Modality-invariant Visual Odometry for Embodied Vision

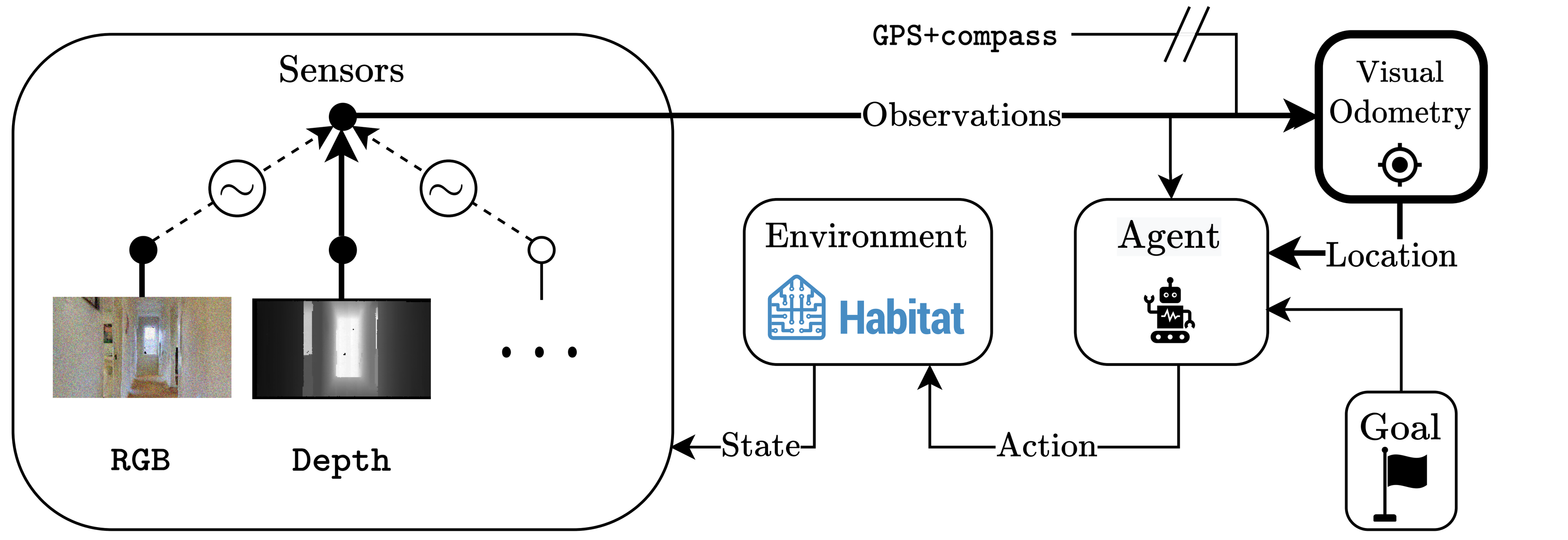
Marius Memmel, Roman Bachmann, Amir Zamir

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

vo-transformer.github.io

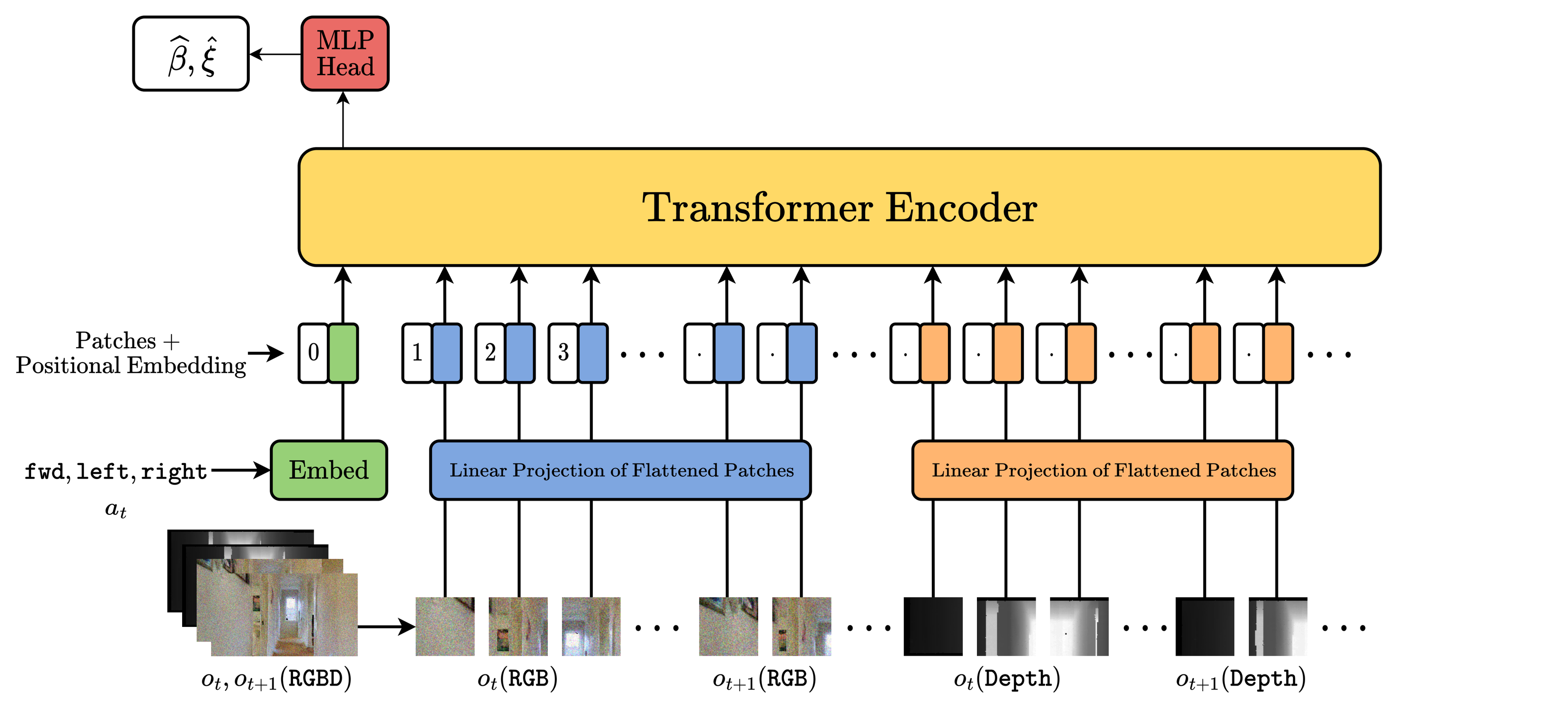
Motivation

- ► We can switch between modalities to localize ourselves. Odometry should too!
- ► Sensors fail, change or are intentionally looped out causing Visual Odometry methods to fail!



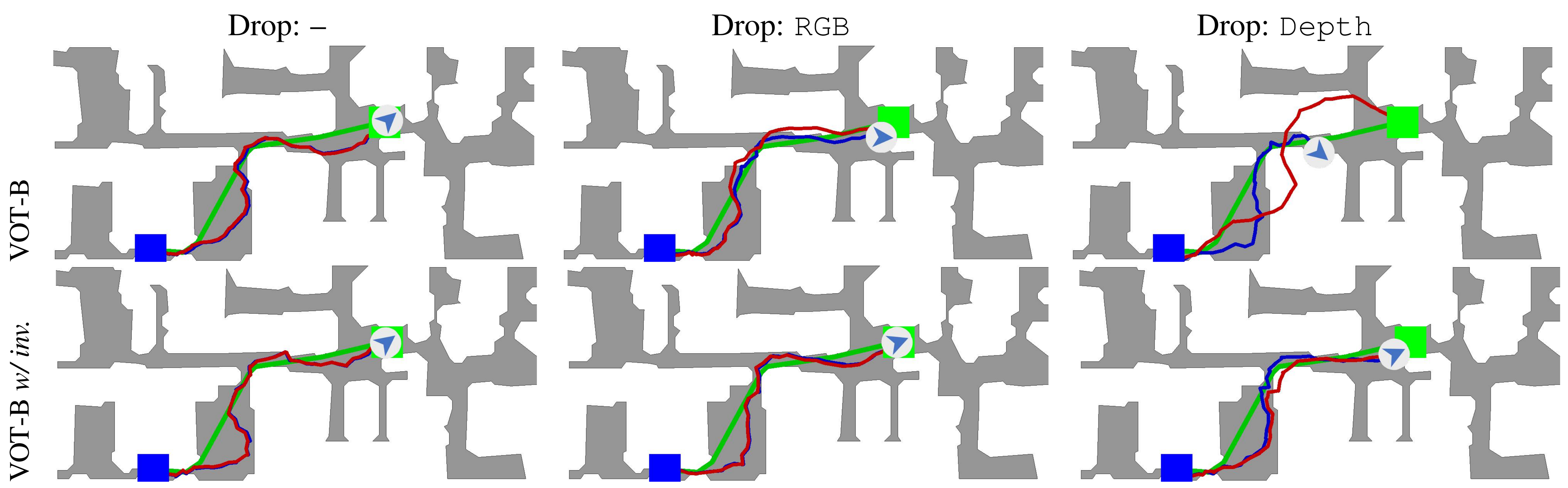
We propose a **modality-agnostic framework** based on the **Vision Transformer** [3] architecture that **deals with optional modalities** without sacrificing performance.

Visual Odometry Transformer



- ► Transformer architecture → agnostic to the number of input tokens and number of modalities
- ► Condition Transformer with action token & MultiMAE [4] pre-training → Reduce data requirements to 5% of previous architectures!
- ► Dropping modalities during training → Explicitly prepare the architecture for test-time modality invariance

Navigation performance under **missing modalities**



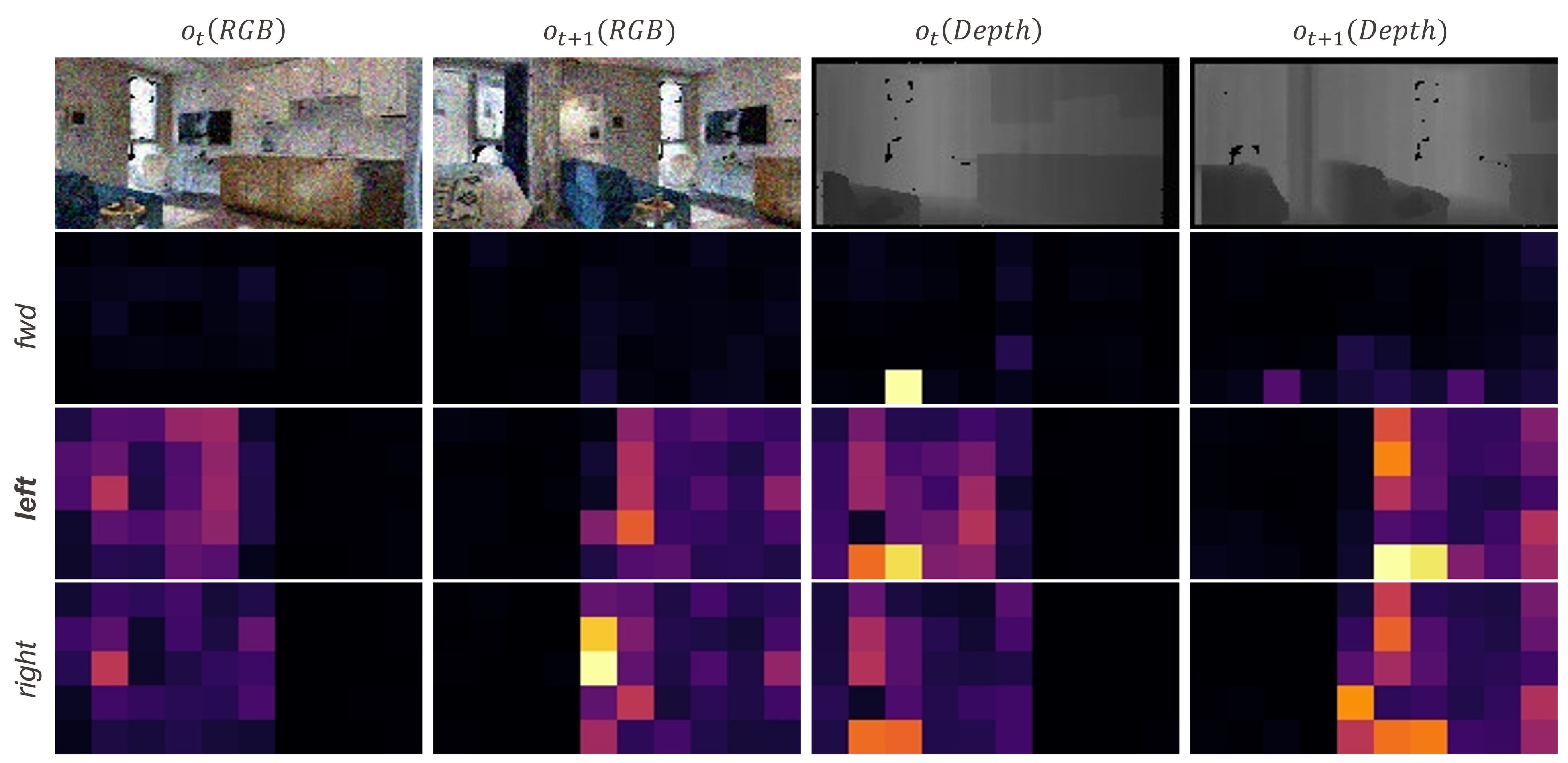
Top-down map of the agent navigating from **start** to **goal**. The plot shows the **shortest path**, the **path taken by the agent**, and the **"imaginary" path the agent took**, i.e., its VO estimate.

w/o explicit invariance training (VOT-B): agent heavily **relies on both modalities** (RGB < Depth) and **fails catastrophically** if either is unavailable (Drop: RGB, or Drop: Depth)

w/ explicit invariance training (VOT-B *w/ inv.*): agent **succeeds** even **when modalities are missing**!

Attention maps

- ► Action token serves as powerful prior
- ► VOT attends to relevant image regions
- ► Action token resolves ambiguities caused by noise and collision



Quantitative results under missing modalities

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
[1]	–	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**
[1]	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**
[1]	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**

- ► ConvNet-based architecture [1,2] can't deal with optional modalities
- ► Explicit invariance training performs on par with single modality model when modalities are dropped
- ► Depth is more informative than RGB for the VO task

Habitat challenge

Highest SSPL training on **only 5% of the data** on Habitat Challenge 2021.

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

Takeaways

- ► VOT is a versatile multi-modal Odometry framework
- ► Dropping modalities during training helps dealing with missing modalities during test time
- ► Action prior and multi-modal pre-training drastically reduce data requirements

References:
[1] Integrating Egocentric Localization for More Realistic Point-Goal Navigation Agents. *Datta et al.* CoRL 2021
[2] 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.* ECCV 2022