

Geometry-Conditioned Visual Place Recognition in Natural Environments

Walter Nedov¹, Saimunur Rahman¹, Kavindie Katuwandeniya¹,
David Hall¹, Kaushik Roy¹, Peyman Moghadam^{1,2}

¹CSIRO Robotics, CSIRO, Australia firstname.lastname@csiro.au

²Queensland University of Technology, Australia firstname.lastname@qut.edu.au

Abstract

Visual Place Recognition (VPR) in natural environments remains challenging due to repetitive vegetation, sparse distinctive landmarks, and substantial appearance and viewpoint variation across traversals. While visual observations of the same place can change considerably, their underlying spatial structure is often more persistent. We exploit this complementary geometric consistency through Depth-Aware Distillation (DAD), which conditions the token representations of a pretrained Vision Foundation Model (VFM) on geometry inferred by a Geometric Foundation Model (GFM), without any depth sensor. Rather than treating geometry as an additional input modality, DAD projects image-aligned depth into the VFM token space and selectively modulates visual representations through channel-wise geometric conditioning. A two-stage teacher-guided learning strategy first anchors the geometry-conditioned representation to the pretrained appearance space, before refining it for place discrimination. Evaluated on the Wild-Cross benchmark, DAD improves average inter-sequence Recall@1 from 61.41% to 66.37% and Recall@5 from 65.86% to 72.49% over a matched appearance-only baseline, with the largest gains under reverse traversal and long-term appearance variation. These results show that GFM-derived geometry can provide a persistent structural prior for VPR when visual appearance becomes unreliable.

1. Introduction

Place Recognition (PR) is the task of determining whether a current sensor observation corresponds to a previously visited place, making it a core capability for autonomous robotic systems operating in GPS-denied/limited environments such as underground spaces, dense forests, and indoor environments [1]. PR has been studied across complementary sensing modalities, including Visual [2, 3], LiDAR [4, 5], and Radar [6], each providing different appearance, geometric, and structural cues for localization. However, reliable PR in natural and unstructured environ-

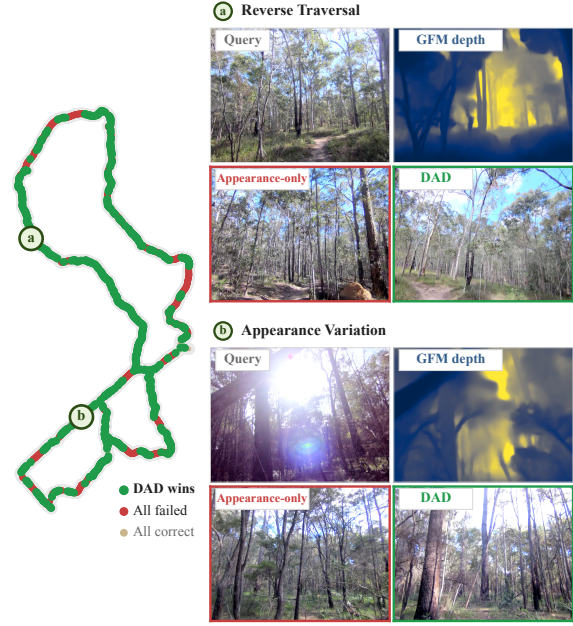


Figure 1: Geometry complements appearance-based VPR in natural environments. DAD corrects appearance-only retrieval failures under reverse traversal and appearance variation by exploiting GFM-derived scene structure.

ments remains particularly challenging due to repetitive vegetation, sparse distinctive landmarks, irregular scene structure, and substantial changes in viewpoint and appearance across revisits [7, 8].

Visual Place Recognition (VPR), in particular, is strongly affected by these conditions because it relies on distinctive and repeatable visual cues. Current VPR methods increasingly build on Vision Foundation Models (VFMs), such as DINOv2, which provide strong semantic representations and achieve high performance in structured urban environments [2, 3]. In such environments, persistent buildings, road markings, signs, and other landmarks provide reliable semantic cues. In contrast, natural environments often contain extensive repetitive vegetation and fewer distinctive landmarks, resulting in perceptual aliasing

between geographically different places. Changes in illumination, weather, season, vegetation growth, and viewpoint can further alter the appearance of the same place across revisits [8, 9], while reverse traversal can produce substantially different observations despite relatively stable scene structure. Fig. 1 illustrates these two representative failure modes, where appearance-only retrieval fails under reverse traversal and severe appearance variation, while estimated depth preserves the coarse geometry of the path and surrounding vegetation. This motivates the use of geometry as a complementary signal for VPR. While LiDAR-based methods demonstrate the robustness of geometric structure, they require additional sensing hardware. Recent Geometric Foundation Models (GFM), such as Depth Anything V2 [10], can instead estimate dense scene geometry directly from RGB images, providing spatially aligned geometric cues that can be integrated with VFM representations.

Motivated by these observations, we introduce Depth-Aware Distillation (DAD), a geometry-conditioned VPR framework that uses depth inferred by a Geometric Foundation Model (GFM) to guide the token representations of a pretrained Vision Foundation Model (VFM), without requiring a dedicated depth sensor. Rather than treating geometry as an additional input modality, DAD projects image-aligned depth into the VFM token space and selectively modulates the visual representation through learnable channel-wise geometric conditioning. To preserve the pretrained appearance representation, DAD adopts a two-stage teacher-guided training strategy that first anchors the geometry-conditioned descriptor to a frozen appearance-based teacher before refining it for place discrimination.

2. Related Work

Visual Place Recognition: VPR is commonly formulated as image retrieval, in which a query image is matched against a database of images through a global descriptor. Classical pipelines aggregated hand-crafted local features using vocabulary-based encodings such as VLAD. Earlier work addressed appearance change through condition-invariant global-to-local image comparison [11]. NetVLAD [12] made feature aggregation differentiable, and subsequent work has increasingly adopted vision foundation models as feature extractors. DINOv2 [13] produces semantically rich patch features that now underpin leading VPR methods. BoQ [14] pools features using learnable queries, SALAD [2] formulates aggregation as optimal transport-based feature-to-cluster assignment, and PairVPR [3] employs masked image modeling to train a pair-classifier decoder that informs its encoder’s retrieval objective. While these models excel in structured domains, their visual features do not readily generalize to unstructured natural scenes [15], motivating the integration of geometric priors into the visual representation space.

Geometric Foundation Models: Parallel to the development of VFMs for semantic understanding, Geometric Foundation Models (GFMs) have emerged that infer 3D scene structure directly from images. Monocular models such as Depth Anything V2 [10] predict dense depth maps from a single RGB frame, while multi-view models such as DUS3R [16], MAS3R [17] VGGT [18] jointly estimate camera poses, depth, and dense point maps using a single feed-forward transformer. These GFMs produce geometric representations such as depth images and point cloud projections that generalize across domains without scene-specific training. Recent work has shown that such features can be transferred to downstream vision models to refine their representations, measurably improving spatial understanding in models that otherwise lack explicit 3D reasoning [19]. Our work takes inspiration from this finding to suggest that GFM features could similarly enrich the representations of appearance-only VFMs for VPR, providing geometric priors that can support retrieval in challenging environments.

Integrating GFM Features into VFMs: Multimodal integration in pretrained foundation models requires a mechanism for incorporating additional signals without disrupting the original representation space. GFM-derived tokens are not naturally aligned with a VFM’s feature space, so naive feature injection can disrupt the pretrained representation without improving its utility for retrieval. RGB-D architectures such as DFormer [20] learn dedicated multimodal backbones for scene understanding, but are not designed to preserve the descriptor space of an existing pretrained VPR model. Joint multimodal training can also suffer from the dominating-modality problem, in which one modality dominates the descriptor space [21]. Alignment-based approaches instead align a new modality to a frozen foundation encoder before fusion, whether across large paired datasets (ViT-Lens [22]) or through token-level distillation over compatible patch tokens that avoids the cross-architecture gap of heterogeneous backbones (TOLiD [23]). DAD applies this alignment principle within a single place descriptor by using a frozen appearance teacher to anchor the geometry-conditioned descriptor to the pretrained VFM retrieval space, as detailed in the following section.

3. Methodology

DAD conditions the representation of a pretrained appearance-based VPR model using geometry inferred by a pretrained GFM. Given an RGB image, the VFM produces appearance tokens, while the GFM estimates an image-aligned depth map that provides geometric information about the scene. The GFM remains fixed throughout training, whereas the VPR model is adapted in stages to incorporate this geometric information without disrupting

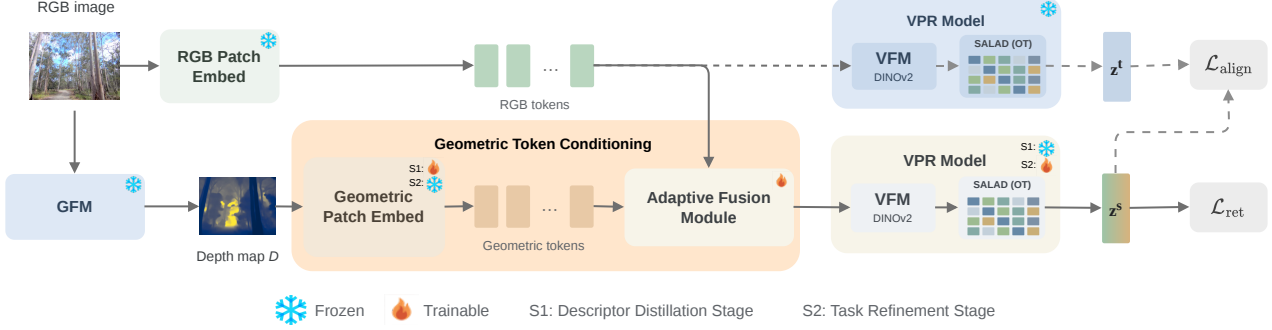


Figure 2: Overview of DAD. GFM-derived geometric tokens condition the VFM appearance representation to produce a global place descriptor. During descriptor distillation, an appearance-only teacher anchors the geometry-conditioned representation through $\mathcal{L}_{\text{align}}$, together with the retrieval objective $\mathcal{L}_{\text{ret}}$. During task refinement, the teacher constraint is removed and the representation is optimized for place discrimination using $\mathcal{L}_{\text{ret}}$.

its established appearance representation. As illustrated in Fig. 2, DAD projects the estimated depth into the VFM token space and selectively integrates it with the appearance tokens before producing a global place descriptor.

DAD adopts a two-stage training strategy. During *descriptor distillation*, the appearance-based VPR model is kept fixed and serves as a teacher, while the newly introduced geometry-conditioning components are learned. This stage establishes compatibility between the geometric information and the existing VPR representation. During *task refinement*, selected higher-level layers of the VPR model are fine-tuned for place discrimination, while the learned geometric embedding is kept fixed.

3.1. Geometry-Conditioned Token Embedding

DAD introduces geometry directly into the token space of the appearance-based VFM while preserving the spatial correspondence between visual and geometric information. Let $E_v(\cdot)$ denote the patch embedding layer of the VFM, which maps an RGB image I to a sequence of appearance

tokens $E_v(I)$. Given the same image, the GFM produces an image-aligned depth map D .

To represent this depth information in the VFM token space, DAD introduces a learnable geometric embedder $E_g(\cdot)$, implemented as a lightweight convolutional projection module. The embedder transforms D into geometric tokens with the same spatial resolution and channel dimension as the appearance tokens:

$$\mathbf{g} = E_g(D) \in \mathbb{R}^{N \times C}, \quad (1)$$

where N denotes the number of spatial patches and C is the VFM token dimension. Although appearance and geometry provide complementary information, directly adding geometric tokens to the appearance representation can disrupt the feature space learned by the VFM. Moreover, the usefulness of geometric information may vary across feature channels. To selectively regulate its contribution, we introduce an *Adaptive Fusion Module* that applies a learnable channel-wise gate:

$$\mathbf{x} = E_v(I) + \alpha \odot E_g(D), \quad (2)$$

where $\alpha \in \mathbb{R}^C$ is a learnable channel-wise gating vector and $\odot$ denotes channel-wise multiplication broadcast across all spatial tokens. As illustrated in Fig. 3, the same gate is shared across spatial tokens, while each feature channel can independently regulate the contribution of geometry. The gate is initialized near zero, with $\alpha = 0.01$ for all channels, so that the initial geometry-conditioned representation remains close to the original appearance representation. During training, the gate learns which channels benefit from geometric conditioning while allowing appearance-dominant channels to remain largely unchanged.

The resulting geometry-conditioned token sequence $\mathbf{x}$ is then processed by the VFM transformer blocks and the VPR aggregation head. Let $T_v(\cdot)$ denote the VFM transformer blocks and $G(\cdot)$ the global descriptor aggregation head, instantiated using SALAD in our experiments. The resulting

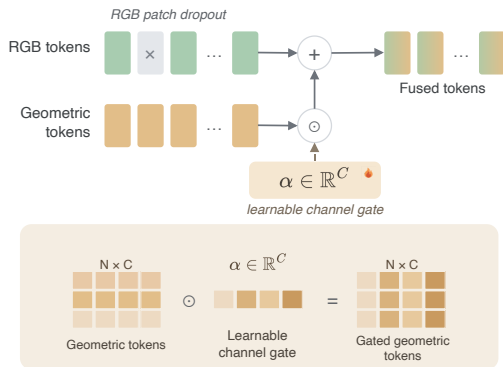


Figure 3: Adaptive Fusion Module. GFM-derived geometric tokens are modulated by a learnable channel-wise gate and fused with the VFM tokens. RGB patch dropout is applied during training to regularize the appearance features.

global place descriptor $\mathbf{z}$ is given by

$$\mathbf{z} = G(T_v(\mathbf{x})). \quad (3)$$

This token-level conditioning has two advantages. First, image-aligned depth preserves spatial correspondence between appearance and geometric tokens, allowing the transformer to jointly reason over both cues. Second, geometry is introduced directly into the existing VFM representation rather than through a separate geometry-specific backbone, retaining the original VPR architecture with only the geometric embedder and channel-wise gate added.

3.2. Descriptor Distillation Stage

The geometry-conditioned representation should introduce complementary structural information without losing the discriminative aspects of the appearance-based descriptor space. To mitigate this, DAD first performs a descriptor distillation stage that learns to incorporate geometric information while maintaining consistency with the established appearance representation.

We formulate this stage using a teacher-student framework. The appearance-only VPR model is kept fixed and serves as the teacher, providing a stable reference descriptor space, while the geometry-conditioned pathway forms the student. Given an RGB image I , the teacher descriptor is

$$\mathbf{z}^t = G(T_v(E_v(I))). \quad (4)$$

While the student descriptor $\mathbf{z}^s$ is obtained from the geometry-conditioned representation defined in Eqs. 2–3. During descriptor distillation, the VFM backbone and aggregation head are kept fixed, and only the geometric embedder $E_g(\cdot)$ and channel-wise gate α are optimized.

The descriptor distillation stage is optimized with two complementary objectives. First, the geometry-conditioned representation should remain discriminative for place retrieval. We therefore employ the multi-similarity retrieval loss [24]:

$$\mathcal{L}_{\text{ret}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left[\frac{1}{\kappa_p} \log \left(1 + \sum_{k \in \mathcal{P}_i} e^{-\kappa_p(S_{ik} - \gamma)} \right) + \frac{1}{\kappa_n} \log \left(1 + \sum_{k \in \mathcal{N}_i} e^{\kappa_n(S_{ik} - \gamma)} \right) \right], \quad (5)$$

where S_{ik} denotes the cosine similarity between the descriptors of samples i and k , and $\mathcal{P}_i$ and $\mathcal{N}_i$ denote the sets of positive and negative samples associated with anchor i , respectively. Second, to constrain the geometry-conditioned descriptor from drifting excessively from the

established appearance-based representation, we introduce a descriptor-level alignment loss between the student and teacher:

$$\mathcal{L}_{\text{align}} = 1 - \frac{\langle \mathbf{z}^s, \mathbf{z}^t \rangle}{\|\mathbf{z}^s\|_2 \|\mathbf{z}^t\|_2}, \quad (6)$$

where $\mathbf{z}^s$ and $\mathbf{z}^t$ denote the student and teacher descriptors, respectively. The cosine alignment objective preserves consistency with the teacher descriptor space while still allowing the geometry-conditioned representation to introduce complementary information for place discrimination. The retrieval and alignment losses jointly define the objective for the descriptor distillation stage ($\mathcal{L}_{\text{distill}}$):

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{ret}} + \lambda \mathcal{L}_{\text{align}}, \quad (7)$$

where λ controls the contribution of the teacher-alignment objective. The retrieval loss encourages the geometry-conditioned representation to exploit complementary geometric information for improved place discrimination, while the alignment loss limits excessive deviation from the appearance-based teacher representation. Together, these objectives enable DAD to learn a geometry-conditioned descriptor that preserves the established retrieval structure while incorporating additional geometric cues.

3.3. Task Refinement Stage

Descriptor distillation optimizes $\mathcal{L}_{\text{distill}}$ to introduce geometric information while constraining the representation to remain consistent with the appearance-based teacher. However, this alignment constraint can limit the extent to which the representation adapts to the discriminative requirements of VPR. We therefore introduce a task refinement stage that relaxes this constraint and optimizes the geometry-conditioned representation directly for place discrimination.

During this stage, the appearance-only teacher is no longer used and the geometry-conditioned pathway is optimized using only the retrieval objective. The geometric embedder $E_g(\cdot)$ is kept fixed to preserve the depth-to-token mapping learned during descriptor distillation. In contrast, the higher-level components of the VPR model, including the final two VFM transformer blocks and the SALAD aggregation head, are fine-tuned to adapt the joint appearance-geometry representation to the target retrieval task. The channel-wise gate α remains trainable, allowing the contribution of geometric information to be recalibrated as the representation becomes more discriminative for VPR. The task refinement objective is therefore

$$\mathcal{L}_{\text{refine}} = \mathcal{L}_{\text{ret}}. \quad (8)$$

By separating representation alignment from task-specific adaptation, the two training stages serve complementary roles. Descriptor distillation introduces geomet-

ric information while maintaining consistency with the established appearance-based descriptor space, whereas task refinement subsequently adapts the geometry-conditioned representation for improved place discrimination.

4. Experimental Setup

4.1. Dataset and GFM Adaptation

We evaluate DAD on the WildCross benchmark [8], which provides over 476,000 sequential RGB frames collected across two natural forests in southeast Queensland, Australia: Karawatha and Venman. Within each location there are four traversal sequences. Sequence 01 is the reference forward traversal, sequence 02 retraces the route in reverse on the same day, sequence 03 follows an alternate extended path, and sequence 04 repeats the reference trajectory after 14 months providing vegetational change due to the passage of time. All frames are paired with GPS/IMU poses and synchronized LiDAR submaps.

We adopt Depth Anything V2 (DA2) [10] with a ViT-B encoder as our GFM. We adapt DA2 to the forest domain using sparse WildCross LiDAR projections as metric-depth supervision, following the depth-adaptation protocol and hyperparameters of WildCross [8]. This adaptation is performed offline, after which the GFM remains frozen and DAD operates on RGB imagery and GFM-estimated depth without requiring LiDAR. Using the adapted checkpoint, we generate dense depth maps at DA2’s native 518-pixel input resolution. Before being provided to DAD, each depth map is resized alongside its corresponding RGB frame and normalized per image to zero mean and unit variance.

4.2. Training Details

Following the WildCross training implementation [8], positive pairs are defined by a GPS distance within 5 m and heading alignment within $\pm 10^\circ$, while negatives must lie beyond 50 m. Each batch contains 128 images sampled from 32 places, with four images per place. Images are resized to 224×224 for training and 322×322 for evaluation.

All augmentations are applied only during training, sampled independently for each image. Horizontal flips are applied jointly to the RGB image and depth map with probability 0.5 to preserve their spatial alignment. We additionally apply RGB-only Gaussian blur with probability 0.3. RGB patch dropout is applied with probability 0.3 by randomly masking 25% of the appearance tokens.

DAD is trained for 40 epochs in two 20-epoch stages using AdamW with weight decay 10^{-4} . In the first stage, only the geometric embedder E_g and gate α are trained at a learning rate of 10^{-4} , using a 5-epoch linear warmup followed by MultiStepLR decay and an alignment weight of $\lambda = 0.05$. The second stage initialises from the first and keeps E_g frozen while fine-tuning the final two DINOv2

transformer blocks and SALAD aggregation head at 10^{-5} , with α remaining trainable at 10^{-4} . This stage uses a 2-epoch warmup with MultiStepLR decay at epoch 15.

4.3. Evaluation Protocol

Following the WildCross protocol [8], each split holds out one sequence from each location for evaluation, while the remaining three sequences per location are used for training. The held-out sequences serve as query sets. For inter-sequence retrieval, the remaining sequences from the same location form the database. For intra-sequence retrieval, the held-out sequence serves as both the query and database, with a 600 s temporal exclusion window. Database images are ranked by cosine similarity between global descriptors, and a retrieval is considered correct if it lies within 25 m of the query. We report Recall@1 and Recall@5 averaged over the four splits. Baselines include a matched appearance-only SALAD model, NetVLAD [12], and MixVPR [25].

5. Results and Discussion

Retrieval Performance The main quantitative results are reported in Tables 1 and 2. DAD achieves the best average performance across both evaluation settings, improving intra-sequence Recall@1 from 61.76 to 65.34. However, the inter-sequence results best demonstrate DAD’s strength in providing robust place recognition across visually ambiguous and inconsistent observations, with Recall@1 increasing from 61.41 to 66.37 and Recall@5 from 65.86 to 72.49 over the matched appearance-only SALAD control.

Appearance-only SALAD					DAD					$\Delta R@1$				
Venman	V-01				V-01					V-01				
		48.5	95.3	95.4		54.7	97.1	96.5			+6.2	+1.8	+1.1	
	V-02	39.3		46.8	31.9	45.3		52.0	39.6		+6.0		+5.2	+7.7
	V-03	87.5	54.2		86.9	89.1	62.2		88.9		+1.6	+8.0		+1.9
V-04	95.5	46.6	87.9		96.1	56.1	88.7				+0.6	+9.4	+0.8	
Database					Database					Database				
Karawatha	K-01									K-01				
		34.6	72.9	94.4		43.3	74.9	92.4			+8.7	+2.0	-1.9	
	K-02	33.7		52.8	45.0	42.9		60.5	52.8		+9.1		+7.7	+7.7
	K-03	56.9	44.1		57.7	57.7	46.7		64.9		+0.8	+2.6		+7.2
K-04	69.7	33.3	55.7		78.4	46.4	66.0				+8.7	+13.1	+10.2	
Database					Database					Database				

Figure 4: Inter-sequence Recall@1 on WildCross for the matched SALAD control, DAD, and their difference. Rows denote query sequences and columns denote database sequences; diagonal cells are excluded.

Fig. 4 decomposes the inter-sequence improvement into

Table 1: Intra-sequence evaluation of fine-tuned VPR methods on each sequence of the *WildCross* benchmark. Metrics are recall at 1 (R@1) and recall at 5 (R@5) with top result bolded. Our method (DAD), highlighted in green is shown to generally improve results within intra-sequence setting.

Method	V-01		V-02		V-03		V-04		K-01		K-02		K-03		K-04		Average	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
NetVLAD [12]	68.95	72.16	69.89	74.05	18.09	22.72	59.90	66.02	83.86	87.53	86.32	91.57	26.82	32.35	56.32	59.96	58.77	63.30
MixVPR [25]	66.16	68.52	68.86	72.78	17.80	25.01	52.90	57.34	84.66	87.89	87.99	89.59	30.40	38.22	60.87	65.04	58.70	63.05
SALAD [2]	69.04	72.68	72.24	80.33	23.30	29.48	58.96	64.65	86.37	89.33	88.98	91.26	34.05	38.83	61.13	62.53	61.76	66.14
DAD (Ours)	70.55	76.87	76.64	80.39	36.46	47.05	65.09	70.89	87.09	89.86	92.47	93.39	33.90	41.85	60.55	62.58	65.34	70.36

Table 2: Inter-sequence evaluation of fine-tuned VPR methods on *WildCross*. Metrics are recall at 1 (R@1) and recall at 5 (R@5) with the top result in bold. Here we see the greatest benefit from our method (DAD), in managing the visually ambiguous/inconsistent images across sequences.

Method	Venman		Karawatha		Average	
	R@1	R@5	R@1	R@5	R@1	R@5
NetVLAD [12]	64.31	67.49	46.94	52.43	55.63	59.96
MixVPR [25]	65.30	68.58	50.24	55.80	57.77	62.19
SALAD [2]	68.54	71.86	54.29	59.86	61.41	65.86
DAD (Ours)	72.17	76.83	60.57	68.15	66.37	72.49

individual query–database pairs following the WildCross evaluation procedure [8]. Most pairs improve with DAD, with the largest Recall@1 gains occurring for K-04→K-02 (+13.1), K-04→K-03 (+10.2), V-04→V-02 (+9.4), and K-02→K-01 (+9.1). These gains concentrate on queries from the reverse-traversal and long-horizon appearance-variation sequences (sequences 02 and 04, respectively). These results suggest that GFM-derived geometry is most useful where appearance alone is ambiguous, while the appearance representation remains broadly intact.

The benefit of DAD providing geometric grounding to the feature space is further reinforced in Fig. 5. Here, we show representative examples of two challenging conditions and one case in which scene structure provides a distinctive cue. In *reverse traversal* (top row), the appearance-only model retrieves a path with superficially similar foliage, whereas DAD returns the correct location despite having an opposing viewing angle thanks to features that focus on the structural components of the tree trunk shape and presence of the path. Under severe *appearance variation* (middle row), glare and shadowing corrupt the RGB observation, while DAD retrieves a target with a consistent forest layout. In the bottom-row scene-structure case, repeated texture and color provide weak discrimination, but the lake boundary and surrounding forest structure distinguish the correct location.

To demonstrate that the benefit of DAD comes from the structural features used for matching images across sequences, Fig. 6 shows a heatmap projection of local patch scores onto the original image, showing which patches contributed most to the final image-level descriptor used for

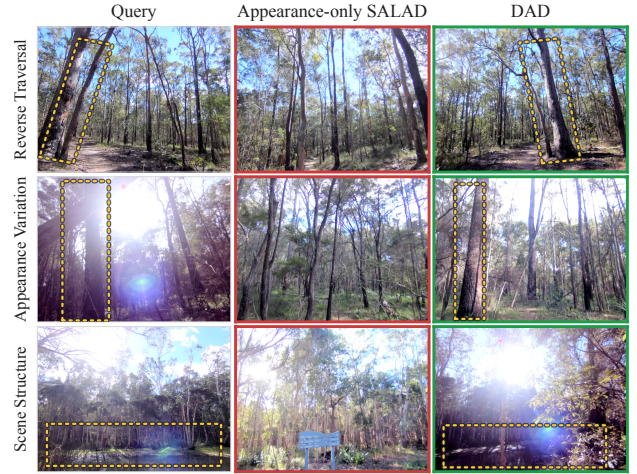


Figure 5: Representative WildCross retrievals illustrating reverse traversal, appearance variation, and distinctive scene structure. Each row shows a query, an incorrect top-ranked appearance-only SALAD retrieval, and the correct DAD retrieval. Red and green borders denote incorrect and correct retrievals; dotted yellow outlines mark corresponding scene structure.

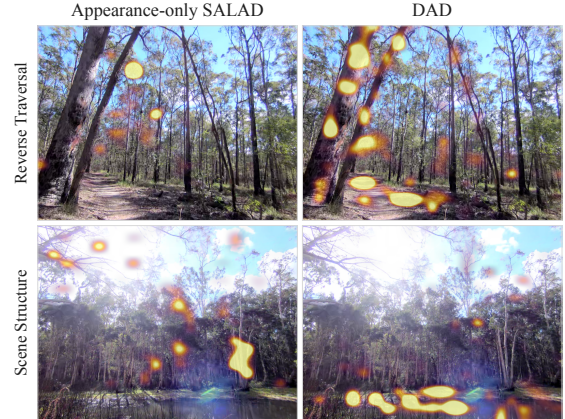


Figure 6: Local descriptor support for the reverse-traversal and scene-structure cases in Fig. 5. Columns compare the matched appearance-only SALAD baseline and DAD on the same query image. Warmer regions indicate greater relative support within each independently normalized map.

Table 3: DAD optimisation and gating ablations on *WildCross*, reported as route-averaged Recall@1.

Configuration	Intra avg. R@1	Inter avg. R@1
DAD (Ours)	65.34	66.37
Appearance-only SALAD control	61.76	61.41
Descriptor Distillation only	59.75	63.89
Single-stage joint optimization	61.11	59.84
Scalar geometric gate ($\alpha \in \mathbb{R}$)	62.64	63.20

final place recognition. Here we see that for a reverse-traversal example (top row) DAD assigns strong support to the distinctive geometry of the foreground trunks and the path boundary, whereas the appearance-only system responds more sparsely to local canopy features. In the scene-structure case (bottom row), DAD concentrates support along the partially occluded lake and its boundary, while the appearance-only descriptor responds primarily to glare, sky, and isolated foliage. These observations support the hypothesis that the geometry-conditioned representation places greater support on stable scene structure for place recognition.

Ablations: Table 3 isolates two central design choices in DAD: staged optimization and channel-wise geometric gating. The appearance-only SALAD model is included as a matched control rather than an ablation. Descriptor distillation alone improves inter-sequence Recall@1 from 61.41 to 63.89, but remains below the control for intra-sequence retrieval, indicating that task refinement is required to obtain consistent gains across both settings. Optimising all DAD components jointly from initialization instead lowers inter-sequence Recall@1 to 59.84, supporting the separation of descriptor distillation and task refinement. Finally, replacing the channel-wise gate $\alpha \in \mathbb{R}^C$ with a scalar $\alpha \in \mathbb{R}$ reduces Recall@1 to 62.64 intra-sequence and 63.20 inter-sequence, supporting channel-specific regulation of the geometric contribution.

6. Conclusion and Future Work

We presented DAD, a geometry-conditioned VPR framework for natural environments that incorporates GFM-derived scene structure into the token representation of a pretrained VFM without requiring a dedicated depth sensor. DAD combines channel-wise geometric conditioning with a two-stage learning strategy: it first aligns the geometry-conditioned descriptor with an appearance-based teacher, then refines the resulting representation for place discrimination. Experiments on WildCross show consistent improvements over the matched appearance-only VPR baseline, increasing average Recall@1 from 61.76% to 65.34% for intra-sequence retrieval and from 61.41% to 66.37% for inter-sequence retrieval. The largest improvements occur under reverse traversal and long-term appearance variation, where visual appearance becomes less reli-

able while scene structure remains comparatively persistent. These results demonstrate that GFM-derived geometry provides a complementary structural prior for improving VPR in challenging natural environments.

References

- [1] P. Yin, J. Jiao *et al.*, “General place recognition survey: Towards real-world autonomy,” *IEEE Transactions on Robotics*, 2025. 1
- [2] S. Izquierdo and J. Civera, “Optimal Transport Aggregation for Visual Place Recognition,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 17 658–17 668. 1, 2, 6
- [3] S. Hausler and P. Moghadam, “Pair-VPR: Place-Aware Pre-Training and Contrastive Pair Classification for Visual Place Recognition With Vision Transformers,” *IEEE Robotics and Automation Letters*, pp. 1–8, 2025. 1, 2
- [4] K. Vidanapathirana, M. Ramezani *et al.*, “LoGG3D-Net: Locally guided global descriptor learning for 3D place recognition,” in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2215–2221. 1
- [5] M. Jung, L. F. T. Fu *et al.*, “ImLPR: Image-based LiDAR place recognition using vision foundation models,” in *9th Annual Conference on Robot Learning*, 2025. 1
- [6] H. Kim, M. Jung *et al.*, “Sherloc: Synchronized heterogeneous radar place recognition for cross-modal localization,” *IEEE Robotics and Automation Letters*, vol. 10, no. 12, pp. 13 264–13 271, 2025. 1
- [7] J. Knights, K. Vidanapathirana *et al.*, “Wild-Places: A large-scale dataset for lidar place recognition in unstructured natural environments,” in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 11 322–11 328. 1
- [8] J. Knights, J. Reid *et al.*, “WildCross: A Cross-Modal Large Scale Benchmark for Place Recognition and Metric Depth Estimation in Natural Environments,” in *IEEE international conference on robotics and automation (ICRA)*, 2026. 1, 2, 5, 6
- [9] M. Leyva-Vallina, N. Strisciuglio, and N. Petkov, “Place recognition in gardens by learning visual representations: Data set and benchmark analysis,” in *Computer Analysis of Images and Patterns*, ser. Lecture Notes in Computer Science, vol. 11678. Springer, 2019, pp. 324–335. 2
- [10] L. Yang, B. Kang *et al.*, “Depth Anything V2,” in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey *et al.*, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 21 875–21 911. 2, 5
- [11] M. Milford, W. Scheirer *et al.*, “Condition-invariant, top-down visual place recognition,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 5571–5577. 2
- [12] R. Arandjelovic, P. Gronat *et al.*, “NetVLAD: CNN Architecture for Weakly Supervised Place Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016. 2, 5, 6

- [13] M. Oquab, T. Darcet *et al.*, “DINOv2: Learning Robust Visual Features without Supervision,” *Transactions on Machine Learning Research*, Jul. 2023. [2](#)
- [14] A. Ali-bey, B. Chaib-draa, and P. Giguère, “BoQ: A Place is Worth a Bag of Learnable Queries,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, Jun. 2024, pp. 17 794–17 803. [2](#)
- [15] N. Keetha, A. Mishra *et al.*, “AnyLoc: Towards Universal Visual Place Recognition,” *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1286–1293, Feb. 2024. [2](#)
- [16] S. Wang, V. Leroy *et al.*, “DUST3R: Geometric 3D vision made easy,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 20 697–20 709. [2](#)
- [17] V. Leroy, Y. Cabon, and J. Revaud, “Grounding image matching in 3D with MAST3R,” in *European Conf. on Computer Vision (ECCV)*, vol. 15130, 2024, pp. 71–91. [2](#)
- [18] J. Wang, M. Chen *et al.*, “Vggt: Visual geometry grounded transformer,” in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 5294–5306. [2](#)
- [19] Y. Yang, M. Lin *et al.*, “Understanding the impact of geometric foundation models on vision-language-action models,” 2026, arXiv:2605.24642 [cs.CV]. [2](#)
- [20] B. Yin, X. Zhang *et al.*, “DFormer: Rethinking RGBD representation learning for semantic segmentation,” in *The Twelfth International Conference on Learning Representations*, 2024. [2](#)
- [21] J. Komorowski, M. Wysoczanska, and T. Trzcinski, “MinkLoc++: Lidar and Monocular Image Fusion for Place Recognition,” in *2021 International Joint Conference on Neural Networks (IJCNN)*. Shenzhen, China: IEEE, Jul. 2021, pp. 1–8. [2](#)
- [22] W. Lei, Y. Ge *et al.*, “VIT-LENS: Towards Omni-modal Representations,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 26 637–26 647. [2](#)
- [23] S. Mahendran, D. Priyasad *et al.*, “TOLiD: Bridging the architecture gap in vision foundation model to lidar pretraining via token lifting for distillation,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2026. [2](#)
- [24] X. Wang, X. Han *et al.*, “Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 5017–5025. [4](#)
- [25] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, “Mixvpr: Feature mixing for visual place recognition,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 2998–3007. [5](#), [6](#)