based Visual Place Recognition (VPR) is a well-established area of research in visual localization [19, 44, 60, 67]. It is important both during mapping for loop

closures [67] as well as for relocalization [51, 58]. The underlying task in both the scenarios remains the same: how to recognize a previously seen place. The state-of-the-art methods in VPR use a global descriptor-based approach which converts an image into a compact vector to enable fast retrieval [2, 3, 6, 29, 32, 56]. The top retrieved hypotheses are often then re-ranked through compute-intensive local feature matching using geometric information [11, 16, 25, 33, 57, 68]. In contrast to previous approaches, we aim to explore image segment level descriptors in this work. This representation falls between point-based local descriptors and the whole-image based global descriptors. Our approach can be considered as 'semi-global', with the proposed segment (and SuperSegment) based descriptors being a 'spatially-reduced' form of whole-image global representation. This is motivated by our hypothesis that to deal with viewpoint variations in VPR with partially-overlapping images, we need a way to partially represent and match them.

2.1 Whole Image Encoders

Earlier works in whole-image representation used methods like Gist [48], BoW (Bag of Word) [63], and VLAD (Vector of Locally Aggregated Descriptors) [30], often defined using hand-crafted features such as SIFT [41]. In recent years, deep learning based methods have demonstrated remarkable performance, with initial successful methods like NetVLAD [3] now rapidly outperformed by better alternatives such as CosPlace [6], MixVPR [2], EigenPlaces [8], TransVPR [68], and more recently AnyLoc [32], SALAD [29] and VLAD-BuFF [35]. All these learning-based methods improve different aspects of representation learning: training datasets [1, 6, 69], objective/loss functions [2, 8], aggregation methods [29, 54, 56, 68], and generalization [32]. Our approach complements these existing methods as we mainly focus on the use of segment-based information, where the segments can be described by any of the image encoders from the aforementioned techniques. In particular, we demonstrate that *both* – an off-the-shelf encoder, e.g., DINOv2-AnyLoc [32, 49] or that finetuned specifically for the VPR task, e.g., DINOv2-NetVLAD [35] – can be used in conjunction with our segment-based approach to further elevate place recognition capability.

2.2 Region/Patch Based Methods

There exist several methods that employ region or patch level information to enhance representational power [13, 14, 34, 40, 52, 71, 73]. However, most of these methods only use this additional information to generate a single (or concatenated) compact vector representation of an image. Other methods such as Patch-NetVLAD [25] create multiple features per image but their primary purpose is to perform local matching based reranking. In contrast to these methods, we aim to use multiple segment descriptors per image to directly retrieve from a database of segments, *without* using any geometric information or reranking. The motivation behind this stems from the very nature of hierarchical VPR pipelines: reranking recall is upper bounded by the coarse retriever. A better coarse retriever can improve reranking performance without needing to rerank from a longer list of top hypotheses. MultiVLAD [4] is similar to our method in the spirit of retrieving multiple features per query image. However, like aforementioned

methods, MultiVLAD defines regions arbitrarily, whereas we use image segments obtained from Segment Anything Model (SAM) [37] which are semantically meaningful.

2.3 Segments-Enhanced Methods

There exist several methods that use semantic segmentation information to improve VPR, as also surveyed in [22]. These methods vary in terms of type of segmentation used and the specific ways in which it is integrated in the VPR pipeline, e.g., planes [17], objects [15,31,45], landmarks [65], outdoor semantics [21,23,46,47], utility/confusion based [33, 38], domain adaptation [26] and even learning to segment for VPR [50]. However, neither these methods aim to perform segment-level retrieval nor do they use open-set segmentation. We also review two concurrent works: MESA [75] and Region-Revisited [61]. Similar to our method, they both use SAM to segment images but for different specific tasks. MESA [75] proposes a graph-based local feature/area matching method to obtain point correspondences. Our method complements this, as we perform coarse retrieval for VPR, which could potentially use MESA for reranking. Regions Revisted [61] delves into the advantages of using SAM masks in conjunction with SLIC [36] to improve semantic segmentation, activity recognition and object *category* retrieval. In contrast, we aim to improve *instance-level* recognition by recognizing specific things belonging to specific places that a robot encounters during a revisit. Similar to our work, [20] creates an image sequence-based topological graph of segments where its segment neighbourhood aggregation is based on average pooling, similar to [61]. In Section 5.3, we show that such segment average pooling deteriorates recognition performance for the VPR task.

2.4 Open-set VPR

Researchers have recently started to shift their focus to open-set, generally-applicable techniques, including that for VPR [31, 32, 45, 62]. FM-Loc [45] uses GPT [9] to recognize object and place labels, whereas [31] uses CLIP [55] for open-set place recognition. LIP-Loc [62] proposes pretraining for cross-modal VPR, but is limited in its zero-shot capabilities. AnyLoc [32] proposes to use DINOv2 with domain-level vocabularies and hard-assignment based VLAD. It achieves state-of-the-art performance particularly on non-streetview datasets, where current VPR-trained methods tend to fail. In this work, we propose a generally-applicable approach which is built on top of models like SAM [37] and DINOv2 [49], and works with both VPR-agnostic [32] and VPR-specific [35] backbone models. We particularly aim for a new paradigm in retrieval based VPR, where we move away from the conventional whole image global descriptors to segments based descriptors and retrieval, which achieves a new state-of-the-art on diverse domains under wide viewpoint variations.

3 Proposed Approach

Despite recent advances in place recognition, viewpoint variations continue to be an open challenge for an embodied agent to recognize the same specific things in its environment. Current methods in visual place recognition tackle this problem by converting

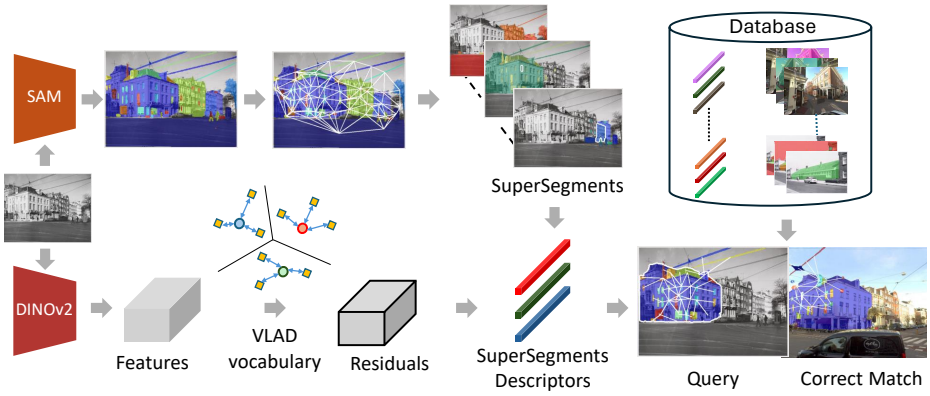

Fig. 1: Overview of our segment-retrieval based VPR pipeline, dubbed SegVLAD: We use open-set segmentor (SAM) to extract segment masks, which are converted into **SuperSegments** using the neighbouring image segments. Using pixel-level DINOv2 descriptors with VLAD-based aggregation, we obtain SuperSegment descriptors, which are matched against a flat index of SuperSegment descriptors obtained from all the images of the entire reference database.

an image *as a whole* into a global descriptor, which does not explicitly deal with the problem of partial visual overlap caused by viewpoint variations. We propose an alternative solution by representing images *partially* with the help of image segments. In the following subsections, we describe our representation and retrieval method, which deviates from the conventional VPR techniques but creates a new capability in terms of recognizing objects/things that constitute a place.

3.1 Problem Formulation

We represent an image as a set of segment descriptors instead of a single global descriptor. For an image I, we obtain binary image segment masks $M \in \{0,1\}^{S\times N}$ and dense pixel-level descriptors $f_p \in \mathcal{R}^D$, where S represents the number of segments per image, D is the descriptor dimension, and $p \in [1, N]$ represents spatial elements across the width (W) and height (H) of the image encoder's output, flattened into $N = W \times H$ for convenience. Figure 1 shows an illustration of our proposed pipeline, as explained in the following subsections.

3.2 Super Segments

Humans are remarkable at visual recognition, where existing studies suggest that we often leverage spatial associations among objects in an environment to represent it internally [5, 28]. This enables us to distinguish between two different scenes through the surrounding context of the objects of interest. In this work, we imbibe this context through the spatial neighbourhood of the image segments. For each image, we construct a graph of segments through their pixel centers using Delaunay Triangulation.

Fig. 2: Neighborhood expansion (Eq. 1) of a window in the leftmost image to the whole building in the rightmost image, progressing from no neighborhood aggregation to a third-order aggregation. This neighborhood expansion is in stark contrast with a typical regular grid- or patch-based approach which may not capture semantically-meaningful SuperSegments.

Fig. 3: Illustration of four SuperSegments obtained from the same image. All four of these spatially overlap with each other, which is different from coarse segmentation methods that do not typically allow overlap across segments.

This provides us with a binary adjacency matrix $A \in \{0,1\}^{S\times S}$ to define the neighborhood for individual segments. We use this adjacency information to expand the context of individual segments to generate new **SuperSegment** masks ($\mathcal{M}$) as below:

$$\mathcal{M}_{S\times N} = {}^{\mathbb{1}}(A^{o}_{S\times S} \cdot M_{S\times N}) \tag{1}$$

where $o \geq 0$ refers to the order for expanding the neighborhood by multiplying the adjacency matrix by itself as $A^{o+1} = A^{o} \cdot A$. This is matrix-multiplied with the original segmentation masks M to expand the neighborhood *at pixel level*. $\mathcal{M}$ is obtained after element-wise binarization (denoted with ${}^{\mathbb{1}}()$) so that all pixels in the SuperSegment mask may only contribute once to the subsequent feature aggregation. In Figure 2, we illustrate the extent of image area covered with different orders of mask expansion. Unlike, a patch or regular grid-based approach, the expanded mask of the window in the leftmost image covers a meaningful entity (building) in the rightmost image. Our approach to creating SuperSegments differs from *coarse* segmentation methods or superpixels in terms of the 'self-overlap'. By expanding neighborhood of each individual segment, we obtain several *partially overlapping* SuperSegments. A coarse segmentor will need to make assumptions about the right sub-segments to be coalesced so that it can enable accurate recognition from a different viewpoint, which could otherwise lead to the same limitation as that of the whole-image descriptors. Figure 3 presents examples of multiple overlapping SuperSegments from the same image.

3.3 SuperSegment Descriptors

In this section, we describe our feature aggregation method to obtain SuperSegment descriptors. Recent state-of-the-art method AnyLoc [32] demonstrated that using off-the-shelf powerful image encoders such as DINOv2 with hard assignment based VLAD

aggregation achieves superior recognition performance. However, AnyLoc does not use segmentation information and only operates at the whole-image level. More recently, [61] showed that average pooling works well for segment-level descriptors, but it didn't consider segment neighborhood information. In this work, we propose a unified formulation for feature aggregation that can easily switch across segments, segment neighborhood and the whole image as well as different aggregation types (see supplementary for details). This simply extends Equation 1 as below:

$$F_{S\times D} = {}^{\mathbb{1}}(A^{o}_{S\times S} \cdot M_{S\times N}) \cdot T_{N\times D} \tag{2}$$

where T represents the features to be aggregated. By replacing A and M with ones matrices, one can obtain the whole-image global descriptor. For methods like Global Average Pooling (GAP), T can be directly used as the output of the image encoder. In our work, we use Hard-VLAD, for which T is the residual feature matrix *per cluster* and is obtained as below with respect to each of the cluster centers c_k:

$$T^{k}_{N_k\times D} = \{\alpha_k(f_p)(f_p - c_k) \mid \alpha_k(f_p) = 1\} \tag{3}$$

where $\alpha_k(f_p) \in \{0, 1\}$ is 1 if f_p belongs to c_k. The cluster centers (vocabulary) can be constructed using the map or the domain [32]. The SuperSegment VLAD descriptors obtained from Eq. 2 are intra-normalized, concatenated across clusters and then finally l2-normalized, following existing works [3, 32].

3.4 Image Retrieval via Segments

Existing global descriptor based VPR techniques produce a single vector per image to search against a database of reference image vectors. In our method, we obtain multiple SuperSegment descriptors per image. We perform retrieval at segment-level, that is, we search for the top matches for each query segment against a flat index of all segments from all the images of the reference database/map. To evaluate in the form of image retrieval-based VPR, we convert the top retrieved segment indices across all segments of a query image into top reference image indices. This is achieved through a weighted frequency measure (i.e., weighted bin/word counting). We first map the top K' retrieved segment indices for each of the query segments $s \in [1, S]$ to their respective reference image indices, denoted with r. Then, for each of the unique retrieved image indices r_j, we accumulate its segment similarity θ and then use the cumulative similarity score $\hat{\theta}$ to rank the image indices to obtain the top image match r^*_j:

$$r^*_j = \underset{r_j}{\text{argmax}}\ \hat{\theta}(r_j); \quad \hat{\theta}(r_j) = \sum_{s=1}^{S}\sum_{k=1}^{K'} \theta_{sk} \cdot \mathbb{1}_{\{r_{sk}=r_j\}} \tag{4}$$

In Section 5.3, we compare our similarity-weighted ranking with other alternatives based on frequency or similarity alone.

4 Experimental Setup

***Datasets**:* VPR datasets are in abundance, as can be found in several benchmarks including VPR-Bench [74], Deep Visual GeoLocalization Benchmark [7], and AnyLoc [32]. In this work, we used a variety of datasets covering both outdoor and indoor environments. Outdoor datasets include Pitts30k [66], AmsterTime [72], Mapillary Street Level Sequences (MSLS) [69], SF-XL [6], VPAir [59], Revisted Oxford5K and Revisited Paris6k [53] . Indoor datasets include Baidu Mall [64], 17Places [76] and InsideOut [27]. Additional datasets-related details are provided in the supplementary.

***Evaluation and Benchmarking**:* We evaluate our method as an image retrieval task using Recall@K metric, where top K' ($= 50$) retrieved segments per query segment are used to obtain top K ($= 5$) images (see Eq. 4). We compare against the most recent and high-performing VPR baseline methods. This includes CosPlace [6], MixVPR [2] and EigenPlaces [8], which are trained on large-scale urban datasets for VPR tasks. We further include two very recent state-of-the-art methods that use DINOv2 as the backbone. These include AnyLoc [32] which uses an *off-the-shelf* DINOv2 model and SALAD [29] which uses a *finetuned* DINOv2 backbone. Given the dichotomy between general-purpose VPR benchmarking of AnyLoc and the typical outdoor-focused benchmarking in SALAD and other methods, we evaluated our method using two different backbones. *a)* **SegVLAD-PreT:** we use the same backbone and aggregation as AnyLoc, i.e., off-the-shelf *pretrained* DINOv2 (ViT-G) backbone with hard VLAD assignment, but the key difference is in the use of SuperSegments for our method as opposed to whole-image description of AnyLoc. *b)* **SegVLAD-FineT:** since our default aggregation method is VLAD, we use a *finetuned* DINOv2 (ViT-B) backbone which is similar to SALAD but replaces its aggregation layer with the original NetVLAD aggregation [3] using 64 clusters, as described in [35]. We use this finetuned backbone with hard VLAD based assignment, similar to AnyLoc. For both these models, we reduce the descriptor dimensions of the VLAD descriptor to 1024 using PCA, as commonly done in previous works [3, 32]. We train PCA transform in a map-specific manner using the database images of the dataset.

5 Results

We first present benchmark comparison of our method against state-of-the-art VPR methods. This is followed by detailed analysis of our proposed aggregation technique. Lastly, we demonstrate results on a downstream task of Object of Interest (OOI) retrieval, showcasing the versatility of our method.

5.1 State-of-the-art comparisons

Table 1 presents Recall@1/5 comparison against state-of-the-art VPR methods on standard outdoor street-view datasets, which are similar to the typical training datasets used for VPR [2,3,6]. Table 2 covers 'out-of-distribution' datasets, inspired by AnyLoc [32], covering indoor environments (Baidu Mall and 17 Places), aerial imagery (VPAir),

Table 1: Recall@1/5 benchmark comparison on outdoor street-view datasets.

Method	Pitts-30K	MSLS SF	MSLS CPH	SF-XL Val	RO5k Med	RO5k Hard	RP6k Med	RP6k Hard
CosPlace	90.4/95.7	93.4/**97.5**	84.9/92.0	94.6/97.6	85.7/87.1	27.1/45.7	94.3/95.7	7.1/15.7
MixVPR	91.5/95.5	91.3/95.9	87.1/92.4	87.8/93.8	68.6/80.0	32.9/54.3	94.3/100	10.0/32.9
EigenPlaces	92.6/96.7	92.6/97.1	87.1/92.8	**96.4/98.2**	85.7/88.6	42.8/57.1	95.7/98.6	4.3/11.4
AnyLoc	87.7/94.7	83.4/94.6	79.9/89.1	84.4/91.9	88.6/92.9	40.0/58.6	97.1/100	11.4/44.3
SALAD	92.6/96.5	91.7/97.1	**92.3/96.1**	93.6/97.3	82.9/90.0	37.1/54.3	95.7/98.6	**14.3/58.6**
SegVLAD-PreT	86.7/94.2	88.4/94.2	81.7/90.7	90.9/96.4	**91.4/95.7**	**60.0/81.4**	94.3/**100**	8.6/48.6
SegVLAD-FineT	**93.2/96.8**	**94.6**/97.1	90.9/95.7	94.9/98.1	87.1/**95.7**	51.4/70.0	95.7/**100**	10.0/48.6

Table 2: Recall@1/5 benchmark comparison on 'out-of-distribution' datasets.

Method	Baidu	AmsterTime	InsideOut	17Places	VPAir
CosPlace	41.6/55.0	47.7/69.8	0.2/2.0	81.3/88.2	4.6/13.7
MixVPR	64.4/80.3	40.2/59.1	0.0/1.8	85.2/90.1	6.8/16.1
EigenPlaces	56.5/72.8	48.9/69.5	0.4/1.4	83.0/90.1	6.5/17.9
AnyLoc	75.2/87.6	50.3/73.0	2.4/8.0	**95.3**/97.3	66.7/79.2
SALAD	74.8/86.5	55.4/75.6	0.6/1.8	82.5/88.2	25.8/38.7
SegVLAD-PreT	**78.5/93.8**	56.8/77.7	4.2/9.4	**95.3/98.0**	**69.8/83.7**
SegVLAD-FineT	68.1/89.0	**58.9/79.3**	**7.4/15.6**	95.1/97.5	35.4/55.3

indoor-to-outdoor viewing (InsideOut), and historical image matching (AmsterTime). Below, we discuss two key aspects of this comparative analysis: *i)* how our segment-based approach compares against whole-image global descriptor based methods and *ii)* how performance trends vary depending on the choices of feature backbone with regards to task-specific (VPR) training.

Aggregating Segments vs Whole Images Table 1 and Table 2 show that our proposed method SegVLAD achieves a new state-of-the-art on the majority of datasets, considering both the backbone variants: PreT and FineT. AnyLoc and SALAD respectively differ from SegVLAD-PreT and SegVLAD-FineT in terms of the aggregation scope (global vs segments). Thus, the superior performance of SegVLAD clearly highlights the role of segments based retrieval over whole-image based approach. On the Baidu Mall dataset – highly-aliased indoor environment – our method improves over AnyLoc by 3% for R@1 and 6% for R@5 in absolute gains. On the InsideOut dataset – matching outdoor images viewed from within indoors – our method leads to a 'meaningful' recall, unlike all other baselines. Overall, these results highlight that even with the use of powerful image encoders (DINOv2), global aggregation struggles to deal with the challenges of matching images across major viewpoint shifts – it is thus the partial image representation and matching (using the same encoder) which is needed to obtain superior recognition performance.

VPR Fine-tuned Encoders + Segments For SegVLAD-FineT, we used a DINOv2 backbone finetuned for the purpose of VPR, mainly to observe the benefit of segments over global descriptor based approach in a *task-specific manner*. Table 1 and Table 2

Table 3: Recall@1 results for various approaches on Object-Instance Retrieval Task

Method	SegVLAD NoNbrAgg	SegVLAD	Segment-to-Global	Global-to-Global
R@1	64.1	**92.7**	30.0	86.4

show that, on the outdoor street-view datasets, SALAD (finetuned DINOv2) generally performs better than AnyLoc (its pretrained counterpart), whereas the latter generally outperforms the former on 'out-of-distribution' datasets. It can be clearly observed that these performance patterns translate well from global- to segment-level results.

5.2 Revisiting Objects of Interest (OOI): Object Instance Retrieval

A typical requirement of an embodied agent is to understand the context of its task through its memory/map information, which is composed of visual and/or semantic cues. For example, navigating to a given object goal requires a robot to visually recognize the goal and not be confused by perceptually-similar items. In this section, we demonstrate our method's ability to retrieve the correct image given just an Object Of Interest (OOI) as a query segment. For this purpose, we use an extended version of the Baidu dataset [70] which annotates OOI as various discriminative areas that can be reliably detected under variable viewpoint and lighting conditions. In total, there are 220 OOI, which cover various things such as logos, brand names, posters, etc., in a highly cluttered mall environment. To cast this dataset in terms of revisiting things, we use the original query images of the Baidu Mall [64] dataset as the database and the images with OOI as the queries. This allows us to evaluate the OOIs directly. This is similar to VPR evaluation of recall in terms of image retrieval but with querying of a specific segment instead of using all the segments of the query image.

We consider four different methods of recognizing known objects in this study. i) `Global-to-Global`: as a baseline method, we use whole images to represent and retrieve, i.e., without using the OOI mask; this resembles object-goal recognition problem for an InstanceImageNavigation task [39]. ii) `Segment-to-Global`: this is the same as the previous setting except that the query image descriptor is aggregated only using the OOI mask; this tests the ability of the image encoder/aggregator to match segment-level descriptor against global descriptors. iii) `SegVLAD` and iv) `SegVLAD NoNbrAgg`, which are our proposed methods but the latter does not use any neighborhood information; this highlights the relevance of spatial context around the OOI for recognition. For SegVLAD, we create a virtual segment mask for the OOI, append it to the other masks of the image, and then perform our neighborhood expansion and feature aggregation, as described in Section 3.

Table 3 reports Recall@1 for different recognition methods. It can be observed that `SegVLAD` outperforms `Global-to-Global` matching by a large margin, which shows that recognizing specific object instances through their images (as in InstanceImageNav) is more prone to failures. It can further be observed from low recall of `SegVLAD NoNbrAgg` that neighborhood aggregation around segments is crucial to capture the required context. Finally, poor recall of `Segment-to-Global` highlights

Table 4: Recall@1/5 for Baidu mall dataset for different aggregation methods and different orders of neighborhood expansion.

Order	SegVLAD	SAP
0	73.1/89.9	**74.6/91.1**
1	77.4/91.7	65.6/87.2
2	76.3/92.4	53.2/81.3
3	**77.7/92.6**	49.8/78.0

Table 5: Recall@1/5 comparison between different methods for ranking images based on segment-level retrieval.

Method	Baidu	AmsterTime
Max Seg	**78.5/93.9**	53.9/70.4
Max Sim	65.2/92.7	34.4/62.4
Ours	**78.5**/93.8	**54.4/76.3**

that matching a part of an image (OOI) with the whole image is not a viable solution for object instance recognition.

5.3 Ablation Studies

Aggregation method & Order of Neighborhood Expansion Previous studies on VPR such as AnyLoc [32] have shown VLAD to be better than other aggregation methods for whole-image based global descriptors. However, in an increasing number of segment-based approaches [12, 20, 24, 61], segment *average* pooling (SAP) is used more commonly. Thus, we compare hard-assignment VLAD against SAP on Baidu dataset. For SAP, we upsample the DINOv2 features to match the resolution of our SAM masks – this upsampling is shown to enhance performance in [61]. For VLAD aggregation, we use our proposed method SegVLAD, where we downsample masks to match with the low resolution of DINOv2. Table 4 shows that SAP performs well for order 0 aggregation (i.e., no neighborhood aggregation) but its performance reduces as the neighborhood expands. On the other hand, SegVLAD has low recall when no neighborhood is considered but benefits significantly even with its immediate neighborhood (order 1). We attribute these inverted trends of SegVLAD and SAP to the very nature of these aggregation methods: as more information becomes available SAP smooths out the overall information content whereas SegVLAD benefits from additional information which gets distributed across its clusters, thus minimizing any possible smoothing effect. It can be observed that R@5 increases for SegVLAD with an increasing order of neighborhood expansion but margins diminish for higher orders. Overall, SegVLAD (order 3) achieves the best results, despite aggregating at a $14\times$ lower resolution than SAP's upsampling based aggregation.

Segment to Image Retrieval Unlike conventional global descriptor retrieval based VPR, we perform retrieval for multiple SuperSegments of the query image. To obtain retrieval output in terms of images (as that is what VPR is typically evaluated on), there exist multiple ways to combine the top segment-level matches across all the query segments. We consider the following alternative options. i) `MaxSeg`: we obtain the best matching segment for each query segment, associate the matched segments to their respective reference image indices, and then rank these image indices based on their

Fig. 4: Qualitative results: Columns respectively represent the query image, correct match of SegVLAD and incorrect match of AnyLoc. Examples from different datasets: AmsterTime, Baidu Mall, Pitts-30K are presented across the rows.

frequency; this method *weakly* resembles an inverted index list based counting of common segments between the query and the reference. ii) `MaxSim`: across the best matching reference segments, we order their image indices based on the segment similarity; this method is similar to that used in MultiVLAD [4]. iii) `Similarity-Weighted Frequency`: this is our proposed method as defined in Section 3.4. Table 5 shows that our proposed method for combining segment-level hypotheses consistently achieves superior results for both the datasets. While MaxSeg achieves a similar performance on Baidu, it suffers a drop in recall for AmsterTime. Both the methods outperform MaxSim at R@1 by a large margin.

Patches vs Segment While open-set segmentation [37] is aimed at a meaningful segregation of visual entities, a simple alternative to our segment-based retrieval is to use uniformly defined regions/patches. Table 6 compares SegVLAD with a patch-based approach, where we consider arbitrary square patch sizes to segment an image. It can be observed that SegVLAD outperforms its patch-based counterparts, where smaller patches perform the worst, and for larger patches, R@1 saturates while R@5 reduces. These results are also in line with similar findings of a recent work [61].

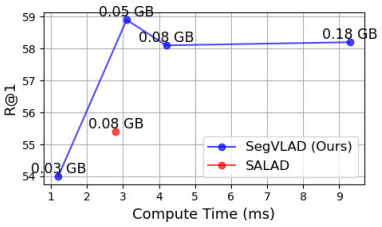

Fig. 5: Recall vs storage/retrieval time on AmsterTime.

Table 6: Patch vs Segments on AmsterTime (Reso. 256×256)

Patch Size	NumSegDb	NumSegIm	R@1/5
16×16	315136	256	35.7/62.1
32×32	78784	64	45.9/72.5
64×64	19696	16	52.0/75.1
128×128	4924	4	53.5/65.6
SegVLAD	129637	105	**56.8/77.7**

5.4 Limitations, Storage, Compute Time & IOU Based Filtering

A key limitation of our method is its large map size, i.e., large storage requirement for segment-level descriptors (see supplementary for further details). In this section, we analyze the resource requirements in terms of database (index) storage and query retrieval time for our method, along with a preliminary study on IOU (Intersection over Union) based filtering of SuperSegments to reduce such costs. We compute IOU between all pairs of SuperSegments in a given image, and remove segments with $IOU(s_i, s_j) > \psi \; \forall \; i \in [1, S], \; j \in [i, S]$, where $\psi \in [0, 1]$ is a threshold on IOU and s_i refers to the list of SuperSegments sorted by their pixel area in a descending order. We only perform this culling on the database segments. We use the outdoor-finetuned DINOv2 backbone for this purpose and compare SegVLAD with SALAD on AmsterTime (see supplementary for additional results on Pitts30K). Figure 5 shows that SegVLAD outperforms SALAD while requiring less storage (annotated on points) and comparable retrieval time (excludes extraction time), using IoU-based filtering threshold ranging from 0.2 to 0.8 (left to right) with a step size of 0.2. In particular, at 0.4 IOU threshold, only 20% of SuperSegments are retained (0.05 GB) while still outperforming the baseline.

5.5 Qualitative Analyses

In this section, we further demonstrate the capabilities of our method through qualitative visualizations. We compare our method against AnyLoc, where the only difference between the two methods is *Global* aggregation/retrieval vs *SuperSegment* aggregation/retrieval. We particularly consider the queries for which our approach successfully retrieved the correct match but AnyLoc failed to do so (additional examples can be found in the supplementary). Figure 4 shows triplets of images in the order of query, correct match (ours), and incorrect match (Anyloc). The segmented part shows one of the correctly matched SuperSegments, displayed as a subgraph in white color overlaid on the corresponding segment masks.

The first row shows a triplet from AmsterTime where our proposed method is able to correctly recognize a subgraph of building across the image pair, whereas Anyloc retrieves an incorrect image of a street-view with buildings, cars, and road, laid out similarly across the image pair. This highlights that a global descriptor can get confused with *perceptually-aliased global context*. In contrast, our SuperSegment based

SegVLAD is not only able to retrieve the correct image but it also correctly finds the mutually-overlapping area. A similar trend follows for the Baidu Mall (middle row) and Pitts30K (last row). In the Baidu Mall example, our approach identifies the piano and the region around it in the query image. It correctly retrieves an image having similar spatial context with the piano. This is akin to how humans use spatial context to identify places. AnyLoc, on the other hand, retrieves an image with similar floor tiles and railings. This example particularly highlights our hypothesis that dissimilarity of non-overlapping regions can dominate the similarity of overlapping regions in global whole-image descriptors. Finally, the Pitts30K example (last row) shows a case of strong viewpoint change. While SegVLAD correctly matches the traffic signal and signboards to retrieve the correct match, AnyLoc retrieves a similar looking image while missing the finer details. This example particularly reinforces the idea of 'revisiting things', as even though the background mountain is common across the triplet, it is the context of the things near the camera/robot which helps in uniquely recognizing a place.

6 Conclusion and Future Work

In this paper, we presented a novel visual place recognition pipeline *SegVLAD* based on image segments-based description and retrieval, which is akin to 'revisting things' as a means to recognize specific instances of what constitute a place. Our proposed *SuperSegments* based image representation and a novel factorization based feature aggregation enables us to effectively represent and retrieve images using our segment similarity-weighted image ranking. Our results show that despite using powerful image encoders such as DINOv2 (pretrained or VPR-finetuned), existing global descriptor based techniques are unable to deal with the challenges of viewpoint variations. In contrast, SegVLAD is able to correctly retrieve images through its ability to match partially-overlapping images with its partial image representation in the form of semi-global subgraphs of segments, i.e., SuperSegments. Thus, our method achieves state-of-the-art results on three diverse datasets (indoor and outdoor) that exhibit strong viewpoint variations on top of other challenges of appearance shift and high perceptual aliasing. Through an additional object instance retrieval study, we demonstrate the unique ability of our method to recognize objects instances within their specific place contexts – an open-set recognition capability that existing VPR methods lack.

Our approach shifts the paradigm in retrieval based VPR research, as the conventional methods predominantly classify into either whole-image global descriptor based coarse retrieval or local feature based geometric reranking. Our approach complements recent concurrent works like MESA [75]; future work can explore a hierarchical VPR pipeline that closely integrates a segment-based coarse retriever with segments-based reranker such as MESA, thus doing away entirely with global whole-image descriptors. Furthermore, segments-based representation with implicitly baked semantics provide a natural way for creating *text-based* interfaces through CLIP [55] and LLMs (Large Language Models) [9], which can be easily integrated with recent efforts in this direction [12, 18, 20, 24, 42, 43, 77].

Acknowledgements

This work was supported by the Centre for Augmented Reasoning (CAR) at the Australian Institute for Machine Learning (AIML), University of Adelaide, Australia. The authors acknowledge the computational support provided by the Indian Institute of Science (IISc), Bengaluru, India, and the International Institute of Information Technology, Hyderabad (IIITH), India. The authors thank Ahmad Khaliq for technical support, Sarah Ibrahimi for sharing the InsideOut dataset, and Martin Humenberger for sharing the OOI annotations for the Baidu Mall dataset.

References

1. Ali-bey, A., Chaib-draa, B., Giguère, P.: Gsv-cities: Toward appropriate supervised visual place recognition. Neurocomputing **513**, 194–203 (2022)
2. Ali-bey, A., Chaib-draa, B., Giguère, P.: Mixvpr: Feature mixing for visual place recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2998–3007 (2023)
3. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5297–5307 (2016)
4. Arandjelovic, R., Zisserman, A.: All about vlad. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1578–1585 (2013)
5. Bar, M.: Visual objects in context. Nature Reviews Neuroscience **5**(8), 617–629 (2004)
6. Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4878–4888 (2022)
7. Berton, G., Mereu, R., Trivigno, G., Masone, C., Csurka, G., Sattler, T., Caputo, B.: Deep visual geo-localization benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5396–5407 (2022)
8. Berton, G., Trivigno, G., Caputo, B., Masone, C.: Eigenplaces: Training viewpoint robust models for visual place recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11080–11090 (October 2023)
9. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
10. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)
11. Cao, B., Araujo, A., Sim, J.: Unifying deep local and global features for image search. In: European Conference on Computer Vision. pp. 726–743. Springer (2020)
12. Chang, M., Gervet, T., Khanna, M., Yenamandra, S., Shah, D., Min, S.Y., Shah, K., Paxton, C., Gupta, S., Batra, D., et al.: Goat: Go to any thing. arXiv preprint arXiv:2311.06430 (2023)
13. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3640–3649 (2016)
14. Chen, Z., Maffra, F., Sa, I., Chli, M.: Only look once, mining distinctive landmarks from convnet for visual place recognition. In: Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on. pp. 9–16. IEEE (2017)

15. Cheng, C., Page, D.L., Abidi, M.A.: Object-based place recognition and loop closing with jigsaw puzzle image segmentation algorithm. In: 2008 IEEE International Conference on Robotics and Automation. pp. 557–562. IEEE (2008)
16. Cummins, M., Newman, P.: Appearance-only slam at large scale with fab-map 2.0. The International Journal of Robotics Research **30**(9), 1100–1123 (2011)
17. Cupec, R., Nyarko, E.K., Filko, D., Kitanov, A., Petrović, I.: Place recognition based on matching of planar surfaces and line segments. The International Journal of Robotics Research **34**(4-5), 674–704 (2015)
18. Folorunsho, S.O.: Semantic segmentation-based approach for autonomous navigation in challenging farm terrains. algorithms **15**, 3 (2024)
19. Garg, S., Fischer, T., Milford, M.: Where is your place, visual place recognition? In: IJCAI (2021)
20. Garg, S., Rana, K., Hosseinzadeh, M., Mares, L., Suenderhauf, N., Dayoub, F., Reid, I.: Robohop: Segment-based topological map representation for open-world visual navigation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA) (2024)
21. Garg, S., Suenderhauf, N., Milford, M.: Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics. In: Proceedings of Robotics: Science and Systems XIV (2018)
22. Garg, S., Sünderhauf, N., Dayoub, F., Morrison, D., Cosgun, A., Carneiro, G., Wu, Q., Chin, T.J., Reid, I., Gould, S., Corke, P., Milford, M.: Semantics for robotic mapping, perception and interaction: A survey. Foundations and Trends® in Robotics **8**(1–2), 1–224 (2020). https://doi.org/10.1561/2300000059, http://dx.doi.org/10.1561/2300000059
23. Gawel, A., Del Don, C., Siegwart, R., Nieto, J., Cadena, C.: X-view: Graph-based semantic multi-view localization. IEEE Robotics and Automation Letters **3**(3), 1687–1694 (2018)
24. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. arXiv preprint arXiv:2309.16650 (2023)
25. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14141–14152 (2021)
26. Hu, H., Qiao, Z., Cheng, M., Liu, Z., Wang, H.: Dasgil: Domain adaptation for semantic and geometric-aware image-based localization. IEEE Transactions on Image Processing **30**, 1342–1353 (2020)
27. Ibrahimi, S., Van Noord, N., Alpherts, T., Worring, M.: Inside out visual place recognition. arXiv preprint arXiv:2111.13546 (2021)
28. Intraub, H.: The representation of visual scenes. Trends in cognitive sciences **1**(6), 217–222 (1997)
29. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition (2023)
30. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3304–3311 (2010)
31. Kassab, C., Mattamala, M., Fallon, M.: Clip-based features achieve competitive zero-shot visual localization. OpenReview preprint arXiv:2306.14846 (2023)
32. Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K.M., Scherer, S., Krishna, M., Garg, S.: Anyloc: Towards universal visual place recognition. IEEE Robotics and Automation Letters (2023)
33. Keetha, N.V., Milford, M., Garg, S.: A hierarchical dual model of environment-and place-specific utility for visual place recognition. IEEE Robotics and Automation Letters **6**(4), 6969–6976 (2021)

34. Khaliq, A., Ehsan, S., Milford, M., McDonald-Maier, K.: A holistic visual place recognition approach using lightweight cnns for severe viewpoint and appearance changes. arXiv preprint arXiv:1811.03032 (2018)
35. Khaliq, A., Xu, M., Hausler, S., Milford, M., Garg, S.: Vlad-buff: Burst-aware fast feature aggregation for visual place recognition. In: European Conference on Computer Vision. Springer (2024)
36. Khorasgani, S.H., Chen, Y., Shkurti, F.: Slic: Self-supervised learning with iterative clustering for human action videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16091–16101 (2022)
37. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
38. Knopp, J., Sivic, J., Pajdla, T.: Avoiding confusing features in place recognition. In: European Conference on Computer Vision. pp. 748–761. Springer (2010)
39. Krantz, J., Lee, S., Malik, J., Batra, D., Chaplot, D.S.: Instance-specific image goal navigation: Training embodied agents to find object instances. arXiv preprint arXiv:2211.15876 (2022)
40. Le, D.C., Youn, C.H.: City-scale visual place recognition with deep local features based on multi-scale ordered VLAD pooling. arXiv preprint arXiv:2009.09255 (2020)
41. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
42. Maalouf, A., Jadhav, N., Jatavallabhula, K.M., Chahine, M., Vogt, D.M., Wood, R.J., Torralba, A., Rus, D.: Follow anything: Open-set detection, tracking, and following in real-time. IEEE Robotics and Automation Letters **9**(4), 3283–3290 (2024)
43. Manglani, S.: Real-time vision-based navigation for a robot in an indoor environment. arXiv preprint arXiv:2307.00666 (2023)
44. Masone, C., Caputo, B.: A survey on deep visual place recognition. IEEE Access **9**, 19516–19547 (2021)
45. Mirjalili, R., Krawez, M., Burgard, W.: Fm-loc: Using foundation models for improved vision-based localization. In: Robotics and Automation (ICRA), 2023 IEEE International Conference on. IEEE (2023)
46. Mousavian, A., Kosecka, J.: Semantic image based geolocation given a map (author's initial manuscript). Tech. rep., George Mason University Fairfax United States (2016)
47. Naseer, T., Oliveira, G.L., Brox, T., Burgard, W.: Semantics-aware visual localization under challenging perceptual conditions. In: IEEE International Conference on Robotics and Automation (ICRA) (2017)
48. Oliva, A.: Gist of the scene. Neurobiology of attention **696**(64), 251–258 (2005)
49. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
50. Paolicelli, V., Tavera, A., Masone, C., Berton, G., Caputo, B.: Learning semantics for visual place recognition through multi-scale attention. In: Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II. pp. 454–466. Springer (2022)
51. Pion, N., Humenberger, M., Csurka, G., Cabon, Y., Sattler, T.: Benchmarking image retrieval for visual localization. In: 2020 International Conference on 3D Vision (3DV). pp. 483–494. IEEE (2020)
52. Puligilla, S.S., Tourani, S., Vaidya, T., Parihar, U.S., Sarvadevabhatla, R.K., Krishna, K.M.: Topological mapping for manhattan-like repetitive environments. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 6268–6274. IEEE (2020)

53. Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Revisiting oxford and paris: Large-scale image retrieval benchmarking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5706–5715 (2018)
54. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1655–1668 (2018)
55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
56. Revaud, J., Almazán, J., Rezende, R.S., Souza, C.R.d.: Learning with average precision: Training image retrieval with a listwise loss. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5107–5116 (2019)
57. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. arXiv preprint arXiv:1812.03506 (2018)
58. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., M, P., J, S., F, K., T, P.: Benchmarking 6dof outdoor visual localization in changing conditions. In: Proc. CVPR (2018)
59. Schleiss, M., Rouatbi, F., Cremers, D.: Vpair – aerial visual place recognition and localization in large-scale outdoor environments (2022)
60. Schubert, S., Neubert, P., Garg, S., Milford, M., Fischer, T.: Visual place recognition: A tutorial. RAM (2023)
61. Shlapentokh-Rothman, M., Blume, A., Xiao, Y., Wu, Y., TV, S., Tao, H., Lee, J.Y., Torres, W., Wang, Y.X., Hoiem, D.: Region-based representations revisited. arXiv preprint arXiv:2402.02352 (2024)
62. Shubodh, S., Omama, M., Zaidi, H., Parihar, U.S., Krishna, M.: Lip-loc: Lidar image pretraining for cross-modal localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 948–957 (2024)
63. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of International Conference on Computer Vision (ICCV). p. 1470. IEEE (2003)
64. Sun, X., Xie, Y., Luo, P., Wang, L.: A dataset for benchmarking image-based localization. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5641–5649 (2017), https://api.semanticscholar.org/CorpusID:20531893
65. Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., Milford, M.: Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. Proceedings of Robotics: Science and Systems XII (2015)
66. Torii, A., Sivic, J., Pajdla, T., Okutomi, M.: Visual place recognition with repetitive structures. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 883–890 (2013). https://doi.org/10.1109/CVPR.2013.119
67. Tsintotas, K.A., Bampis, L., Gasteratos, A.: The revisiting problem in simultaneous localization and mapping: A survey on visual loop closure detection. IEEE Transactions on Intelligent Transportation Systems **23**(11), 19929–19953 (2022)
68. Wang, R., Shen, Y., Zuo, W., Zhou, S., Zheng, N.: Transvpr: Transformer-based place recognition with multi-level attention aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13648–13657 (2022)
69. Warburg, F., Hauberg, S., López-Antequera, M., Gargallo, P., Kuang, Y., Civera, J.: Mapillary street-level sequences: A dataset for lifelong place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2626–2635 (2020)
70. Weinzaepfel, P., Csurka, G., Cabon, Y., Humenberger, M.: Visual localization by learning objects-of-interest dense match regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5634–5643 (2019)

71. Xin, Z., Cai, Y., Lu, T., Xing, X., Cai, S., Zhang, J., Yang, Y., Wang, Y.: Localizing discriminative visual landmarks for place recognition. In: 2019 International conference on robotics and automation (ICRA). pp. 5979–5985. IEEE (2019)
72. Yildiz, B., Khademi, S., Siebes, R.M., Van Gemert, J.: Amstertime: A visual place recognition benchmark dataset for severe domain shift. In: 2022 26th International Conference on Pattern Recognition (ICPR). IEEE (Aug 2022). https://doi.org/10.1109/icpr56361.2022.9956049, http://dx.doi.org/10.1109/ICPR56361.2022.9956049
73. Yu, J., Zhu, C., Zhang, J., Huang, Q., Tao, D.: Spatial pyramid-enhanced netvlad with weighted triplet loss for place recognition. IEEE Trans. Neural Netw. Learn. Syst. **31**(2), 661–674 (2019)
74. Zaffar, M., Garg, S., Milford, M., Kooij, J., Flynn, D., McDonald-Maier, K., Ehsan, S.: Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. International Journal of Computer Vision pp. 1–39 (2021)
75. Zhang, Y., Zhao, X.: Mesa: Matching everything by segmenting anything. arXiv preprint arXiv:2401.16741 (2024)
76. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
77. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in Neural Information Processing Systems **36** (2024)