

Towards Universal Place Recognition

Jay Karhade

CMU-RI-TR-24-31

June 19, 2024



The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:

Prof. Sebastian Scherer, *chair*

Prof. Michael Kaess

Dr. Wenshan Wang

Zhao Shibo

*Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Robotics.*

Copyright © 2024 Jay Karhade. All rights reserved.

To my parents and grandparents.

Abstract

Place Recognition is a key component for building a robust and reliable SLAM system that enables robot autonomy and global localization in complex scenarios for disaster-response and search-and-rescue tasks. Despite recent advances leveraging large-scale training and improved training techniques for visual, LIDAR and thermal place recognition, current systems remain fragile and are heavily engineered towards specific environments, limiting in-the-wild deployment. At the same time, the recent success of vision foundation models have shown impressive generalized and open-vocabulary behaviour in diverse environments for visual tasks.

Building on these key insights, we first demonstrate AnyLoc - a universal solution to Visual Place Recognition that works across diverse structured and unstructured environments without any re-training or fine-tuning. Despite being self-supervised and without any VPR-specific training, we show that aggregating these features helps us achieve up to $4\times$ significantly higher performance than state-of-the-art VPR systems. Furthermore, these features reveal distinct semantic domains corresponding to datasets from similar environments, helping us further improve performance.

We further develop MultiLoc and show that these features can be distilled into other modalities, namely LIDAR and thermal enabling cross-modal place recognition even in challenging environments. We evaluate our approach by repurposing existing public datasets for visual-LIDAR-thermal place recognition datasets. For the first time we show that we can achieve zero-shot cross-modal place recognition between unseen modalities at test time. The experiments and analysis in this thesis lays a foundation for building VPR solutions that may be deployed anywhere, anytime, across anyview and on any-sensor.

Acknowledgments

At the outset, I would like to thank my advisor, Basti. Without his continuous guidance and mentorship, my love for perception would not have been kindled. When I joined the M.S. Robotics Program, I knew little about perception and field robotics. In these 2 years, I have gained an extra-ordinary amount of knowledge and diverse viewpoints through my discussions with him, and learnt to identify problems that with a real-world impact. I am grateful for the academic freedom he has nourished at AirLab, and the resources that I've been provided with to carry out my research. I admire his ability to look at the bigger picture as a researcher, and I hope to continue learning this in the future. Basti - thanks for taking a chance on me 2 years back!

I would like to express my gratitude towards my MSR thesis committee members, Prof. Michael Kaess, Dr. Wenshan Wang and Shibo Zhao. Their constant feedback on my research has been extremely constructive. I have learnt some amazing perspectives from Michael on SLAM, and I have looked up to his research even before joining CMU. I am incredibly grateful for Wenshan for being an amazing mentor in the AirLab. Her feedback on my thesis, as well as the countless research discussions I have had with her have influenced my thoughts as a researcher. I would also like to thank Shibo, who has been a great mentor, lab-mate and friend. I have had incredible discussions with him on the kind of research worth pursuing and I am inspired by his dedication, kindness and constant support throughout my MSR studies.

This thesis is the culmination of the two years I have spent so far in the AirLab, and it would have not been possible without the wonderful peers I have met there. This thesis would be incomplete without mentioning Nikhil - who has been a great friend over the past 2 years. His push on research collaborations has been extremely helpful, and I've had some amazing research discussions, late-night coding sessions with him. To say that my academic networking is majorly attributed to working with him would be an understatement, and I look forward to more research collaborations in the coming years! (P.S. thanks for being a driver to some nice places!) I would like to thank Yao He and Nayana for being awesome lab-mates and friends and work together on some very fun projects! I have had also some amazing interactions with other labmates including Jay P. , Brady, Conner, Matt, Yifei and Yuheng both in and outside the lab.

I would like to express my gratitude towards Sourav Garg, K.M. and Avneesh. AnyLoc and this thesis would not be possible without collaborating with them. I have learnt a lot from all of you and I hope to continue collaborating with you all in the future. I am also fortunate to have collaborated with Jonathan and Deva on SplaTAM - their opinions have shaped my research opinions as I enter my doctoral studies.

In the last 2 years at the Robotics Institute, I have met some extremely gifted people and have the privilege of calling them friends. Aman has been the most awesome running, biking, kayaking and adventure buddy and an extremely fun person throughout the program and I am so glad that I met him early on in the program. (He ran his first half-marathon this May!!). I am grateful to have met Bharath (GBhai) who has been a gem of a person and also someone who has greatly shaped my academic and personal outlook to life. I haven't witnessed anyone with more patience, understanding and kindness. He and Aman have been a constant partner in mischief throughout. I am glad that I met Prachi and I have grown to admire her dedication and thirst for research. I've had some amazing late-night discussions with her over the past year and I cannot wait to see what she does next at UIUC! Pittsburgh would not have been the same without the three of you. I would also like to thank Adi and Andrew for being awesome friends - the trip to Puerto Rico was amazing! I have had the privilege of knowing some amazing folks from the MSCV and MRSD programs. Achleshwar, Richa and Chris have been amazing friends since the first semester and I've had a blast of a time with them. The entire MSCV cohort has been extremely supportive - I would name all of you if only there was enough space on this page! Shout-out to Bhuvan and Roshan who have been great labmates, friends, and gym-bros.

I would like to also thank my past mentors - notably Prof. Rajesh Tripathy, Prof. Marcelo Ang, and Dr. Zhu Haiyue for believing in me back then and igniting in me the spark for computer vision and robotics, as well as their continuous support during my Master's and PhD applications. I would not be here without them.

Finally, I would like to acknowledge the everlasting support of my Mom and Dad. You guys are the most amazing parents anyone could ever have, and I'm incredibly grateful to you both for all the sacrifices you have made. Mom, thank you for always being there, and showing me tough love when I needed it. Dad, thanks for being a constant pillar of our family. Finally, I would like to mention my grandparents. Ajoba, thank you for all the

life-lessons you've taught. Aiji, thank you for the everlasting love - I wish you could have read this.

There are many more people who could not be named here who have been an integral part of my journey. You know who you are - and I am indebted to you for your support!

Funding

This work was supported by ARL grant W911QX20D0008/W911QX22F0078(TO6). Parts of this work used Bridges-2 at PSC through allocation cis220039p from the ACCESS program, which is supported by NSF grants 2138259, 2138286, 2138307, 2137603, and 213296.

Contents

1	Introduction	1
1.1	Key Contributions	2
2	Place Recognition for Multi-Robot SLAM	3
2.1	System Overview	4
2.2	Place Recognition with Overlap-Transformer Rotary Embedding . . .	6
2.2.1	Rotation and Permutation Invariance in Overlap Transformer	9
2.2.2	Rotary Position Encoding Enhanced Overlap Transformer . .	10
2.2.3	Conclusion	11
2.3	Acknowledgement	11
3	AnyLoc	13
3.1	Problem Introduction	14
3.2	Related Work	16
3.2.1	Foundation Models	16
3.2.2	Visual Place Recognition	17
3.3	Method	18
3.4	Choosing Foundation Models	18
3.4.1	DINO and DINOv2	19
3.4.2	CLIP - Contrastive Language Image Pre-training	19
3.4.3	MAE - Masked AutoEncoders	19
3.4.4	SAM - Segment Anything Model	19
3.5	Choosing Feature Extraction	20
3.6	Choosing Feature Aggregation	20
3.7	Choosing Vocabulary Design	22
3.8	Datasets	23
3.8.1	Structured Environments	23
3.8.2	Unstructured Environments	25
3.9	Baselines and Evaluation Metrics	26
3.10	Experiments and Analysis	27
3.10.1	State-of-the-art Comparison	28
3.10.2	Vocabulary Analysis	31
3.10.3	Insights into <i>AnyLoc</i> Design	33
3.10.4	Self-supervised vs VPR-supervised ViT	34

3.11	Conclusion	35
3.12	Acknowledgement and Contribution Statement	35
4	MultiLoc	37
4.1	Introduction	37
4.2	Related Work	38
4.2.1	Non-visual Place Recognition	39
4.2.2	Multi-Modal Place Recognition	39
4.2.3	Multi-Modal Foundation Models	40
4.3	Method	40
4.4	Datasets and Metrics	41
4.5	Experiments and Analysis	42
4.6	Limitations and Conclusions	47
5	Conclusion and Future Work	49
	Bibliography	51

When this dissertation is viewed as a PDF, the page header is a link to this Table of Contents.

List of Figures

2.1	Overview of map from our multi-robot SLAM system.	3
2.2	An overview of the sensor pack used in SubT-MRS dataset. It is equipped with a Xavier processing unit with hardware time synchronization for multimodal sensors including LIDAR, fisheye cameras, thermal cameras, depth cameras (option), and an IMU.	4
2.3	Each robot builds a local map using superodometry and maintains LIDAR-IRIS/Scan-Context descriptors of keyframes. When another robot is in proximity, the robots communicate their stored descriptors and try to search for inter-robot loop closures through a 2-stage verification protocol, after which the inter-robot transformations are calculated. Information to the base-stations is relayed whenever possible and a globally-consistent map is displayed.	5
2.4	The figure demonstrates the largest set of pairwise consistent transformations being used to close the inter-robot loop while the outlier place-recognition matches are discarded after PCM. PCM leverages the pairwise consistency between measurements of 2 different robots to identify the largest self-consistent set of keypose transformations, and are over a certain threshold. This makes PCM robust to false loop closures.	6
2.5	Erroneous Loop Closure	7
2.6	Environment causing erroneous loop-closure: Long and feature-less corridors make it challenging for both Lidar and Visual Place Recognition to reliably detect loop closures.	7
2.7	Overlap-Transformer produces the same descriptor even if the point cloud is permuted or a point-cloud with similar values but different geometry is used.	8
2.8	Adding rotary positional encoding to the Overlap-Transformer preserves rotational invariance and produces a different descriptor if the point-cloud is different or permuted, even though it may have similar values.	8

3.1	AnyLoc enables <i>universal</i> visual place recognition (VPR) across a massively diverse set of environments (<i>anywhere</i>), temporal changes (<i>anytime</i>), and a wide range of viewpoint variations (<i>anyview</i>). AnyLoc achieves this by aggregating per-pixel features extracted from large-scale pretrained models (<i>foundation models</i>), <i>without any training or finetuning</i> . In the PCA panels (<i>middle</i>), 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 <i>domains</i> encompassing datasets with similar properties, marked with the same color. Using these <i>domains</i> 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 (<i>right</i>), across structured (urban outdoors, indoors) and unstructured (underwater, aerial, subterranean, visually degraded) environments.	13
3.2	Point correspondences (as markers) & similarity maps show the robustness of foundation model features to various VPR challenges : (<i>top</i>) text and scale change, (<i>middle</i>) perceptually aliased features and viewpoint shift, and (<i>bottom</i>) 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.	21
3.3	Qualitative ablation comparing the absolute-scale similarity maps of features from different DINOv2 ViT-G <i>layers</i> and <i>facets</i> . Layer 31 value facet has the sharpest contrast in the similarity map, which is crucial for robustness against distractors within an image	22
3.4	Hawkins Retrieval Visualizations	28
3.5	Laurel Caverns Retrieval Visualizations	29
3.6	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.	32
3.7	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.	32
4.1	We show that binding LIDAR and thermal modalities to features to vision foundation models is effective to achieve <i>zero-shot</i> cross-modal place recognition.	37

4.2	We distill image features into other modalities through student-teacher training where the teacher(vision foundation model) is frozen, and the student(modality-specific model) is updated through an InfoNCE[61] loss.	40
4.3	Qualitative Retrievals for Visual-Thermal Place Recognition on a night-time MS2 sequence	44
4.4	Qualitative Retrievals for Thermal-LIDAR Place Recognition on a night-time MS2 sequence	44
4.5	Qualitative Retrievals for Visual-LIDAR Place Recognition on a night-time MS2 sequence	45
4.6	Qualitative Retrievals for Visual-Thermal Place Recognition on the Idyll-Wild Sequence	46
4.7	Qualitative Retrievals for Visual-Thermal Place Recognition on the Big-Bear Sequence	47

List of Tables

3.1	Unstructured Environments used in Evaluation	25
3.2	State-of-the-art Baselines used for Comparison	26
3.3	Performance comparison on Benchmark Structured Environments . .	27
3.4	Performance Comparison on Unstructured Environments	27
3.5	Effect of vocabulary type on R@1 for <i>AnyLoc-VLAD-DINOv2</i>	30
3.6	Analysing intra-domain transferability of <i>AnyLoc-VLAD-DINOv2</i> vo- cabularies	31
3.7	Analysis comparing the Recall@1 & Descriptor Dimensionality across varying aggregation methods	34
3.8	Analysis comparing the Recall@1 of VPR-trained ViTs to Self-supervised ViTs	34
4.1	Cross-Modal Place Recognition MS2 dataset - Rainy Sequences . . .	43
4.2	Cross-Modal Place Recognition on MS2 dataset - Night-time Sequences	43
4.3	Visual-Thermal Place Recognition CART dataset	46

Chapter 1

Introduction

As robot-teams are increasingly deployed in unstructured and previously unseen environments to carry out critical tasks such as search and rescue, disaster-response etc, it becomes increasingly important to have a general and multi-modal perception system to achieve resilient autonomy. This has been increasingly seen through recent challenges such as the DARPA Subterranean Challenge[21], which have propelled the need for robots to operate in such environments. At the center-stage of such robot-teams is building a robust SLAM(Simultaneous Localization and Mapping) system that works across different perceptual aliasing and degradation. Recent advances such as SuperOdometry[98] have laid the foundation for odometry and mapping systems that utilize multiple sensors and are robust to various environmental degradation commonly present in these environments such as smoke, lighting changes and texture-less regions such as long-corridors. However such systems are still prone to drift over long regions, which have prompted the need to incorporate systems that perform place-recognition for loop-closure.

While Place Recognition is a fundamental and well-studied component in localization systems for over 2 decades, most place recognition systems are uni-modal, and trained for specific environments - hindering them from being deployed on multi-robot platform in the wild. This naturally begs the need for a place-recognition system that can work out-of-the-box *anytime, anywhere, anyview and across any-sensor*. To this end, the work discussed in this thesis takes a step towards building such a system. Specifically we leverage recent advances in large-scale pre-trained models

(foundation models) and demonstrate that these models have learnt rich discriminative features, which is highly desirable for the task of place-recognition.

1 introduces the theme of this thesis and provides a motivation of the problem. 2 provides an overview of the multi-robot SLAM system which we built, and the limitations and challenges faced in place recognition, setting the context for the research carried out in this thesis. Subsequently, 3 describes how foundation model features are extremely effective towards providing state-of-the-art visual place recognition performance across diverse environments without assuming VPR-specific training. We further show that these features can be distilled into different modalities, enabling zero-shot uni-modal and cross-modal place recognition in 4. Finally, we conclude the thesis and talk about the future directions for place-recognition and localization systems in 5.

1.1 Key Contributions

The key-contributions of this thesis can be summarized as follows:

- We build an real-time online multi-robot system with improved online lidar place recognition, and demonstrate by deploying it in scenarios with perceptual degradation.
- We then propose a new method AnyLoc, that demonstrates state-of-the-art(SOTA) performance for visual place recognition in a zero-shot manner across diverse environments.
- Finally, we show that features from visual foundation models can be distilled into other modalities, enabling zero-shot cross-modal place-recognition, namely Visual, LIDAR and Thermal sensors.

Chapter 2

Place Recognition for Multi-Robot SLAM

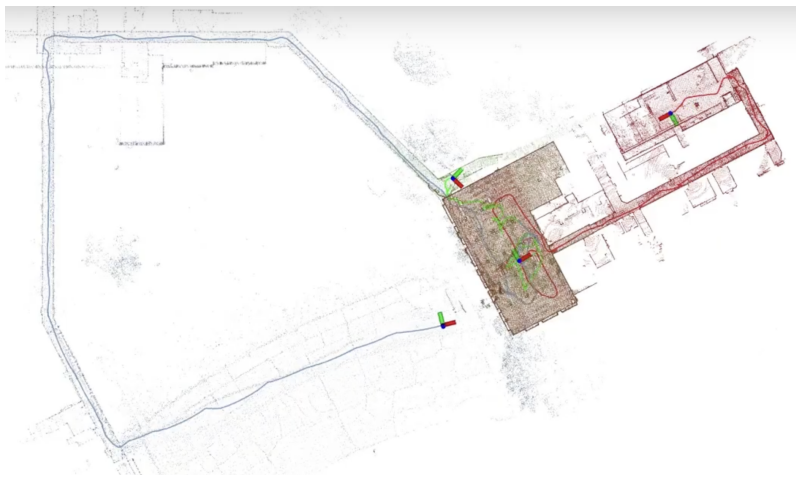


Figure 2.1: Overview of map from our multi-robot SLAM system.

In this section, we first describe the built multi-robot SLAM system that was scaled to a group of 3 robots. The multi-robot SLAM system is distributed, performed online LIDAR place recognition and works in communication-constrained environments. We also discuss the subsequent improvements proposed to a learning-based LIDAR place recognition system for robustness under perceptual degradation.

2.1 System Overview

Let $X = x_1, x_2, \dots, x_n$ be a group of n -robots collaboratively performing SLAM for an environment. Each robot x_i builds a local map consisting of m robot key frame poses $x_{i1}, x_{i2}, \dots, x_{im}$. We wish to find all transformations $T_{ij}, i \neq j$ between robots i and j , such that we can fuse the obtained local-maps into a global map.

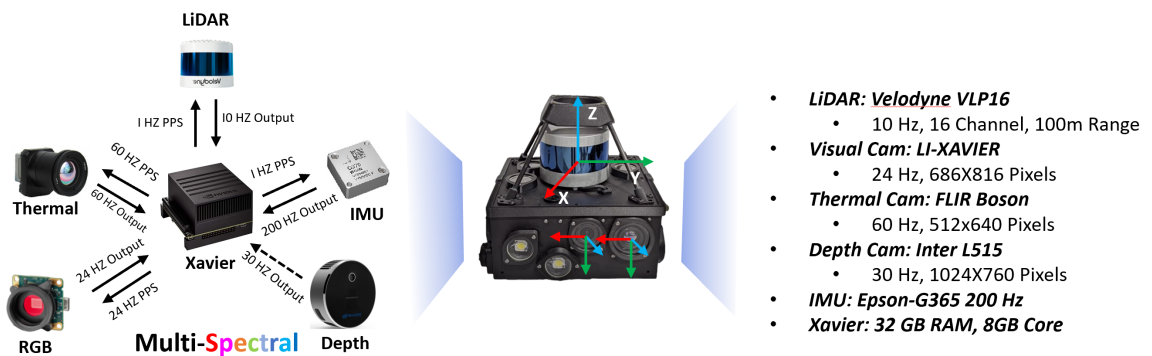


Figure 2.2: An overview of the sensor pack used in SubT-MRS dataset. It is equipped with a Xavier processing unit with hardware time synchronization for multimodal sensors including LIDAR, fisheye cameras, thermal cameras, depth cameras (option), and an IMU.

Our system is visible in 2.3 and consists of payloads mounted on mobile-robots which have been highlighted in 2.2. We adopt the pipeline from DCL-SLAM. In particular, we use a distributed approach to multi-robot SLAM as illustrated in Figure 2.1, where each robot maintains a local map of the environment, and calculates its global transform w.r.t to a master robot.

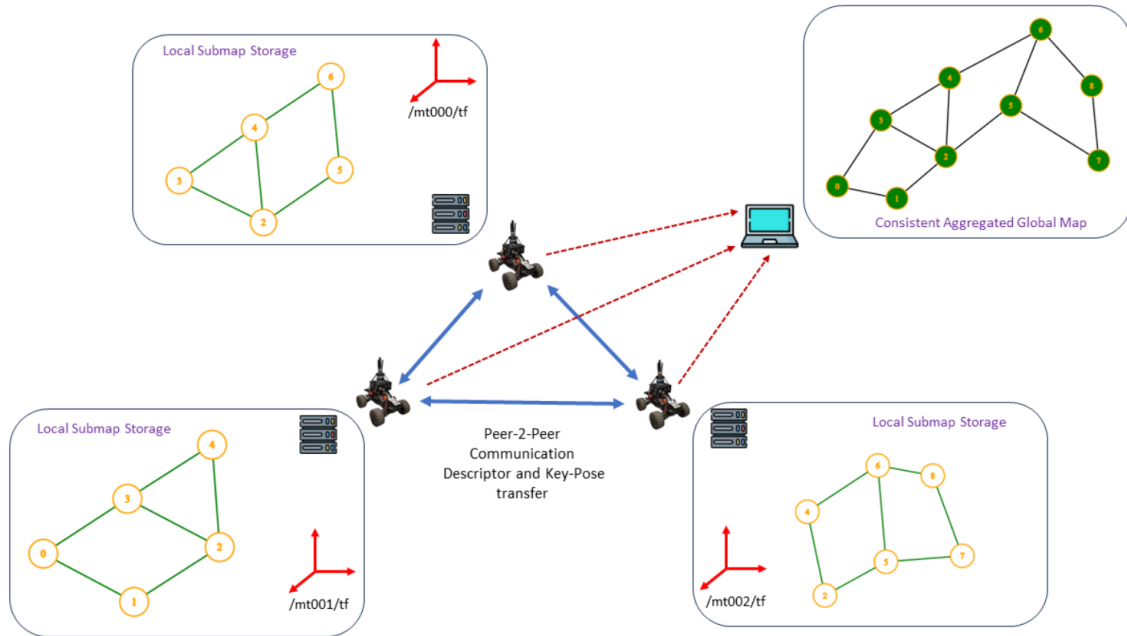


Figure 2.3: Each robot builds a local map using superodometry and maintains LIDAR-IRIS/Scan-Context descriptors of keyframes. When another robot is in proximity, the robots communicate their stored descriptors and try to search for inter-robot loop closures through a 2-stage verification protocol, after which the inter-robot transformations are calculated. Information to the base-stations is relayed whenever possible and a globally-consistent map is displayed.

Specifically, this global alignment is achieved by performing inter-loop closures through LIDAR Place Recognition(LPR). Furthermore, each robot also performs intra-loop closures to prevent drift of local maps. To detect this loop-closures, the well-established Scan-Context descriptor[39] is used. However, due to the perceptual aliasing and sparsity of LIDAR-Scans, we empirically observe that this tends to produce spurious loop-closures, necessitating the need for outlier-rejection. To this end, we use the Pair-wise Consistency Maxmization [56]. 2.4 demonstrates the an example of false loop-closures being rejected, while accepting consistent loop-closure detections for calculating inter-robot transformation.

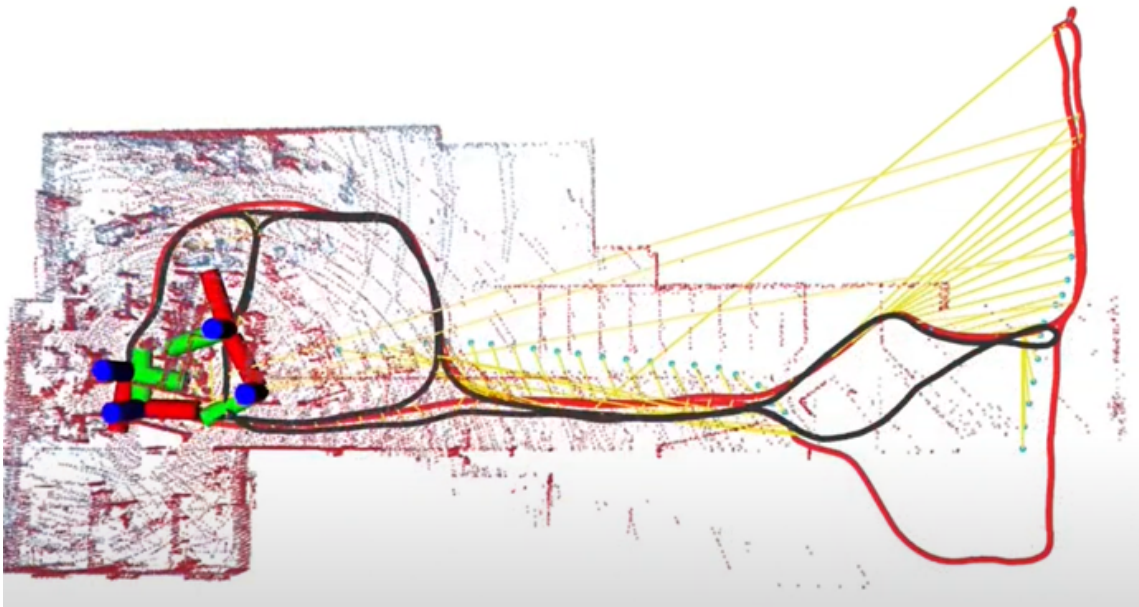


Figure 2.4: The figure demonstrates the largest set of pairwise consistent transformations being used to close the inter-robot loop while the outlier place-recognition matches are discarded after PCM. PCM leverages the pairwise consistency between measurements of 2 different robots to identify the largest self-consistent set of keypose transformations, and are over a certain threshold. This makes PCM robust to false loop closures.

2.2 Place Recognition with Overlap-Transformer Rotary Embedding

While classical global-descriptor methods like Scan-Context[39] and its variants ranging from ScanContext++[40] to LIDAR-IRIS[86] have been popularly used and adopted due to its simple aggregation method as well as ability to provide a coarse alignment, these often fail in sparse-pointclouds and geometrically-aliased environments as also indicated during our system’s deployment.

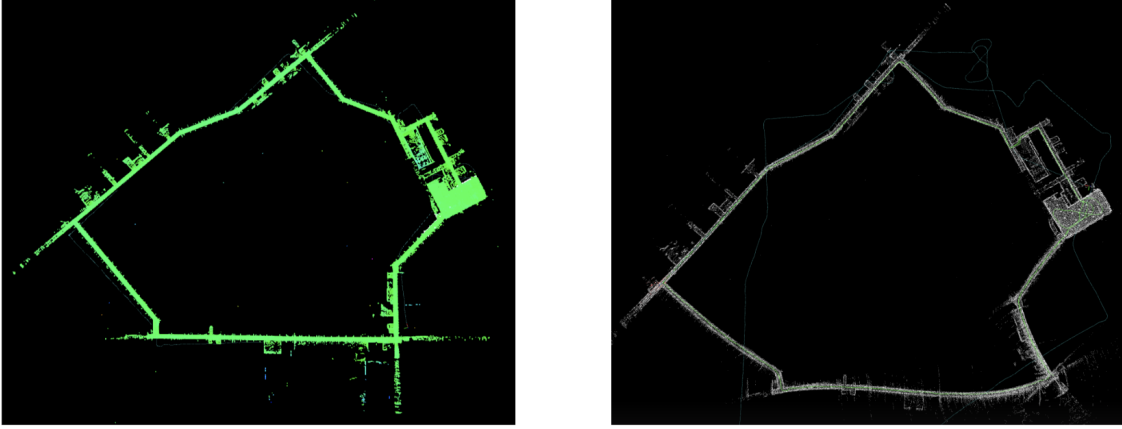


Figure 2.5: Erroneous Loop Closure



Figure 2.6: Environment causing erroneous loop-closure: Long and feature-less corridors make it challenging for both Lidar and Visual Place Recognition to reliably detect loop closures.

More recently, learning based methods have become popular and provide better descriptors. OverlapNet[19] proposes a learning-based method by learning to predict the overlap between two pairs in the training process. OverlapTransformer[?] and SeqOT[50] further extend this by adopting an attention-scheme to enhance discriminative features. Furthermore, by modifying the CNN from OverlapNet and equivariance properties of transformers, OverlapTransformer and SeqOT also ensure yaw-angle rotation invariance in their final descriptor generation.

However, Overlap-Transformer overlooks the problem of avoiding permutation-invariance which results in the same output descriptor even if the input range-images are different and performs a convolution approach, adding extra-computation to avoid this. To this end, we propose a simple solution by using Rotary Positional Encodings(RoPE)[77] and show that they preserve the yaw-angle invariance properties, while simultaneously avoiding permutation-invariance and show the advantage of these result.

2. Place Recognition for Multi-Robot SLAM

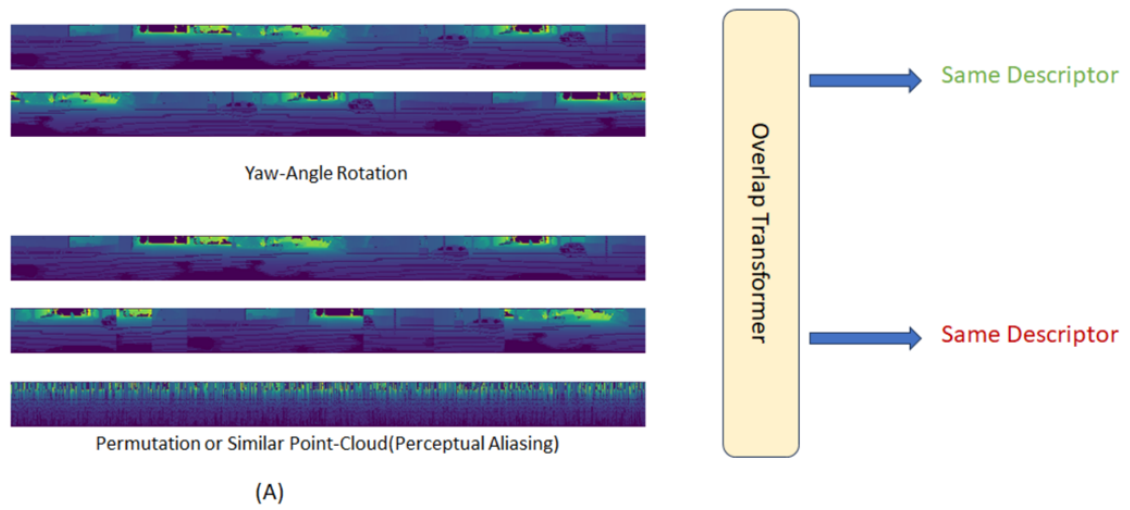


Figure 2.7: Overlap-Transformer produces the same descriptor even if the point cloud is permuted or a point-cloud with similar values but different geometry is used.

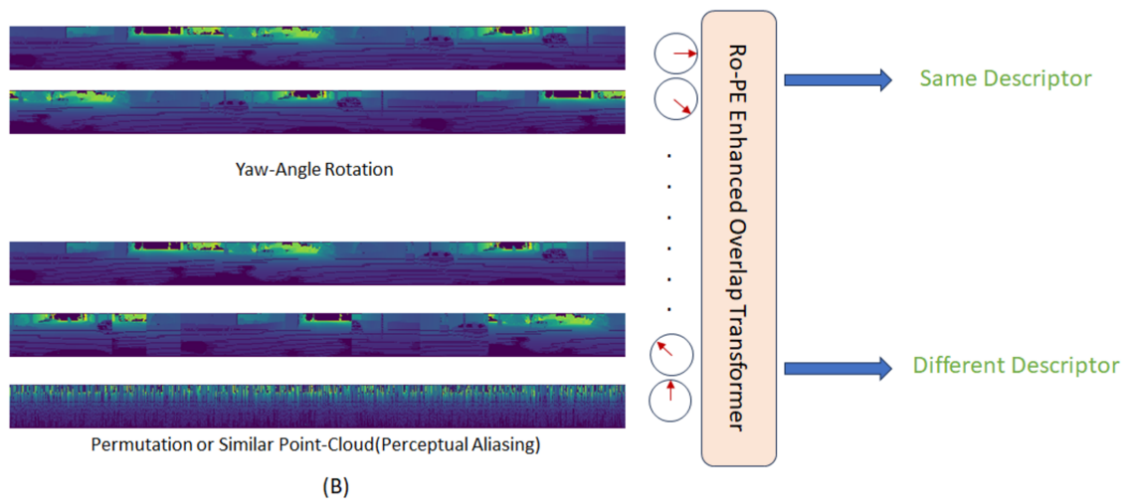


Figure 2.8: Adding rotary positional encoding to the Overlap-Transformer preserves rotational invariance and produces a different descriptor if the point-cloud is different or permuted, even though it may have similar values.

2.2.1 Rotation and Permutation Invariance in Overlap Transformer

Range-images and equi-rectangular projections are yaw-angle equivariant, i.e, a yaw-rotation of the LIDAR will perform a circular shift of the columns. Denoting the columns of the range image as $C = \{c_1, c_2, \dots, c_n\}$, performing a rotation R on the LIDAR will result in the following column order on the range-image.

$$R\{c_1, c_2, \dots, c_n\} = \{c_i, c_{i+1}, \dots, c_n, c_1, \dots, c_{i-1}\} \quad (2.1)$$

Overlap-Transformer uses range-images as input to get a feature-volume from a modified form of OverlapNet wherein the convolution filters are only performing convolution in the vertical dimension, and not the width dimension to avoid any discretization error. For an input range image of size $h \times w \times 1$, the size of the output feature volume is given by $1 \times w \times c$ where c is the number of encoded channels from the range image encoder.

The transformer module is then used to enhance the output volume features. The transformer module preserves yaw-angle equivariance since transformers are permutation-equivariant and any permutation will only change the order sequence of the tokens, and not affect individual tokens. If we represent the input to the transformer module as a set $C = \{c_1, c_2, \dots, c_n\}$ and the obtained output set as the set $O = \{o_1, o_2, \dots, o_n\}$ performing any permutation P on the input does not change the output set.

$$O' = F(PC) = PO = O \quad (2.2)$$

Performing a soft-NetVLAD operation on this set outputs an invariant descriptor. Although OverlapTransformer can provide rotationally-equivariant and rotationally-invariant descriptors from the transformer layer output and the NetVLAD layer output respectively, the method is also permutationally-invariant, which is not desirable as permuting the columns of a range-image corresponds to a different LIDAR-Scan. This poses a challenge especially in perceptually aliased environments and sparse-LIDAR scans.

2.2.2 Rotary Position Encoding Enhanced Overlap Transformer

We now prove that a simple modification to the OverlapTransformer’s transformer module by adding rotary positional encoding(RoPE) avoids the highlighted problem. RoPE explicitly encodes relative position between two input tokens to a transformer.

Unlike Absolute(APE) or Learned Positional Encoding(LPE)[74] that concatenates or adds a position embedding to the existing token and encodes absolute position before passing it to the self-attention layers, RoPE uses a rotation matrix to calculate the inner product between a query vector/token at position m q_m and key vector k_n , as the following relation :

$$q_m^T k_n = (R_{\Theta, m}^d W_q x_m)^T (R_{\Theta, n}^d W_k x_n) = x^T W_q R_{\Theta, n-m} W_k x_n \quad (2.3)$$

where $R_{\Theta, m}$ and $R_{\Theta, n}$ are rotation matrices whose angle is defined by the token location $\Theta = N^{-2(i-1)/d}, i \in [1, 2, \dots, d/2]$.

Given the output feature volume of size $1 \times w \times c$ from the OverlapNet leg, we take the input to the transformer-encoder as w number of tokens with a dimension c . Furthermore, since the LIDAR ranges in a sweep has uniform angular-spacing , we set $N = w$ in RoPE.

It can be shown that any rotation to the sequence will not affect the attention mechanism between the query-key pair since all key-query pairs’ positions undergo the rotation R , and the output sequence is simply a rotation-equivariant output sequence:

$$(R_s q_m)^T R_s k_n = (R_s R_{\Theta, m}^d W_q x_m)^T (R_s R_{\Theta, n}^d W_k x_n) \quad (2.4)$$

$$(R_s q_m)^T R_s k_n = x^T W_q R_{\Theta, n+s-m-s} W_k x_n \quad (2.5)$$

$$(R_s q_m)^T R_s k_n = x^T W_q R_{\Theta, n-m} W_k x_n = q_m^T k_n \quad (2.6)$$

At the same time, a permutation P to the columns of the LIDAR range image produces a different output :

$$(P_s q_m)^T P_s k_n = (P_s R_{\Theta, m}^d W_q x_m)^T (P_s R_{\Theta, n}^d W_k x_n) \neq (q_m)^T k_n \quad (2.7)$$

Since the output sequence set is different incase of a permutation, adding RoPE will always produce different descriptors after the VLAD aggregation layer is obtained even under high perceptual aliasing and similar values.

$$\{z_1, z_2, z_3, \dots, z_n\} = \{z_1^r, z_2^r, z_3^r, \dots, z_n^r\} \neq \{z_1^p, z_2^p, z_3^p, \dots, z_n^p\} \quad (2.8)$$

In contrast, APE or LPE do not preserve rotational equivariance since a rotation to the input scan changes the input tokens after concatenating with the APE/RPE. Hence they have not been used in OverlapTranformer.

$$\{z_1, z_2, z_3, \dots, z_n\} \neq \{z_1^r, z_2^r, z_3^r, \dots, z_n^r\} \neq \{z_1^p, z_2^p, z_3^p, \dots, z_n^p\} \quad (2.9)$$

Furthermore, using token residual concatenation is suboptimal to avoid permutation invariance. Given the input tokens from range-image $C = \{c_1, c_2, \dots, c_n\}$, the residuals are $R = \{(c_1 - c_2), (c_2 - c_3), \dots, (c_n - c_1)\}$. The modified input to the OverlapTransformer can then be represented as $E = \{(c_1 \oplus (c_1 - c_2)), (c_2 \oplus (c_2 - c_3)), \dots, (c_n \oplus (c_n - c_1))\}$. This method while avoiding permutation invariance for a single scan, results in different residual encoding for different scans and does not capture relative attention, both of which are used when using RoPE.

2.2.3 Conclusion

We show that a simple modification to the OverlapTransformer by adding Rotary Positional Encoding improves robustness of OverlapTransformer, especially under perceptual-aliasing having similar patterns or semantic degradation by preserving yaw-angle invariance, but avoiding permutation invariance. In the future, we aim to evaluate the performance of our method and deploy on our robots for challenging environments.

2.3 Acknowledgement

This work resulted in a first-authored workshop paper at IROS 2023 and was presented in the Last-mile delivery workshop.

2. *Place Recognition for Multi-Robot SLAM*

Chapter 3

AnyLoc

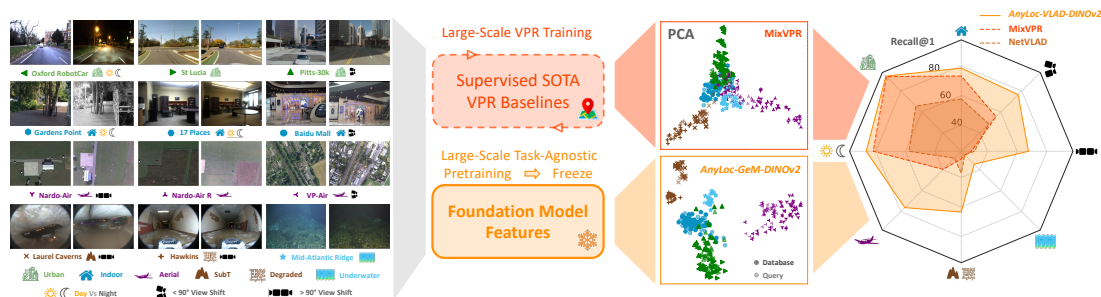


Figure 3.1: **AnyLoc** enables *universal* visual place recognition (VPR) across a massively diverse set of environments (*anywhere*), temporal changes (*anytime*), and a wide range of viewpoint variations (*anyview*). AnyLoc achieves this by aggregating per-pixel features extracted from large-scale 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 $4\times$ 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.

From our multi-robot system, we observe that LiDAR alone in many cases is insufficient to detect loop-closures, especially in environments like narrow and long corridors, where perceptual aliasing and feature-less environments cause false loop-

closure detections. In these scenarios, performing Visual Place Recognition (VPR) using onboard cameras emerges as a natural solution. In this chapter, we discuss our proposed method AnyLoc that performs Visual Place Recognition out-of-the-box in diverse unstructured environments *without any VPR-specific training*, serving as the perfect substrate to build a universal place recognition system.

3.1 Problem Introduction

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 (??). 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* [12]). 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 them with prevailing local-aggregation

approaches in the VPR literature [5, 8, 25, 75], resulting in all of the aforementioned desirable characteristics of a *universal* VPR solution.

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* [12]). 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 them with prevailing local-aggregation approaches in the VPR literature [5, 8, 25, 75], 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 [62]) and unsupervised aggregation methods (like VLAD [36] & GeM [66]) 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.

3.2 Related Work

3.2.1 Foundation Models

Foundation models [12] perform a wide array of tasks without the need for finetuning or additional re-training. This is mainly due to the way they are trained on a vast amount of data, which can include multiple modalities. These models tend to vary based on the type of supervision, where the primary categories are self-supervised [16, 62], weakly-supervised using multiple modalities [30, 67], and supervised [41]. Amongst self-supervised models [63, 75], there are primarily two broad categories, i.e., Joint-Embedding or Contrastive learning based methods [16, 62] and Reconstruction or Masked Image Modeling based methods [32].

Many recent approaches have explored the open-set properties of these foundation models for robotics in the context of planning and control, where the models have shown impressive open-set reasoning and interaction abilities [10, 13, 81?]. Similarly, there has been recent work exploring the properties of a self-supervised Vision Transformer (ViT) (DINO [16]) for dense visual descriptor extraction [3]. It has been showcased that these self-supervised ViT features encode rich semantic information across different object categories with fine spatial granularity, allowing them to be used as powerful dense visual descriptors for a wide range of applications, including part co-segmentation and keypoint correspondences. More recently, ConceptFusion [35] proposed a zero-shot approach to align features computed across regions without losing the open-set properties of foundation models, enabling spatial reasoning applications. Although there has been a wide variety of work exploring the use of foundation models in various applications, to the best of our knowledge, no current work explores the properties of these models in the context of “places”.

3.2.2 Visual Place Recognition

VPR is often cast as an image retrieval problem [24] that comprises two phases. In the indexing phase, a reference map 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 [8, 49, 64, 73, 95]. In this work, we focus on global descriptors which offer the best tradeoff between accurate matching and search efficiency [24, 36, 70]. This is in contrast to local descriptor methods, which are computationally intensive to match, particularly over larger databases.

Researchers have explored various training objectives [7, 27, 47, 92], aggregation techniques [5, 18, 66], and transfer learning [9, 31, 44] 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 [82], weakly-supervised contrastive learning in NetVLAD [5] led to substantial improvements over classical hand-crafted features. Following suit, the Google-Landmark V1 (1 million images) and V2 datasets [90] (5 million images) enabled training DeLF [60] and DeLG [15] 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 [85, 88, 102]. More recently, CosPlace [7] coupled classification-based learning with the San Francisco XL dataset comprising 40 million images having GPS & heading. The current SOTA, MixVPR [2], proposed an MLP-based feature mixer, trained on the GSV-Cities dataset [1] – 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 3.1, 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 [44, 80] and camera viewpoint [23, 26]. 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.

3.3 Method

Given a database of images represented by $\{I_{d1}, I_{d2} \dots I_{dn}\}$ of the trajectory or map that the agent traverses and a query image for a given place I_{qi} , our aim is to learn holistic and compact representations of places that can be used for retrieval. We observe that general general-purpose foundation model features exhibit excellent visual and semantic consistency, which is extremely useful for image-retrieval. However these features show sub-optimal performance when used as-is for place recognition.

Hence, we examine and notice that the performance of AnyLoc is greatly influenced by the choice of :

- Foundation Model
- Feature Extraction
- Feature Aggregation
- Construction of Database Vocabularies

3.4 Choosing Foundation Models

We compare 5 foundation models in our experiments namely, DINO, DINOv2, MAE, CLIP and SAM. Each of these models use a unique pre-training strategy for training Vision Transformers, allowing us to investigate the effect this has in the learnt representations of ViTs.

3.4.1 DINO and DINOv2

DINO [16] and DINO-v2 [62] are a family of foundation models that learn robust out-of-the-box visual features using joint-embedding prediction through self-distillation. DINO is trained using global supervision, while DINOv2 also uses patch-wise supervision.

3.4.2 CLIP - Contrastive Language Image Pre-training

CLIP [67] learns to align visual and language embeddings by pre-training a language encoder and a visual encoder to predict image-text pair representations over large batch-sizes. We generate place representations using the image-encoder.

3.4.3 MAE - Masked AutoEncoders

Masked Auto-Encoders [32] are a scalable self-supervised method to pre-train vision transformers by masking random patches of an input image, and using an asymmetric encoder-decoder architecture to predict the masked input patches. These models display visual features that give SOTA performance when completely fine-tuned. To get place representations, we discard the decoder and use the encoder’s representation only.

3.4.4 SAM - Segment Anything Model

Segment-Anything Model [41] is a ViT segmentation model composed of a heavy-weight image encoder and a lightweight promptable decoder. The model is first supervised with ground-truth segmentation annotations. The training is done through a data engine with the model being trained with manual annotations, followed by model-assisted annotation and finally model automated annotation. We use the image-encoder’s representation to get place representations.

Our experiments indicated that DINO and DINOv2 provide the best feature representation for VPR, followed by CLIP and SAM. These findings are corroborated in [62, 63, 75], highlighting the benefits of learning long-range global patterns captured by joint embedding methods. Furthermore, we observed that the performance of

MAE was much lower than the above models - which can be explained since MAE features are often interpreted as initialization to a full task-specific fine-tuning of the model and do not have out-of-the box visual semantic representations unlike the other foundation models. Hence we adopt DINO and DINOv2 as our backbone architecture for AnyLoc.

3.5 Choosing Feature Extraction

Feature extraction from Vision-Transformers can be done in different ways. While DINO and DINOv2 typically extract the CLS token for classification, we find that CLS token features provide sub-optimal performance compared to extracting per-patch intermediate features from the different layers of DINO. Furthermore, we examine these per-patch features for different facets in a ViT layer, i.e, key-facet, value-facet, query-facet and token-facet which corresponds to feature-maps taken after the subsequent feature representations as well examine this effect for different layers. We observe that:

- The token facet displays the clearest discrimination when comparing point-wise features between 2 images, as compared to the key,value or query facet as visible in Fig. 3.
- The early layers of DINO and DINO-v2 capture positional-information of points, while later layers capture semantic information.
- DINOv2 suffers from artefacts, unlike DINO. This was corroborated and subsequently addressed in works like [20].

3.6 Choosing Feature Aggregation

We examine the effect of feature aggregation techniques to obtain a global place descriptor. In addition to using the CLS token, we consider a comprehensive set of aggregation techniques, namely, Global Average Pooling (GAP) [6], Global Max Pooling (GMP) [68], Generalized Mean Pooling (GeM) [66], and the soft & hard assignment variants of VLAD [36].

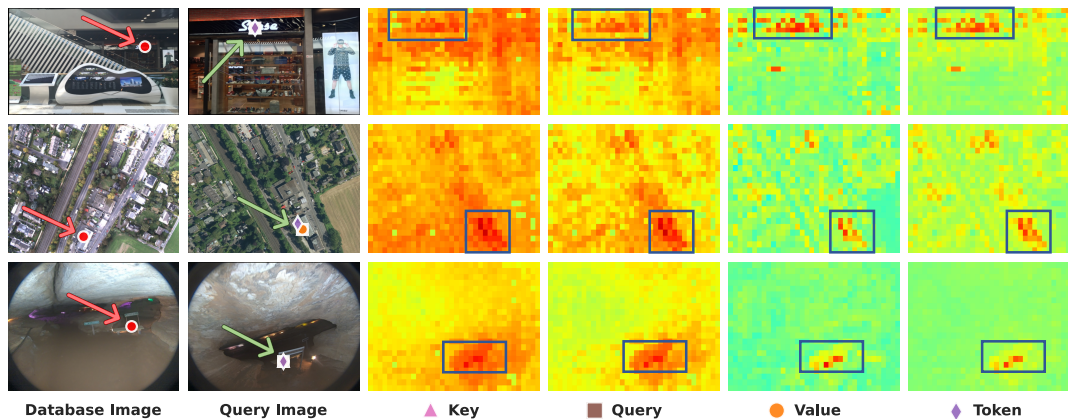


Figure 3.2: Point correspondences (as markers) & similarity maps show the **robustness of foundation model features to various VPR challenges**: (*top*) text and scale change, (*middle*) perceptually aliased features and viewpoint shift, and (*bottom*) low illumination combined with opposing viewpoint. The **value** facet has the highest contrast between the background and the matched points, which is vital for discarding distractors within an image.

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:

$$F_G = \left(\sum_{i=1}^{H \times W} f_i^p \right)^{\frac{1}{p}} \quad (3.1)$$

where $p = 1$, $p = 3$, and $p \rightarrow \infty$ 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:

$$F_{V_k} = \sum_{i=1}^{N \times H \times W} \alpha_k(f_i)(f_i - c_k) \quad (3.2)$$

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 [4], we perform intra-normalization, concatenation, and inter-normalization to obtain the final VLAD descriptor F_V .

3. AnyLoc

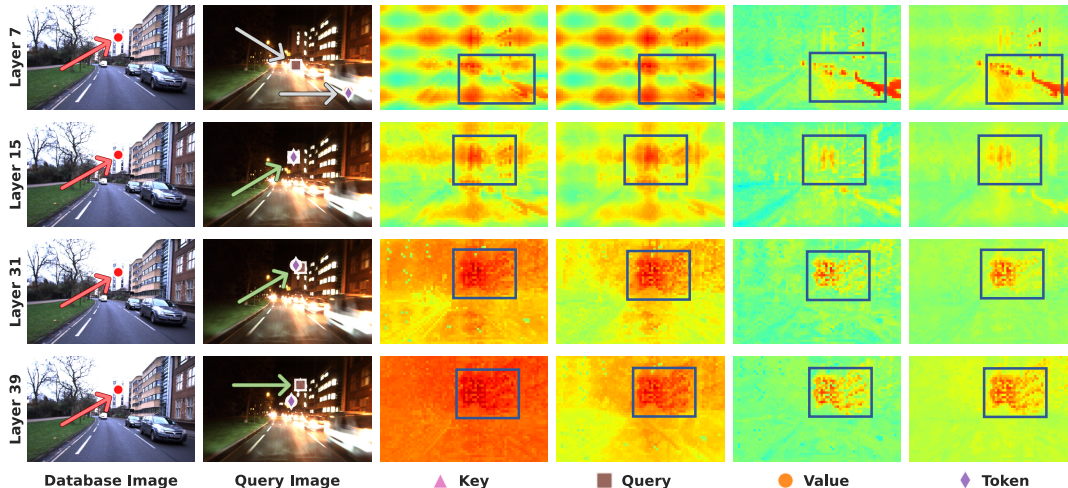


Figure 3.3: Qualitative ablation comparing the absolute-scale similarity maps of features from different DINOv2 ViT-G *layers* and *facets*. **Layer 31 value facet has the sharpest contrast** in the similarity map, which is **crucial for robustness against distractors within an image**.

3.7 Choosing Vocabulary Design

When performing any vocabulary-based aggregation like VLAD on the model, we are required to construct a vocabulary(cluster centers) that can capture the distinct semantic properties from the database image features. In our study, we consider different ways of forming the vocabulary, inspired from previous approaches as well as taking into account the separation of AnyLoc feature representations through PCA that gives rise to certain "domains". Traditionally, the database vocabularies are constructed by either using database images obtained from the robot trajectory, or from a pre-built map. Recent methods have also explored the construction of these vocabularies using multiple large-scale datasets. In AnyLoc, given the diversity of datasets we also consider building a vocabulary using all datasets (global vocabulary), using unstructured datasets(unstructured vocabulary) and structured datasets(structured vocabulary).

Furthermore, from the PCA visualizations of AnyLoc features, we observe the emergence of distinct domains in the latent space that share similar semantic features. These include Urban, Indoor, Aerial, SubT, Degraded, and Underwater domains.

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. Hence, we also construct domain-specific vocabularies based on these emergent domains.

3.8 Datasets

To demonstrate AnyLoc’s robust performance *anywhere*, *anyview* and *anytime*, we use a variety of datasets that capture unprecedented diversity in terms of environments, appearance viewpoint, temporal and long-term changes. Broadly, we classify our datasets into structured and unstructured datasets. Structured environments consist of organized areas featuring human-made structures, which are often encountered in autonomous driving and indoor robotics applications. In contrast, unstructured environments are obtained from robots deployed in-the-wild such as forests, subterranean, aerial and underwater environments.

3.8.1 Structured Environments

We evaluate our proposed approach on six benchmark indoor and outdoor datasets: Baidu Mall [78], Gardens Point [29, 79], 17 Places [69], Pittsburgh-30k [5], St Lucia [89], Oxford RobotCar [53]. These VPR datasets encompass a wide variety of challenging situations, including drastic viewpoint shifts, perceptual aliasing, and substantial visual appearance change, as follows:

Baidu Mall This visual localization dataset consists of images captured within a mall with varying camera poses. The dataset provides groundtruth location and 3D pose of an image, making it suited for both 6-Degrees of Freedom (DoF) Localization and VPR testing. We use the entire dataset consisting of 2292 query images & 689 reference images for evaluation. This mall dataset presents interesting and challenging properties, including perceptually aliased structures, distractors for VPR (such as people), and semantically rich information, such as billboards and signs.

3. *AnyLoc*

Gardens Point This dataset contains two traverses through the Gardens Point campus of Queensland University of Technology (QUT) captured at different times of the day, i.e., day and night. Both the database and query traverses contain 200 images, respectively. The drastic lighting changes and transitions from indoor to outdoor scenarios make it a difficult VPR dataset.






17 Places This indoor dataset consists of traverse collected within buildings at York University (Canada) and Coast Capri Hotel (British Columbia). The reference and query traverses consist of 406 images. The high clutter, change in lighting conditions, and semantically rich information make this dataset interesting.

Pittsburgh-30k This benchmark VPR dataset consists of images collected at various locations and poses throughout downtown Pittsburgh. We use the test split consisting of 10,000 database images and 6816 query images. This dataset is challenging due to the presence of drastic viewpoint shifts, a large variety of geometric structures such as buildings, and distractors such as cars and pedestrians.

St Lucia This dataset consists of daytime traverses collected using a stereo camera pair on a car, where the traverses span a total distance of 9.5 km. The reference traverse consists of 1549 images, while the query traverse consists of 1464 images. A large number of loop closure events, reverse traverses, shadows, and vegetation make this dataset challenging.

Oxford RobotCar This dataset consists of Oxford City traverses, which showcase shifts in seasonal cycles and daylight. We use a subsampled version of the Overcast Summer and Autumn Night traverses, similar to HEAPUtil [38]. The original traverses are subsampled with an approximate spacing of 5 meters to obtain a total of 213 frames in the summer traverse and 251 frames in the autumn night traverse with a total distance spanning 1.5 Km. This dataset presents a challenging shift in visual appearance caused by the time of day and seasonal shifts.

Table 3.1: Unstructured Environments used in Evaluation

Dataset	N_{Db}	N_Q	Traj. Span	Loc. Radius	Type
Hawkins [99]	65	101	282 m	8 m	
Laurel Caverns [99]	141	112	102 m	8 m	
Nardo-Air	102	71	700 m / 1 km^2	60 m	
VP-Air [72]	12.7k	2.7k	100 km	3 frames	
Mid-Atlantic Ridge [11]	65	101	18 m	0.3 m	

3.8.2 Unstructured Environments

Hawkins This dataset is an indoor mapping of an abandoned multi-floor hospital in Pittsburgh, where it is particularly challenging due to long corridors with visually-degraded features [99]. In particular, we use a long corridor spanning 282 m with a localization radius of 8 m, where the database and query images are collected from 2 opposing viewpoints (forward & backward direction). The database and query set contain 65 and 101 images, respectively.

Laurel Caverns This subterranean dataset consists of images collected using a handheld payload [99]. The low illumination scenarios and lack of rich visual features make this dataset particularly challenging. The opposing viewpoint of the database and query images adds additional complexity to the strong distribution shift. We use a 102 m trajectory with a localization radius of 8 m, where the database and query sets contain 141 and 112 images, respectively.

Nardo-Air This is a GNSS-denied localization dataset collected using a 100° FoV downward-facing camera on board a hexacopter flying at 10 m/s and an altitude of 50 m across a grass-strip runway named Nardo. The reference database comprises 102 images obtained from a Google Maps TIF satellite image, while the query set contains 71 drone-collected imagery. The perceptual aliasing at the end of the runway and non-typical vegetative features combined with a long time shift make this dataset challenging. The -R variant of this dataset indicates rotation where the drone imagery is rotated to match the satellite image orientation. We use a 700 m trajectory

Table 3.2: State-of-the-art Baselines used for Comparison

Method	Backbone	Training Dataset	Supervision
NetVLAD [5, 8]	ResNet-18	Pitts-30k	VPR - Contrastive
CosPlace [7]	ResNet-101	SF-XL	VPR - Classification
MixVPR [2]	ResNet-50	GSV-Cities	VPR - Contrastive
1-4 CLIP [34, 67]	ViT-bigG-14	Laion 2B	Image-Caption Pairs
DINO [16]	ViT-S8	ImageNet	Self-Supervised
DINOv2 [62]	ViT-G14	LVD-142M	Self-Supervised

spanning across a square kilometer area, where the localization radius is 60 m.

VP-Air This aerial VPR dataset consists of 2,706 database-query image pairs and 10,000 distractors collected at 300 m altitude with a downward-facing camera on an aircraft [72]. The dataset spans over 100 km, encompassing various challenging landscapes such as urban regions, farmlands, and forests. We use a localization radius of 3 frames.

Mid-Atlantic Ridge We construct this dataset using the overlapping sequences of an underwater visual localization dataset [11]. It presents OOD challenges including seabed objects, low illumination, and appearance shifts over a long time period (2015 vs. 2020). The dataset contains 65 database images and 101 query images, where the trajectory spans 18 m and the localization radius is 0.3 m.

3.9 Baselines and Evaluation Metrics

As an evaluation metric, we adopt Recall@ K for our quantitative analysis. Recall@ K is a widely adopted metric in VPR [95]. For a pre-defined localization radius, Recall@ K is the ratio of correctly retrieved queries to the total number of queries, where the checked retrievals are within the top K predictions. Furthermore, all experiments are seeded to 42 and done on the same platform (NVIDIA RTX 3090) to ensure consistency and reproducibility of experiments.

We benchmark AnyLoc against SOTA VPR algorithms, namely, Cos-Place, NetVLAD and MixVPR since each of these baselines capture the broad spectrum of training strategies, feature aggregation techniques as well as are representative of

Table 3.3: Performance comparison on Benchmark Structured Environments













Methods	 Baidu Mall		 Gardens Point		 17 Places		 Pitts-30k		 St Lucia		 Oxford		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
NetVLAD [5]	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 [7]	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 [2]	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
1-15 CLIP-CLS [67]	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 [16]	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 [62]	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
<i>AnyLoc-GeM-DINOv2</i>	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
<i>AnyLoc-VLAD-DINO</i>	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
<i>AnyLoc-VLAD-DINO-PCA</i>	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
<i>AnyLoc-VLAD-DINOv2</i>	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
<i>AnyLoc-VLAD-DINOv2-PCA</i>	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 3.4: Performance Comparison on Unstructured Environments

Methods	 Hawkins		 Laurel Caverns		 Nardo-Air		 Nardo-Air R		 VP-Air		 Mid-Atlantic Ridge		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
NetVLAD [5]	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 [7]	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 [2]	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
1-15 CLIP-CLS [67]	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 [16]	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 [62]	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
<i>AnyLoc-GeM-DINOv2</i>	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
<i>AnyLoc-VLAD-DINO</i>	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
<i>AnyLoc-VLAD-DINOv2</i>	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

the current large-scale training trend. Furthermore, we introduce 3 newer baselines, which use the CLS token from CLIP, DINO and DINO-v2 as the feature descriptor. A comprehensive summary of these baselines can be found in 3.2.

3.10 Experiments and Analysis

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.

3.10.1 State-of-the-art Comparison

Structured Environments

3.3 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 (49k to 512).

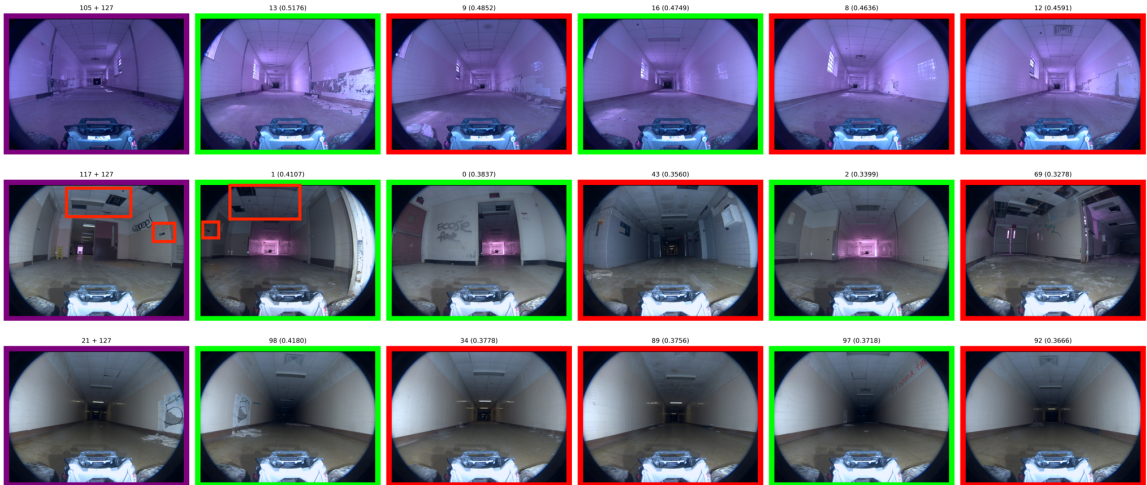


Figure 3.4: Hawkins Retrieval Visualizations

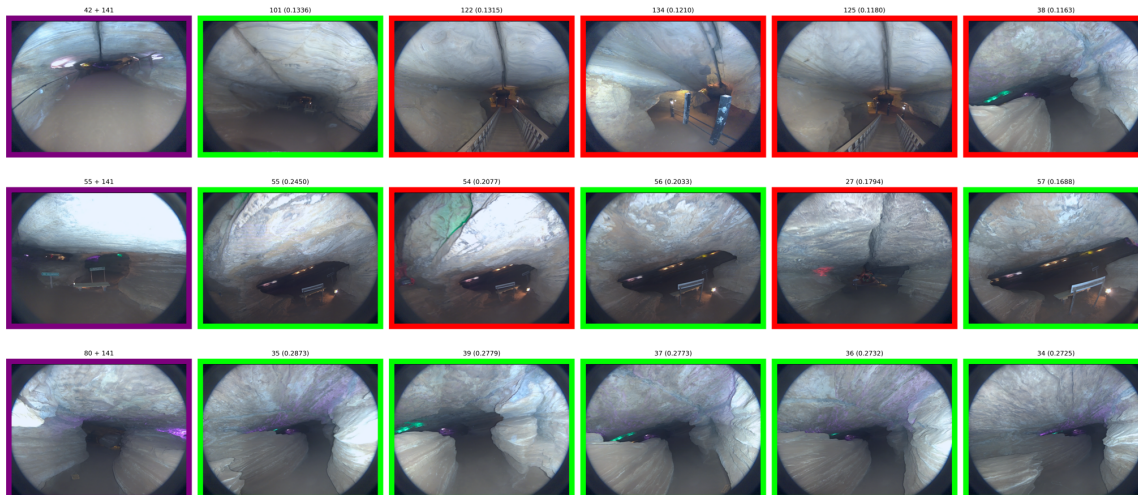


Figure 3.5: Laurel Caverns Retrieval Visualizations




Unstructured Environments

3.4 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). We visualize example retrievals in unstructured environments in 3.4 and 3.5.

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, 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

Table 3.5: 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

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%.

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.

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.

Table 3.6: 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

3.10.2 Vocabulary Analysis

Vocabulary Source

3.5 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 3.10.2, the subsets for different domains are obtained through a qualitative PCA visualization (see 3.1), 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.

Consistency

3.6 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

3. AnyLoc

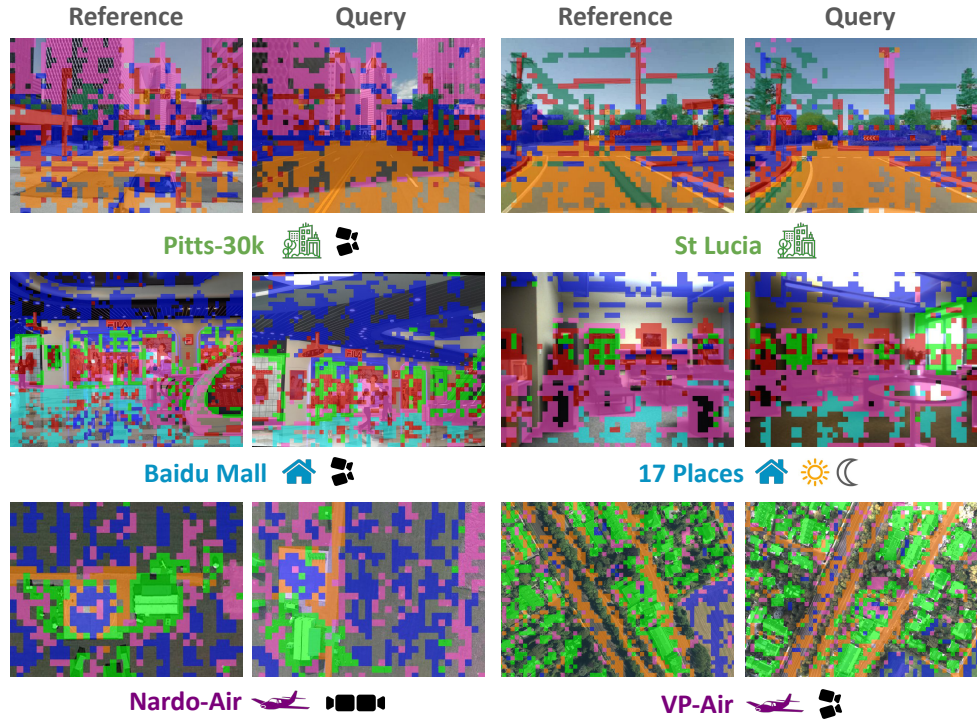


Figure 3.6: VLAD cluster assignment visualizations of the reference-query pairs highlight the **intra-domain consistency** of the domain-specific vocabulary. Similar colors across images of a specific domain indicate matched clusters.

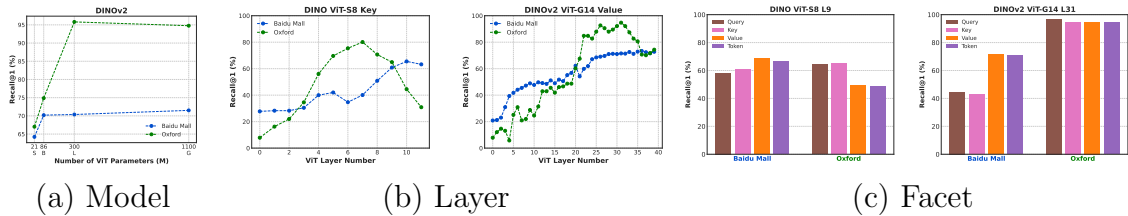


Figure 3.7: **Design Choices for *AnyLoc-VLAD***: (a) Performance scales with the model size but saturates at ViT-L. (b) Performance peaks at intermediate layers instead of the final layer for both DINO & DINOv2. (c) On average, **key & value** perform the best respectively for DINO & DINOv2.

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 3.6, 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 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 3.5.

3.10.3 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:

ViT Architecture

Fig. 3.7 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.

ViT Layers & Facets

Fig. 3.7 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, we choose 9 and 31 as our operating layers for DINO and DINOv2, respectively.






In Fig. 3.7 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

3. AnyLoc

Table 3.7: Analysis comparing the Recall@1 & Descriptor Dimensionality across varying aggregation methods

Aggregation Methods	DINO			DINOv2		
	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 3.8: 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 <i>AnyLoc-VLAD-DINO</i>	72.9	79.6	47.8	52.7	41.6
ViT-B <i>AnyLoc-VLAD-DINOv2</i>	77.0	82.6	53.6	60.2	35.6
ViT-G <i>AnyLoc-VLAD-DINOv2</i>	78.0	92.3	62.9	63.4	34.6

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.

Aggregation Methods

In 3.7, we compare the various unsupervised local feature aggregation techniques 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.

3.10.4 Self-supervised vs VPR-supervised ViT

3.8 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 [62]. In particular, we compare a ViT trained specifically for VPR (i.e., CosPlace [7]) 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 $4\times$ 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.

3.11 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.

3.12 Acknowledgement and Contribution Statement

AnyLoc[37] was published at IEEE RA-L 2023 and ICRA 2024 as a co-first authored publication with Nikhil Keetha and Avneesh Mishra. The contribution is given below:

3. *AnyLoc*

Nikhil Keetha conceived the idea and led the project. Responsible for initial code development, writing major sections of the paper, and producing figures, tables & videos.

Avneesh Mishra implemented vital components, including the foundation model feature extraction and modular scripts, to run experiments at a large scale. Responsible for running the ablation experiments, writing the first draft of the results section, and producing qualitative visualizations & the Hugging Face demo.

Jay Karhade scaled the evaluation to a diverse suite of unstructured environments, implemented the vocabulary ablations, and performed explorations into various foundation models, including SAM. Responsible for the website, retrieval visualizations, and diverse suite of interactive demos.

Krishna Murthy was actively involved in brainstorming and critical review throughout the project. Responsible for the exploration of self-supervised visual foundation models. Wrote & proofread sections of the paper.

Sebastian Scherer pushed us towards evaluating the practicality of current VPR systems in unstructured environments and developing a universal VPR system. Suggested a vital paper restructuring to ensure the critical message and insights are easily parsable. Sebastian provided compute resources for initial explorations and ablations.

Madhava Krishna was involved in initial brainstorming discussions and provided feedback throughout the development. Suggested revisions for sections of the paper. Madhav also provided most of the compute for the experiments conducted in this work.

Sourav Garg provided resourceful visual place recognition perspectives and critical thoughts in the brainstorming sessions, which led to clear insights into the applicability of foundation model features for VPR. Wrote and proofread sections of the paper.

The authors thank Ivan Cisneros & Yao He for collecting the Nardo-Air dataset. Lastly, the authors thank Deepak Pathak, Murtaza Dalal, Ananye Agarwal, Aditi Raghunathan, and Tuomas Sandholm for feedback on an initial version of the work. We also thank the members of CMU AirLab for their insightful discussions throughout this project.

Chapter 4

MultiLoc



Figure 4.1: We show that binding LIDAR and thermal modalities to features to vision foundation models is effective to achieve *zero-shot* cross-modal place recognition.

Continuing the paradigm of a universal place recognition method, we introduce Multi-Loc, a method that enables zero-shot cross-modal place recognition the performance of foundation models(DINO and DINO-v2) by binding modality-specific encoders through knowledge distillation.

4.1 Introduction

In AnyLoc, we observed that using visual features from foundation models (DINO and DINO-v2) provide SOTA performance over even supervised methods without receiving any VPR-specific training. While AnyLoc offers robust performance in

diverse domains, it is limited to the visual modality. Unfortunately, performance of visual modalities degrades under conditions such as darkness, illumination and smoke.

Other modalities, namely LIDAR and thermal modalities are robust to these changes. Currently, place-recognition for these modalities is limited due to lack of training data, and multi-modal/cross-modal place recognition techniques require paired triplet datasets. While LIDAR and thermal modalities often contain paired images, paired data among other non-visual modalities is scarce, limiting generalization for these modalities.

To this end, we propose a new paradigm to enable uni-modal and cross-modal place recognition by distilling foundation model features. Similar to Image-Bind, we bind encoders of different modalities to DINO and DINO-v2 features. Our approach, named Multi-Loc enables a simple and scalable approach towards general place recognition that transfers across modalities. We represent these modalities such that they can utilize vision transformers, followed by alignment of the modality-specific representations with image representations generated from DINO. Since all modalities are aligned to a common image-representation that produces robust semantic features, we observe robust uni-modal and cross-modal performance even across modalities that do not have paired data. Our contributions can be summarized as follows:

- We align LIDAR-image data and thermal-image data and show that these distilled feature representations are strong uni-modal and cross-modal place descriptors, without requiring place-recognition specific training.
- Because these modalities are distilled with a common representation, we show that our method enables zero-shot LIDAR-thermal place recognition, without requiring any paired thermal-LIDAR data.
- We observe that these fine-tuned models are scalable and generalize to unseen and unstructured environments, enabling *any-place*, *any-view*, *any-time* and *any-sensor* place recognition.

4.2 Related Work

A number of recent works have made advances towards place recognition in different modalities. We broadly group and discuss them in 3 subsections as below, particularly

for thermal and LIDAR modalities.

4.2.1 Non-visual Place Recognition

A number of methods have been proposed for LIDAR place recognition ranging from classical to learning based pipelines. [39] introduced a simple yet highly effective scan-based LIDAR place recognition algorithm, which was subsequently extended in [22, 40, 84]. [83] extended the hugely popular NetVLAD paradigm to point-clouds through the clever use of the point-net architecture. [86] proposes an elegant ring-based descriptor for LIDAR place recognition. More recently, learning based methods have been proposed that train descriptors based on point-cloud overlap [19, 50, 51, 52]. Other methods such as [96] propose a localization pipeline based on instance segmentation and learning. While a number of thermal-slam and odometry methods [59, 87, 97] have been developed, fewer works target thermal place recognition, [54] demonstrated thermal sensors are effective for day-night changes compared to visual sensors. More recently [46] proposed a GAN-based approach for thermal place recognition using image translation techniques. Despite impressive performance, these methods are often limited by the scale of LIDAR and thermal data, and cannot leverage priors from visual encoders that produce strong image descriptors.

4.2.2 Multi-Modal Place Recognition

Multi-modal and cross-modal place recognition between images and LIDAR has been widely studied. [17, 58] propose a fine localization of image queries in a pre-built LIDAR map. Following a retrieval paradigm, [94, 100, 101] propose cross-modal place recognition techniques between pinhole/360 images to LIDAR scans from LIDAR maps. To fully utilize the advantage of multi-modal inputs, [42, 43] propose fusion-strategies between point-clouds and images. [54] also proposes a multi-modal approach to use both visual and thermal images for place recognition. More recently [91, 93] propose visual-thermal geo-localization pipelines by generating fake thermal images and domain adaptation techniques to boost visual-thermal satellite localization.

4.2.3 Multi-Modal Foundation Models

Building on the success of vision foundation models, multi-modal foundation models have shown impressive performance by binding other modalities through knowledge distillation with image encoders. [33, 48, 55, 65, 71] show that effective LIDAR pre-training can be achieved by distilling vision foundation models into LIDAR backbones and demonstrate improved performance on LIDAR segmentation and classification tasks. More recently, [28] first showed that zero-shot retrieval across over 6 modalities can be achieved by aligning modality-specific representations to visual representations. We follow the footsteps of Image-Bind’s paradigm for zero-shot alignment of thermal-LIDAR place recognition.

4.3 Method

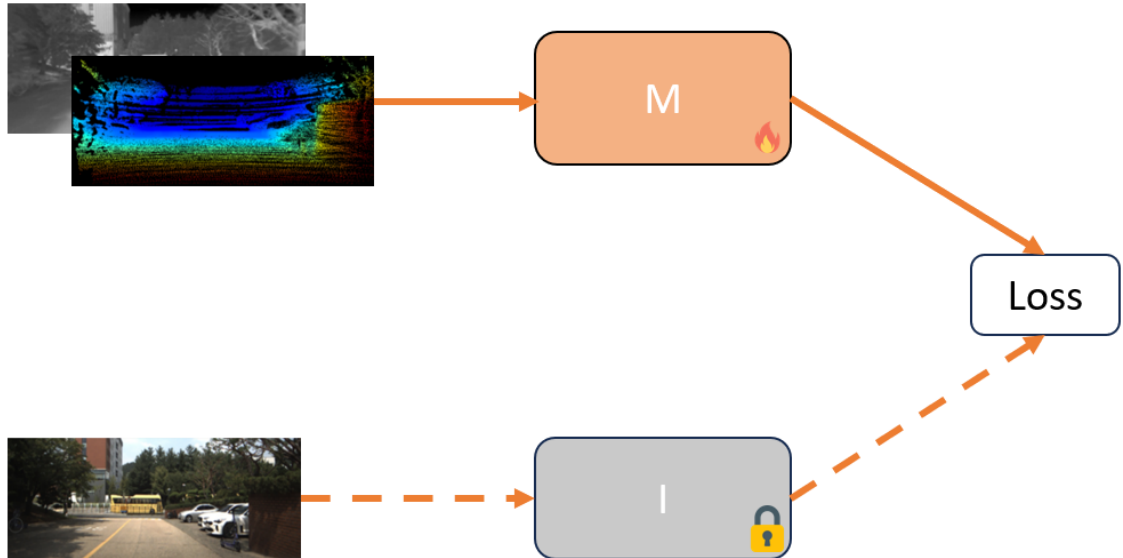


Figure 4.2: We distill image features into other modalities through student-teacher training where the teacher(vision foundation model) is frozen, and the student(modality-specific model) is updated through an InfoNCE[61] loss.

We now describe our proposed method to enable zero-shot cross-modal place recognition. While we discuss the case for visual-thermal-LIDAR place recognition, this can be trivially extended to additional modalities as well. Specifically, given

paired image-LIDAR and paired image-thermal data, we use knowledge distillation techniques to align LIDAR encoders f_L and thermal encoders f_T with pre-trained image encoders f_I .

We choose to represent LIDAR scans as projected range images in the camera frame, enabling scalability using simple Vision-Transformer(ViT) encoders without bells and whistles. Furthermore, we empirically find that densification of LIDAR projections through morphological operations scans better results.

We keep the image ViT encoder frozen and update the LIDAR/thermal encoder modalities through knowledge distillation. To perform knowledge distillation, we experiment with two loss functions - batched MSE and batched contrastive Info-NCE loss.

$$L_{(I,M)} = \sum_{i=1,\dots,N} (q_i - k_i)^2 \quad (4.1)$$

where $q_i = f(I_i)$ and $k_i = f_M(M_i)$

$$L_{(I,M)} = \sum_{i=1,\dots,N} -\log \frac{\exp(q_i^T k_i / \tau)}{\exp(q_i^T k_i / \tau) + \sum_{i \neq j} \exp(q_i^T k_j / \tau)} \quad (4.2)$$

where M represents a specific modality - LIDAR or thermal. In our experiments, we observe that the batched contrastive the InfoNCE loss provides better performance and is subsequently used for all training.

4.4 Datasets and Metrics

To demonstrate MultiLoc’s performance, we require datasets that are representative of diverse environments, and contain synchronized LIDAR-Visual, LIDAR-thermal and LIDAR-Visual-Thermal data for ease of comparison. Unfortunately, no such place-recognition datasets are publicly available. Hence, we re-purpose 2 publicly datasets for place recognition - the MS2 dataset[76] and CART dataset[45].

MS2 Dataset The Multi-Spectral Stereo (MS2) dataset[76] at locations near KAIST and captures over 195k synchronized visual-LIDAR-thermal data pairs across residential, university, and other urban and suburban environments. Furthermore,

this data has been collected across different weather conditions including clear, cloudy and rainy weather conditions.

CART Dataset The Caltech Aerial RGB-Thermal (CART) dataset[45] contains paired RGB-Thermal data captured in-the-wild across California. These locations are spread across diverse terrain including lakes, mountains, forests, deserts and streams and the coast-line, and the trajectories are obtained from aerial-mounted, hand-held and still payloads. In particular, we subsample and use the image-thermal data from "Idyll-Wild", "Big-bear" and "Duck Ocean" as test sequences.

Similar to AnyLoc and other place recognition papers, we adopt Recall@K for our quantitative analysis. For a pre-defined localization radius, Recall@K is the ratio of correctly retrieved queries to the total number of queries, where the checked retrievals are within the top K predictions. To ensure consistency and reproducibility, all experiments are carried out on the same platform (NVIDIA A-100 GPU) and same random seed.

4.5 Experiments and Analysis

For our pre-trained encoder, we distill different ViT versions of DINO, DINO-v2, and CLIP models. Due to changes in thermal and LIDAR sensor properties between datasets, we distill dataset specific-models. Particularly, we use a pre-trained ViTS-16 for CLIP, ViT-B16 for DINO and ViTS-14 for DINOv2 as our encoders, and keep the same architecture across modality-specific encoders.

We first evaluate MultiLoc’s performance on the MS-2 dataset. We train our thermal and LIDAR encoders only on day-time sequences. For evaluation, we use previously unseen sequences with rainy conditions and night-time sequences.

A breakdown of the data sequences can be found in [table 4.1](#).

From [4.1](#) and [4.2](#) We observe that DINOv2 features are much more robust compared to CLIP feature descriptors. Across all sequences, we observe that distilled DINOv2 features consistently outperform distilled CLIP features. Across both rainy sequences, we observe that DINOv2 descriptors improve average Recall1 by approximately 30% compared to CLIP descriptors. In Night-time sequences, we observe over 35% improvements for Visual-Thermal place recognition, indicating that

Table 4.1: Cross-Modal Place Recognition MS2 dataset - Rainy Sequences

Methods	Road-3(Clear-Sky)		Residential(Clear-Sky)		Road-2(Rainy)		Residential(Rainy)	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Visual-Thermal								
CLIP	0.7451	0.93333	0.68277	0.89496	0.43019	0.58639	0.43862	0.69162
DINOv2	0.93333	1	0.87685	0.98522	0.69895	0.86824	0.83533	0.95509
Visual-LIDAR								
CLIP	0.47059	0.76863	0.25945	0.39916	0.18237	0.28883	0.14371	0.41018
DINOv2	0.81961	0.98824	0.5021	0.74475	0.27923	0.45637	0.51198	0.74701
Thermal-LIDAR								
CLIP	0.55294	0.85098	0.22269	0.45693	0.13613	0.28185	0.27695	0.54341
DINOv2	0.9098	0.98824	0.66912	0.85819	0.28185	0.49127	0.67665	0.88174

Table 4.2: Cross-Modal Place Recognition on MS2 dataset - Night-time Sequences

Methods	Road-3(Night)		Residential(Night)	
	R@1	R@5	R@1	R@5
Visual-Thermal				
CLIP	0.15608	0.36376	0.25591	0.53937
DINOv2	0.55159	0.81349	0.53937	0.77165
Visual-LIDAR				
CLIP	0.09524	0.25529	0.2126	0.50787
DINOv2	0.37831	0.6455	0.42126	0.71654
Thermal-LIDAR				
CLIP	0.2619	0.50265	0.23228	0.6063
DINOv2	0.59524	0.83862	0.52756	0.82677

DINOv2 features transfer well to thermal encoders, and do not merely overfit to the trajectory sequences.

A similar trend is seen in Visual-LIDAR place recognition on these sequences, showing that DINOv2 descriptors transfer to both modalities. However, we observe significantly lower Visual-LIDAR recalls as compared to visual-thermal place recognition. We believe this is because of the sparse nature of the LIDAR scans, compared to dense thermal images that have been taken from a VLP-16 LIDAR scan, and this is further evident from the LIDAR densification ablation study.

Finally, for the first time we show that 2 modalities(LIDAR and thermal) can be aligned at test-time through a bridge-modality(RGB). It is interesting to note that this zero-shot place recognition performance is in-fact, higher than RGB-LIDAR place recognition. We believe this is because similar features get highlighted in the

4. MultiLoc

input LIDAR and thermal thermal modalities, compared to RGB features which can suffer from visual distractors and other sensor aliasing that LIDAR and thermal sensors are robust to.



Figure 4.3: Qualitative Retrievals for Visual-Thermal Place Recognition on a night-time MS2 sequence

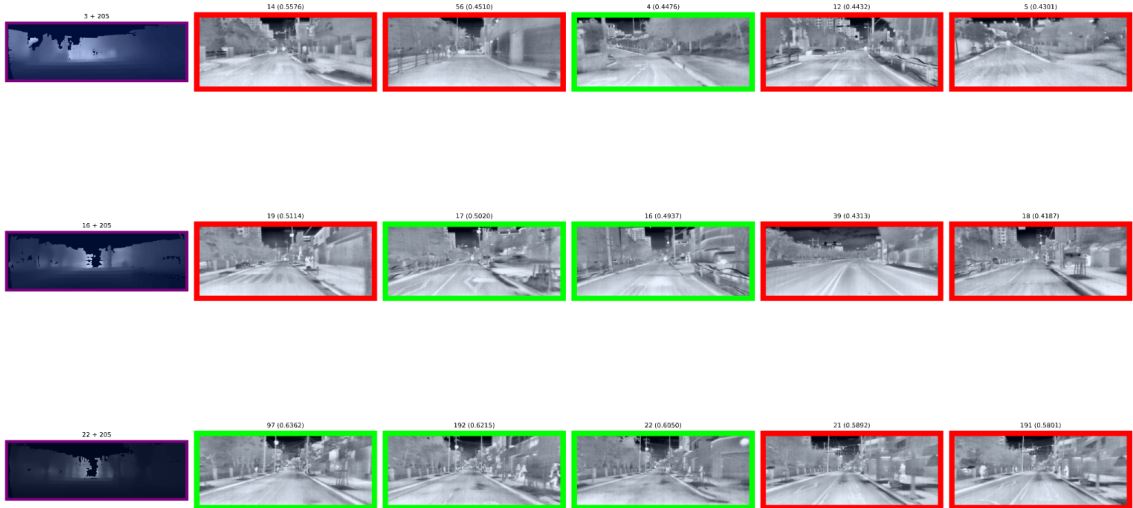


Figure 4.4: Qualitative Retrievals for Thermal-LIDAR Place Recognition on a night-time MS2 sequence

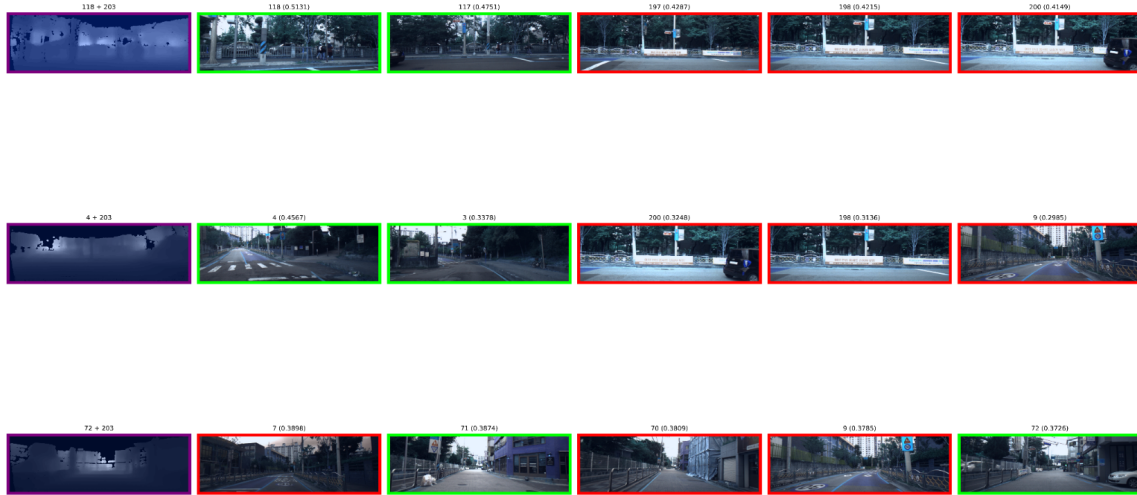




Figure 4.5: Qualitative Retrievals for Visual-LIDAR Place Recognition on a night-time MS2 sequence

This is further evidenced from the night-time sequence recalls, where we see that LIDAR-thermal recalls have slightly-higher/similar recalls compared to visual-thermal recalls and significantly higher recalls than visual-LIDAR recalls. Since LIDAR and thermal images do not suffer from illumination changes, they achieve higher recalls compared to cross-modal place recognition involving the visual modality.

We further test the effective-ness of MultiLoc on challenging unstructured in the wild environments by re purposing UAV and hand-held sequences of the CART dataset. We train the thermal-LIDAR backbone by using sub sampled synchronized pairs from trajectories in environments, namely Joshua Tree (desert), Duck Ocean, Kentucky River (river-side environment), Colorado-River sequences. For evaluation, we use Idyll-Wild and Big-Bear Lake as our test sequence. We observe that DINOv2 features are much more robust compared to CLIP feature descriptors. On the Idyll-wild sequence, DINOv2 features improve Recall1 and Recall5 performance by over 16% and 23% for visual-thermal place recognition. This is similar on the Big-bear Lake where we observe over 25% and close to 30% improvement on Recall1 and Recall5. These retrievals are also visualized in 4.6 and 4.7.

4. MultiLoc

Table 4.3: Visual-Thermal Place Recognition CART dataset

Methods	 Idyll Wild		 Big-Bear Lake	
	R@1	R@5	R@1	R@5
Visual-Thermal				
CLIP	0.32093	0.57674	0.25854	0.49756
DINOv2	0.51628	0.84186	0.5122	0.78537
Thermal-Visual				
CLIP	0.30233	0.54884	0.2878	0.57561
DINOv2	0.46977	0.77209	0.56098	0.90732

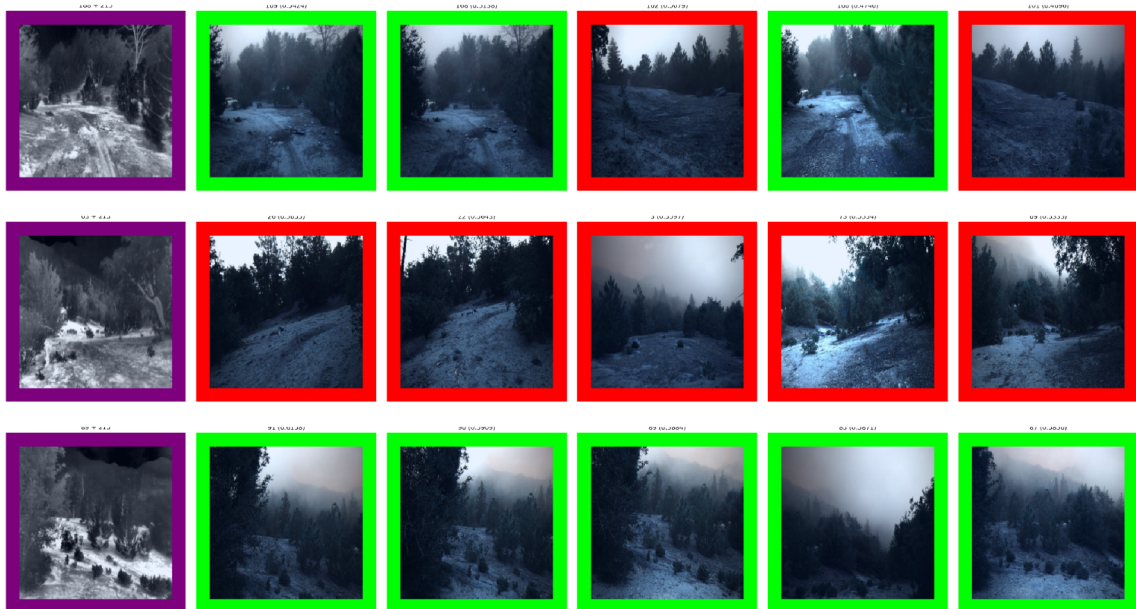


Figure 4.6: Qualitative Retrievals for Visual-Thermal Place Recognition on the Idyll-Wild Sequence



Figure 4.7: Qualitative Retrievals for Visual-Thermal Place Recognition on the Big-Bear Sequence

4.6 Limitations and Conclusions

In this chapter, we presented MultiLoc, a simple way of extending image foundation model features to other modalities and showing that zero-shot cross modal performance can be achieved through such alignment. However, MultiLoc has several limitations - since it is distilled with DINOv2-CLS features, it is sub-optimal compared to the extremely robust performance like AnyLoc. Furthermore, there is a noticeable drop in LIDAR-based place recognition. Furthermore, this approach still requires training for "similar" environments and similar sensor configurations. We hope that future work can address these limitations.

4. MultiLoc

Chapter 5

Conclusion and Future Work

In this thesis, we take a step towards place recognition pipelines that are generalized across sensors and environmental conditions. We identify exciting future directions to enable robust real-time in-the-wild place recognition for robots:

Uncertainty for Place Recognition While our approach demonstrates SOTA Place-Recognition, it fails to capture uncertainties in the retrieval process. Current mechanisms often employ heuristics to reject uncertain retrievals or typically rely on robust back-end solvers [56, 57]. Developing uncertainty quantification mechanisms that are both agnostic and tightly-coupled with place recognition remain a valuable future direction.

Multi-Modal Place Recognition While our approach shows applicability in cross-modal place recognition, the knowledge distillation inherently prevents the advantages of different sensory modalities, and it would be very interesting to develop ways that improve robustness through fusion of multiple modalities such that one modality does not dominate the other modality.

Real-time performance Our proposed techniques require significant computational resources. While we demonstrate that the AnyLoc descriptor sizes can be reduced, saving memory, replacing large ViT backbones with more efficient backbones such as [14] can enable deployment on compute-constrained devices.

5. Conclusion and Future Work

Bibliography

- [1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguere. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 2022. [3.2.2](#)
- [2] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Mixvpr: Feature mixing for visual place recognition. In *WACV*, 2023. [3.2.2](#), [3.2](#), [3.3](#), [3.4](#)
- [3] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv:2112.05814*, 2021. [3.2.1](#)
- [4] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *CVPR*, 2013. [3.6](#)
- [5] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, 2016. [3.1](#), [3.2.2](#), [3.8.1](#), [3.2](#), [3.3](#), [3.4](#)
- [6] Artem Babenko and Victor Lempitsky. Aggregating deep convolutional features for image retrieval. *arXiv:1510.07493*, 2015. [3.6](#)
- [7] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geolocalization for large-scale applications. In *CVPR*, 2022. [3.2.2](#), [3.2](#), [3.3](#), [3.4](#), [3.10.4](#)
- [8] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *CVPR*, 2022. [3.1](#), [3.2.2](#), [3.2](#)
- [9] Gabriele Moreno Berton, Valerio Paolicelli, Carlo Masone, and Barbara Caputo. Adaptive-attentive geolocalization from few queries: A hybrid approach. In *WACV*, pages 2918–2927, 2021. [3.2.2](#)
- [10] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. *EMNLP*, 2020. [3.2.1](#)
- [11] Clémentin Boittiaux, Claire Dune, et al. Eiffel tower: A deep-sea underwater dataset for long-term visual localization. *IJRR*, 2022. [3.1](#), [3.8.2](#)
- [12] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut,

- Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021. [3.1](#), [3.2.1](#)
- [13] Anthony Brohan et al. Rt-1: Robotics transformer for real-world control at scale. In *arXiv:2212.06817*, 2022. [3.2.1](#)
- [14] Han Cai, Chuang Gan, and Song Han. Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. *arXiv preprint arXiv:2205.14756*, 2022. [5](#)
- [15] Bingyi Cao, André Araujo, and Jack Sim. Unifying deep local and global features for image search. In *ECCV*, 2020. [3.2.2](#)
- [16] Mathilde Caron, Hugo Touvron, Ishan Misra, et al. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. [3.2.1](#), [3.4.1](#), [3.2](#), [3.3](#), [3.4](#)
- [17] Daniele Cattaneo, Matteo Vaghi, Augusto Luis Ballardini, Simone Fontana, Domenico G Sorrenti, and Wolfram Burgard. Cmrnet: Camera to lidar-map registration. In *2019 IEEE intelligent transportation systems conference (ITSC)*, pages 1283–1289. IEEE, 2019. [4.2.2](#)
- [18] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15789–15798, 2021. [3.2.2](#)
- [19] X. Chen, T. Labe, A. Milioto, T. Rohling, O. Vysotska, A. Haag, J. Behley, and C. Stachniss. OverlapNet: Loop Closing for LiDAR-based SLAM. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020. [2.2](#), [4.2.1](#)
- [20] Timothee Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. [3.5](#)
- [21] Kamak Ebadi, Lukas Bernreiter, Harel Biggie, Gavin Catt, Yun Chang, Arghya Chatterjee, Christopher E Denniston, Simon-Pierre Deschenes, Kyle Harlow, Shehryar Khattak, et al. Present and future of slam in extreme underground environments. *arXiv preprint arXiv:2208.01787*, 2022. [1](#)
- [22] Yongzhi Fan, Xin Du, Lun Luo, and Jizhong Shen. Fresco: Frequency-domain scan context for lidar-based place recognition with translation and rotation invariance. In *2022 17th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 576–583. IEEE, 2022. [4.2.1](#)
- [23] Sourav Garg, Niko Suenderhauf, and Michael Milford. Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics. *RSS*, 2018. [3.2.2](#)
- [24] Sourav Garg, Tobias Fischer, and Michael Milford. Where is your place, visual place recognition? *IJCAI*, 2021. [3.2.2](#)

- [25] Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *ICLR*, 2023. [3.1](#)
- [26] Abel Gawel, Carlo Del Don, Roland Siegwart, Juan Nieto, and Cesar Cadena. X-view: Graph-based semantic multi-view localization. *IEEE RAL*, 2018. [3.2.2](#)
- [27] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. 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*, pages 369–386. Springer, 2020. [3.2.2](#)
- [28] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. [4.2.3](#)
- [29] A Glover. Gardens point day and night, left and right. *Zenodo DOI*, 10, 2014. [3.8.1](#)
- [30] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021. [3.2.1](#)
- [31] Lukas Haas, Silas Alberti, and Michal Skreta. Learning generalized zero-shot learners for open-domain image geolocation. *arXiv:2302.00275*, 2023. [3.2.2](#)
- [32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. [3.2.1](#), [3.4.3](#)
- [33] Georg Hess, Adam Tonderski, Christoffer Petersson, Kalle Åström, and Lennart Svensson. Lidarclip or: How i learned to talk to point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7438–7447, 2024. [4.2.3](#)
- [34] Gabriel Ilharco, Mitchell Wortsman, et al. Openclip. In *Zenodo*, 2021. doi: 10.5281/zenodo.5143773. [3.2](#)
- [35] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, et al. Conceptfusion: Open-set multimodal 3d mapping. *RSS*, 2023. [3.2.1](#)
- [36] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*. IEEE, 2010. [3.1](#), [3.2.2](#), [3.6](#)
- [37] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 2023.

3.12

- [38] Nikhil Varma Keetha, Michael Milford, and Sourav Garg. A hierarchical dual model of environment-and place-specific utility for visual place recognition. *IEEE RAL*, 6(4):6969–6976, 2021. [3.8.1](#)
- [39] Giseop Kim and Ayoung Kim. Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4802–4809. IEEE, 2018. [2.1](#), [2.2](#), [4.2.1](#)
- [40] Giseop Kim, Sunwook Choi, and Ayoung Kim. Scan context++: Structural place recognition robust to rotation and lateral variations in urban environments. *IEEE Transactions on Robotics*, 38(3):1856–1874, 2021. [2.2](#), [4.2.1](#)
- [41] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv:2304.02643*, 2023. [3.2.1](#), [3.4.4](#)
- [42] Jacek Komorowski, Monika Wysoczańska, and Tomasz Trzcinski. Minkloc++: lidar and monocular image fusion for place recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. [4.2.2](#)
- [43] Haowen Lai, Peng Yin, and Sebastian Scherer. Adafusion: Visual-lidar fusion with adaptive weights for place recognition. *IEEE Robotics and Automation Letters*, 7(4):12038–12045, 2022. [4.2.2](#)
- [44] Yasir Latif, Ravi Garg, Michael Milford, and Ian Reid. Addressing challenging place recognition tasks using generative adversarial networks. In *ICRA*, 2018. [3.2.2](#)
- [45] Connor Lee, Matthew Anderson, Nikhil Raganathan, Xingxing Zuo, Kevin Do, Georgia Gkioxari, and Soon-Jo Chung. Cart: Caltech aerial rgb-thermal dataset in the wild. *arXiv preprint arXiv:2403.08997*, 2024. [4.4](#), [4.4](#)
- [46] Dong-Guw Lee, Hyeonjae Gil, Seungsang Yun, Jeongyun Kim, and Ayoung Kim. Night-to-day thermal image translation for deep thermal place recognition. *Intelligent Service Robotics*, 16(4):403–413, 2023. [4.2.1](#)
- [47] María Leyva-Vallina, Nicola Strisciuglio, and Nicolai Petkov. Data-efficient large scale place recognition with graded similarity supervision. In *CVPR*, 2023. [3.2.2](#)
- [48] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. *Advances in Neural Information Processing Systems*, 36, 2024. [4.2.3](#)
- [49] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox,

- Peter Corke, and Michael J Milford. Visual place recognition: A survey. *T-RO*, 2015. [3.2.2](#)
- [50] Junyi Ma, Xieyuanli Chen, Jingyi Xu, and Guangming Xiong. Seqot: A spatial-temporal transformer network for place recognition using sequential lidar data. *IEEE Transactions on Industrial Electronics*, 2022. doi: 10.1109/TIE.2022.3229385. [2.2](#), [4.2.1](#)
- [51] Junyi Ma, Jun Zhang, Jintao Xu, Rui Ai, Weihao Gu, and Xieyuanli Chen. Overlaptransformer: An efficient and yaw-angle-invariant transformer network for lidar-based place recognition. *IEEE Robotics and Automation Letters*, 7(3): 6958–6965, 2022. [4.2.1](#)
- [52] Junyi Ma, Guangming Xiong, Jingyi Xu, and Xieyuanli Chen. Cvtnet: A cross-view transformer network for lidar-based place recognition in autonomous driving environments. *IEEE Transactions on Industrial Informatics*, 2023. doi: 10.1109/TII.2023.3313635. [4.2.1](#)
- [53] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *IJRR*, 36(1):3–15, 2017. [3.8.1](#)
- [54] William Maddern and Stephen Vidas. Towards robust night and day place recognition using visible and thermal imaging. In *Proceedings of the RSS 2012 Workshop: Beyond laser and vision: Alternative sensing techniques for robotic perception*, pages 1–6. University of Sydney, 2012. [4.2.1](#), [4.2.2](#)
- [55] Anas Mahmoud, Jordan SK Hu, Tianshu Kuai, Ali Harakeh, Liam Paull, and Steven L Waslander. Self-supervised image-to-point distillation via semantically tolerant contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7102–7110, 2023. [4.2.3](#)
- [56] Joshua G Mangelson, Derrick Dominic, Ryan M Eustice, and Ram Vasudevan. Pairwise consistent measurement set maximization for robust multi-robot map merging. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 2916–2923. IEEE, 2018. [2.1](#), [5](#)
- [57] Daniel McGann, John G Rogers, and Michael Kaess. Robust incremental smoothing and mapping (risam). In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4157–4163. IEEE, 2023. [5](#)
- [58] Jinyu Miao, Kun Jiang, Yunlong Wang, Tuopu Wen, Zhongyang Xiao, Zheng Fu, Mengmeng Yang, Maolin Liu, Jin Huang, Zhihua Zhong, et al. Poses as queries: End-to-end image-to-lidar map localization with transformers. *IEEE Robotics and Automation Letters*, 9(1):803–810, 2023. [4.2.2](#)
- [59] Tarek Mouats, Nabil Aouf, Lounis Chermak, and Mark A Richardson. Thermal stereo odometry for uavs. *IEEE Sensors Journal*, 15(11):6335–6347, 2015. [4.2.1](#)

- [60] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, 2017. [3.2.2](#)
- [61] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. [\(document\)](#), [4.2](#)
- [62] Maxime Oquab, Timothée Darcet, et al. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023. [3.1](#), [3.2.1](#), [3.4.1](#), [3.4.4](#), [3.2](#), [3.3](#), [3.4](#), [3.10.4](#)
- [63] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoon Yun. What do self-supervised vision transformers learn? In *ICLR*, 2023. [3.2.1](#), [3.4.4](#)
- [64] Noé Pion, Martin Humenberger, Gabriela Csurka, Yohann Cabon, and Torsten Sattler. Benchmarking image retrieval for visual localization. In *3DV*. IEEE, 2020. [3.2.2](#)
- [65] Gilles Puy, Spyros Gidaris, Alexandre Boulch, Oriane Siméoni, Corentin Sautier, Patrick Pérez, Andrei Bursuc, and Renaud Marlet. Three pillars improving vision foundation model distillation for lidar. In *CVPR*, 2024. [4.2.3](#)
- [66] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE T-PAMI*, 41(7):1655–1668, 2018. [3.1](#), [3.2.2](#), [3.6](#)
- [67] Alec Radford, Jong Wook Kim, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. [3.2.1](#), [3.4.2](#), [3.2](#), [3.3](#), [3.4](#)
- [68] Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual instance retrieval with deep convolutional networks. *ITE TMTA*, 2016. [3.6](#)
- [69] Raghavender Sahdev and John K Tsotsos. Indoor place recognition system for localization of mobile robots. In *2016 13th CRV*, pages 53–60. IEEE, 2016. [3.8.1](#)
- [70] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *CVPR*, 2018. [3.2.2](#)
- [71] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9891–9901, June 2022. [4.2.3](#)
- [72] Michael Schleiss, Fahmi Rouatbi, and Daniel Cremers. Vpair–aerial visual place recognition and localization in large-scale outdoor environments. *ICRA 2022*

- Aerial Robotics Workshop arXiv:2205.11567*, 2022. [3.1](#), [3.8.2](#)
- [73] Stefan Schubert, Peer Neubert, Sourav Garg, Michael Milford, and Tobias Fischer. Visual place recognition: A tutorial. *RAM*, 2023. [3.2.2](#)
- [74] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations, 2018. [2.2.2](#)
- [75] Shashank Shekhar, Florian Bordes, Pascal Vincent, and Ari Morcos. Objectives matter: Understanding the impact of self-supervised objectives on vision transformer representations. *arXiv:2304.13089*, 2023. [3.1](#), [3.2.1](#), [3.4.4](#)
- [76] Ukcheol Shin, Jinsun Park, and In So Kweon. Deep depth estimation from thermal image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1043–1053, 2023. [4.4](#), [4.4](#)
- [77] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2022. [2.2](#)
- [78] Xun Sun, Yuanfan Xie, Pei Luo, and Liang Wang. A dataset for benchmarking image-based localization. In *CVPR*, 2017. [3.8.1](#)
- [79] Niko Sünderhauf, Sareh Shirazi, Feras Dayoub, Ben Upcroft, and Michael Milford. On the performance of convnet features for place recognition. In *IROS*, 2015. [3.8.1](#)
- [80] Li Tang, Yue Wang, Qianhui Luo, Xiaqing Ding, and Rong Xiong. Adversarial feature disentanglement for place recognition across changing appearance. In *ICRA*, 2020. [3.2.2](#)
- [81] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):25–55, 2020. doi: 10.1146/annurev-control-101119-071628. [3.2.1](#)
- [82] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *CVPR*, 2013. [3.2.2](#)
- [83] Mikaela Angelina Uy and Gim Hee Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4470–4479, 2018. [4.2.1](#)
- [84] Han Wang, Chen Wang, and Lihua Xie. Intensity scan context: Coding intensity and geometry relations for loop closure detection. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2095–2101. IEEE, 2020. [4.2.1](#)
- [85] Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. Transvpr: Transformer-based place recognition with multi-level attention aggregation. In *CVPR*, 2022. [3.2.2](#)

- [86] Ying Wang, Zezhou Sun, Cheng-Zhong Xu, Sanjay E Sarma, Jian Yang, and Hui Kong. Lidar iris for loop-closure detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5769–5775. IEEE, 2020. [2.2](#), [4.2.1](#)
- [87] Yu Wang, Haoyao Chen, Yufeng Liu, and Shiwu Zhang. Edge-based monocular thermal-inertial odometry in visually degraded environments. *IEEE Robotics and Automation Letters*, 8(4):2078–2085, 2023. [4.2.1](#)
- [88] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *CVPR*, 2020. [3.2.2](#)
- [89] Michael Warren, David McKinnon, Hu He, and Ben Upcroft. Unaided stereo vision based pose estimation. In *ACRA*, volume 47, page 60. Citeseer, 2010. [3.8.1](#)
- [90] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *CVPR*, 2020. [3.2.2](#)
- [91] Jiahong Xiao, Daniel Tortei, Eloy Roura, and Giuseppe Loianno. Long-range uav thermal geo-localization with satellite imagery. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5820–5827. IEEE, 2023. [4.2.2](#)
- [92] Jiahong Xiao, Gao Zhu, and Giuseppe Loianno. Visual geo-localization with self-supervised representation learning. *arXiv preprint arXiv:2308.00090*, 2023. [3.2.2](#)
- [93] Jiahong Xiao, Ning Zhang, Daniel Tortei, and Giuseppe Loianno. Sthn: Deep homography estimation for uav thermal geo-localization with satellite imagery. *arXiv preprint arXiv:2405.20470*, 2024. [4.2.2](#)
- [94] Peng Yin, Lingyun Xu, Ji Zhang, Howie Choset, and Sebastian Scherer. i3dloc: Image-to-range cross-domain localization robust to inconsistent environmental conditions. *arXiv preprint arXiv:2105.12883*, 2021. [4.2.2](#)
- [95] Mubariz Zaffar, Sourav Garg, Michael Milford, et al. Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. *IJCV*, pages 1–39, 2021. [3.2.2](#), [3.9](#)
- [96] Lintong Zhang, Tejaswi Digumarti, Georgi Tinchev, and Maurice Fallon. Instaloc: One-shot global lidar localisation in indoor environments through instance learning. *arXiv preprint arXiv:2305.09552*, 2023. [4.2.1](#)
- [97] Shibo Zhao, Peng Wang, Hengrui Zhang, Zheng Fang, and Sebastian Scherer. Tp-tio: A robust thermal-inertial odometry with deep thermalpoint. In *2020*

- IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4505–4512. IEEE, 2020. [4.2.1](#)
- [98] Shibo Zhao, Hengrui Zhang, Peng Wang, Lucas Nogueira, and Sebastian Scherer. Super odometry: Imu-centric lidar-visual-inertial estimator for challenging environments. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8729–8736. IEEE, 2021. [1](#)
- [99] Shibo Zhao, Damanpreet Singh, et al. Subt-mrs: A subterranean, multi-robot, multi-spectral and multi-degraded dataset for robust slam. *arXiv:2307.07607*, 2023. [3.1](#), [3.8.2](#), [3.8.2](#)
- [100] Zhipeng Zhao, Huai Yu, Chenwei Lyv, Wen Yang, and Sebastian Scherer. Attention-enhanced cross-modal localization between 360 images and point clouds. *arXiv preprint arXiv:2212.02757*, 2022. [4.2.2](#)
- [101] Zhipeng Zhao, Huai Yu, Chenwei Lyu, Wen Yang, and Sebastian Scherer. Attention-enhanced cross-modal localization between spherical images and point clouds. *IEEE Sensors Journal*, 2023. [4.2.2](#)
- [102] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *CVPR*, 2023. [3.2.2](#)