Vilnius University

An investigation of deep imitation learning for mobile robot navigation

Shubham Juneja
Supervisor: Dr. Virginijus Marcinkevičius
Semester 8

Vilnius University

Plan of studies & implementation summary

Study year	Exams		Conference participations		Publications	
	Planned	Completed	Planned	Completed	Planned	Completed
I (2020/2021)	2	2	1	1		
II (2021/2022)	2	2				
III (2022/2023)					1	1
IV (2023/2024)			1	1	1	(Under review)

Vilnius University

Report of activity plan

Exams		Conference Participation		Publications	
Planned	**Status**	**Planned**	**Status**	**Planned**	**Status**
Machine Learning	Passed with 9/10	All Sensors 2021, Nice, France	Paper accepted and presented at All sensors 2021 conference at Nice, France. On the 20th of July, 2021.	Journal paper at IEEE Access Journal. Title: “Visual Place Recognition Pre-Training for End-to-End Trained Autonomous Driving Agent”	Published
Research methods and methodology of informatics and computer engineering	Passed with 9/10	AI SyS 2024, Venice, Italy	Paper accepted and presented at AI SyS 2024 conference at Venice, Italy. On the 29th of September, 2024.	Journal paper for Baltic Journal of Moden Computing. Title: “DINO	Under-review
Fundamentals of informatics	Passed with 7/10				
Optimisation	Passed with 7/10				

Vilnius University

Workshops

Workshop
MOKSLINIŲ REZULTATŲ PUBLIKAVIMAS PAGAL FORMALAUS VERTINIMO REIKALAVIMUS
MOKSLINĖS INFORMACIJOS IŠTEKLIAI, PAIEŠKA, IR ĮRANKIAI
MENDELEY PRAKTINIS UŽSIĖMIMAS
DeepLearn 2022 Summer School

Stages of research and dissertation preparation

Vilnius University

	Name of task	Duration	Notes
4.	Preparation and consideration of the doctoral dissertation in the department.	May 2024 Onwards	Awaiting second journal paper review. Writing dissertation.

Vilnius University

Research Object and Aim

Research object:

- The research object is imitation learning-based autonomous driving techniques with focus on exploration in pre-training methods and their effect over the problem of co-variate shift.

Research aim:

- To develop, implement and research an autonomous navigation system driving algorithms for mobile robots based on imitation learning and deep neural networks, that explore pre-training techniques and enhance generalisation over seen and unseen environmental settings.

Objectives of Research

1. Performing of a study of the state-of-the-art methods in imitation learning based end-to-end autonomous driving.
2. Identifying the common cause of the problems in the studied state-of-the-art, i.e., under-exploration of pre-training techniques.
3. Proposing pre-training algorithms that enhance the state-of-the-art end-to-end autonomous driving, followed by software level implementation of the proposed algorithms.
4. Evaluation and comparison of the implemented versions of the proposed algorithms.

Scientific Novelty

1. Extending the **under-explored research on pre-training** methods for end-to-end trained neural networks for autonomous driving.
2. This thesis proposes **visual place recognition as a pre-training task** for autonomous driving. It also empirically shows how such pre-training out performs the commonly used pre-training technique.
3. Similarly another pre-training method, **self-distillation with no labels (DINO) pre-training**, is proposed and shown to be effective with the support of experiments.
4. Additionally, this thesis also proposes **monocular depth estimation as a pre-training task** for driving, followed by, **use of the perceiver architecture** as another method for training end-to-end autonomous driving methods.

Practical Significance

1. The experiments done using pre-training methods, namely visual place recognition and DINO, show higher resistance to changes in the environment when deployed in simulation environments. This means that such practices can **lead the way to reliable driving in environment** that are not exposed to learner and hence lessening training data requirements.
2. The experiments also show **faster convergence** to higher performance whenever the proposed methods are trained. This shows lessening of expensive GPU hours requirements.
3. The thesis also makes **training code** for training autonomous driving methods **publicly available** and mentions other important repositories.
4. The proposed used of the perceiver opens horizons on further application of **multi-modal methods**.

Vilnius University

Statements To Be Defended

1. The area of imitation learning based autonomous driving lacks exploration in techniques which pre-train visual encoders.
2. Pre-training the visual encoder over the task of visual place recognition using triplet loss instead of the commonly used classification task on the ResNet architecture, enhances the driving performance of imitation learning based autonomous driving system over multiple metrics.
3. Pre-training the visual encoder over ImageNet dataset using self-distillation with no labels (DINO) method instead of the commonly used classification task on the ResNet architecture, produces richer features for imitation learning based autonomous driving, that enable better driving performance as per multiple metrics.

Works Produced

- In journals:
  - Juneja, S., Daniušis, P., \& Marcinkevičius, V. (2023). **Visual place recognition pre-training for end-to-end trained autonomous driving agent**. IEEE access, 11, 128421-128428.
  - Juneja, S., Daniušis, P., \& Marcinkevičius, V. (2024). **DINO Pre-training for Vision-based End-to-end Autonomous Driving**. Baltic Journal of Modern Computing. [Under Review]
- In conference proceedings:
  - Juneja, S., Marcinkevičius, V., \& Daniušis, P. **Combining Multiple Modalities with Perceiver in Imitation-based Urban Driving.** All Sensors 2021. 18th July, 2021. Nice, France.
  - Juneja, S., Daniušis, P., \& Marcinkevičius, V. (2024). **Monocular Depth Estimation Pre-training for Autonomous Driving.** AI Sys 2024. 30th September, 2024. Venice, Italy.

Research

Vilnius University

Autonomous Driving

- Autonomous driving refers to the technology that enables vehicles to operate without human intervention by sensing their environment and navigating safely.
- These systems rely on sensors, cameras, artificial intelligence, and advanced algorithms to make real-time decisions.
- **Modular** autonomous driving systems break down the driving task into distinct modules such as perception, planning, and control; requiring for every module to co-ordinate and be programmed individually.
- **End-to-end** learning is an approach in autonomous driving where the driving policy is learned directly from sensory inputs to control outputs. This method uses neural networks to map raw data (like images from cameras) directly to steering angles and throttle commands, simulating a human-like perception and decision-making process.

Learning to imitate

Vilnius University

- In imitation learning:
  - Given: Demonstrations
  - Goal: Train a policy (model) to mimic demonstrations
- Being a form of machine learning, data is collected, models are optimized, accuracies are evaluated.

Vilnius University

About the problem to solve

- Learning sensorimotor skills to drive and navigate based on visual input.
- It can be done with traditional methods such as SLAM, but it would require expensive sensors and extensive programming.
- The idea of imitation learning promises to solve this problem by learning from human demonstrations.
- Yet, it remains unsolved due the unpredictability of the real world causing the problem of covariate shift.
- To compare the ability between methods Leaderboard benchmark has been established.
- Leaderboard benchmark uses CARLA simulator to seed vehicles in different parts of a map and tests the ability of reaching from point A to B, under different sets of conditions.

Vilnius University

Visual Place Recognition Pre-Training for End-to-End Trained Autonomous Driving Agent

- All end-to-end driving benchmarks test under varying weather conditions.
- Performance in the state-of-the-art seem to generalise worse for previously unseen weather settings.
- Current algorithms train for different weather conditions during the imitation learning process, where it is implied but not explicit.

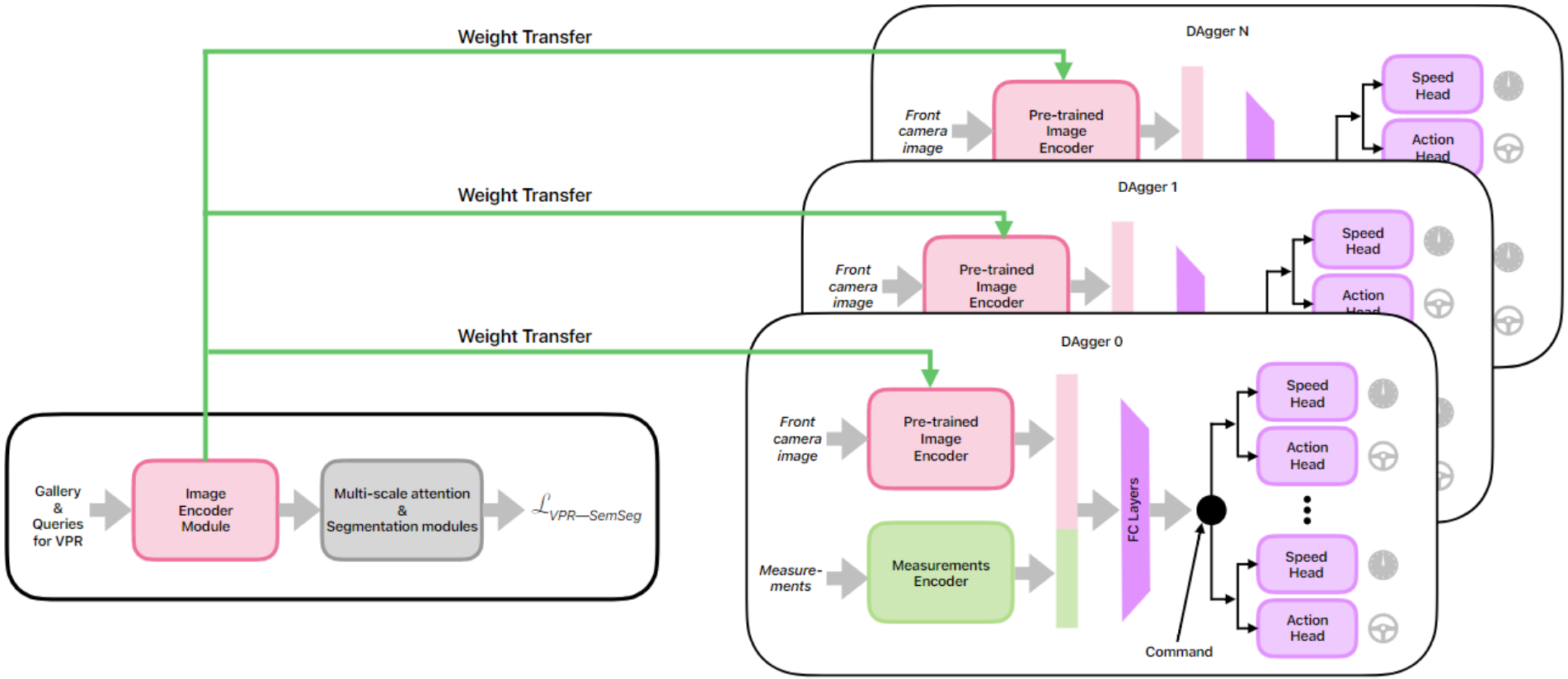
Vilnius University

Proposed method

- Currently used encoders in end-to-end driving architectures are ImageNet pre-trained ResNet encoders.
- We propose, pretraining the encoder for the task of visual place recognition at first.
- Followed by carrying out imitation learning to learn the task of driving.
- This way we attempt to achieve weather invariance.

Vilnius University

Proposed method



Vilnius University

Visual Place Recognition

- Visual place recognition is an area which deals with making recognition of places invariant to weather and lighting conditions.
- Data is collected at different times of the day, with the help of GPS coordinates to be able to sample closer locations and further locations as per requirement.

(a) Wet noon

(b) Clear sunset

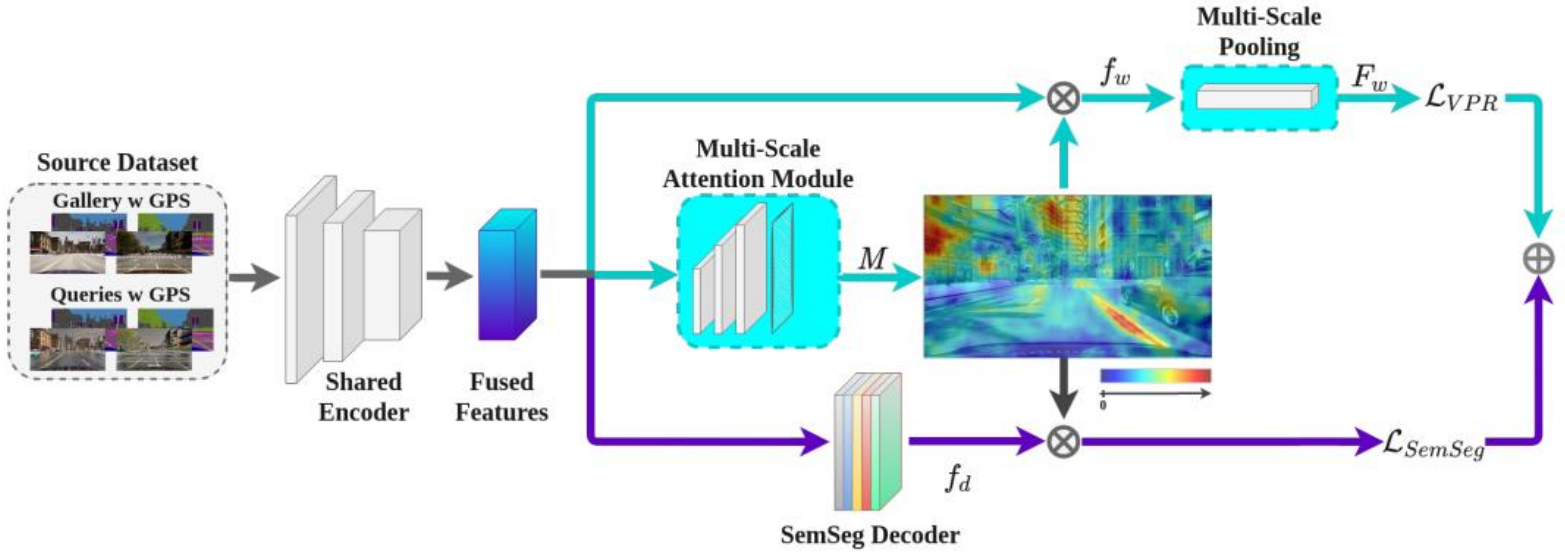
(c) Soft rain sunset

(d) Wet sunset

Vilnius University

VPR Method used: SegVPR

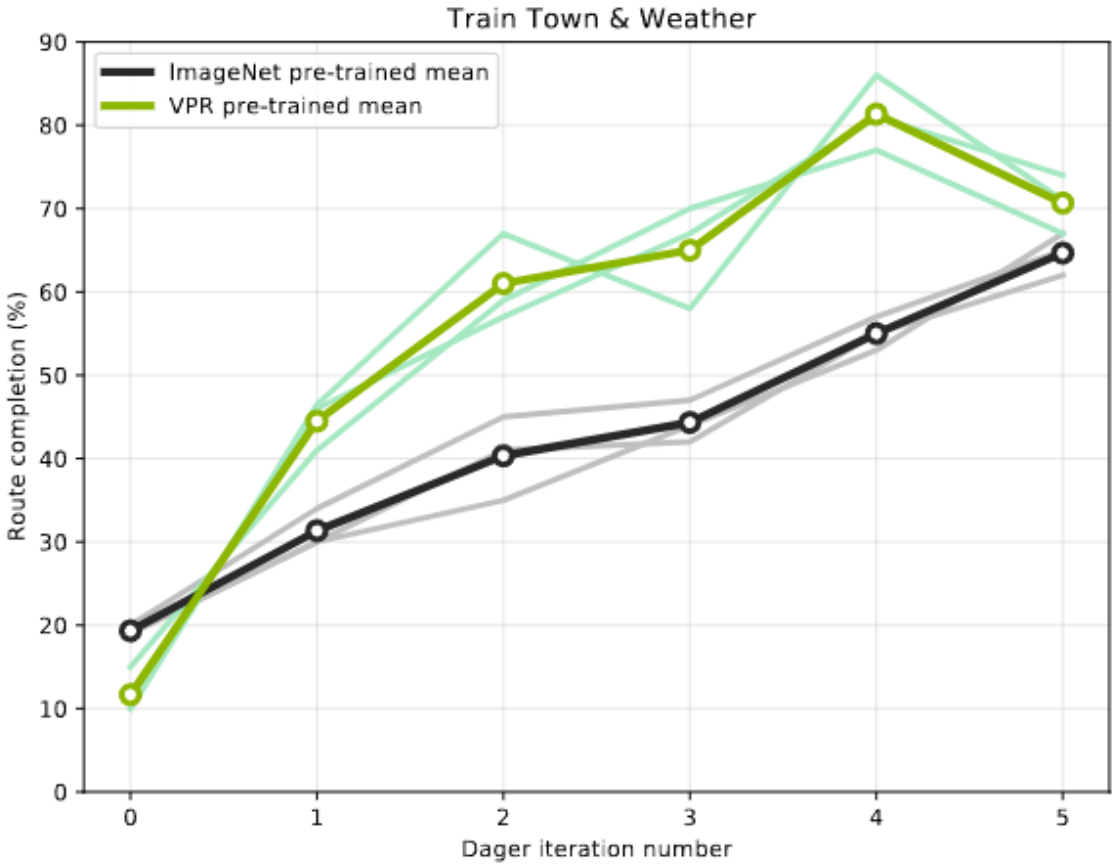
- SegVPR learns global embeddings from visual and semantic context of data.
- It uses semantic segmentation to dynamically guide the task of recognising places.

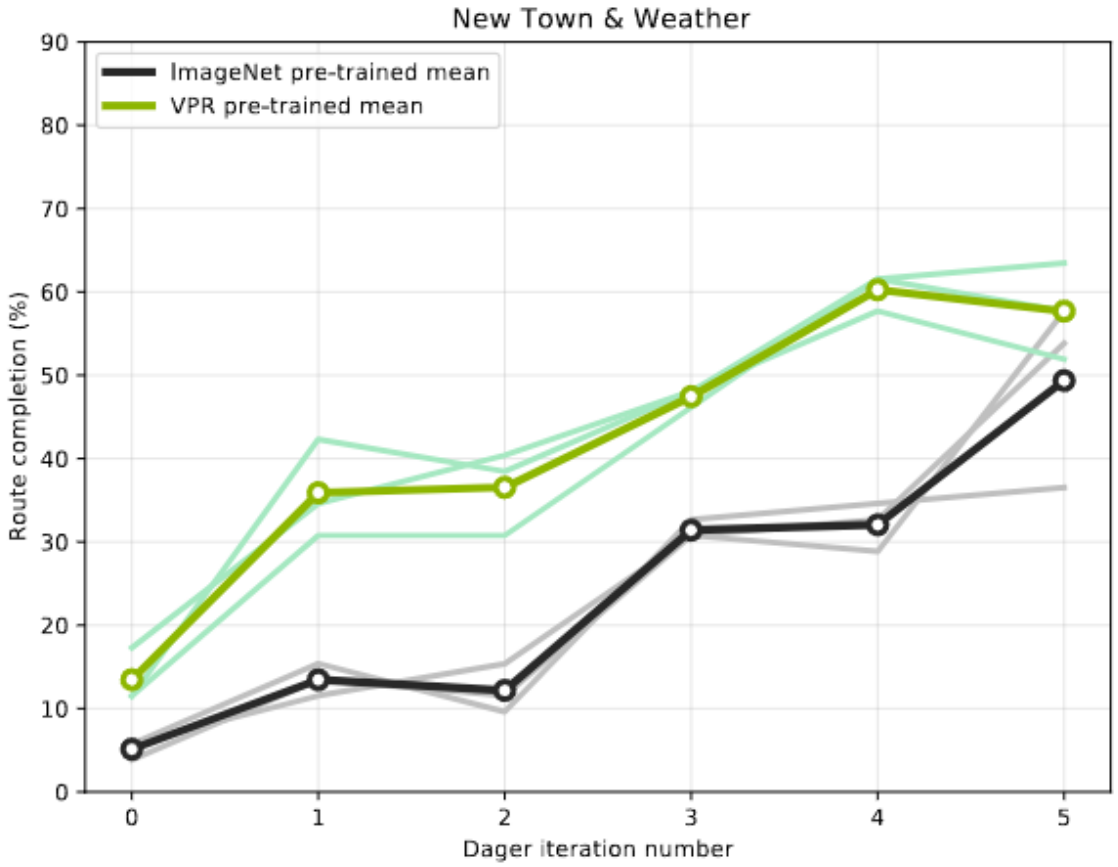


Vilnius University

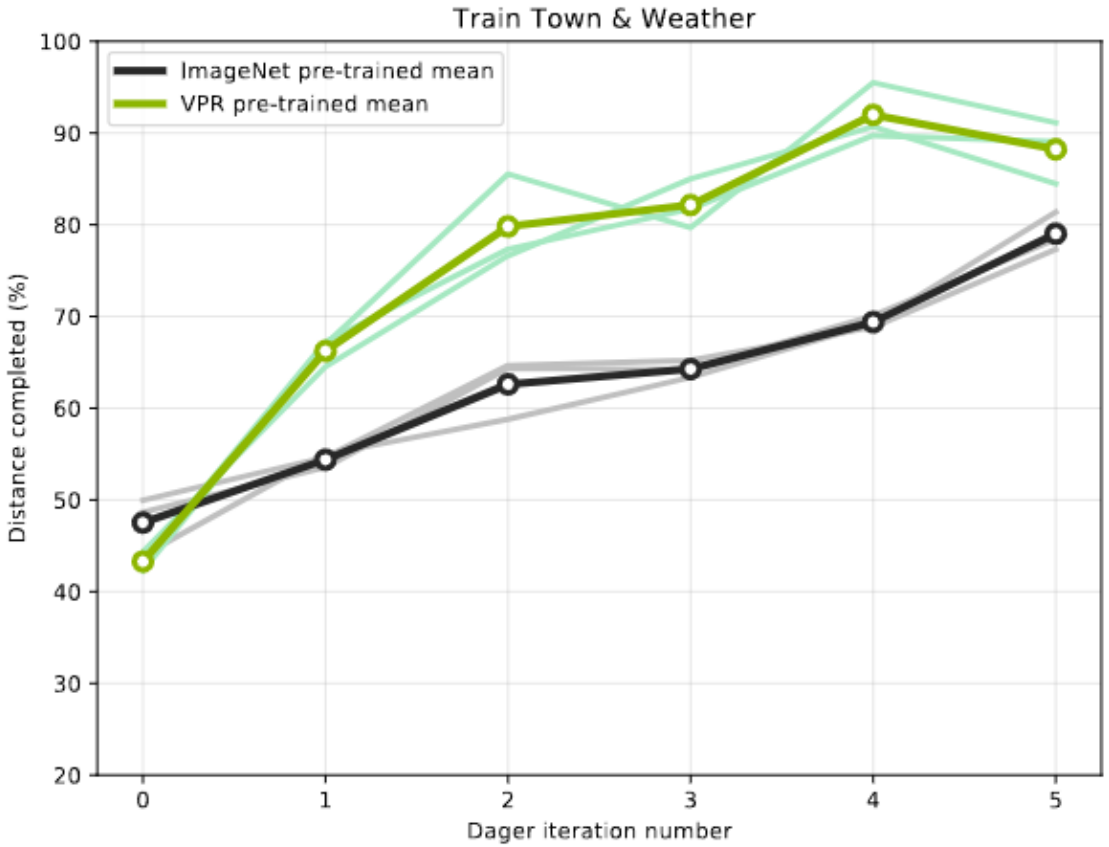
Results

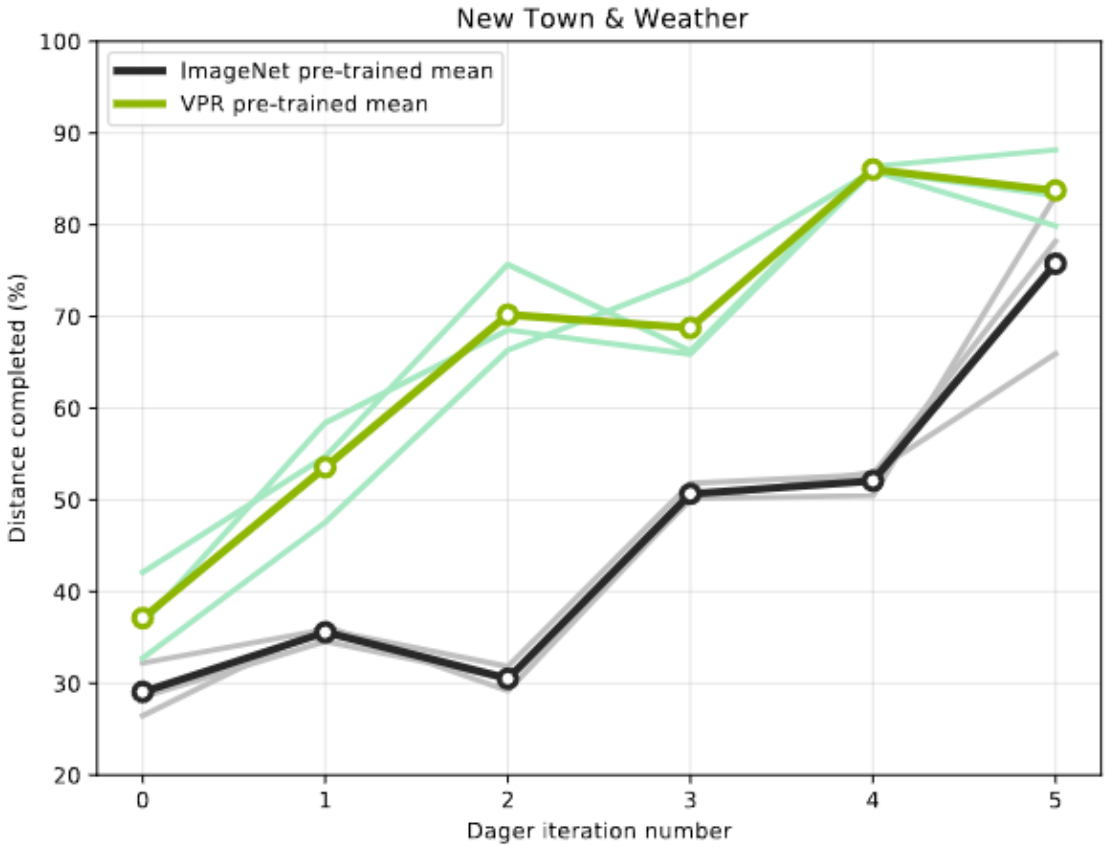
Vilnius University





Vilnius University





Vilnius University

DINO Pre-training for Vision-based End-to-end Autonomous Driving

- Supervised learning methods may consist of a strong bias that is brought by guidance of labels.
- We hypothesise that the presence of strong image-level supervision could be reducing the concept of an image to a very narrow understanding.
- On the other hand, self-supervised learning offers advantages of making the learning more general.
- With use of self-supervised learning the models learn not just a single concept with regards to the image, but they explicitly learn:
  - semantic segmentation of an image
  - scene layout
  - object boundaries

Vilnius University

DINO Pre-training for Vision-based End-to-end Autonomous Driving

- We propose using the self-distillation with no labels (DINO) method for pre-training a visual encoder.
- Followed by training an architecture containing the pre-trained visual encoder over the task of autonomous driving.
- We inherit the settings from the previous research to be able to compare results.
- We later compare the results against baseline methods, over multiple metrics

Self-distillation with no labels (DINO)

- DINO uses two networks for pre-training, in a student-teacher form.
- This method updates one network faster than the other and hence enabling self-distillation.
- Additionally, the two methods view different parts of the image while training, using the method termed as multi-crop training.
- During training, the following loss function is optimised for each of the netwroks:

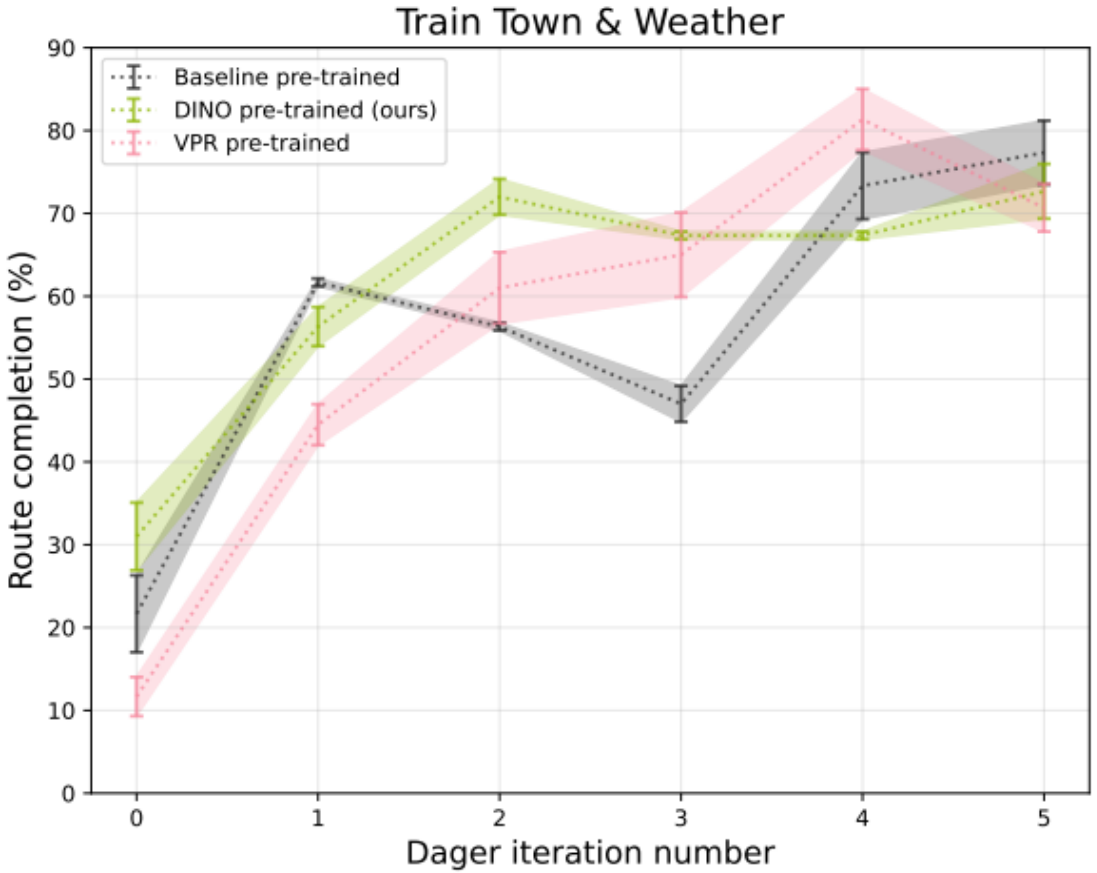
$$P_s(x)^{(i)} = \frac{\exp\left(g_\theta(x)^{(i)}/\tau_s\right)}{\sum_{k=1}^{K}\exp\left(g_\theta(x)^{(k)}/\tau_s\right)}$$

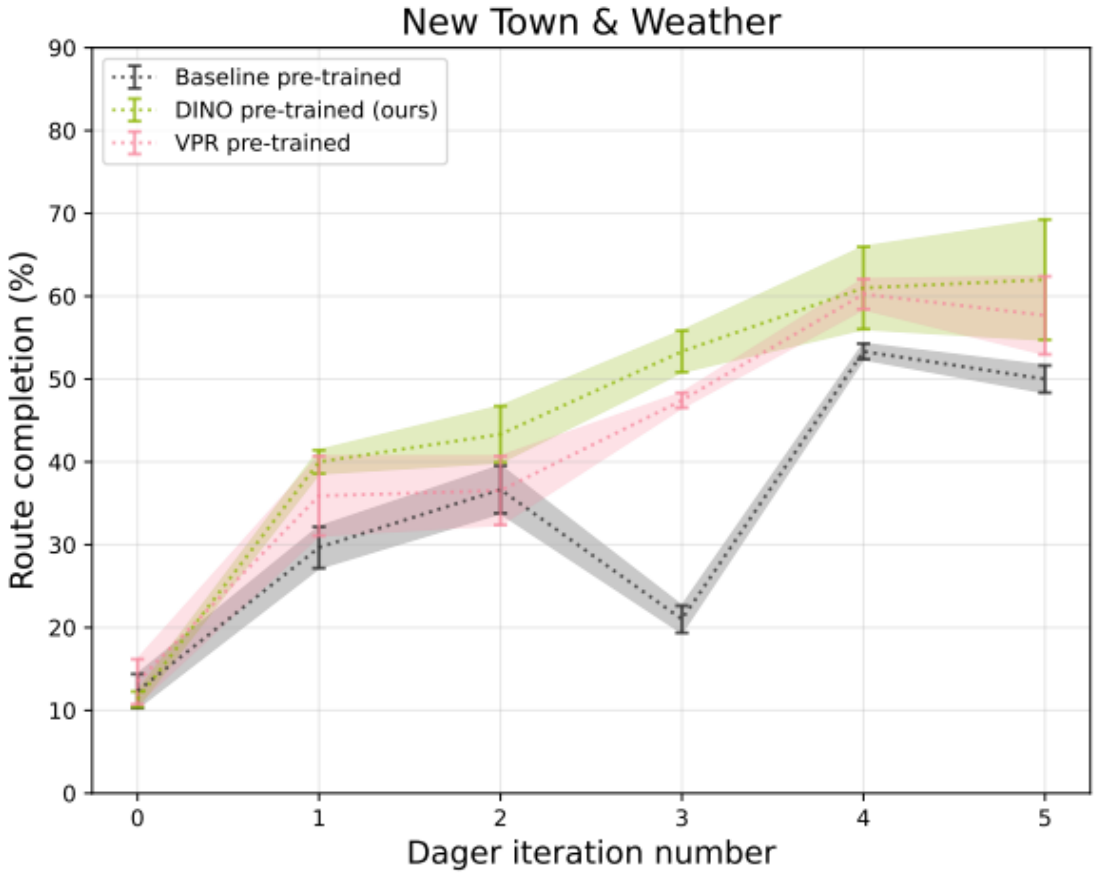
- The presented equation belongs one of the networks.

Vilnius University

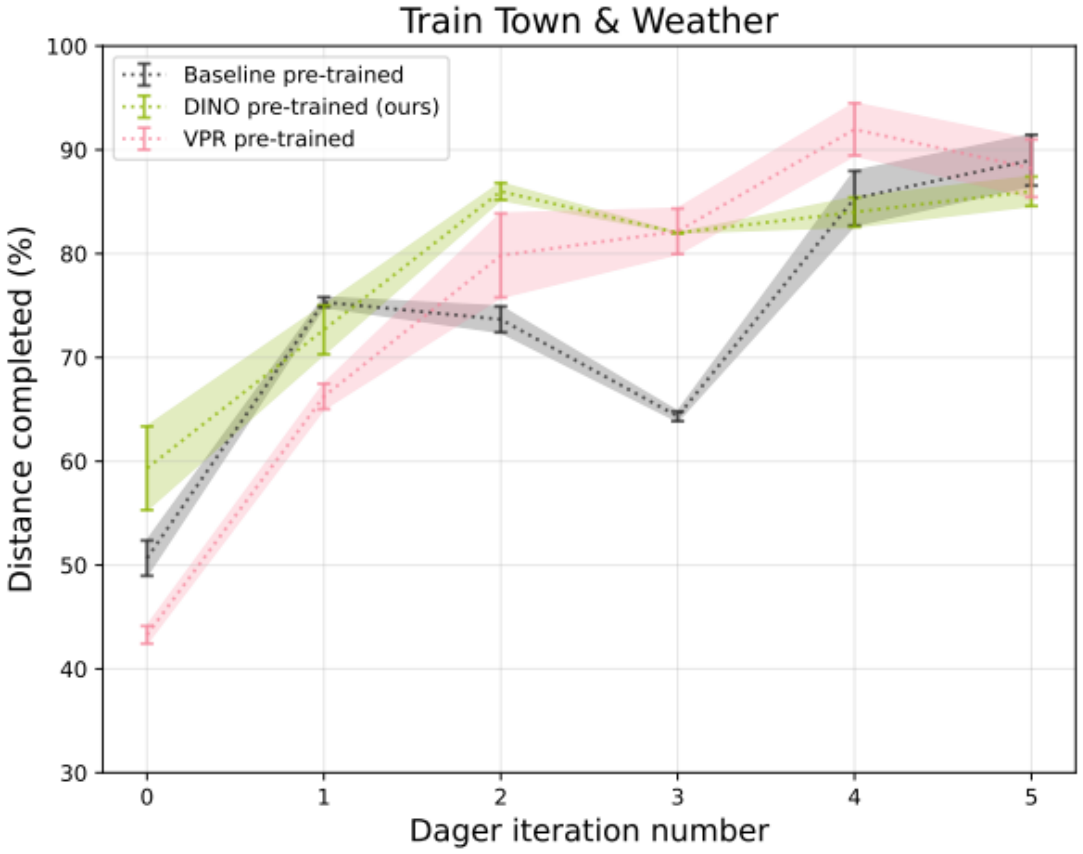
Results

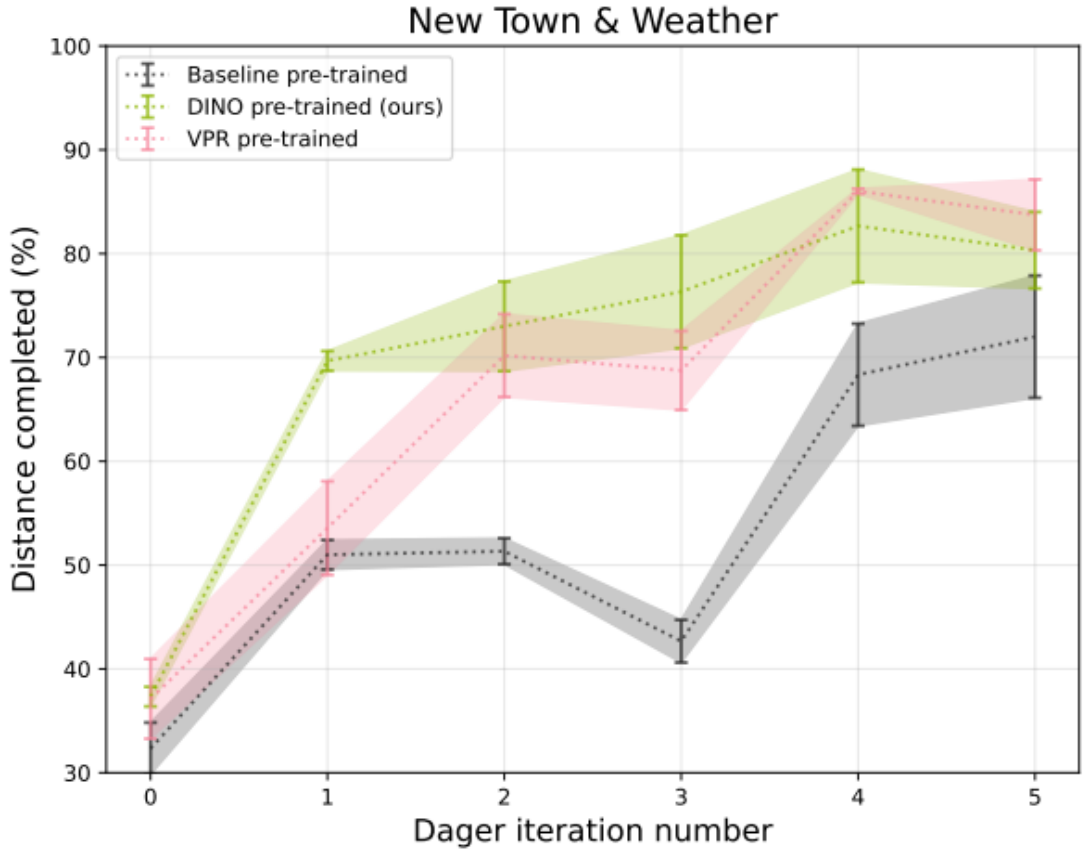
Vilnius University





Vilnius University





Vilnius University

Conference Papers

Vilnius University

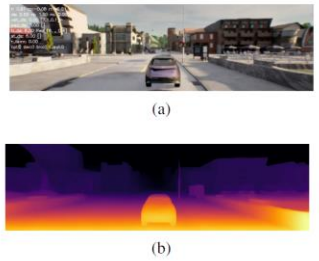
Combining multiple modalities with Perceiver in IL based learning

- We present a study pointing out how end-to-end methods rely on a single modality while lacking the performance compared to traditional autonomous driving methods which take a modular approach.
- Therefore, we propose a method to enrol more than one modality in the learner.
- We propose the use of a perceiver architecture in the learner as this architecture shows capability of learning with varying number and types of modalities as input data.
- Since the published paper is a idea paper, no experiments were presented.

Monocular Depth Estimation Pre-training for Imitation-based Autonomous Driving

Vilnius University

- We propose pre-training an encoder over the task of monocular depth estimation using the Depth Anything method.
- Thereon, using the pre-trained encoder to train over the task of autonomous driving.
- We propose using the architecture from recent methods that explore pre-training [1] and compare with a baseline approach.
- The baseline model could be the most common form of pre-training, i.e., classification-based pre-training over the ImageNet dataset, as in [1].



(a)

(b)

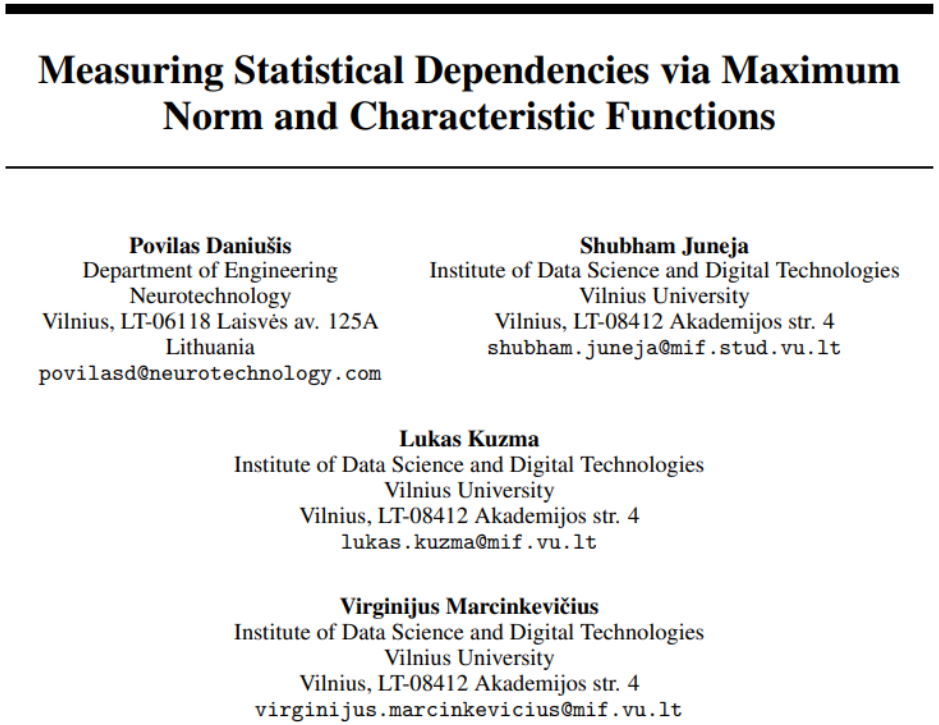
Fig. 1. (a) An RGB image from CARLA simulator. (b) Same RGB image's estimated depth with depth anything [9].

[1] S. Juneja, P. Daniusis, and V. Marcinkevicius, "Visual place recognition pre-training for end-to-end trained autonomous driving agent," IEEE Access, vo 128 421–128 428, 2023.

Vilnius University

Publication work in progress:

- Under review in a journal



Measuring Statistical Dependencies via Maximum Norm and Characteristic Functions

Povilas Daniušis
Department of Engineering Neurotechnology
Vilnius, LT-06118 Laisvės av. 125A
Lithuania
povilasd@neurotechnology.com

Shubham Juneja
Institute of Data Science and Digital Technologies
Vilnius University
Vilnius, LT-08412 Akademijos str. 4
shubham.juneja@mif.stud.vu.lt

Lukas Kuzma
Institute of Data Science and Digital Technologies
Vilnius University
Vilnius, LT-08412 Akademijos str. 4
lukas.kuzma@mif.vu.lt

Virginijus Marcinkevičius
Institute of Data Science and Digital Technologies
Vilnius University
Vilnius, LT-08412 Akademijos str. 4
virginijus.marcinkevicius@mif.vu.lt

Vilnius
University

Thank you