TITLE

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Programme
by Research

by

NAME
ROLL NUMBER
EMAIL ID

International Institute of Information Technology
(Deemed to be University)
Hyderabad - 500 032, INDIA
MONTH YEAR

Copyright © NAME, YEAR

All Rights Reserved

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **"TITLE"** by **NAME**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Prof. NAME

To SOMEONE

Acknowledgements

Acknowledgements goes here ...

Abstract

Abstract goes here ...

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Robotic Systems

Write in the end. Add the following in this section

- Components of an autonomous robot system: Environment + Perception + Localization and map building + Cognition, path planning + Motion control. Highlight Localization and map building (as ”contributed area”).
- Parts of a localization (SLAM) system and where VPR plays a role
- Image retrieval as a part of VPR systems. Elaborate the space/place of VPR (very brief of [8]).

1.1.1 General Autonomous Agents

A brief on AGI for autonomous robots. Open set works with Foundation Models (that work in any setting) are trending: Drive Anywhere [23], MUVO [2], GAIA [14].

1.1.2 Localization and Mapping

Info on SLAM systems

1.1.3 Visual Place Recognition

VPR and image retrieval

1.2 Foundation Models

Brief on Foundation Models. Two paragraphs maximum.

1.3 Contribution

List the contributions of the work in this thesis

Chapter 2

Foundation Models

All the basics of Vision Foundation Models required for understanding this thesis.
Foundation models (virtually all AI modes in general) have the following components

- *Model architecture*: MLP, convolution, transformers. Also MLP mixer [21], ConvNext [18], transformer variants (CCT) [10], etc.
- *Dataset*: type (labelled for supervised, unlabelled for unsupervised or self-supervised), size (large), augmentations, data processing pipelines.
- *Objective, training strategy and Loss function*: formulation of training procedure to guide the model output. Distillation [13], representation learning, MAE [11], contrastive losses (aligning modalities), knowledge transfer (student-teacher), MoCo [12, 6], SwAV [3], SimCLR [4, 5], BYOL [9], etc.
- *Optimizer*: usually Adam [17] (doesn't need explanation)

2.1 Vision Transformers

ViT [7] and DeIT [22]

2.2 SSL Concepts

Start with a short summary of the SSL cookbook [1].

Some of the above along with requirements for DINOv2: iBOT [24], LayerScale and Stochastic Depth [15], KoLeo regularizer [19], SwiGLU activation [20], Sinkhorn-Knoop centering [3] (SwAV).

2.3 DINO and DINOv2 details

Architecture, data, training, etc.

Chapter 3

AnyLoc: Foundation model features for VPR

Description of AnyLoc [16]

Chapter 4

Future Scope

What else can be done ahead for AnyLoc.

- Results with PCA seem promising, more model optimizations could give better results (with higher throughput/faster speed)
- Integration into a full SLAM system

Chapter 5

Conclusions

Something

Related Publications

1. Keetha, N.V., *Mishra, A.*, Karhade, J., Jatavallabhula, K., Scherer, S.A., Krishna, M., & Garg, S. (2023). AnyLoc: Towards Universal Visual Place Recognition. *IEEE Robotics and Automation Letters*, 9, 1286-1293. doi: 10.1109/LRA.2023.3343602 (arXiv: 2308.00688)

Bibliography

[1] Randall Balestriero et al. "A Cookbook of Self-Supervised Learning". In: *ArXiv* abs/2304.12210 (2023). URL: https://api.semanticscholar.org/CorpusID:258298825.

[2] Daniel Bogdoll, Yitian Yang, and J. Marius Zollner. "MUVO: A Multimodal Generative World Model for Autonomous Driving with Geometric Representations". In: *ArXiv* abs/2311.11762 (2023). URL: https://api.semanticscholar.org/CorpusID:265295410.

[3] Mathilde Caron et al. "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments". In: *ArXiv* abs/2006.09882 (2020). URL: https://api.semanticscholar.org/CorpusID:219721240.

[4] Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *ArXiv* abs/2002.05709 (2020). URL: https://api.semanticscholar.org/CorpusID:211096730.

[5] Ting Chen et al. "Big Self-Supervised Models are Strong Semi-Supervised Learners". In: *ArXiv* abs/2006.10029 (2020). URL: https://api.semanticscholar.org/CorpusID:219721239.

[6] Xinlei Chen et al. "Improved Baselines with Momentum Contrastive Learning". In: *ArXiv* abs/2003.04297 (2020). URL: https://api.semanticscholar.org/CorpusID:212633993.

[7] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ArXiv* abs/2010.11929 (2020). URL: https://api.semanticscholar.org/CorpusID:225039882.

[8] Sourav Garg, Tobias Fischer, and Michael Milford. "Where is your place, Visual Place Recognition?" In: *ArXiv* abs/2103.06443 (2021). URL: https://api.semanticscholar.org/CorpusID:232185215.

[9] Jean-Bastien Grill et al. "Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning". In: *ArXiv* abs/2006.07733 (2020). URL: https://api.semanticscholar.org/CorpusID:219687798.

[10] Ali Hassani et al. "Escaping the Big Data Paradigm with Compact Transformers". In: *ArXiv* abs/2104.05704 (2021). URL: https://api.semanticscholar.org/CorpusID:233210459.

[11] Kaiming He et al. "Masked Autoencoders Are Scalable Vision Learners". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 15979–15988. URL: https://api.semanticscholar.org/CorpusID:243985980.

[12] Kaiming He et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 9726–9735. URL: https://api.semanticscholar.org/CorpusID:207930212.

[13] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. "Distilling the Knowledge in a Neural Network". In: *ArXiv* abs/1503.02531 (2015). URL: https://api.semanticscholar.org/CorpusID:7200347.

[14] Anthony Hu et al. "GAIA-1: A Generative World Model for Autonomous Driving". In: *ArXiv* abs/2309.17080 (2023). URL: https://api.semanticscholar.org/CorpusID:263310665.

[15] Gao Huang et al. "Deep Networks with Stochastic Depth". In: *European Conference on Computer Vision*. 2016. URL: https://api.semanticscholar.org/CorpusID:6773885.

[16] Nikhil Varma Keetha et al. "AnyLoc: Towards Universal Visual Place Recognition". In: *IEEE Robotics and Automation Letters* 9 (2023), pp. 1286–1293. URL: https://api.semanticscholar.org/CorpusID:260351368.

[17] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2014). URL: https://api.semanticscholar.org/CorpusID:6628106.

[18] Zhuang Liu et al. "A ConvNet for the 2020s". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 11966–11976. URL: https://api.semanticscholar.org/CorpusID:245837420.

[19] Alexandre Sablayrolles et al. "Spreading vectors for similarity search". In: *arXiv: Machine Learning* (2018). URL: https://api.semanticscholar.org/CorpusID:62841605.

[20] Noam M. Shazeer. "GLU Variants Improve Transformer". In: *ArXiv* abs/2002.05202 (2020). URL: https://api.semanticscholar.org/CorpusID:211096588.

[21] Ilya O. Tolstikhin et al. "MLP-Mixer: An all-MLP Architecture for Vision". In: *Neural Information Processing Systems*. 2021. URL: https://api.semanticscholar.org/CorpusID:233714958.

[22] Hugo Touvron et al. "Training data-efficient image transformers & distillation through attention". In: *International Conference on Machine Learning*. 2020. URL: https://api.semanticscholar.org/CorpusID:229363322.

[23] Tsun-Hsuan Wang et al. "Drive Anywhere: Generalizable End-to-end Autonomous Driving with Multi-modal Foundation Models". In: *ArXiv* abs/2310.17642 (2023). URL: https://api.semanticscholar.org/CorpusID:264490392.

[24] Jinghao Zhou et al. "iBOT: Image BERT Pre-Training with Online Tokenizer". In: *ArXiv* abs/2111.07832 (2021). URL: https://api.semanticscholar.org/CorpusID:244117494.