

Depth-Aware Distillation: Lightweight depth features for localization in natural environments

Walter Nedov^{1,2}, Saimunur Rahman¹, Kavindie Katuwandeniya¹, David Hall¹, Kaushik Roy¹, Peyman Moghadam^{1,3}
¹CSIRO Robotics, ²University of Queensland and ³Queensland University of Technology
Brisbane, Australia

Introduction

Problem

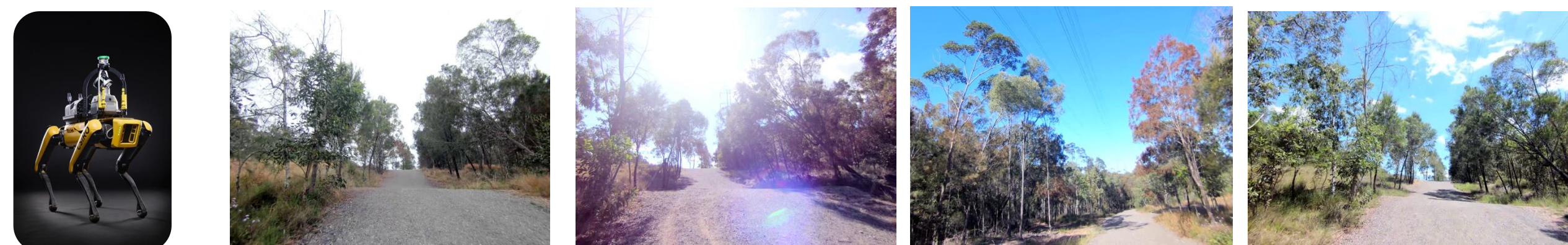
Unstructured forests provide a highly challenging environment for autonomous mapping systems due to their unstructured nature. Visual place recognition (VPR) systems particularly struggle with false negatives due to appearance variations when revisiting locations.

Proposed Solution

Our solution is depth-aware distillation (DAD), a lightweight fine-tuning strategy for adding depth to the feature space of VPR systems while keeping the original RGB encodings.

Challenging Data

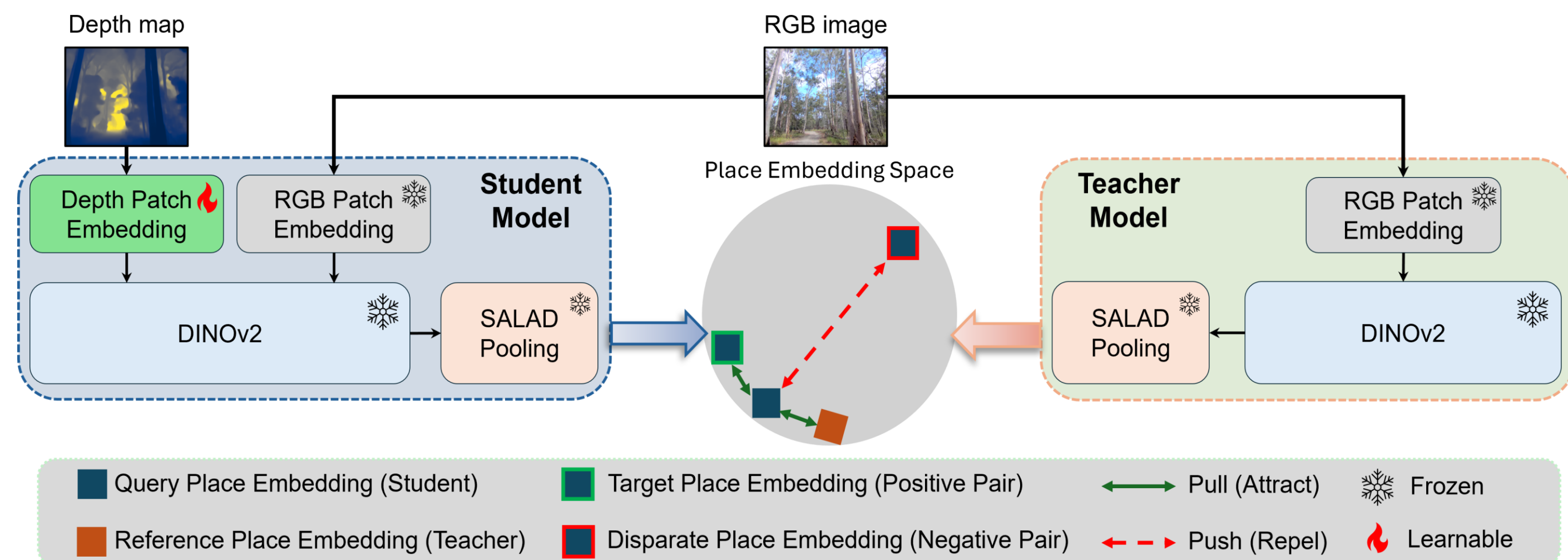
We test the validity of our method for localization in natural environments using the new WildCross Dataset [1]. Designed to benchmark field-robotics tasks like VPR in natural environments, there are many instances of the same place looking visually very different.



Results

DAD increases VPR performance, reducing errors from lighting inconsistencies and reverse revisits.

Method — Depth-Aware Distillation (DAD)



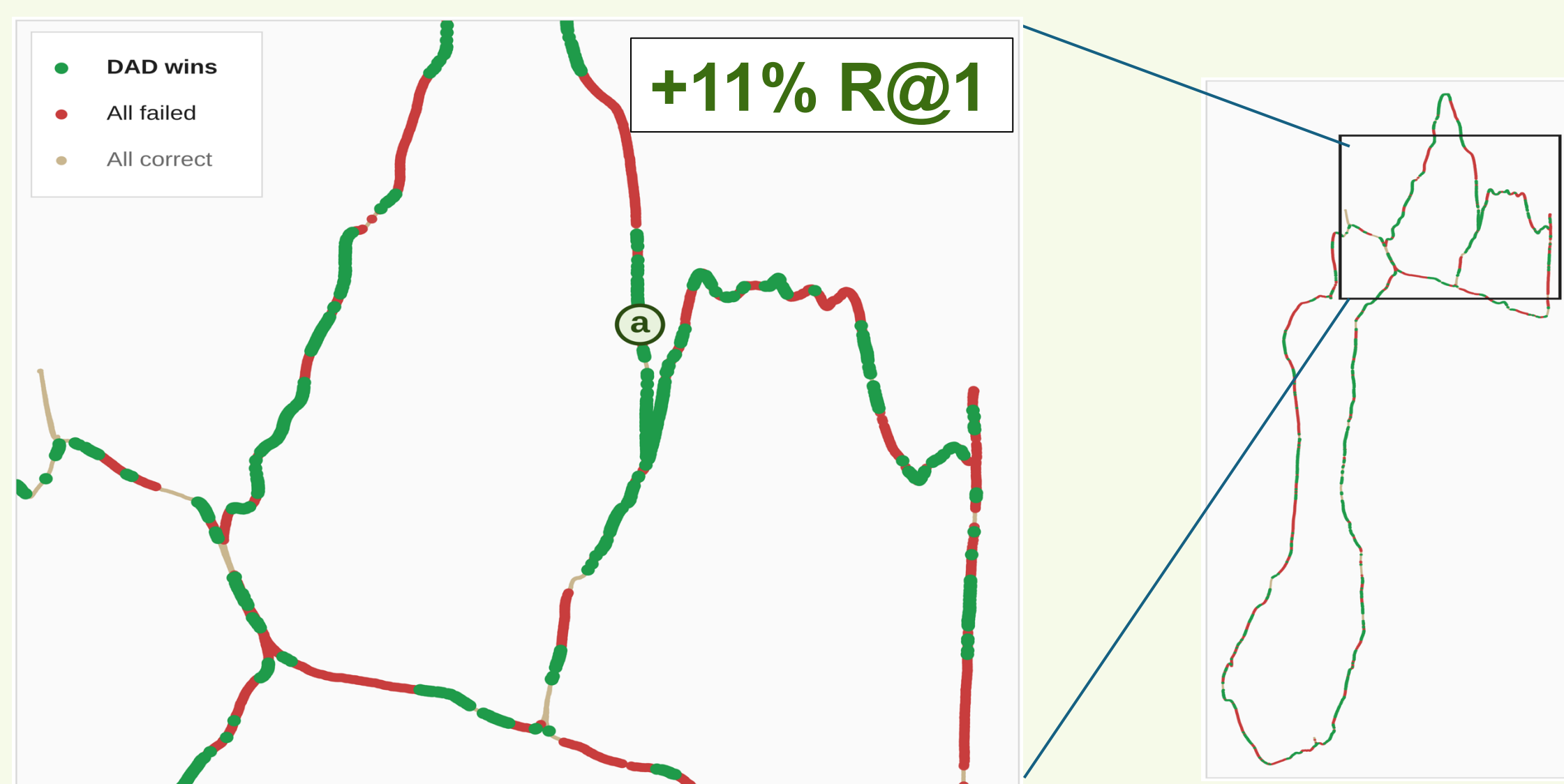
Depth-Aware Distillation (DAD) for Robust Visual Place Recognition

- Teacher-Student Architecture:** A frozen DINOv2 [2] teacher model extracts robust appearance-based descriptors to guide a depth-augmented student. Monocular depth maps extracted from the pre-trained Depth Anything v2 model [4] are used to learn a depth patch embedding and an adaptive mixing coefficient that fuses structural geometry with visual texture while preserving strong RGB representations.
- Embedding Space Alignment:** Synchronizes embeddings via contrastive loss on SALAD descriptors [3], pulling matching pairs together and pushing non-matching pairs apart.

Where does depth (DAD) help?

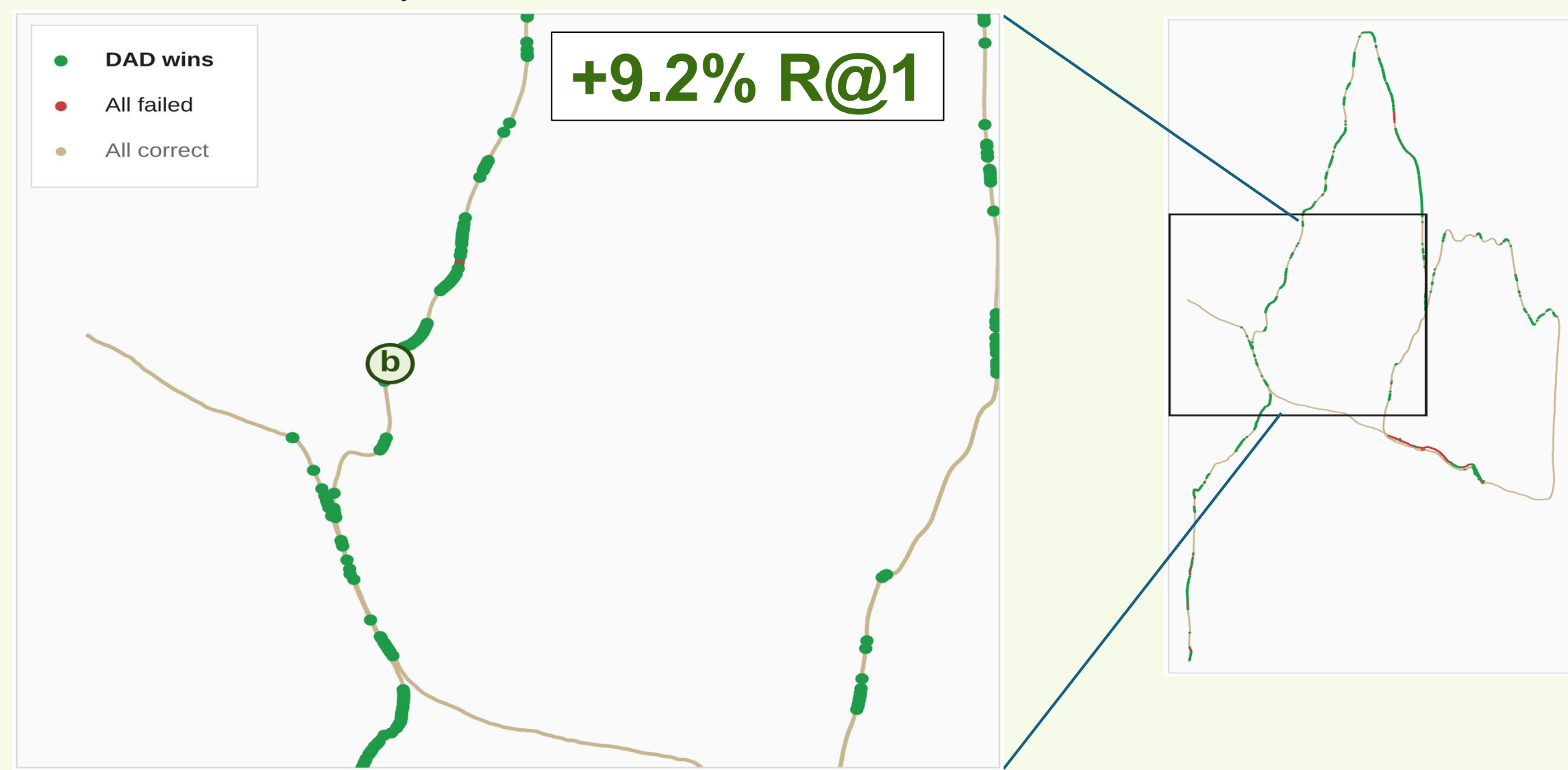
Reverse Revisit

Query K-02, Database K-01



Appearance Variation

Query K-04, Database K-01



Case studies

a Reverse Revisit

Query Image



DAD Retrieved



Baseline Retrieved



Reverse Revisit: DAD finds the correct match; Baseline does not. Red ovals show the same tree from opposite directions. Baseline matches a similar-looking tree elsewhere on the trail.

SALAD Assignment

DAD (Bottom) vs. Appearance (Top)



DAD focuses attention onto viewpoint-robust features, such as the path and tree shapes

b Appearance Variation



Even when the RGB images are dissimilar, similar depth maps act as a feature enhancer

Results

Method	Intra-Sequence [1]																Inter-Sequence [1]							
	V-01		V-02		V-03		V-04		K-01		K-02		K-03		K-04		Average		Venman		Karawatha		Average	
	R1	R5	R1	R5	R1	R5	R1	R5	R1	R5	R1	R5	R1	R5	R1	R5	R1	R5	R1	R5	R1	R5	R1	R5
Baseline	50.64	57.01	44.60	49.06	13.05	19.58	43.72	48.53	78.48	83.50	54.71	68.16	26.67	32.78	54.93	59.31	45.85	52.24	57.49	64.49	41.27	50.14	49.38	57.32
Baseline fine-tuned	68.10	72.40	72.20	76.80	23.60	31.02	59.80	64.40	86.30	88.50	84.40	90.00	30.00	38.20	59.80	62.00	60.50	65.40	66.60	71.01	49.68	55.81	58.14	63.41
DAD (Ours)	64.00	68.66	70.12	75.67	28.05	37.78	60.65	68.39	86.63	87.89	88.52	93.38	30.27	35.86	58.35	59.85	60.82	65.94	68.35	73.70	57.04	65.67	62.69	69.68

Baseline: Pre-trained RGB Patch Embedding, DINOv2 backbone and SALAD pooling layer; **Baseline fine-tuned:** Fine-tuned RGB Patch Embedding, Pre-trained DINOv2 backbone and SALAD pooling layer;
DAD: Fine-tuned Depth Patch Embedding, Pre-trained RGB Patch Embedding, Pre-trained DINOv2 backbone and SALAD pooling layer.

Cross-sequence VPR R@1 Comparison

	Baseline fine-tuned				DAD				Δ (DAD - Baseline fine-tuned)			
	V-01	V-02	V-03	V-04	V-01	V-02	V-03	V-04	V-01	V-02	V-03	V-04
Venman Query												
V-01		48.7	95.1	95.5		50.9	93.5	94.9		+2.2	-1.6	-0.6
V-02	38.1		45.5	29.7	44.8		52.2	37.0	+6.6		+6.6	+7.3
V-03	86.6	49.8		85.6	86.6	47.1		85.6	-0.0	-2.8		+0.0
V-04	94.4	44.0	86.6		95.1	45.5	87.2		+0.7	+1.5	+0.6	
Karawatha Query												
K-01		31.3	65.8	92.5		37.6	73.3	93.4		+6.3	+7.5	+0.9
K-02	27.3		48.1	37.3	38.3		57.5	48.3	+11.0		+9.4	+11.0
K-03	54.0	38.5		54.6	56.3	41.5		58.1	+2.4	+3.1		+3.4
K-04	66.7	27.1	53.0		75.9	41.6	62.6		+9.2	+14.6	+9.6	

We note that the largest inter-sequence gains from DAD come from reverse revisits (01 vs 02 sequences) and between sequences with high visual disparity such as K-04 which has different weather conditions to all other K sequences

Conclusions & Future Work

- Unstructured natural environments present unique challenges for VPR systems
- Depth estimation models provide a 'free' complementary signal for VPR.
- Our DAD approach provides a lightweight method adding depth information while preserving original descriptors.
- DAD provides robustness to visual discrepancies and reverse revisits. Conditions common in field deployment.

Future work will target (1) using LiDAR point assisted metric depth images, (2) evaluate DAD on more recent vision foundation models, (3) scale the training with larger datasets.

Ablations

Why not just augment RGB?

Depth Patch ft.	Methods			Results (Average)	
	Baseline Patch ft.	Horizontal Flip Aug.		Inter-Sequence R@1	Intra-Sequence R@1
	✓			58.1	60.5
	✓			49.5	60.0
✓				59.1	59.2
✓		✓		62.7	60.8
✓	✓			58.9	59.5

References

- [1] J. Knights, J. Reid, K. Roy, D. Hall, M. Cox, and P. Moghadam, "Wild-Cross: A Cross-Modal Large Scale Benchmark for Place Recognition and Metric Depth Estimation in Natural Environments," in IEEE ICRA, 2026.
- [2] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, et al., "DINOv2: Learning robust visual features without supervision," arXiv preprint arXiv:2304.07193, 2023.
- [3] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in IEEE/CVF CVPR, pp. 17658–17668, 2024.
- [4] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," NeurIPS, vol. 37, pp. 21875–21911, 2024.