EPFL

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

THU-PM-090

# Modality-invariant Visual Odometry for Embodied Vision

vo-transformer.github.io



Marius Memmel, Roman Bachmann, Amir Zamir

EPFL

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

# Modality-invariant Visual Odometry for Embodied Vision

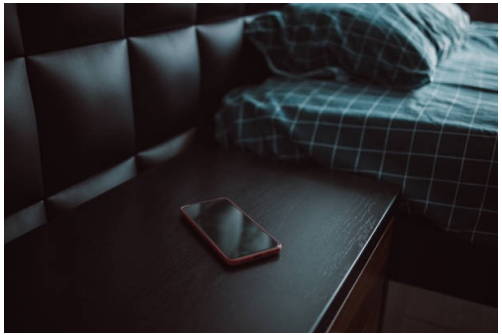
vo-transformer.github.io

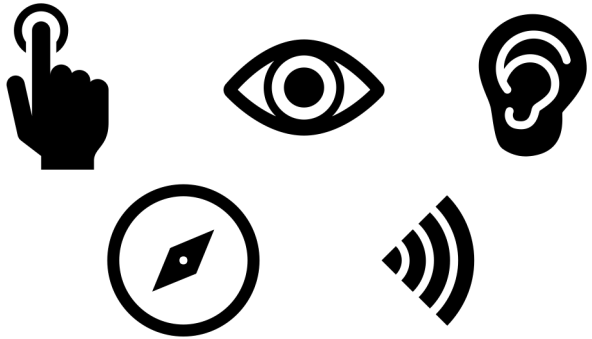


Marius Memmel, Roman Bachmann, Amir Zamir

# We can switch between modalities to localize. Odometry models should too!





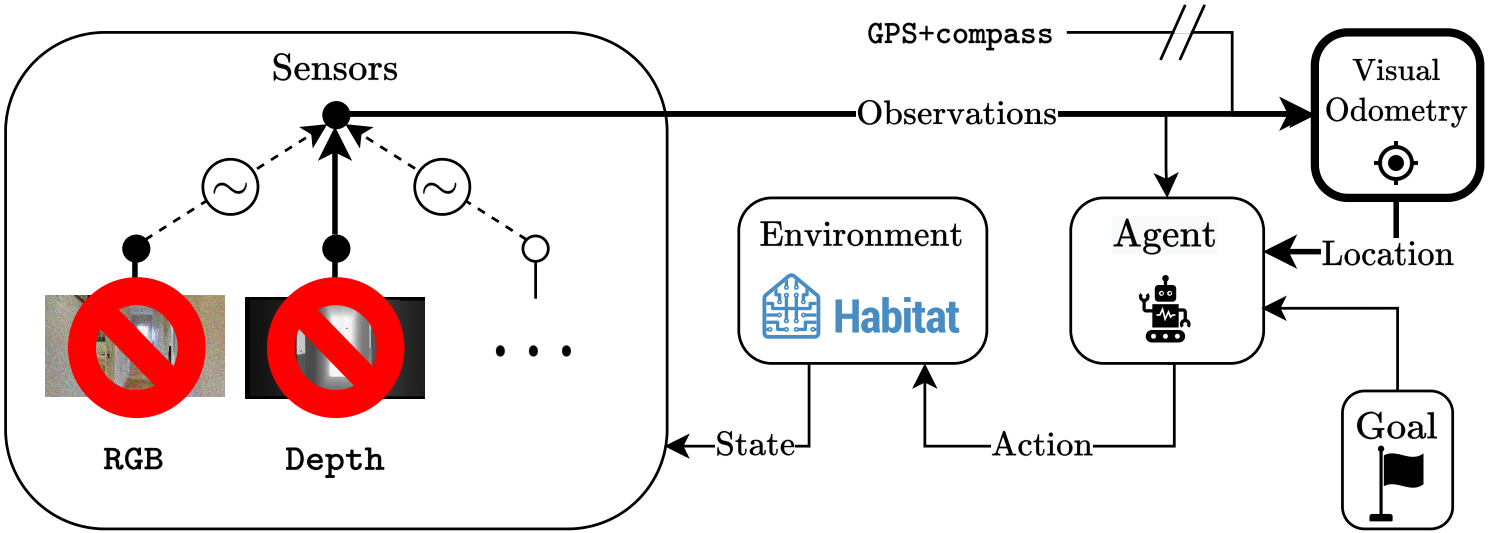




EPFL

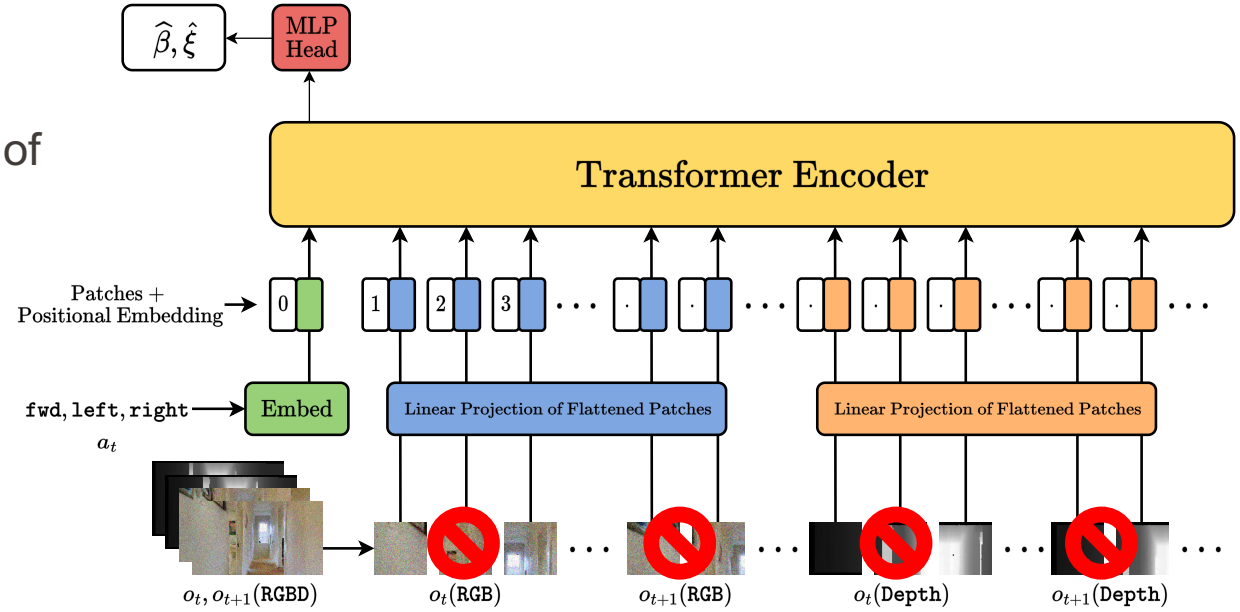
# Problem

- Multi-modal Point-goal Navigation with *'optional'* modalities



# Visual Odometry Transformers (VOT)

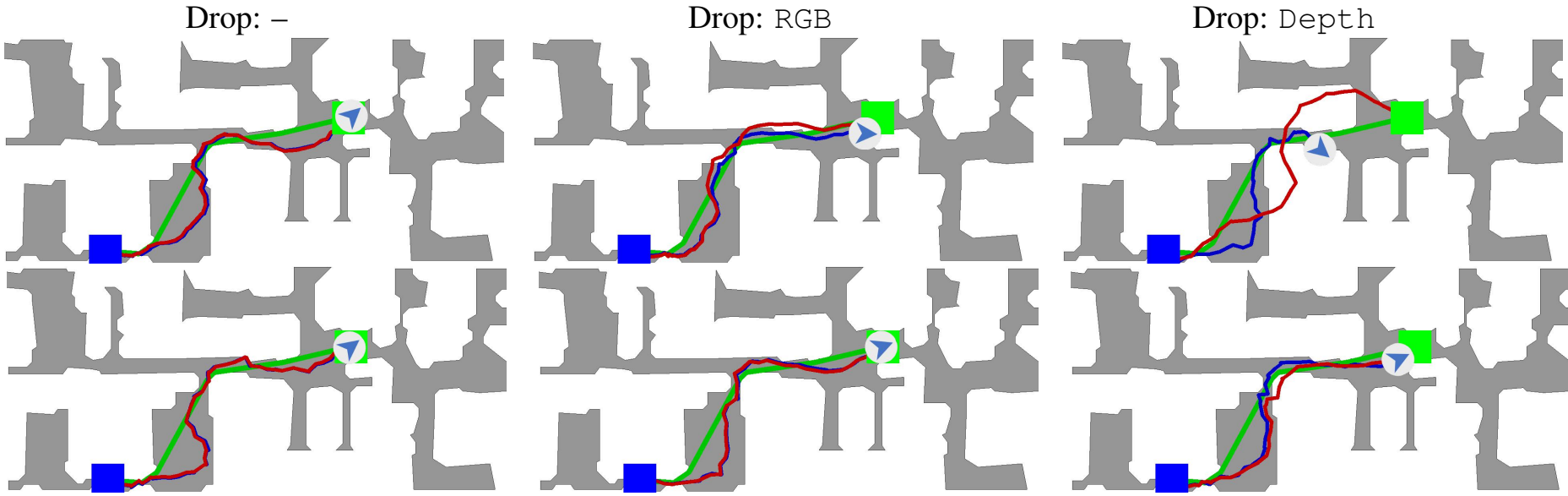
- Transformers are agnostic to the number of input tokens/modalities
- Action token &multi-modal pre-training (MultiMAE [1])
- Explicit modality-invariance training






[1] MultiMAE: Multi-modal Multi-task Masked Autoencoders. Bachmann et al. 2022.

# Dropping modalities during deployment



# Explicit modality-invariance training

| Method | Drop | $S \uparrow$ | SPL↑ | SSPL↑ | $d_g \downarrow$ |
|---|---|---|---|---|---|
| VOT `RGB` | – | 59.3 | 45.4 | 66.7 | 66.2 |
| VOT `Depth` | – | 93.3 | 71.7 | 72.0 | 38.0 |
| [12] | – | 64.5 | 48.9 | 65.4 | 85.3 |
| VOT | – | 88.2 | 67.9 | 71.3 | 42.1 |
| VOT w/ *inv.* | – | **92.6** | **70.6** | **71.3** | **40.7** |
| [12] | `RGB` | 0.0 | 0.0 | 5.4 | 398.7 |
| VOT | `RGB` | 75.9 | 58.5 | 69.9 | 59.5 |
| VOT w/ *inv.* | `RGB` | **91.0** | **69.4** | **71.2** | **37.0** |
| [12] | `Depth` | 0.0 | 0.0 | 5.4 | 398.7 |
| VOT | `Depth` | 26.1 | 20.0 | 58.7 | 148.1 |
| VOT w/ *inv.* | `Depth` | **60.9** | **47.2** | **67.7** | **72.1** |

| Rank | Participant team | S | SPL | SSPL | $d_g$ |
|---|---|---|---|---|---|
| 1 | **MultiModalVO (VOT)** (ours) | 93 | 74 | 77 | 21 |
| 2 | VO for Realistic PointGoal [35] | 94 | 74 | 76 | 21 |
| 3 | inspir.ai robotics | 91 | 70 | 71 | 70 |
| 4 | VO2021 [64] | 78 | 59 | 69 | 53 |
| 5 | Differentiable SLAM-net [24] | 65 | 47 | 60 | 174 |

- SOTA on Habitat PointNav Challenge [1] with only 5% of the training data!



[1] Are We Making Real Progress in Simulated Environments? Measuring the Sim2Real Gap in Embodied Visual Navigation. Kadian et al. 2019.

# Attention Maps



- VOT attends to relevant regions in the image

# Modality-invariant Visual Odometry for Embodied Vision

Marius Memmel, Roman Bachmann, Amir Zamir

- **Versatile multi-modal** odometry framework that can deal with *'optional'* modalities
- Dropping modalities during training leads to **modality-invariance** during test time
- **Action prior** and **multi-modal pre-training** drastically reduce data requirements

vo-transformer.github.io

JUNE 18–22, 2023
CVPR
VANCOUVER, CANADA