Dilated Superpixel Aggregation for Visual Place Recognition

Zichao Zeng, June Moh Goo, and Jan Boehm

***Abstract*—Visual Place Recognition (VPR) is a fundamental task in robotics and computer vision, enabling systems to identify locations seen in the past using visual information. Previous state-of-the-art approach focuses on encoding and retrieving semantically meaningful supersegment representations of images to significantly enhance recognition recall rates. However, we find that they struggle to cope with significant variations in viewpoint and scale, as well as scenes with sparse or limited information. Furthermore, these semantic-driven supersegment representations often exclude semantically meaningless yet valuable pixel information. In this work, we present Sel-V and MuSSel-V, two efficient variants within the segment-level VPR paradigm that replace heavy and fragmented supersegments with lightweight, visually compact and complete dilated superpixels for local feature aggregation. The use of superpixels preserves pixel-level details while reducing computational overhead. A multi-scale extension further enhances robustness to viewpoint and scale changes. Comprehensive experiments on twelve public benchmarks show that our approach achieves a better trade-off between accuracy and efficiency than existing segment-based methods. These results demonstrate that lightweight, non-semantic segmentation can serve as an effective alternative for high-performance, resource efficient visual place recognition in robotics. The code will be available in https://zichaozeng.github.io/MuSSel-V/.**

***Index Terms*—Localization, Vision-Based Navigation, Visual Place Recognition, Superpixel, Aggregation**

I. INTRODUCTION

VISUAL Place Recognition (VPR), which is closely related to image-based localisation [1] or visual geo-localisation [2], seeks to estimate the location of a query image by identifying the most relevant reference image from a geographically annotated image database. VPR is typically formulated as an image retrieval task, using global or local descriptors for localisation by nearest-neighbour matching [3], [4]. Last decade, researchers have focused on learning or fine-tuning image encoders to ensure that global descriptors exhibit invariance to appearance variations [3], viewpoint changes [3], [5], [6], and scene clutter [7]. Vector of Locally Aggregated Descriptors (VLAD) [8] and its advanced variants [3], [9], [10] have gained popularity by aggregating local features into compact global representations. Simultaneously, the feature extraction methods for VPR have transitioned from traditional handcrafted techniques [8] to deep learning based approaches [3], and more recently, to pre-trained or fine-tuned foundation models [7], [11], [12]. Compact global descriptors enable efficient retrieval and robustness to viewpoint changes but often lack spatial information, making them prone to perceptual aliasing [1], [12], [13].

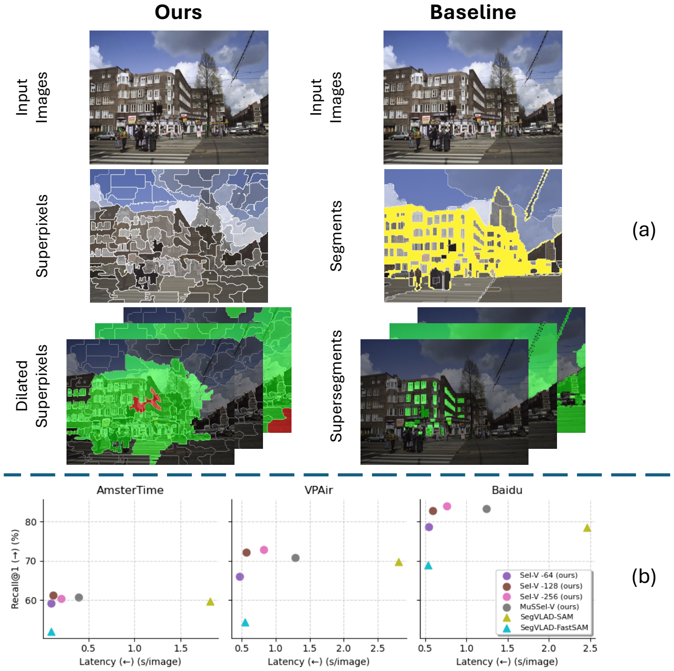


Fig. 1. (a) Comparison of our segment-level method with a baseline approach for VPR. SegVLAD [13] based on SAM [14] discards pixels shown in yellow, while our method employs dilated superpixels for a more compact representation. (b) Recall@1 and Latency for Sel-V and MuSSel-V (Ours) vs. SegVLAD on AmsterTime, VPAir, and Baidu, showing better recall and faster compute for our methods.

A promising solution is the direct use or fusion of local descriptors. This approach primarily revolves around two-stage VPR frameworks, where geometric verification of local feature matches is employed to re-rank candidate results [4], [15], [16]. Numerous approaches have been developed to characterise images through different elements, including segments [17], [18], linear and planar structures [19], objects [20], [21], or partitioned sections [15], bridging spatial information and global descriptors. Despite advancements, most methods focus on improving retrieval accuracy through re-ranking rather than addressing the limitations of global descriptors. In contrast, MultiVLAD extracts VLADs at multiple image scales instead of relying on a single global descriptor [9]. Building on this, SegVLAD uses the Segment Anything Model (SAM) [14] to generate segment-level local descriptors, achieving state-of-the-art performance in VPR [13]. However, SegVLAD discards many seemingly insignificant but crucial segments,

Manuscript received: August 8, 2025; Revised: November 2, 2025; Accepted: November 23, 2025.

This paper was recommended for publication by Editor Pascal Vasseur upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported by the Engineering and Physical Sciences Research Council through an industrial CASE studentship with Ordnance Survey (Grant number EP/W522077/1 and EP/X524840/1). *(Corresponding author: Zichao Zeng.)*

The authors are with Department of Civil, Environmental and Geomatic Engineering, University College London, WC1E 6BT London, U.K. zichao.zeng.21@ucl.ac.uk; june.goo.21@ucl.ac.uk; j.boehm@ucl.ac.uk

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

resulting in a notable loss of approximately 20% of pixels on representative datasets (see Figure 1). Moreover, its reliance on a heavy segmentation model increases computational overhead, making the pipeline less efficient.

In this paper, we propose an efficient variant under segment-level VPR paradigm using **S**uperpix**el**-based **V**LAD (**Sel-V**). We replace heavy segmenters based on transformer models with light-weight superpixel methods to segment images by *visual homogeneity* based on low-level features (colour or texture). Next, similar to SegVLAD [13], superpixels are expanded by growing to neighbouring superpixels, which form cohesive dilated superpixels. Subsequently, local features based on pre-trained or fine-tuned DINOv2 are aggregated according to the dilated superpixels. In addition, we provide a **Mu**lti-**S**cale **S**uperpix**el** **V**LAD (**MuSSel-V**), which is more robust under different environments and scale variations. Both our proposed approaches **Sel-V** and **MuSSel-V** are much faster than the current leading segment-level VPR method - SegVLAD. Our main contributions are:

- We present **Sel-V**, an efficient variant within the segment-level VPR paradigm that uses *dilated superpixels*, enabling richer local feature aggregation and faster computation. Its multi-scale extension, **MuSSel-V**, further enhances robustness to viewpoint and scale variations.
- We evaluate our methods on 12 benchmarks spanning diverse scenarios, showing clear improvements over State-of-the-art (SOTA) methods, particularly on aerial, seasonally varying, and long time span datasets where the proposed dilated superpixel aggregation better preserves complete local information under large appearance changes.
- Compared to previous segment-level VLAD methods, our approach is more efficient, balancing visual compactness and descriptor completeness without relying on heavy transformer-based models in the segmentation stage.

II. Related Work

A. Visual place recognition (VPR)

Hand-crafted methods, for feature extraction, were frequently used in early VPR approaches, typically aggregated using methods like Gist [22], Bag of Words (BoW) [23], and VLAD [8]. While these methods demonstrated effectiveness in controlled scenarios, they struggled with challenges such as viewpoint variations, illumination changes, and large-scale place retrieval [24]. The use of deep learning drastically improved VPR, enabling more discriminative and invariant representations. Pioneering works like NetVLAD [3] established a strong foundation for feature aggregation in neural networks, inspiring the development of more advanced models. Recent methods, including CosPlace [2], MixVPR [25], EigenPlaces [5], and TransVPR [16], have further refined representation learning through improved dataset design, objective functions, and aggregation mechanisms. More recently, advanced methods [7], [10]–[13], [26] have replaced traditional feature extraction backbones with foundation models such as CLIP [27] and DINO [28], [29], significantly enhancing model generalisation and robustness across diverse environments.

B. Region-based aggregation

Some methods have shifted focus to the regional level to enhance the representation with region-level information, aggregating images into single or connected segment-level compact representations [30], [31]. Other approaches [4], [12], [15], [16] create multiple features for each image, performing local match-based reordering. Some of the latest contributions use multiple segment descriptors per image to retrieve them directly from a segment database without reordering. The motivation is that improving the coarse retriever itself directly enhances performance, including subsequent reordering [13]. The idea of MultiVLAD [9] is to retrieve multiple features per query image, based on arbitrarily defined regions. Alternatively, SegVLAD [13] uses semantically meaningful image segments obtained from SAM [14]. Segment-level aggregation and retrieval again boost performance. However, some semantically meaningless segments — often discarded by semantic segmentation frameworks like SAM — play a crucial role in achieving robust place recognition across varying scales. VPR often benefits more from robust, low-level features and spatial groupings that capture local textures and structural cues rather than detailed semantic categories [32]. Moreover, the segments of SAM often contain non-segmented gaps, which can introduce additional uncertainty in determining object boundaries, particularly in cases where the model struggles to assign confidence to ambiguous regions. SAM generates multiple high-quality segmentation masks per image, making it computationally expensive and time-consuming [33], [34]. Some effective methods, such as FastSAM [35], provide faster versions, but increase uncertainty and sacrifice the accuracy of masks, resulting in less favourable VPR [13].

C. Superpixel mechanisms

A superpixel is a group of neighbouring pixels with *visually homogeneous* characteristics such as colour, brightness, or texture, creating a consistent area within an image. By combining pixels into significant regions, the image representation transitions from millions of separate pixels to merely a few hundred or thousand superpixels. This simplification not only reduces computational demands but also yields a semantically enriched, mid-level representation. Such representation is advantageous for multiple computer vision applications, including medical imaging, remote sensing, and robotics. Several algorithms have been developed to generate superpixels efficiently. Commonly used superpixel algorithms include SLIC [36], which uses modified k-means to group pixels by colour and spatial proximity. SEEDS [37] stands out with its block-wise exchange mechanism, enabling rapid initial boundary adjustment followed by fine, detailed updates for superior overall segmentation. Unlike SAM, superpixel algorithms such as SLIC or SEEDS operate using lightweight clustering or iterative refinement techniques, which are significantly faster and more memory-efficient [36], [37].

III. Proposed Framework

Building on previous studies [7], [9]–[13], [26], we retain a universal foundation model as the backbone and introduce

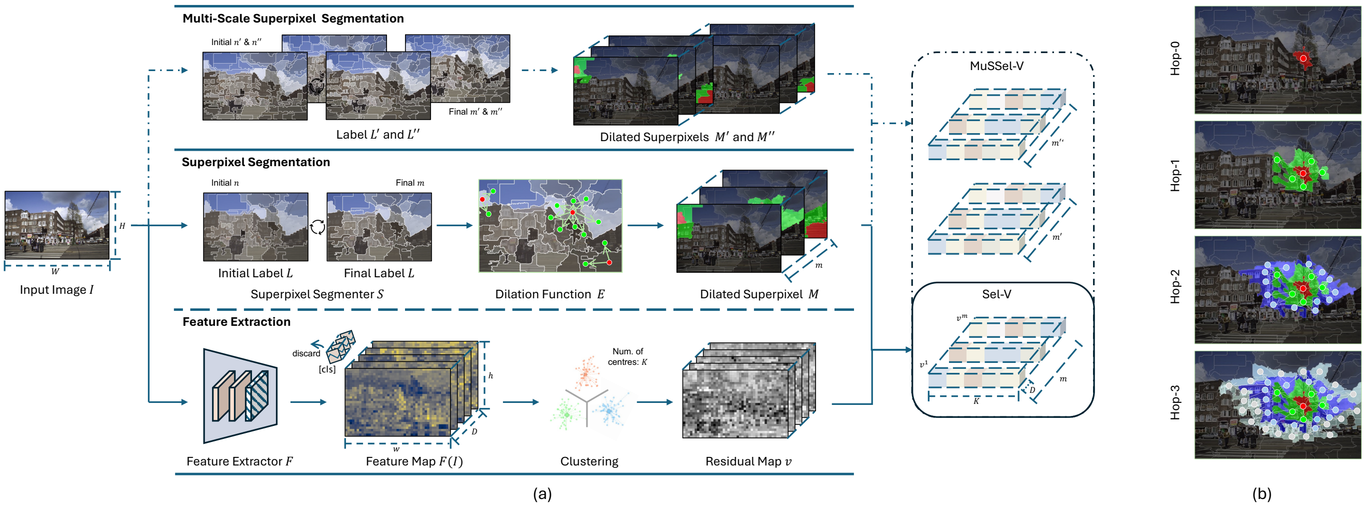


Fig. 2. (a) Pipeline of our approach Sel-V and its multi-scale variant MuSSel-V. (b) Dilation function from hop-0 to hop-3.

Sel-V, a segment-level VLAD for VPR. The segmenter uses superpixels based on image attributes (colour, luminance, texture) instead of rectangular cutting or learned features [14], [35], efficiently generating cohesive segments. Inspired by SegVLAD [13], we introduce mechanisms to dilate the superpixels.

To enhance generalisation and robustness, we propose **MuSSel-V**, a hyperparameter-free multi-scale version inspired by previous works [9], [15]. For **Sel-V**, the retrieval strategy uses the similarity-weighted ranking method [13] to evaluate the similarity between the query image and the database image. For **MuSSel-V**, similarity scores of different scales are accumulated at equal weight.

A. Problem Definition

We deviate from methods which directly design a function $\mathcal{V}$ to act as an image representation generator. Our approach is similar to SegVLAD in that we design a classifier S, a feature extractor F, and a segment-level descriptor function V shown in Figure 2a. Given an image I of size $H \times W$, the segmenter S^n generates a superpixel labelling matrix $L \in \{0, 1, \ldots, m-1\}^{H \times W}$. Subsequently, dilated superpixels $M_{m \times H \times W}$ are generated based on the neighbourhood matrix R. In parallel, the extractor extracts a pixel-level feature map $F(I) \in \mathbb{R}^{h \times w \times D}$ of dimension D. Local feature descriptors v (residuals) are then obtained by assigning pixel features to their nearest cluster centres and computing the differences (residuals) between them. In the end, the dilated superpixel representation function V^s aggregates the local feature descriptors v based on the dilated superpixels M. A set of dilated superpixel VLADs $\mathcal{V}(I) = [V^1, V^2, ..., V^m]$, each with the same vector size, is introduced for image I.

For the reference, the function $\mathcal{V}$ extracts dilated superpixel representations $\mathcal{D} = \{\mathcal{V}r_1, \mathcal{V}r_2, \mathcal{V}_{r_3}, ...\}$ offline, where $\mathcal{V}_{r_1} = [V^1, V^2, ..., V^{m_{r_1}}]$. For a query image, the descriptor $\mathcal{V}_q = [V^1, V^2, ..., V^{m_q}]$ is computed online by concatenating all VLADs. At retrieval time, similarity is measured using Euclidean distance between query and database segments. The combined score of the `Top-50` most similar segment matches determines the ranking of database images.

B. Feature Extractor

We adopt ViT-based *DINOv2* as our feature extraction backbone. ViT contains many layers, and each layer has multiple facets (queries, keys, values, and tokens) from which features can be extracted. We follow AnyLoc [7] and select the middle `[layer]` from *DINOv2* to extract `[value]` tokens and discard `[CLS]` tokens. We focus on pixel-level features to allow fine matching rather than extracting global features for each image (i.e., one global feature vector-`[CLS]` token for the entire image).

F_{DINOv2} initially divides an input image $I^{H \times W}$ into $p \times p$ patches, with $p = 14$. These patches are sequentially encoded, resulting in the output tokens $F(I) \in \mathbb{R}^{h \times w \times D}$, where $h = H/p$, and $w = W/p$.

C. Multi-Scale and Dilated Superpixel

We adopt superpixels to partition the image. This method utilises low-level features to split the image into distinct, compact, and visually homogeneous segments represented in a single label matrix. Using superpixels significantly reduces computational and storage costs. We choose SEEDS [37] as the core method of our framework, which primarily relies on image colour histograms and spatial constraints. We define the superpixel segmenter $L = S^n(I)$ and the segmentation process as follows:

$$L \in \{0, 1, \ldots, m-1\}^{H \times W}, m \in (1, n] \tag{1}$$

where the image I is partitioned into n expected superpixels. During SEEDS's iterative process adaptive region merging and additional refinement steps are performed. As a consequence, the actual number of superpixels, denoted as m, can be lower than the expected number n.

For **Sel-V**, n controls the output resolution of I, where $n = 128$, with the specific choice varying based on task requirements and dataset conditions. For **MuSSel-V**, we extract three superpixel label maps at different scales: L_n, $L_{n'}$, and $L_{n''}$, where $n, n', n'' \in \{64, 128, 256\}$. Next, based on the label map L, we establish a neighbourhood relationship matrix to determine the dilation of superpixels.

To maintain comparability with SegVLAD, we also employ Delaunay Triangulation as our expansion function E^o. The neighbourhood relationship matrix R is defined as:

$$R = E^o(L_i), R \in \{0,1\}^{m \times m} \tag{2}$$

where $R_{i,j} = 1$ indicates that superpixel i is adjacent to superpixel j, while $R_{i,j} = 0$ otherwise. The parameter o controls the hop (order) of neighbourhood expansion. A lower value of o restricts connections to immediate neighbours, while a higher value incorporates broader contextual information.

Since F_{DINOv2} is extracted at a lower spatial resolution $h \times w$ than the superpixel labels $L \in \mathbb{R}^{H \times W}$, we downsample each label map to the F_{DINOv2} grid before VLAD. Each patch inherits the label of any superpixel that occupies at least one pixel within its receptive field ("existence assignment"), ensuring that all patches are associated with a valid segment. This mapping aligns the dense local features with the label space while maintaining the spatial consistency of segment boundaries.

Given L and R, we construct dilated superpixels M, which expand each superpixel based on its neighbourhood connectivity shown in Figure 2b. The dilatation process is formulated as:

$$M_{m \times H \times W} = \mathbb{I}(R \times L_{\text{one-hot}}) \tag{3}$$

$$M_i = \sum_{j=0}^{m-1} R_{i,j} \times \mathbb{I}_{(L==j)} \tag{4}$$

where $L_{\text{one-hot}}$ is the one-hot encoding of superpixel labels, and $\mathbb{I}(R \times L_{\text{one-hot}})$ is a binarisation function ensuring non-overlapping contributions in feature aggregation. Specifically, when dilation causes multiple neighbouring segments to overlap on the same patch, the binarisation operator assigns the patch to each dilated superpixel at most once, preventing duplicated counting from overlapping neighbourhoods. $\mathbb{I}(L == j)$ is a binary mask indicating pixels belonging to superpixel j, and $R_{i,j}$ determines whether j falls within the expanded neighbourhood of i. This ensures that the dilated superpixels maintain compactness while effectively aggregating local contextual information. For **MuSSel-V**, this expansion is applied independently to L_n, $L_{n'}$, and $L_{n''}$, producing M, M', and M''.

D. Segment-Level Aggregated Descriptor

After obtaining the feature map F and the dilated superpixels M, we aggregate local features using a VLAD-based encoding scheme tailored to the segment-level representation.

We construct a visual dictionary with K cluster centres, assigning each local feature to its closest cluster based on cosine similarity. For each dilated superpixel M_s, we compute the accumulated residuals for cluster k as v_k^s, capturing the difference between local features and their nearest clusters. To construct a dilated superpixel-wise VLAD descriptor, we concatenate residual vectors for all VLAD-clusters covered by a dilated superpixel mask:

$$V^s = [v_1^s, v_2^s, ..., v_K^s] \tag{5}$$

where K is the number of clusters. This process uses Hard-VLAD, assigning each feature only to the nearest visual word.

For an image I, the final image descriptor is obtained by concatenating the superpixel-level VLAD vectors:

$$\mathcal{V}(I) = [V^1, V^2, ..., V^{m_t}] \tag{6}$$

where m denotes the total number of dilated superpixels in the image. For Sel-V, m_t is m from a defined scale. For MuSSel-V, to incorporate multi-scale superpixel representations, we compute and aggregate residual vectors within each dilated superpixel across three scales, resulting in $M_t = M \cup M' \cup M''$.

E. Matching and Ranking

For retrieval, following [7], [13], we employ the FAISS flat index [38] to match each query superpixel descriptor against all superpixels extracted from the reference database. All descriptors are L2-normalised before indexing, so Euclidean distance in FAISS (or the equivalent inner-product search) is consistent with cosine similarity. The flat index provides exact nearest-neighbour retrieval and ensures fair comparison across methods. To evaluate retrieval performance at the image level, we aggregate the retrieved superpixel matches into reference image indices using a weighted frequency measure. This allows us to transition from local superpixel matches to global image-level retrieval.

For **Sel-V**, for each query image I_q, we retrieve the top-$\mathcal{K}$ ($\mathcal{K} = 50$) dilated superpixel matches for each of its dilated superpixels s across all reference images. Each retrieved match corresponds to a reference superpixel belonging to an image in the reference database. These matches are then mapped to their respective reference image indices r_j, where we aggregate the similarity scores θ to compute a cumulative image similarity score:

$$\hat{\theta}\left(r_j\right) = \sum_{s \in M_t} \sum_{i=1}^{\mathcal{K}} \theta_{si} \times \mathbb{I}_{\{r_{si}=r_j\}} \tag{7}$$

where $M_t = M$ in **Sel-V** which represents the set of dilated superpixels in the query image, θ_{si} is the similarity score of the i-th retrieved superpixel match for superpixel s, and $\mathbb{I}_{\{r_{si}=r_j\}}$ is an indicator function that counts matches belonging to reference image r_j. For **MuSSel-V**, we extend the retrieval process to multiple superpixel scales, i.e. $M_t = M \cup M' \cup M''$. Finally, we rank the reference images based on their similarity scores and select the top-ranked image match.

IV. Experiments

A. Implementation Details

a) Datasets: In this work, we build upon several recent studies and benchmark evaluations. Our experiments are conducted on a diverse range of datasets spanning multiple domains. These datasets cover a broad spectrum of scenarios

TABLE I
COMPARISON OF RECALL@N PERFORMANCE AGAINST BASELINES IN DIVERSE DOMAINS ACROSS INDOOR, CAVE AND AERIAL SCENARIOS. WE COMPARE SEL-V AND MUSSEL-V WITH PRE-TRAINED DINOV2 (VIT-G) WITHOUT TUNING AGAINST TWO RECENT ADVANCED BASELINES [7], [13] WITH THE SAME BACKBONE.

Method	Segment	Tuned	17Places H		17Places E		Baidu		VPAir		Laurel		Hawkins	
			R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
AnyLoc	-	-	65.0	80.5	95.3	97.3	75.2	87.6	66.7	79.2	**61.6**	90.2	**65.2**	94.1
SegVLAD	✓	-	64.8	80.3	95.3	**98.0**	78.5	93.8	69.8	83.7	28.6	65.2	52.5	94.1
MuSSel-V	✓	-	**65.3**	79.6	**95.6**	97.8	**83.3**	**95.0**	70.9	85.8	51.8	**91.1**	61.9	**97.5**
Sel-V	✓	-	**65.3**	**80.8**	95.3	97.8	82.8	**95.0**	**72.1**	**86.0**	54.5	**91.1**	63.6	95.8

TABLE II
COMPARISON OF RECALL@N PERFORMANCE AGAINST BASELINES IN OUTDOOR ENVIRONMENTS WHEN USING FINE-TUNING. WE COMPARE SEL-V AND MUSSEL-V WITH FINE-TUNED DINOV2 (VIT-B) AGAINST ANYLOC [7] WITH PRE-TRAINED DINOV2 AND FIVE POPULAR BASELINES [5], [11], [13], [25] TRAINED OR FINE-TUNED ON STREET VIEWS.

Method	Segment	Tuned	AmsterTime		SF-XL Val		Pitts-30k		SPED		MSLS-C		Nordland	
			R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
AnyLoc	-	-	50.3	73.0	84.4	91.9	87.7	94.7	85.3	94.4	42.2	53.5	16.1	25.4
CosPlace	-	✓	47.7	69.8	94.6	97.6	90.4	95.7	80.1	89.6	67.2	78.0	44.2	59.7
Mix VPR	-	✓	40.2	59.1	87.8	93.8	91.5	95.5	85.2	92.1	64.0	75.9	58.4	74.6
EigenPlaces	-	✓	48.9	69.5	**96.4**	**98.2**	92.6	96.7	69.9	82.9	67.4	77.1	54.4	68.8
SALAD	-	✓	55.4	75.6	93.6	97.3	92.6	96.5	**92.1**	96.2	75.0	88.0	76.0	89.2
SegVLAD	✓	✓	59.8	77.2	94.9	98.1	93.2	96.8	91.3	95.2	**78.0**	88.4	65.3	77.6
MuSSel-V	✓	✓	60.8	79.9	94.8	98.1	**93.4**	**96.9**	91.3	96.2	77.2	88.8	72.0	87.1
Sel-V	✓	✓	**61.3**	**80.2**	95.4	98.0	93.3	**96.9**	91.6	**96.5**	77.3	**89.2**	**81.8**	**93.3**

and demonstrate significant variations in intra-dataset characteristics. These datasets include indoor scenes under different conditions 17Places [39] with Easy and Hard tolerance, Baidu [40] and Hawkins [41], aerial images VPAir [42], cave scenes Laurel [41] and also images of various changing outdoor environments AmsterTime [6], Pitts30K [43], SF-XL [2], MSLS-C (MSLS Challenge) [44], SPED [45], and Nordland [46].

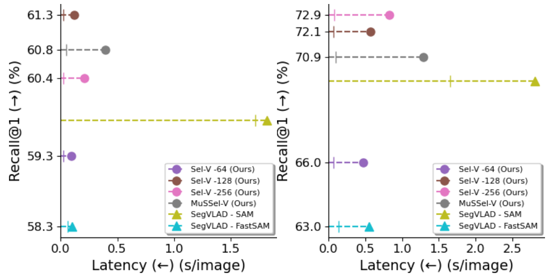
b) Architecture: To ensure a comprehensive evaluation, we follow SegVLAD's benchmarking protocol and compare our method across two different backbone configurations:

- Utilising an off-the-shelf pre-trained DINOv2 (ViT-G) backbone following AnyLoc [7] for feature extraction.
- Employing a fine-tuned DINOv2 (ViT-B) backbone fine-tuned on street views provided by SegVLAD [13] integrated SALAD [11] and NetVLAD [3].

These two frameworks allow us to benchmark our approach against foundation model-based methods in general VPR benchmarks, as well as against training/fine-tuning-based methods in city-scale urban scene benchmarks.

We use SEEDS [37] for image segmentation with initial superpixel counts of 128 for Sel-V, while MuSSel-V combines all three scales. All other modules follow SegVLAD [13] for fair comparison. Our VLAD clustering uses 32 cluster centres. Our dilated superpixel encoding employs Delaunay Triangulation with an expansion order of 3. The original 49,152-dimensional vectors are reduced to 1024 via PCA, following prior works [7], [13]. Cluster centres are built only from the reference map dataset [7] without external data. All experiments use the same random seed (42), AMD Ryzen Threadripper PRO 5975WX CPU, and NVIDIA RTX A6000 GPU for consistency.

c) Evaluation: We use Recall@N as the main evaluation metric, common in VPR benchmarks. It assesses the fraction of queries with a correct match among the top-N results. We also compare our method with other segment-level VLAD methods for both offline preprocessing costs and online query processing costs.



(a) AmsterTime. (b) VPAir.

Fig. 3. Comparison of Sel-V and MuSSel-V (circular markers) with baseline method [13] using SAM and FastSAM (triangular markers). We show Recall@1 performance over latency including segmentation (first segment), extraction and retrieval (second segment) time across two datasets.

B. State-of-the-Art Comparison

a) Baselines: We compare our approach with SOTA VPR methods. The baselines can be categorised into two groups. The first group consists of recent advanced methods that leverage visual foundation models, particularly utilising DINOv2 [29] as the backbone. Among them, AnyLoc [7] employs an off-the-shelf DINOv2 model without additional task-specific training. Another key baseline in our comparison is SegVLAD [13], which is highly relevant to our approach as it applies segment-level VLAD for VPR. We apply SegVLAD with pre-trianed DINOv2 for diverse domains. The second group comprises methods specifically trained or fine-tuned for the VPR task on large-scale urban datasets, including CosPlace [2], MixVPR [25], EigenPlaces [5], SALAD [11], and SegVLAD [13] with fine-tuned DINOv2. These approaches

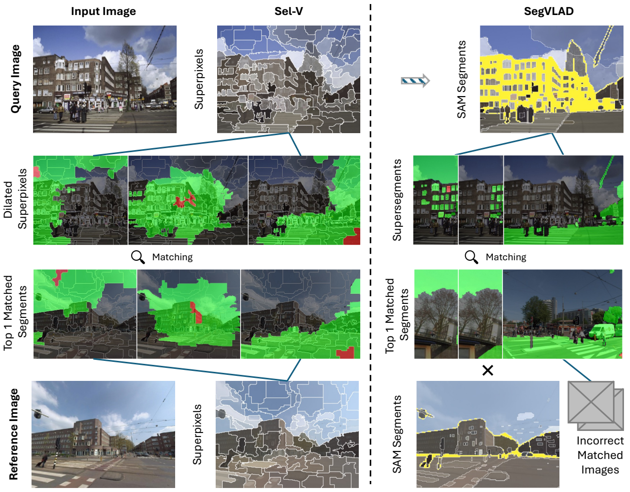


Fig. 4. Qualitative results. A challenging example from AmsterTime demonstrates the advantages of our method (left), which utilises compact dilated superpixels comprising multiple visually homogeneous regions to enhance robustness in retrieval. Meanwhile, a semantically inspired segmentation method (right) experiences pixel loss (shown in yellow) which can lead to matching errors.

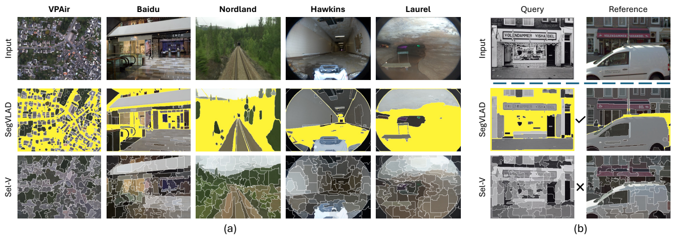


Fig. 5. (a) Qualitative results across different datasets. SegVLAD exhibit significant pixel loss (highlighted in yellow), leading to reduced feature representation and potential retrieval errors, while Sel-V preserves all local details. (b) A failure case where SegVLAD using SAM is able to extract the signboard in the query image in a single large segment. Sel-V on the other hand splits it into many smaller segments. Similar effects can be observed in the reference image with the occluding object (car). This leads to Sel-V missing the correct match.

are optimised for city-scale place recognition and serve as strong task-specific baselines.

b) Discussion of results: As shown in Table I, Sel-V and MuSSel-V outperform other methods in R@1/5 on complex indoor datasets like 17Places (Hard), Baidu, and VPAir. Even the simplified Sel-V achieves notable gains on these datasets. MuSSel-V also shows consistent improvements. Both Sel-V and MuSSel-V achieve near-perfect performance on 17Places (Easy). This validates the effectiveness of our dilated superpixel representation when using an off-the-shelf foundation model. Although our method falls short in R@1 in challenging environments like Laurel and Hawkins, it achieves substantial gains in R@5, suggesting that global representations might be preferable in simpler environments. Notably, both Sel-V and MuSSel-V outperform SegVLAD, achieving gains of well over 20% on Laurel. Gains on Hawkins are lower but still significant. Furthermore, we compared our method with other baselines [2], [5], [11], [25], and the results remain consistent with those of SegVLAD [13] - zero-shot inference substantially outperforms these task-specific approaches.

In outdoor scenarios shown in Table II, Sel-V and MuSSel-V using fine-tuned DINOv2 show good performance as well. On AmsterTime, which involves diverse scales, temporal variations, and grayscale-to-RGB transitions, Sel-V performs best in R@1 and R@5 by 1.5%/3%. For Nordland, featuring seasonal changes but uniform scenes, it gains 5.8%/4.1%. On the Pitts benchmark, both methods achieve top-tier results. On SF-XL, a dataset with significant viewpoint variations, our methods, despite not being trained on it, perform second only to SF-XL-trained EigenPlaces. For SPED and MSLS-C, while our R@1 ranks second, R@5 shows slight improvements. These results demonstrate the strong generalisation capability of our dilated superpixel representation, ensuring consistent performance across diverse conditions.

C. Comparison Against Segment-Level Methods

Our method demonstrates significant improvements over existing segment-level representation methods, with Sel-V achieving an average gain of 5.1%/4.2% in R@1/5, while MuSSel-V improves by 4.2%/3.8%. This suggests that our compact superpixel-based segmentation may offer advantages over more scattered, semantically guided segments. We performed paired t-tests over five runs (without a fixed random seed) to assess statistical reliability on the VPAir and AmsterTime datasets. The observed improvements are statistically significant ($p < 0.05$), confirming the robustness of our results.

a) Computing time: We chose superpixels to generate segments as they are computationally efficient. We illustrated AmsterTime and VPAir in Figure 3, repeating each experiment 14 times and deleting the two fastest and slowest results for reliability. Sel-V is significantly faster than SegVLAD on these datasets. In addition to the default setting, where the initial number of superpixels is 128, we also provide two alternative variants (with initial superpixel counts of 64 and 256) for a more complete comparison. MuSSel-V still achieves speed-ups of 4.33 and 2.04 in AmsterTime and VPAir. For comparison, we include a more efficient foundational segmenter, FastSAM [35]. FastSAM offers a segmentation speed on par with our simplest superpixel based variant (Sel-V-64) but has a significantly lower recall. Similar results occur in other datasets. Overall, the computational efficiency of our method comes from low-level feature-based superpixel segmentation, which reduces computational complexity compared to a transformer-based architecture (SAM). We note that segmentation for Sel-V and MuSSel-V is achieved using only CPU, while SAM-based methods use GPU. Superpixel methods applying GPU acceleration may further improve efficiency.

b) Qualitative analysis: Figure 4 and 5a qualitatively demonstrates the robustness of our method in different domains. In contrast to SegVLAD, our approach, like traditional VLAD, retains all pixels for aggregation. Specifically, SegVLAD discards 19.57%, 22.80%, and 19.05% of pixels on the AmsterTime, VPAir, and Baidu datasets, respectively. Our dilated superpixels are more compact compared to the supersegments of SegVLAD, enabling a more effective aggregation of local information. Furthermore, we observe that the dilated superpixels are more uniform in size.

In addition to image-to-image matching, we conducted tests on segment-level VLAD matching using the Amster-

TABLE III
CHOICE OF INITIAL NUMBER OF SUPERPIXELS.

Method	Scale(s)	AmsterTime	Baidu	VPAir
		Recall@1/5	Recall@1/5	Recall@1/5
Sel-V	16	58.0/77.2	4.5/21.6	58.3/76.7
Sel-V	32	58.3/77.1	5.8/20.6	58.6/76.3
Sel-V	64	59.3/78.9	78.6/92.3	66.0/82.1
Sel-V	128	**61.3/80.2**	82.8/95.0	72.1/86.0
Sel-V	256	60.4/79.4	**84.0**/94.9	**72.9/86.5**
MulSSel-V	64+128	60.7/79.2	82.7/94.8	69.8/84.3
MulSSel-V	128+256	60.7/79.0	83.3/**95.1**	72.0/86.0
MulSSel-V	64+128+256	60.8/79.9	83.3/95.0	70.9/85.8

TABLE IV
EFFECT OF SUPERPIXEL AND DILATION. LAST TWO ROWS REPRESENT SEL-V AND MUSSEL-V. *RESULTS ARE AVERAGED OVER THREE SCALES (64, 128, AND 256). $^+$RESULTS ARE BASED ON A MULTI-SCALE (64+128+256) APPROACH.

Segmenter	Dilation	AmsterTime	VPAir	Baidu
		Recall@1/5	Recall@1/5	Recall@1/5
SAM	×	38.4/65.5	53.0/70.3	74.4/89.8
FastSAM	×	41.3/63.9	54.4/72.4	67.1/85.8
Grid-based*	×	43.1/68.9	62.6/76.3	79.0/92.3
Grid-based$^+$	×	46.1/70.9	64.0/77.7	80.1/93.1
Superpixel*	×	48.8/73.7	65.2/78.8	80.7/92.7
Superpixel$^+$	×	48.4/73.3	65.3/78.8	80.1/93.1
Superpixel*	Random	59.0/77.8	66.3/82.4	76.0/91.4
Superpixel$^+$	Random	59.2/77.2	69.0/82.3	78.0/93.1
Superpixel*	Neigh.	60.3/79.5	70.3/84.9	81.8/94.1
Superpixel$^+$	Neigh.	**60.8/79.9**	**70.9/85.8**	**83.3/95.0**

Time dataset as an example. Sel-V achieves a Recall@1/5 of 40.31%/66.78% for dilated superpixel matching, while SegVLAD reaches 34.47%/56.65% for supersegment matching. These evaluation (unlike global image-level matching, which is affected by pixel loss) suggest that dilated superpixels better aggregate homogeneous local features and yield fewer spurious matches than global image-level methods. Nevertheless, in samples with rich semantic content, Sel-V may perform less effectively (Figure 5b).

D. Ablation Studies

a) Effect of superpixel: The initial number of superpixels n is a critical hyperparameter that directly influences both the granularity of the representation and the overall performance on the VPR task. To balance precision and efficiency, we empirically evaluate a range of values $k \in \{16, 32, 64, 128, 256\}$ (Table III). We observe that performance is significantly degraded when using very small n values (e.g., 16 or 32), likely due to the overly coarse segmentation losing essential local details. In contrast, initial numbers n with 128 and 256 consistently yield the best performance across benchmarks. Therefore, we adopt $n = 128$ as the default setting, as it offers a good trade-off between accuracy and efficiency. Additionally, we include $n = 64$ and $n = 256$ as variants for ablation purposes. For MuSSel-V, we compare combinations of different granularities, specifically $(64 + 128)$, $(128 + 256)$, and $(64 + 128 + 256)$. Among these, the combination $(64 + 128 + 256)$ demonstrates the most robust and stable performance across tasks, and is thus selected as the default configuration for MuSSel-V.

We excluded the influence of neighbouring masks to compare our superpixel-based VLAD method directly with SAM-based segmented VLAD and grid-based patch VLAD. As shown in the top part of Table IV, visually homogeneous superpixels significantly outperform both semantically meaningful segments and uniform grid patches. Interestingly, our results show that grid-based patches achieve higher recall rates than semantically meaningful segments without neighbouring mask influence, suggesting that complete masks are more effective for VPR capturing holistic visual information. Furthermore, SLIC [36] was implemented as an alternative. It demonstrated recall for VPR on par with SEEDS, while incurring a slightly greater latency — still markedly lower than that of SAM.

b) Effect of dilation: As shown in the bottom part of Table IV, dilated superpixels significantly boost VPR performance for both Sel-V and MuSSel-V, aligning with SegVLAD's findings [13]. They outperform both non-dilated superpixels and randomly expanded superpixels without neighbourhood relationships. While both the multi-scale approach in MuSSel-V and the dilation of superpixels change the extent of the segments, they work complimentary. The multi-scale method changes the spatial extent while maintaining visual homogeneity of each superpixel. The neighbourhood matrix groups multiple superpixels into clusters, ignoring homogeneity among them (see Figure 4). We also tested other dilation methods, including connectivity flow and fixed-radius dilation, and observed similar results to Delaunay triangulation. Our findings on the dilation order (neighbouring hops) were consistent with SegVLAD. Therefore, we adopted the same order and dilation method.

c) Sel-V vs. MuSSel-V: The comparison between Sel-V and MuSSel-V highlights the effectiveness of multi-scale versus fixed-scale methods in VPR tasks. Recall@1 scores for Sel-V variants (64, 128, 256) are 76.73%, 78.87%, and 77.64%, with Recall@5 scores at 91.15%, 92.22%, and 91.73%. MuSSel-V achieves 77.66% Recall@1 and 91.75% Recall@5, slightly underperforming Sel-V-128 but outperforming Sel-V-64 and Sel-V-256. MuSSel-V's multi-scale approach helps in matching images with scale differences. For example, it enables to match a building in 128-scale with 256-scale in the reference image, showing robustness to scale variations. While we can fix an optimal scale for each dataset, our multi-scale approach effectively eliminates this hyperparameter and provides better generalisation.

V. CONCLUSION

In this paper, we introduced **Sel-V**, a segment-level representation method using dilated superpixels based on low-level visual features, and **MuSSel-V**, a scale-adaptive extension for enhanced robustness. By aggregating features within dilated superpixels into VLAD representations, our approach reduces information loss and creates more compact and effective segments. Results on 12 benchmarks show that Sel-V and MuSSel-V outperform SOTA methods in both recall performance and computational efficiency compared to other segment-level VLAD methods, demonstrating their effectiveness for VPR tasks. With these promising results, future work will include on-robot validation, analysis of the VLAD cluster number, and studies on flat-index scalability and DINO-layer dependence.

References

[1] C. Masone and B. Caputo, "A survey on deep visual place recognition," *IEEE Access*, vol. 9, pp. 19 516–19 547, 2021.

[2] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4878–4888.

[3] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.

[4] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 726–743.

[5] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Eigenplaces: Training viewpoint robust models for visual place recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 11 080–11 090.

[6] B. Yildiz, S. Khademi, R. M. Siebes, and J. Van Gemert, "Amstertime: A visual place recognition benchmark dataset for severe domain shift," in *Proc. Int. Conf. Pattern Recognit.* IEEE, 2022, pp. 2749–2755.

[7] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," *IEEE Robot. Autom. Lett.*, 2023.

[8] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, 2010, pp. 3304–3311.

[9] R. Arandjelovic and A. Zisserman, "All about vlad," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1578–1585.

[10] A. Khaliq, M. Xu, S. Hausler, M. Milford, and S. Garg, "Vlad-buff: burst-aware fast feature aggregation for visual place recognition," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2024, pp. 447–466.

[11] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 17 658–17 668.

[12] F. Lu, L. Zhang, X. Lan, S. Dong, Y. Wang, and C. Yuan, "Towards seamless adaptation of pre-trained models for visual place recognition," *arXiv preprint arXiv:2402.14505*, 2024.

[13] K. Garg, S. S. Puligilla, S. Kolathaya, M. Krishna, and S. Garg, "Revisit anything: Visual place recognition via image segment retrieval," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2024, pp. 326–343.

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.

[15] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14 141–14 152.

[16] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13 648–13 657.

[17] H. Hu, Z. Qiao, M. Cheng, Z. Liu, and H. Wang, "Dasgil: Domain adaptation for semantic and geometric-aware image-based localization," *IEEE Trans. Image Process.*, vol. 30, pp. 1342–1353, 2020.

[18] N. V. Keetha, M. Milford, and S. Garg, "A hierarchical dual model of environment-and place-specific utility for visual place recognition," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 6969–6976, 2021.

[19] R. Cupec, E. K. Nyarko, D. Filko, A. Kitanov, and I. Petrović, "Place recognition based on matching of planar surfaces and line segments," *Int. J. Robot. Res.*, vol. 34, no. 4-5, pp. 674–704, 2015.

[20] C. Cheng, D. L. Page, and M. A. Abidi, "Object-based place recognition and loop closing with jigsaw puzzle image segmentation algorithm," in *Proc. IEEE Int. Conf. Robot. Autom.* IEEE, 2008, pp. 557–562.

[21] R. Mirjalili, M. Krawez, and W. Burgard, "Fm-loc: Using foundation models for improved vision-based localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2023, pp. 1381–1387.

[22] A. Oliva, "Gist of the scene," in *Neurobiology of attention*. Elsevier, 2005, pp. 251–256.

[23] Sivic and Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.* IEEE, 2003, pp. 1470–1477.

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

[25] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "Mixvpr: Feature mixing for visual place recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 2998–3007.

[26] F. Lu, X. Lan, L. Zhang, D. Jiang, Y. Wang, and C. Yuan, "Cricavpr: Cross-image correlation-aware representation learning for visual place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16 772–16 782.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Int. Conf. Mach. Learn.* PmLR, 2021, pp. 8748–8763.

[28] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9650–9660.

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

[30] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2017, pp. 9–16.

[31] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes," *IEEE Trans. Robot.*, vol. 36, no. 2, pp. 561–569, 2019.

[32] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2015, pp. 4297–4304.

[33] L. Taipe, J. Bardales, K. Pena-Pena, G. Comina, and M. Segovia, "A hybrid approach incorporating superpixels for diabetic foot lesion segmentation using yolov5 and sam," in *Proc. IEEE Int. Symp. Biomed. Imaging*, 2024, pp. 1–4.

[34] M. Cai, X. Liu, Z. Xiong, and X. Chen, "Biosam: Generating sam prompts from superpixel graph for biological instance segmentation," *IEEE J. Biomed. Health Inform.*, vol. 29, no. 1, pp. 273–284, 2025.

[35] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, "Fast segment anything," *arXiv preprint arXiv:2306.12156*, 2023.

[36] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[37] M. Van den Bergh, X. Boix, G. Roig, and L. Van Gool, "Seeds: Superpixels extracted via energy-driven sampling," *Int. J. Comput. Vis.*, vol. 111, pp. 298–314, 2015.

[38] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[39] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, 2017.

[40] X. Sun, Y. Xie, P. Luo, and L. Wang, "A dataset for benchmarking image-based localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7436–7444.

[41] S. Zhao, D. Singh, H. Sun, R. Jiang, Y. Gao, T. Wu, J. Karhade, C. Whittaker, I. Higgins, J. Xu *et al.*, "Subt-mrs: A subterranean, multi-robot, multi-spectral and multi-degraded dataset for robust slam," *arXiv preprint arXiv:2307.07607*, vol. 1, 2023.

[42] M. Schleiss, F. Rouatbi, and D. Cremers, "Vpair–aerial visual place recognition and localization in large-scale outdoor environments," *arXiv preprint arXiv:2205.11567*, 2022.

[43] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 883–890.

[44] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2626–2635.

[45] M. Zaffar, S. Garg, M. Milford, J. Kooij, D. Flynn, K. McDonald-Maier, and S. Ehsan, "Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *Int. J. Comput. Vis.*, vol. 129, no. 7, pp. 2136–2174, 2021.

[46] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," in *Proc. IEEE Int. Conf. Robot. Autom.* Citeseer, 2013, p. 2013.