AnyLoc: Towards Universal Visual Place Recognition

https://anyloc.github.io/

Nikhil Keetha*[1], Avneesh Mishra*[2], Jay Karhade*[1], Krishna Murthy Jatavallabhula[3], Sebastian Scherer[1], Madhava Krishna[2], and Sourav Garg[4]

[1]CMU, [2]IIIT Hyderabad, [3]MIT, [4]University of Adelaide

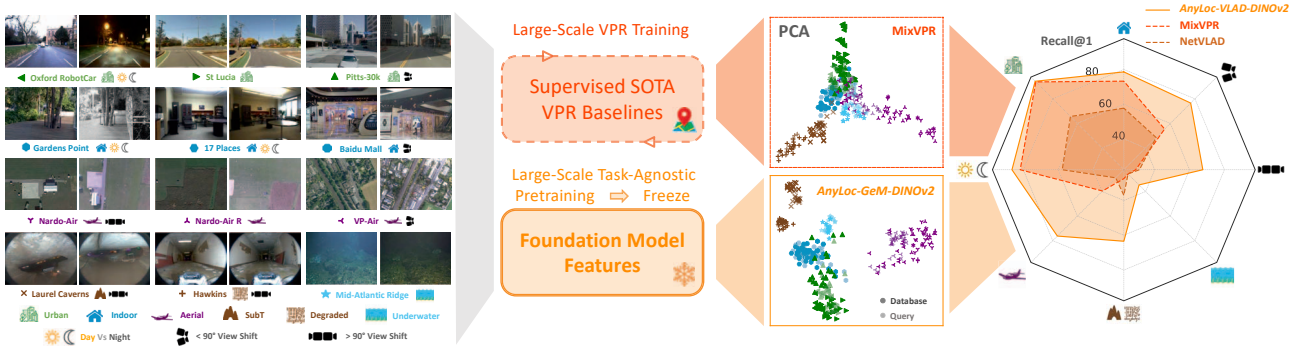


Fig. 1. ***AnyLoc*** enables *universal* visual place recognition (VPR) across a massively diverse set of environments (*anywhere*), temporal changes (*anytime*), and a wide range of viewpoint variations (*anyview*). *AnyLoc* achieves this by aggregating per-pixel features extracted from large-scale pretrained models (*foundation models*), *without any training or finetuning*. In the PCA panels (*middle*), notice how the features from MixVPR — a state-of-the-art method trained specifically for VPR — concentrate to a small region of the feature space, losing discriminative ability. On the other hand, *AnyLoc* uncovers distinct *domains* encompassing datasets with similar properties, marked with the same color. Using these *domains* to construct vocabularies for unsupervised VLAD aggregation enables *AnyLoc* to achieve up to 4× higher Recall@1, as seen in the polygonal areas in the radar chart (*right*), across structured (urban outdoors, indoors) and unstructured (underwater, aerial, subterranean, visually degraded) environments.

***Abstract*— Visual Place Recognition (VPR) is vital for robot localization. To date, the most performant VPR approaches are *environment- and task-specific*: while they exhibit strong performance in structured environments (predominantly urban driving), their performance degrades severely in unstructured environments, rendering most approaches brittle to robust real-world deployment. In this work, we develop a *universal* solution to VPR – a technique that works across a broad range of structured and unstructured environments (urban, outdoors, indoors, aerial, underwater, and subterranean environments) without any re-training or finetuning. We demonstrate that general-purpose feature representations derived from off-the-shelf self-supervised models *with no VPR-specific training* are the right substrate upon which to build such a universal VPR solution. Combining these derived features with *unsupervised feature aggregation* enables our suite of methods, *AnyLoc*, to achieve up to 4× significantly higher performance than existing approaches. We further obtain a 6% improvement in performance by characterizing the semantic properties of these features, uncovering unique *domains* which encapsulate datasets from similar environments. Our detailed experiments and analysis lay a foundation for building VPR solutions that may be deployed *anywhere*, *anytime*, and across *anyview*. We encourage the readers to explore our project page and interactive demos: https://anyloc.github.io/.**

I. Introduction

Visual Place Recognition (VPR) is a fundamental capability for robot state estimation and is widely applied in robotic systems such as autonomous cars, other uncrewed (aerial, terrestrial, and underwater) vehicles, and wearable devices. Despite significant advancements in VPR over the years, achieving out-of-the-box applicability across a diverse set of scenarios remains challenging; this is critical to bootstrap a mobile robot *anywhere*, *anytime*, and across *anyview*.

State-of-the-art (SOTA) approaches are *specifically trained* for VPR and exhibit strong performance on environments similar to those found in the training dataset (for instance, urban driving). However, when the same methods are deployed in an environment where the extracted visual features differ substantially (such as underwater or aerial), their performance drops sharply (Fig. 1). In this context, we address the question, "**How can one design a universal VPR solution?**" This entails generating place representations from a *general* model, which is pre-trained in an embodiment-, task- and environment-agnostic manner and can be readily adjusted to its *specific* deployment environment. Specifically, a *universal* VPR solution must be applicable *anywhere* (seamlessly operates across any environment, including aerial, subterranean, and underwater), *anytime* (robust to temporal changes in the scene, such as day-night or seasonal variations, or to transient objects), and across *anyview* (robust to perspective viewpoint variations, including diametrically opposite views).

We rethink the VPR problem from the lens of (visual) feature representations derived from large-scale pretrained models (coined *foundation models* [1]). We show that, despite not being trained for VPR, these models encode rich visual features that serve as the right substrate upon which a *universal* VPR solution may be built. Our approach, termed ***AnyLoc***, involves a careful selection of models and visual features with the *right* invariance properties and blends

*Equal Contribution

them with prevailing local-aggregation approaches in the VPR literature [2]–[5], resulting in all of the aforementioned desirable characteristics of a *universal* VPR solution.

Our key takeaways are as follows:

- *AnyLoc* emerges as a new baseline VPR method that works universally across 12 datasets exhibiting massive diversity along the axes of *place*, *time*, and *perspective*;
- Self-supervised features (such as DINOv2 [6]) and unsupervised aggregation methods (like VLAD [7] & GeM [8]) are *both* crucial for strong VPR performance. Applying these aggregation techniques on per-pixel features offers substantial performance gains over the direct use of per-image features from off-the-shelf models.
- Characterizing the semantic properties of the aggregated local features uncovers distinct *domains* in the latent space, which can further be used to enhance VLAD vocabulary construction; in turn boosting performance.

We evaluate *AnyLoc* on an extensive and diverse range of datasets (urban, indoors, aerial, underwater, subterranean) across challenging VPR conditions (day-night and seasonal variations, opposing viewpoints), establishing a strong baseline for future research towards universal VPR solutions.

II. VPR: Overview, Trends & Limitations

VPR – Problem definition: VPR is often cast as an image retrieval problem [9] that comprises two phases. In the *indexing* phase, a **reference map (image database)** is gathered from a robot's onboard camera when traversing through an environment. In the *retrieval* phase, given a **query image**—captured during a future traverse—VPR entails retrieving the closest match to this query image in the reference map. There exists a variety of VPR methods and alternative problem formulations [3], [10]–[13]; in this work, we focus on **global descriptors**, which offer the best tradeoff between accurate matching and search efficiency [7], [9], [14]. This is in contrast to local descriptor methods, which are computationally intensive to match, particularly over larger databases.

Researchers have explored various training objectives [15]–[18], aggregation techniques [2], [8], and transfer learning [19]–[21] to improve global descriptor-based VPR. High performance of most of these modern approaches can be attributed to **large-scale training on VPR-specific data**. Powered by deep learning and the Pitts-250k dataset [22], weakly-supervised contrastive learning in NetVLAD [2] led to substantial improvements over classical hand-crafted features. Following suit, the Google-Landmark V1 (1 million images) and V2 datasets [23] (5 million images) enabled training DeLF [24] and DeLG [25] for large-scale image retrieval. Likewise, the Mapillary Street-Level Sequences (MSLS) dataset, containing 1.6 million *street* images, substantially boosted VPR performance by tapping orders of magnitude larger data from urban and suburban settings [26]–[28]. More recently, CosPlace [18] coupled classification-based learning with the San Francisco XL dataset comprising 40 million images having GPS & heading. The current SOTA, MixVPR [29], proposed an MLP-based feature mixer, trained on the GSV-Cities dataset [30] – a curated large-scale dataset with 530,000 images spanning 62,000 places worldwide.

This trend of scaling up VPR training is mostly driven by easily-available positioning data for outdoor environments, which leads to SOTA performance in urban settings, but **does not generalize to indoor and unstructured environments.** As shown in Fig. 1, the PCA projections of descriptors extracted by SOTA methods concentrate to a narrow region in the feature space, diminishing their discriminative abilities in environments outside the training distribution. Apart from environment-specificity, prior methods have tackled *specific* challenges in isolation, such as extreme temporal variations in scene appearance [21], [31] and camera viewpoint [32], [33]. This data- and task-specificity of current VPR approaches limits their out-of-the-box applicability, which may be mitigated by task-agnostic learning. Hence, in this work, we analyze the design space of VPR using web-scale self-supervised visual representations and **develop a *universal* solution that does not assume any VPR-specific training**.

III. AnyLoc: Towards Universal VPR

To the best of our knowledge, our approach, *AnyLoc*, is the first VPR solution that exhibits *anywhere*, *anytime*, and *anyview* capabilities (see Fig. 1). *AnyLoc* is guided by two crucial insights (see Section V for details) that emerged when exploring the design space of VPR solutions through the lens of foundation model features: (a) *existing VPR solutions are task-specific and perform poorly when evaluated in environments outside the training distribution*; and (b) *while per-pixel features from off-the-shelf foundation models [6], [34], [35] demonstrate remarkable visual and semantic consistency [5], [36]–[38], the per-image features are suboptimal when used as-is for VPR*. Therefore, a careful investigation is needed to transfer these per-pixel invariances to the image level for recognizing *places*, where recent approaches in this direction are only limited to small-scale indoor settings or vision-language use-cases [39], [40]. In this context, for designing *AnyLoc*, we investigate the following questions:

A. What foundation models are best suited to VPR?
B. How do we extract VPR-suited local features from these general-purpose models?
C. How do we aggregate *local* features to describe *places*?
D. How to construct vocabularies that generalize across datasets?

A. Choice of Foundation Model

There exist three broad classes of **self-supervised foundation models that extract task-agnostic visual features**: (a) joint embedding methods (DINO [34], DINOv2 [6]), (b) contrastive learning methods (CLIP [35]), and (c) masked autoencoding approaches (MAE [41]). Joint embedding methods need a stable training recipe; DINO is trained on ImageNet [42] through global image-level self-supervision, while DINOv2 is trained on a much larger, carefully-curated

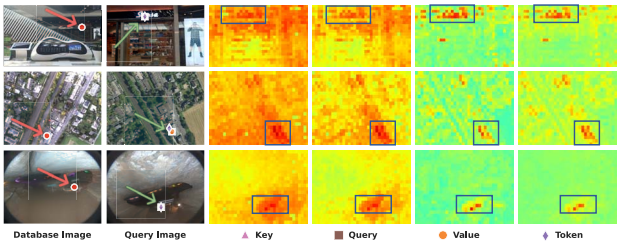


Fig. 2. Point correspondences (as markers) & similarity maps show the **robustness of foundation model features to various VPR challenges**: (*top*) text and scale change, (*middle*) perceptually aliased features and viewpoint shift, and (*bottom*) low illumination combined with opposing viewpoint. The `value` facet has the highest contrast between the background and the matched points, which is vital for discarding distractors within an image.

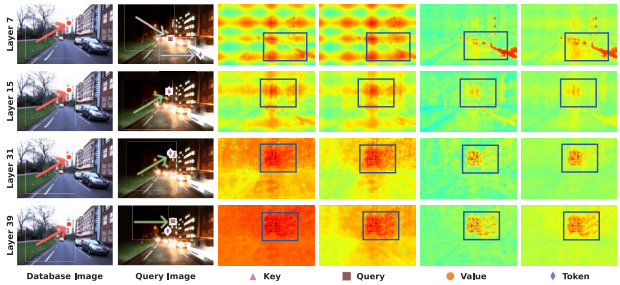


Fig. 3. Qualitative ablation comparing the absolute-scale similarity maps of features from different DINOv2 ViT-G *layers* and *facets*. **Layer 31** `value` **facet has the sharpest contrast** in the similarity map, which is **crucial for robustness against distractors within an image.**

dataset with joint image-/token-level losses. These methods offer the highest level of performance; followed by contrastive learned approaches like CLIP [35], which is trained on millions of *aligned* image-text pairs. In our initial experiments, we found all these models to perform better than MAE [41], which only has token-level self-supervision. These findings are corroborated in [5], [6], [36], highlighting the benefits of learning long-range global patterns captured by joint embedding methods. Therefore, *AnyLoc* employs DINO & DINOv2 vision transformers for extracting features.

B. Choice of Features

Another important design choice is how we extract visual features from these pretrained vision transformers (ViT) [43]. Rather than extract per-image features[1] (i.e., one feature vector for the entire image), we observe that per-pixel features enable fine-grained matching and result in superior performance. Each layer in the ViT has multiple *facets* (`query`, `key`, `value`, and `token`) from which features may be extracted. Following [38], we extract features from intermediate layers across the ViT and discard the `CLS` token. In Fig. 2, we illustrate this **applicability of the dense ViT features for VPR** by assessing the robustness of local feature correspondences. We select a point on a database image, match it with all (per-pixel) features from the query image, and plot heatmaps indicating the likelihood these points correspond. Notice how the correspondences are robust even in the presence of semantic text and scale change (first row), perceptual aliasing and viewpoint shift (second row), and low illumination combined with opposing viewpoint (third row).

Comparing the similarity maps in Fig. 2, notice how the `value` **facet exhibits the largest contrast** between the matched points and the background, which is **crucial for robustness against distractors within an image**. Upon further analysis *across layers* (Fig. 3), we observe an interesting trend. The earlier layers of the ViT (top rows), especially the `key` and `query` facets, exhibit a high positional encoding bias, while **the** `value` **facet of deeper layers has the sharpest contrast** in the similarity map. We further justify our selection of layer & facet quantitatively in Section V-C.2.

[1]In a ViT, per-image features are encoded in a special token, `CLS`, and interpreted as a summary of the image content.

C. Choice of Aggregation Technique

The next design choice to make towards our VPR pipeline entails selecting an *aggregation technique* that determines how local features are grouped together to describe sections of an image and, eventually, an environment. While prior work has used the `CLS` token directly for image retrieval [6], [40], [44], we observed contradictory trends under a *universal* retrieval setting (i.e., retraining or finetuning is prohibited). We **comprehensively explore multiple unsupervised aggregation techniques**: Global Average Pooling (GAP) [45], Global Max Pooling (GMP) [46], Generalized Mean Pooling (GeM) [8], and the soft & hard assignment variants of VLAD [7].

For an input image of size $H \times W$, and a per-pixel feature $f_i \in \mathbb{R}^D$, we define a global descriptor as:

$$F_G = \left(\sum_{i=1}^{H \times W} f_i^p \right)^{\frac{1}{p}} \tag{1}$$

where $p = 1$, $p = 3$, and $p \to \infty$ represent GAP, GeM, and GMP respectively.

For VLAD variants, we cluster all the features from the database images to obtain N cluster centers. This forms our *vocabulary*. The global VLAD descriptor is then calculated as the sum of residuals per cluster center k, as below:

$$F_{V_k} = \sum_{i=1}^{N \times H \times W} \alpha_k(f_i)(f_i - c_k) \tag{2}$$

where $\alpha_k(x_i)$ is 1 if f_i is assigned to cluster k and 0 otherwise. In the soft-assignment variant of VLAD, $\alpha_k(f_i)$ indicates the assignment probability and lies between 0 and 1. Following [47], we perform intra-normalization, concatenation, and inter-normalization to obtain the final VLAD descriptor F_V.

D. Choice of Vocabulary

For vocabulary-based aggregation techniques, we construct our vocabulary with the goal of **characterizing the distinct semantic properties of globally pooled local features across diverse environments**. Prior work based on VLAD has either used a global vocabulary based on representative places & features [7], a reference map-specific one [47], or a learnt [2] vocabulary based on the training

TABLE I

UNSTRUCTURED ENVIRONMENTS USED IN EVALUATION

Dataset	N_{Db}	N_Q	Traj. Span	Loc. Radius	Type
Hawkins [48]	65	101	282 m	8 m	
Laurel Caverns [48]	141	112	102 m	8 m	
Nardo-Air	102	71	700 m / 1 km^2	60 m	
VP-Air [49]	12.7k	2.7k	100 km	3 frames	
Mid-Atlantic Ridge [50]	65	101	18 m	0.3 m	

TABLE II

STATE-OF-THE-ART BASELINES USED FOR COMPARISON

Method	Backbone	Training Dataset	Supervision
NetVLAD [2], [3]	ResNet-18	Pitts-30k	VPR - Contrastive
CosPlace [18]	ResNet-101	SF-XL	VPR - Classification
MixVPR [29]	ResNet-50	GSV-Cities	VPR - Contrastive
CLIP [35], [58]	ViT-bigG-14	Laion 2B	Image-Caption Pairs
DINO [34]	ViT-S8	ImageNet	Self-Supervised
DINOv2 [6]	ViT-G14	LVD-142M	Self-Supervised

dataset. These approaches work well where domain- or map-specific data is abundant and task-specific training is feasible. However, a more scalable approach is to leverage the open-set semantic attributes encoded in the foundation model features to determine the appropriate domain and feature vocabulary. Hence, we use vocabulary-independent global descriptors (DINOv2-GeM) and their (unsupervised) PCA projection to define vocabularies for VLAD aggregation.

From Fig. 1, we observe that **projecting the global descriptors using PCA uncovers distinct *domains* in the latent space**, which characterizes datasets having similar properties, namely: **Urban**, **Indoor**, **Aerial**, **SubT**, **Degraded**, and **Underwater**. Further demonstrating discriminative robustness, although the **SubT** and **Degraded** domains have similar imagery types, they are dispersed to distinct regions, whereas the visually degraded indoor domain is concentrated relatively close to the indoor collection. Lastly, we can observe that the projected features for the query images are close to the projected features of their respective database images[2]. Hence, using the PCA-based segregation, we construct the visual vocabularies for VLAD in a domain-specific manner (further justified in Section V-B.1).

IV. EXPERIMENTAL SETUP

A. Datasets

There exist several VPR datasets where the composition of benchmarks is influenced by either the end task, i.e., urban data for Geo-localization [3] or the evaluation aspects of viewpoint variability [12]. We evaluate on datasets from both structured and unstructured environments, offering unprecedented diversity in terms of environments (*anywhere*), coupled with a range of temporal (*anytime*) and camera viewpoint[3] (*anyview*) variations. We define structured environments as organized places composed of human-built structures that are commonplace in applications such as autonomous driving and indoor robotics. These represent the typical images collected and shared by humans on the web. On the other hand, unstructured environments represent in-the-wild scenarios where the objects and types of images encountered are not commonly observed.

1) Structured Environments: We used six **benchmark** indoor and outdoor datasets, exhibiting challenges like drastic viewpoint shifts, perceptual aliasing, and substantial visual appearance change. This includes Baidu Mall [51], Gardens Point [52], [53], 17 Places [54], Pittsburgh-30k [22], St Lucia [55], and Oxford RobotCar [56], where the ground truth localization radius is 10 meters, 2 frames, 5 frames, 25 meters, 25 meters, and 25 meters, respectively. For Oxford RobotCar, we use a subsampled version of the Overcast Summer and Autumn Night traverses, following HEAPUtil [57].

2) Unstructured Environments: While our structured environments enable us to benchmark with respect to existing VPR techniques, to **truly assess robustness and versatility**, we evaluate on a number of unstructured environments, including aerial, underwater, visually degraded, and subterranean environments[4]. Table I provides an overview of these unstructured datasets, which exhibit challenging distribution shifts, visually degraded long corridors, satellite & aerial imagery covering various landscapes, low illumination, and seasonal variations. Nardo-Air R aligns the orientation of drone imagery with the satellite map.

B. Benchmarking & Evaluation

We use Recall@K [12] as the evaluation metric (a higher recall score indicates superior performance). All experiments use the same random seed (`42`) and GPU hardware (NVIDIA RTX 3090) for consistency and reproducibility.

1) State-of-the-art Baselines: We evaluate *AnyLoc* against a variety of VPR methods such that it encompasses variations in terms of VPR-specific training, global image representation, type of supervision, backbone models, and the scale and nature of training data. We include three *specialized* baselines, which are trained for the VPR task on large-scale urban datasets, and three new baselines that use the `CLS` token of the foundation models, as summarized in Table II.

2) AnyLoc - Nomenclature and Model Specifications: All names are of the form *AnyLoc*-`aggregation`-`model`, where `aggregation` is one of *VLAD*, *GeM*; and `model` is one of *DINO*, *DINOv2*. For *AnyLoc-VLAD-DINO*, we use the ViT-S8 layer 9 `key` facet features and 128 clusters for VLAD. Likewise, for *AnyLoc-GeM* and *AnyLoc-VLAD-DINOv2*, we use ViT-G14 layer 31 `value` facet features, with 32 clusters for VLAD.

[2]The PCA transform is computed solely from the database images, and does not make use of the query images, for fair analysis.

[3]The viewpoint shifts range from $< 90°$ with minimal (Oxford, St Lucia) and moderate shifts (Pitts30K, Baidu) to $> 90°$ with extreme shifts (orthogonal in Nardo-Air and opposite in Hawkins, Laurel). $> 90°$ criterion for the opposite-viewpoint datasets refers to a $180°$ orientation change in observing a place from a nearby but not the same 3D position [9], [32].

[4]Models such as DINOv2 and CLIP are trained on web-scale datasets, and consequently, will likely have seen structured environments similar to those in Table III. Therefore, the true test for these models is their performance on unstructured environments, which are highly unlikely to have featured in any of the training subsets for these models.

TABLE III

PERFORMANCE COMPARISON ON BENCHMARK STRUCTURED ENVIRONMENTS

	Baidu Mall		Gardens Point		17 Places		Pitts-30k		St Lucia		Oxford		Average	
Methods	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
NetVLAD [2]	53.1	70.5	58.5	85.0	61.6	77.8	86.1	92.7	57.9	73.0	57.6	79.1	62.5	79.7
CosPlace [18]	41.6	55.0	74.0	94.5	61.1	76.1	90.4	**95.7**	99.6	99.9	95.3	99.5	77.0	86.8
MixVPR [29]	64.4	80.3	91.5	96.0	63.8	78.8	**91.5**	95.5	**99.7**	**100**	92.7	99.5	83.9	91.7
CLIP-CLS [35]	56.0	71.6	42.5	74.5	59.4	77.6	55.0	77.2	62.7	80.7	46.6	60.7	53.7	73.7
DINO-CLS [34]	48.3	65.1	78.5	95.0	61.8	76.4	70.1	86.4	45.2	64.0	20.4	46.6	54.1	72.3
DINOv2-CLS [6]	49.2	64.6	71.5	96.0	61.8	78.8	78.3	91.1	78.6	89.7	47.1	58.1	64.4	79.7
AnyLoc-GeM-DINOv2	50.1	70.6	88.0	97.5	63.6	79.6	77.0	87.3	76.9	89.3	92.2	97.9	74.6	87.0
AnyLoc-VLAD-DINO	61.2	78.3	95.0	98.5	63.8	78.8	83.4	92.0	88.5	94.9	82.2	99.0	79.0	90.2
AnyLoc-VLAD-DINO-PCA	62.3	81.2	91.5	99.5	63.3	78.8	82.8	90.8	87.6	94.3	82.7	96.3	78.4	90.1
AnyLoc-VLAD-DINOv2	**75.2**	87.6	95.5	**99.5**	**65.0**	80.5	87.7	94.7	96.2	98.8	**99.5**	**100**	**86.5**	93.5
AnyLoc-VLAD-DINOv2-PCA	74.9	**89.4**	**96.0**	**99.5**	64.8	**81.0**	86.9	93.8	96.4	99.5	96.9	**100**	86.0	**93.9**

TABLE IV

PERFORMANCE COMPARISON ON UNSTRUCTURED ENVIRONMENTS

	Hawkins		Laurel Caverns		Nardo-Air		Nardo-Air R		VP-Air		Mid-Atlantic Ridge		Average	
Methods	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
NetVLAD [2]	34.8	71.2	39.3	71.4	19.7	39.4	60.6	85.9	6.4	17.7	25.7	53.5	31.1	56.5
CosPlace [18]	31.4	59.3	24.1	47.3	0	1.4	91.6	**100**	8.1	14.2	20.8	40.6	29.3	43.8
MixVPR [29]	25.4	60.2	29.5	67.0	32.4	42.2	76.1	98.6	10.3	18.3	25.7	60.4	33.2	57.8
CLIP-CLS [35]	33.0	67.0	36.6	66.1	42.2	70.4	62.0	97.2	36.6	52.8	25.7	51.5	39.4	67.5
DINO-CLS [34]	46.6	84.8	41.1	57.1	57.8	90.1	84.5	**100**	24.0	38.4	27.7	49.5	47.0	70.0
DINOv2-CLS [6]	28.0	62.7	40.2	65.2	73.2	88.7	71.8	91.6	45.2	59.9	24.8	48.5	47.2	69.4
AnyLoc-GeM-DINOv2	53.4	83.9	58.9	86.6	**76.1**	83.1	57.8	97.2	38.3	53.8	14.8	49.5	49.9	75.7
AnyLoc-VLAD-DINO	48.3	84.8	57.1	79.5	43.7	54.9	**94.4**	**100**	17.8	28.7	**41.6**	**66.3**	50.5	69.0
AnyLoc-VLAD-DINOv2	**65.2**	**94.1**	**61.6**	**90.2**	**76.1**	**94.4**	85.9	**100**	**66.7**	**79.2**	34.6	61.4	**65.0**	**86.5**

V. EXPERIMENTS, RESULTS, AND ANALYSES

We first evaluate *AnyLoc* against SOTA VPR techniques and report results across structured & unstructured environments, viewpoint shifts, and temporal appearance variations. We further present a comparative analysis of the specialized baselines and variants directly using the CLS token (i.e., per-image features). We then present a detailed vocabulary analysis followed by insights into the design of *AnyLoc*. Lastly, we demonstrate the benefits of self-supervised ViTs by contrasting them with existing VPR-trained ViTs.

A. State-of-the-art Comparison

1) Structured Environments: Table III highlights the **general applicability of the *AnyLoc* methods on structured environments**, in particular, the **Indoor** and **Urban** domains. *AnyLoc-VLAD-DINOv2* achieves the highest recall across all the **Indoor** datasets while outperforming MixVPR (the second best) and CosPlace by 5% and 20% on average (R@1). Interestingly, foundation models' CLS descriptors (while being inferior to our method) are competitive with baselines such as CosPlace and NetVLAD, e.g., CLIP outperforms them respectively by 15% and 3% on Baidu Mall. Through our proposed use of feature aggregation for foundation models, we observe that simply using GeM pooling over DINOv2 features (i.e., *AnyLoc-GeM-DINOv2*) significantly improves performance over the DINOv2 CLS token. This is further improved by *AnyLoc-VLAD*, which beats all prior approaches on these datasets. In the **Urban** case – which well aligns with the training distribution of the baselines supervised specifically for VPR on urban data – we observe that *AnyLoc-VLAD* is inferior by 3-4% on daytime conditions of Pitts30k and St Lucia, but it achieves state-of-the-art for day-night variations on Oxford. We further showcase that a PCA-Whitening of the *AnyLoc-VLAD* descriptors using the domain-specific database enables similar SOTA performance while having a 100× smaller embedding size (49*k* to 512).

2) Unstructured Environments: **Table IV highlights the fragility of the specialized baselines and shows that *AnyLoc* outperforms all the baselines by a *large* margin in these challenging unstructured environments**. Even the CLS methods outperform VPR-specialized baselines, e.g., DINOv2-CLS exceeds MixVPR by 41% on Nardo-Air and 35% on VP-Air under strong viewpoint variations. The *AnyLoc* methods consistently outperform both the specialized and the CLS baselines, where the best performers in the respective categories, i.e., MixVPR and DINOv2-CLS, lag behind *AnyLoc-VLAD* by 32% and 18% on average (R@1).

3) Temporal & Viewpoint Changes: We further demonstrate the **robustness of *AnyLoc* for *anytime* and *anyview* VPR**. We evaluate multiple datasets where revisiting a place at different time intervals leads to variations in scene appearance (*anytime*). In comparison to the SOTA VPR baselines,

TABLE V

EFFECT OF VOCABULARY TYPE ON R@1 FOR *AnyLoc-VLAD-DINOv2*

Vocabulary Type	Indoor	Urban	Aerial
Global	77.0	93.9	57.1
Structured	77.0	93.3	56.4
Unstructured	74.8	89.0	75.8
Map-Specific	78.0	92.3	62.9
Domain-Specific	**78.6**	**94.4**	**76.2**

TABLE VI

ANALYSING INTRA-DOMAIN TRANSFERABILITY OF *AnyLoc-VLAD-DINOv2* VOCABULARIES

Vocabulary Dataset	Evaluation Dataset	Map-Specific Recall@1	Vocab-Transfer Recall@1
Baidu Mall (0.7k)	17 Places (0.4k)	**64.5**	63.8
	Gardens Point (0.2k)	**98.0**	94.5
VP-Air (2.7k)	Nardo-Air (0.1k)	57.8	**64.8**
	Nardo-Air R (0.1k)	70.4	**88.7**
Pitts-30k (10k)	Oxford (0.2k)	94.8	**99.0**

MixVPR/CosPlace, we observe the following gains using *AnyLoc-VLAD* on different temporal changes: 5/11% on day-night cycles affecting outdoors (Oxford), indoors (17 Places), and mixture (Gardens Point); 9/8% on seasonal shifts (Oxford); 21/28% on long period jumps (2022 vs. 2023 for Nardo-Air, 2015 Vs. 2020 for the Mid-Atlantic Ridge). A similar trend is observed for viewpoint shifts (*anyview*), where we test on datasets that vary both in terms of the *view-type*, e.g., street vs aerial, and the *shift-type*. *AnyLoc-VLAD* outperforms MixVPR/CosPlace on orientation-based shifts by 21/30% and extreme 90°/180° shifts by 39/49%.

4) Specialized Baselines: The average recall of NetVLAD, CosPlace, and MixVPR confirms the **general trend of better performance in task-specific baselines with an increasing scale of urban training data**, combined with innovations in learning objective (CosPlace) and learnable aggregation (MixVPR). Additionally, we observe one peculiar failure case of CosPlace on the Nardo-Air dataset. No correct matches were found under the combined effect of out-of-distribution (aerial) and extreme viewpoint (90 degrees) shifts. Visual inspection revealed that all queries incorrectly matched to a handful of reference images having similar orientation of fields and roads.

5) `CLS` vs. Aggregation (AnyLoc): **When the foundation models are used with local feature aggregation instead of the `CLS` token, we observe significant performance jumps**: DINOv2-based *AnyLoc-GeM* and *AnyLoc-VLAD* outperform DINOv2-`CLS` by 9%/2% and 23%/18% respectively on structured/unstructured environments. Furthermore, the average recall of the `CLS` token-based global descriptors (CLIP, DINO & DINOv2) indicates their superiority to specialized baselines on unstructured environments.

B. Vocabulary Analysis

1) Vocabulary Source: **Table V shows how the vocabulary source used for VLAD influences recall, where domain-specific vocabulary leads to the best recall.** We construct multiple VLAD vocabularies using different subsets of the 12 datasets used in this work and report average recall per domain. As described in Section III-D, the subsets for different domains are obtained through a qualitative PCA visualization (see Fig. 1), which is quantitatively justified through the results presented here. The other vocabulary sources that we compare against are: *Global* using all 12 datasets; *Structured* using 3 indoor and 3 urban datasets; *Unstructured* using the complement set of structured; and *Map-specific* using only the reference database of a particular dataset. In the aerial domain, domain-specific achieves 13% over map-specific and 19% over global vocabulary.

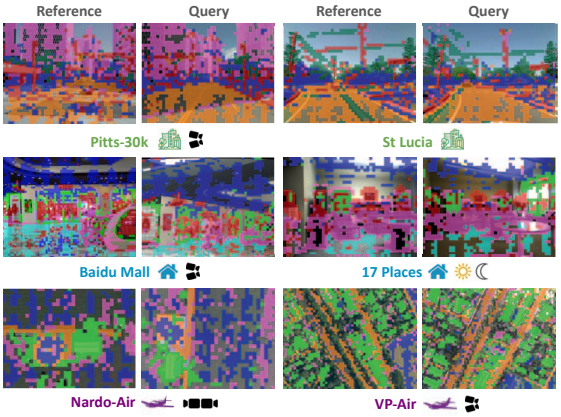
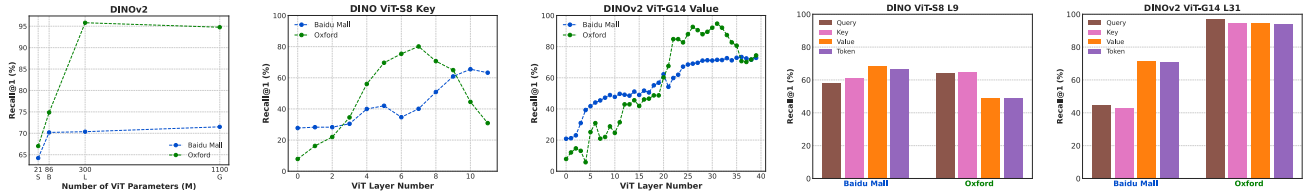


Fig. 4. VLAD cluster assignment visualizations of the reference-query pairs highlight the **intra-domain consistency** of the domain-specific vocabulary. Similar colors across images of a specific domain indicate matched clusters.

2) Consistency: Fig. 4 showcases the **robust intra-domain consistency of the domain-specific vocabulary**, further justifying the high performance of *AnyLoc-VLAD*. Specifically, we visualize the cluster assignments (with $K = 8$) for the local features using the domain-specific vocabulary. In the **Urban** domain, the roads, pavements, buildings, and vegetation are consistently assigned to the same cluster across changing conditions and places. For the **Indoor** domain, we can observe intra-domain consistency for the floor & ceiling, while there is intra-place consistency for the text signs and furniture. For the **Aerial** domain, it can be observed that roads, vegetation, and buildings are assigned to unique clusters across both the rural and urban images.

We further demonstrate that this **robust consistency within a domain enables us to deploy *AnyLoc-VLAD* in target environments with small reference databases (maps) that lack information richness**. For datasets belonging to a given domain, we pick the largest reference database to form the vocabulary and evaluate on other datasets from that domain. In Table VI, for **Aerial** and **Urban** domains, we can observe that 7-18% higher R@1 can be achieved when using a larger source of vocabulary as compared to just using the target dataset's own smaller map, thus demonstrating the transferability of vocabularies within the same domain. For the **Indoor** domain, the drop in performance is either due to



(a) Model (b) Layer (c) Facet

Fig. 5. **Design Choices for *AnyLoc-VLAD***: (a) Performance scales with the model size but saturates at ViT-L. (b) Performance peaks at intermediate layers instead of the final layer for both DINO & DINOv2. (c) On average, `key` & `value` perform the best respectively for DINO & DINOv2.

TABLE VII
ANALYSIS COMPARING THE RECALL@1 & DESCRIPTOR DIMENSIONALITY ACROSS VARYING AGGREGATION METHODS

	DINO			DINOv2		
Aggregation Methods	**Baidu ↑**	**Oxford ↑**	**Dim ↓**	**Baidu ↑**	**Oxford ↑**	**Dim ↓**
Global Average Pool (GAP)	29.6	28.8	**384**	41.6	78.5	**1536**
Global Max Pool (GMP)	34.9	38.2	**384**	64.4	74.9	**1536**
Generalized Mean Pool (GeM)	34.7	47.6	**384**	50.1	92.2	**1536**
Soft Assignment VLAD	33.8	28.3	49152	40.3	82.2	49152
Hard Assignment VLAD	**60.9**	**64.9**	49152	**71.5**	**94.8**	49152

TABLE VIII
ANALYSIS COMPARING THE RECALL@1 OF VPR-TRAINED VITS TO SELF-SUPERVISED VITS

Method	**Indoor**	**Urban**	**Aerial**	**SubT & D**	**Underwater**
ViT-B CosPlace	62.9	80.7	26.3	26.5	18.8
ViT-B CosPlace-VLAD	68.5	82.9	38.4	37.5	23.8
ViT-S *AnyLoc-VLAD-DINO*	72.9	79.6	47.8	52.7	**41.6**
ViT-B *AnyLoc-VLAD-DINOv2*	77.0	82.6	53.6	60.2	35.6
ViT-G *AnyLoc-VLAD-DINOv2*	**78.0**	**92.3**	**62.9**	**63.4**	34.6

a relatively limited size of the largest reference database or the large diversity across datasets, e.g., shops in Baidu Mall compared to offices in the other two datasets. Nevertheless, when using this unified diverse vocabulary from all the datasets in the indoor domain, the overall recall is better than using map-specific vocabularies, as shown in Table V.

C. Insights into AnyLoc Design

We present insights on varying parameters within *AnyLoc*, using two datasets, Baidu Mall & Oxford, which are representative of the typical VPR challenges:

1) ViT Architecture: Fig. 5 a showcases that **larger DINOv2 ViT backbones lead to better performance**, where the performance tends to saturate at ViT-L (300 million parameters). Since, on average, ViT-G performs better than ViT-L, we use ViT-G for DINOv2. For DINO, we use ViT-S, which is the only available architecture.

2) ViT Layers & Facets: Fig. 5 b shows that **peak performance is achieved through deeper layers, somewhere between the middle and the last layer**. For a smaller ViT architecture (DINO ViT-S on the left), it can be observed that middle layers have higher performance on Oxford. This can be attributed to their higher positional encoding bias, which is helpful under no viewpoint shift across reference-query pairs. Hence, aligning with the findings presented in Section III-B, we choose 9 and 31 as our operating layers for DINO and DINOv2, respectively.

In Fig. 5 c, the **`key` & `value` facets consistently achieve high recall for DINO & DINOv2 respectively**. Although `query` and `key` facets perform better on Oxford when using DINO (left), this gap diminishes when using DINOv2 (right). The performance difference between the `query` & `value` gets inverted from Baidu to Oxford; indicating a high positional bias in the `query` & `key`, leading to poor performance under the significant viewpoint shift in Baidu.

3) Aggregation Methods: In Table VII, we compare the various unsupervised local feature aggregation techniques as discussed in Section III-C and observe that **hard assignment-based VLAD works the best**. We can further see that the vocabulary-free methods provide an optimal trade-off between performance and storage, where GeM pooling tends to do the best. Also, we observed that hard assignment is typically 1.4 times faster than soft assignment.

D. Self-supervised vs VPR-supervised ViT

Table VIII shows that the **high performance of *AnyLoc-VLAD* is not a consequence of simply using a large ViT but an outcome of self-supervised training on large-scale curated data, which leads to generality in the underlying features** [6]. In particular, we compare a ViT trained specifically for VPR (i.e., CosPlace [18]) against those based on self-supervision (i.e., DINO & DINOv2). For the VPR-supervised CosPlace, we include the authors' GeM pooling-based ViT-B model along with its adapted version that uses a VLAD layer ($K = 128$) on top of ViT-B's 6th layer (which performed better than other layers). For self-supervised methods, we include *AnyLoc-VLAD* variants: DINO ViT-S, DINOv2 ViT-B and ViT-G. All VLAD-based methods in these comparisons use map-specific vocabulary. Comparing ViT-B-based methods, we can observe that even though CosPlace's overall performance improves with VLAD, *AnyLoc-VLAD-DINOv2* outperforms it by 8-13%. Interestingly, even ViT-S based *AnyLoc-VLAD-DINO* outperforms ViT-B-based CosPlace-VLAD by 4-18% while using 4× fewer parameters. The only exception to these trends is in the urban domain, where CosPlace-VLAD outperforms ViT-S and ViT-B based *AnyLoc-VLAD*, which is justified by CosPlace's VPR-specific training on urban data. Despite this, *AnyLoc-VLAD-DINOv2* ViT-G surpasses all other methods.

VI. Conclusion

This paper introduces *AnyLoc* – a significant step towards *universal* VPR. Driven by the limitations of *environment*- and *task-specific* VPR techniques, and the fragility of per-image features extracted from foundation models, we propose to blend the per-pixel features computed by these models with unsupervised feature aggregation techniques like VLAD and GeM. Through our benchmarking and analyses on a diverse suite of datasets, we shed light on the brittleness of current large-scale urban-trained VPR approaches and show that *AnyLoc* outperforms the previous state-of-the-art by up to $4\times$. This work stretches the applicability scope of VPR and, in turn, robot localization to *anytime*, *anywhere* & under *anyview*, which is crucial to enable downstream capabilities, such as robot navigation in the wild.

Acknowledgments

This work was supported by ARL grant W911QX20D0008/W911QX22F0078(TO6). The authors thank Ivan Cisneros & Yao He for collecting the Nardo-Air dataset. We also thank the members of CMU AirLab for their insightful discussions throughout this project.

References

[1] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *arXiv:2108.07258*, 2021.
[2] R. Arandjelovic *et al.*, "Netvlad: Cnn architecture for weakly supervised place recognition," in *CVPR*, 2016.
[3] G. Berton *et al.*, "Deep visual geo-localization benchmark," in *CVPR*, 2022.
[4] Q. Garrido *et al.*, "On the duality between contrastive and non-contrastive self-supervised learning," in *ICLR*, 2023.
[5] S. Shekhar *et al.*, "Objectives matter: Understanding the impact of self-supervised objectives on vision transformer representations," *arXiv:2304.13089*, 2023.
[6] M. Oquab *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv:2304.07193*, 2023.
[7] H. Jégou *et al.*, "Aggregating local descriptors into a compact image representation," in *CVPR*. IEEE, 2010.
[8] F. Radenović *et al.*, "Fine-tuning cnn image retrieval with no human annotation," *IEEE T-PAMI*, vol. 41, no. 7, pp. 1655–1668, 2018.
[9] S. Garg *et al.*, "Where is your place, visual place recognition?" *IJCAI*, 2021.
[10] S. Lowry *et al.*, "Visual place recognition: A survey," *T-RO*, 2015.
[11] N. Pion *et al.*, "Benchmarking image retrieval for visual localization," in *3DV*. IEEE, 2020.
[12] M. Zaffar *et al.*, "Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *IJCV*, pp. 1–39, 2021.
[13] S. Schubert *et al.*, "Visual place recognition: A tutorial," *RAM*, 2023.
[14] T. Sattler *et al.*, "Benchmarking 6dof outdoor visual localization in changing conditions," in *CVPR*, 2018.
[15] Y. Ge *et al.*, "Self-supervising fine-grained region similarities for large-scale image localization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 369–386.
[16] J. Xiao *et al.*, "Visual geo-localization with self-supervised representation learning," *arXiv preprint arXiv:2308.00090*, 2023.
[17] M. Leyva-Vallina *et al.*, "Data-efficient large scale place recognition with graded similarity supervision," in *CVPR*, 2023.
[18] G. Berton *et al.*, "Rethinking visual geo-localization for large-scale applications," in *CVPR*, 2022.
[19] L. Haas *et al.*, "Learning generalized zero-shot learners for open-domain image geolocalization," *arXiv:2302.00275*, 2023.
[20] G. M. Berton *et al.*, "Adaptive-attentive geolocalization from few queries: A hybrid approach," in *WACV*, 2021, pp. 2918–2927.
[21] Y. Latif *et al.*, "Addressing challenging place recognition tasks using generative adversarial networks," in *ICRA*, 2018.
[22] A. Torii *et al.*, "Visual place recognition with repetitive structures," in *CVPR*, 2013.
[23] T. Weyand *et al.*, "Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval," in *CVPR*, 2020.
[24] H. Noh *et al.*, "Large-scale image retrieval with attentive deep local features," in *ICCV*, 2017.
[25] B. Cao *et al.*, "Unifying deep local and global features for image search," in *ECCV*, 2020.
[26] F. Warburg *et al.*, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *CVPR*, 2020.
[27] R. Wang *et al.*, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," in *CVPR*, 2022.
[28] S. Zhu *et al.*, "R2former: Unified retrieval and reranking transformer for place recognition," in *CVPR*, 2023.
[29] A. Ali-bey *et al.*, "Mixvpr: Feature mixing for visual place recognition," in *WACV*, 2023.
[30] ——, "Gsv-cities: Toward appropriate supervised visual place recognition," *Neurocomputing*, 2022.
[31] L. Tang *et al.*, "Adversarial feature disentanglement for place recognition across changing appearance," in *ICRA*, 2020.
[32] S. Garg *et al.*, "Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics," *RSS*, 2018.
[33] A. Gawel *et al.*, "X-view: Graph-based semantic multi-view localization," *IEEE RAL*, 2018.
[34] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021.
[35] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021.
[36] N. Park *et al.*, "What do self-supervised vision transformers learn?" in *ICLR*, 2023.
[37] K. M. Jatavallabhula *et al.*, "Conceptfusion: Open-set multimodal 3d mapping," *RSS*, 2023.
[38] S. Amir *et al.*, "Deep vit features as dense visual descriptors," *arXiv:2112.05814*, 2021.
[39] R. Mirjalili *et al.*, "Fm-loc: Using foundation models for improved vision-based localization," *arXiv:2304.07058*, 2023.
[40] C. Kassab *et al.*, "Clip-based features achieve competitive zero-shot visual localization," *OpenReview preprint arXiv:2306.14846*, 2023.
[41] K. He *et al.*, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022.
[42] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.
[43] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.
[44] A. El-Nouby *et al.*, "Training vision transformers for image retrieval," *arXiv:2102.05644*, 2021.
[45] A. Babenko *et al.*, "Aggregating deep convolutional features for image retrieval," *arXiv:1510.07493*, 2015.
[46] A. S. Razavian *et al.*, "Visual instance retrieval with deep convolutional networks," *ITE TMTA*, 2016.
[47] R. Arandjelovic *et al.*, "All about vlad," in *CVPR*, 2013.
[48] S. Zhao *et al.*, "Subt-mrs: A subterranean, multi-robot, multi-spectral and multi-degraded dataset for robust slam," *arXiv:2307.07607*, 2023.
[49] M. Schleiss *et al.*, "Vpair–aerial visual place recognition and localization in large-scale outdoor environments," *ICRA 2022 Aerial Robotics Workshop arXiv:2205.11567*, 2022.
[50] C. Boittiaux *et al.*, "Eiffel tower: A deep-sea underwater dataset for long-term visual localization," *IJRR*, 2022.
[51] X. Sun *et al.*, "A dataset for benchmarking image-based localization," in *CVPR*, 2017.
[52] A. Glover, "Gardens point day and night, left and right," *Zenodo DOI*, vol. 10, 2014.
[53] N. Sünderhauf *et al.*, "On the performance of convnet features for place recognition," in *IROS*, 2015.
[54] R. Sahdev *et al.*, "Indoor place recognition system for localization of mobile robots," in *2016 13th CRV*. IEEE, 2016, pp. 53–60.
[55] M. Warren *et al.*, "Unaided stereo vision based pose estimation," in *ACRA*, vol. 47. Citeseer, 2010, p. 60.
[56] W. Maddern *et al.*, "1 year, 1000 km: The oxford robotcar dataset," *IJRR*, vol. 36, no. 1, pp. 3–15, 2017.
[57] N. V. Keetha *et al.*, "A hierarchical dual model of environment-and place-specific utility for visual place recognition," *IEEE RAL*, vol. 6, no. 4, pp. 6969–6976, 2021.
[58] G. Ilharco *et al.*, "Openclip," in *Zenodo*, 2021.