# Modality-invariant Visual Odometry for Indoor Navigation

Marius Memmel, Amir Zamir

EPFL

TECHNISCHE UNIVERSITÄT DARMSTADT

Code, paper, visuals

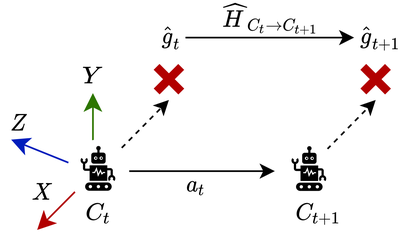
CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

## Problem

- *GPS+compass* are **not reliable for localization** in indoor environments [1]
- Visual Odometry (VO) is a sufficient alternative, but models can be **dominated by a single modality** in a multi-modal (*RGBD*) setting
- **Erroneous sensors** causes **catastrophic failure** of VO models

## Visual Odometry for Indoor Navigation [1,2]

- Estimate transformation between agent's coordinate systems



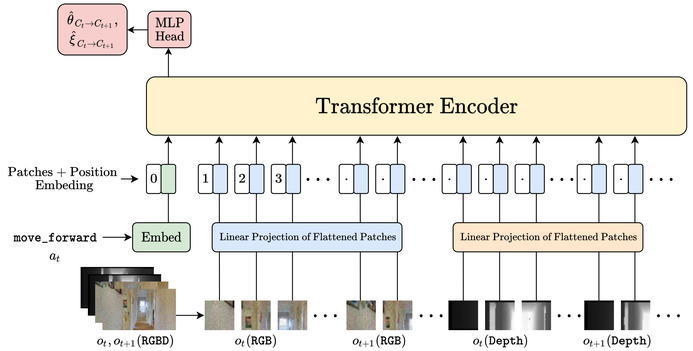

$$\widehat{H}_{C_t \to C_{t+1}} = \begin{bmatrix} \widehat{R}_{C_t \to C_{t+1}} & \widehat{\boldsymbol{\xi}}_{C_t \to C_{t+1}} \\ 0 & 1 \end{bmatrix}$$

$$\text{with} \quad \widehat{R}_{C_t \to C_{t+1}} = \begin{bmatrix} \cos(\widehat{\beta}_{C_t \to C_{t+1}}) & -\sin(\widehat{\beta}_{C_t \to C_{t+1}}) \\ \sin(\widehat{\beta}_{C_t \to C_{t+1}}) & \cos(\widehat{\beta}_{C_t \to C_{t+1}}) \end{bmatrix} \in SO(2)$$

$$\langle \widehat{\beta}_{C_t \to C_{t+1}}, \widehat{\boldsymbol{\xi}}_{C_t \to C_{t+1}} \rangle = f_\phi(\boldsymbol{o}_t, \boldsymbol{o}_{t+1})$$

- Predict transformation parameters from observations $\boldsymbol{o}_t, \boldsymbol{o}_{t+1}$ using $f_\phi$
- When $f_\phi$ is ConvNet, input size is fixed
- ▶ Dropping any modality leads to catastrophic failure!

## Solution: **Visual Odometry Vision Transformers** (VOT)




- *Vision Transformers* are **flexible** *w.r.t.* **input modalities** even during test time due to the independence of their weights on the sequence length [3]
- Multi-modal **pre-training** (*e.g. MultiMAE* [4])
- Attention mechanism to **identify importance** of regions and modalities
- Pass **action prior** as separate token

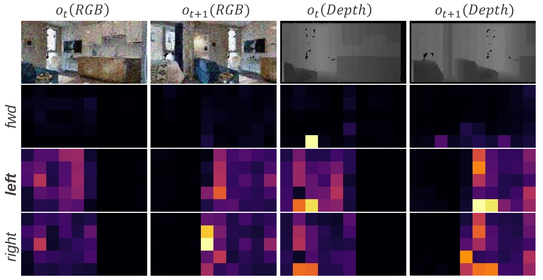
## PointNav Results

- dropping modalities (drp) lets ConvNet approaches converge to a *blind* VO model
- VOT maintains sufficient localization capabilities, *c.f., SoftSPL* (SSPL) [1]

| method | obs | drp | $S\uparrow$ | SPL↑ | SSPL↑ | |
|---|---|---|---|---|---|---|
| *blind* | – | – | 0.00 | 0.00 | 5.40 | |
| *oracle* | – | – | 97.89 | 74.80 | 73.10 | |
| [17] | RGB-D | – | 64.50 | 48.90 | 65.40 | |
| [17] | RGB-D | RGB | 0.00 | 0.00 | 5.40 | |
| [17] | RGB-D | Depth | 0.00 | 0.00 | 5.40 | |
| VOT (ours) | RGB | – | 59.30 | 45.40 | 66.70 | 1) |
| VOT (ours) | Depth | – | 93.30 | 71.70 | 72.00 | 2) |
| VOT (ours) | RGB-D | – | 88.20 | 67.90 | 71.30 | 3) |
| VOT (ours) | RGB-D | RGB | 75.90 | 58.50 | 69.90 | 4) |
| VOT (ours) | RGB-D | Depth | 26.10 | 20.00 | 58.70 | 5) |

## Attention Maps

(VOT trained on RGBD and pre-trained with a MultiMAE)




- Action taken: *left*, Injected actions: *fwd, left, right*
- Embed *fwd* causes the attention to focus on the center
- Embed *left* and *right* move attention towards regions of the image that would be consistent across time steps *t,t+1* in case of rotation

## Limitations & Future Work

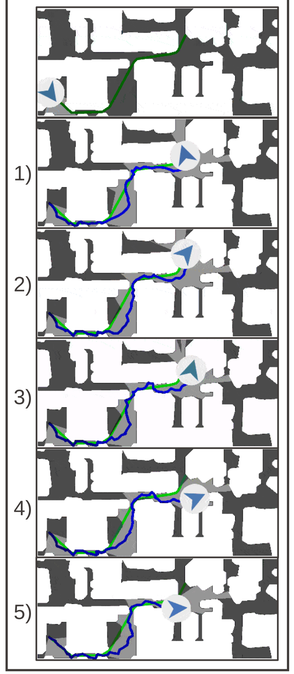
- ▶ Full robustness may require explicit measures (randomly drop modalities during training)
- ▶ Dropping modalities causes inconsistencies in the localization (fine-tune the policy to adapt)
- ▶ Only two modalities (extend by semantic segmentation)

## Takeaways

- ▶ VOT can function when modalities are missing
- ▶ VOT implicitly learns some invariance to modalities
- ▶ Action prior primes model to attend relevant regions
- ▶ Pre-training reduces data requirements to 25%

## PointNav Paths

- ▶ VOT helps agents reach the goal 1) - 3)
- when dropping modalities, agent still makes progress towards the goal 4) 5)
- however, those agents get stuck in narrow passages
- ▶ don't reach the goal in *T* time steps




References: [1] Integrating Egocentric Localization for More Realistic Point-Goal Navigation Agents. *Datta et al.* CoRL 2021
[2,17] The Surprising Effectiveness of Visual Odometry Techniques for Embodied PointGoal Navigation. *Zhao et al.* ICCV 2021
[3] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Dosovitskiy et al.* ICLR 2021
[4] MultiMAE: Multi-modal Multi-task Masked Autoencoders. *Bachmann et al.* arXiv preprint 2022