



# What does Embodied Intelligence mean?

## Lessons Learned from Drone Racing

*Antonio Loquercio*

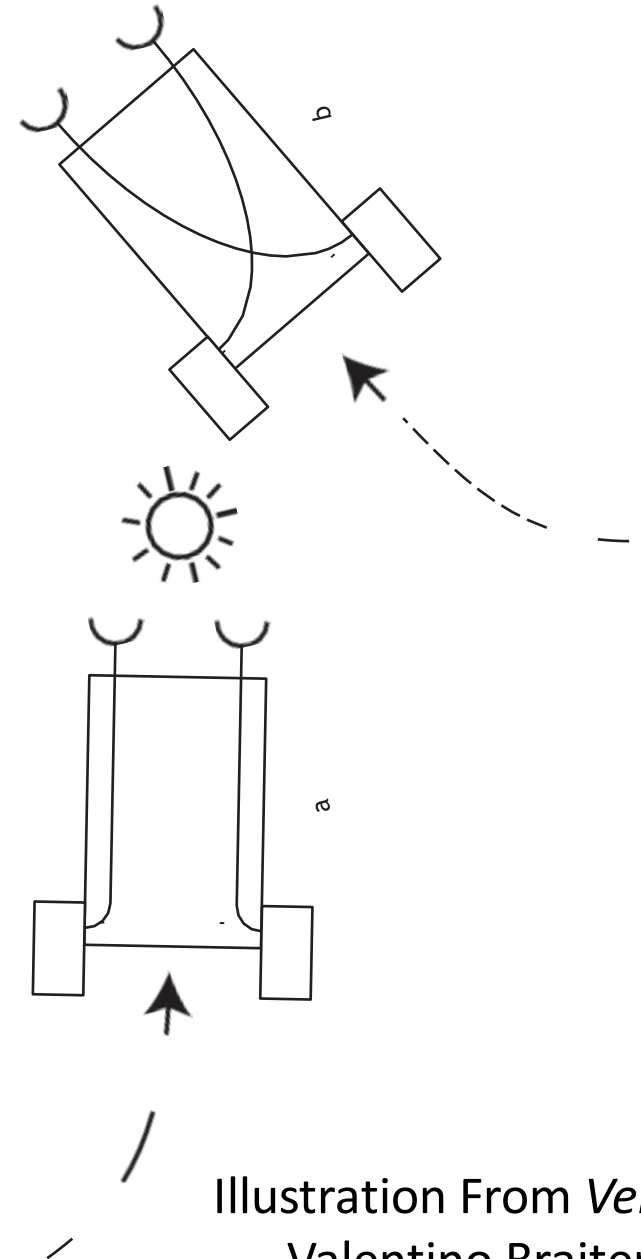
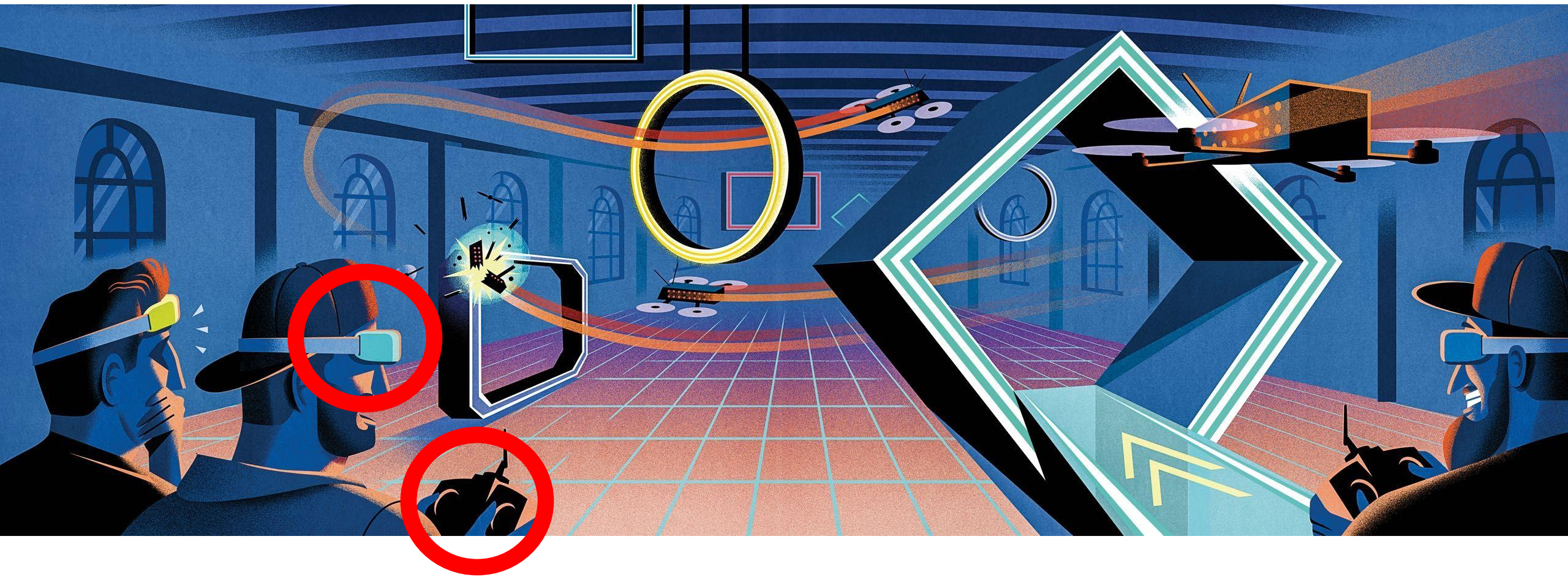


Illustration From *Vehicles* by  
Valentino Braitenberg







# World Championship Qualifiers



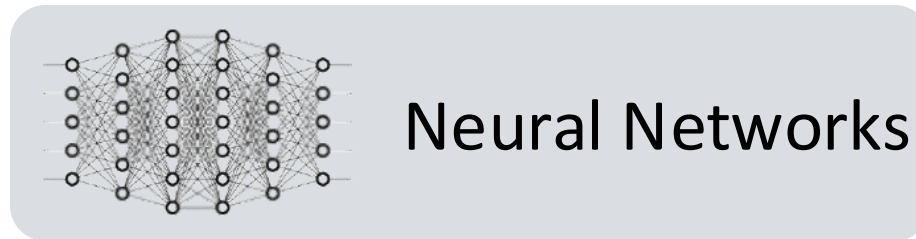
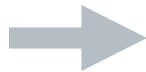
	Name	3 laps (seconds)
1	MinChan 'MCKFPV' Kim	27.057
2	Konstantin 'KostaFPV' Sonnentag	28.771
3	Levi 'Leviathann' Johnson	29.229056
4	Silas 'Propsicle' Aaron	29.329408
5	Marvin 'MARV_FPV' Schäpper	29.748
6	Mason 'Hyper' Lively	29.81888
7	Jacob 'JakeHammer' Capobres	30.010368
8	Evan 'headsupfpv' Turner	30.019584
9	Ashton 'Drobotracer' Gamble	30.400992
10	Sebastian 'SebaFPV' Espinal	30.44

~2s difference

~1s difference



# Racing is not a good fit for Imitation Learning

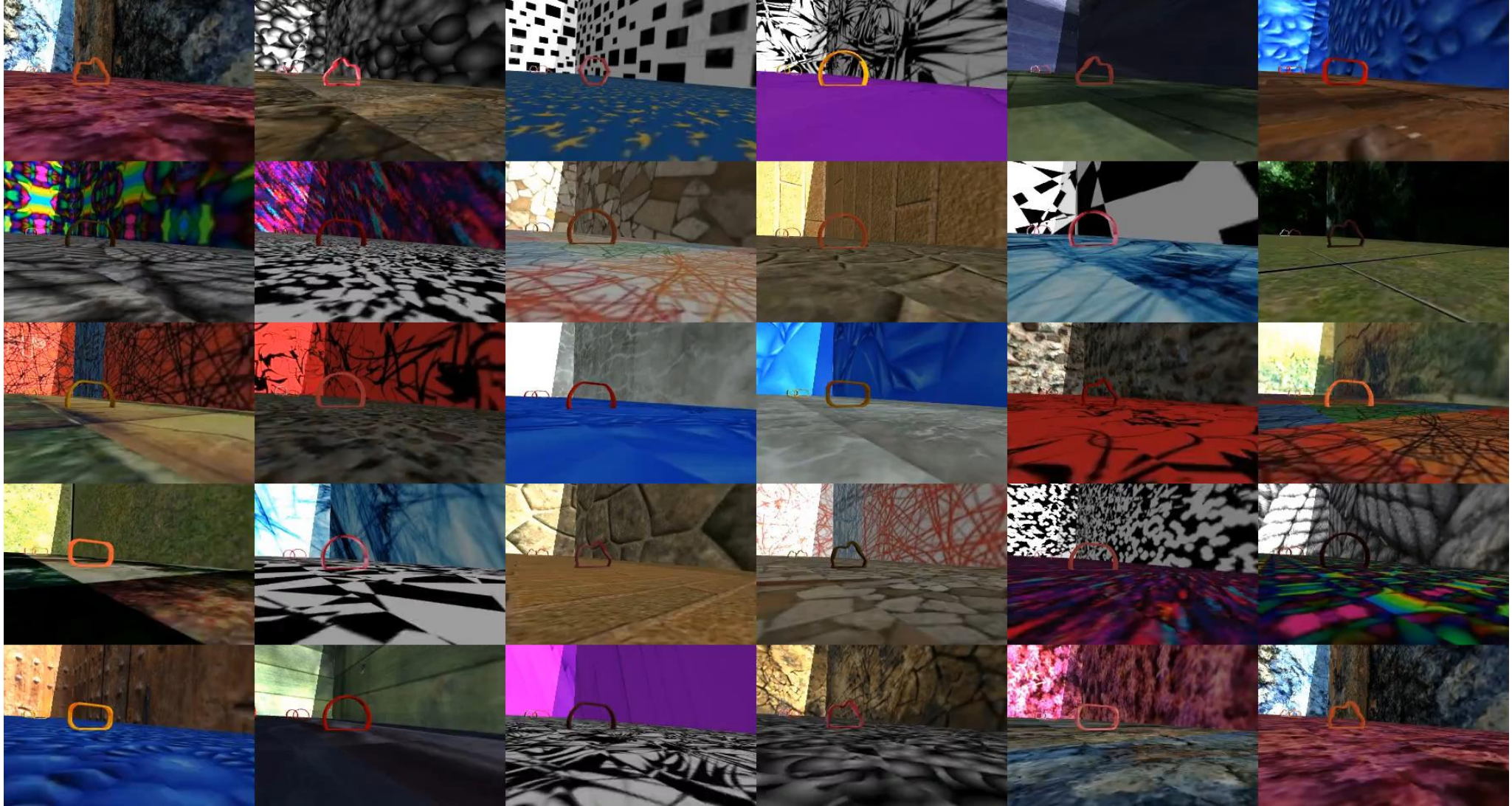


Body Rates  
Thrust



Body Rates  
Thrust

# Learning End-To-End Control For Drone Racing

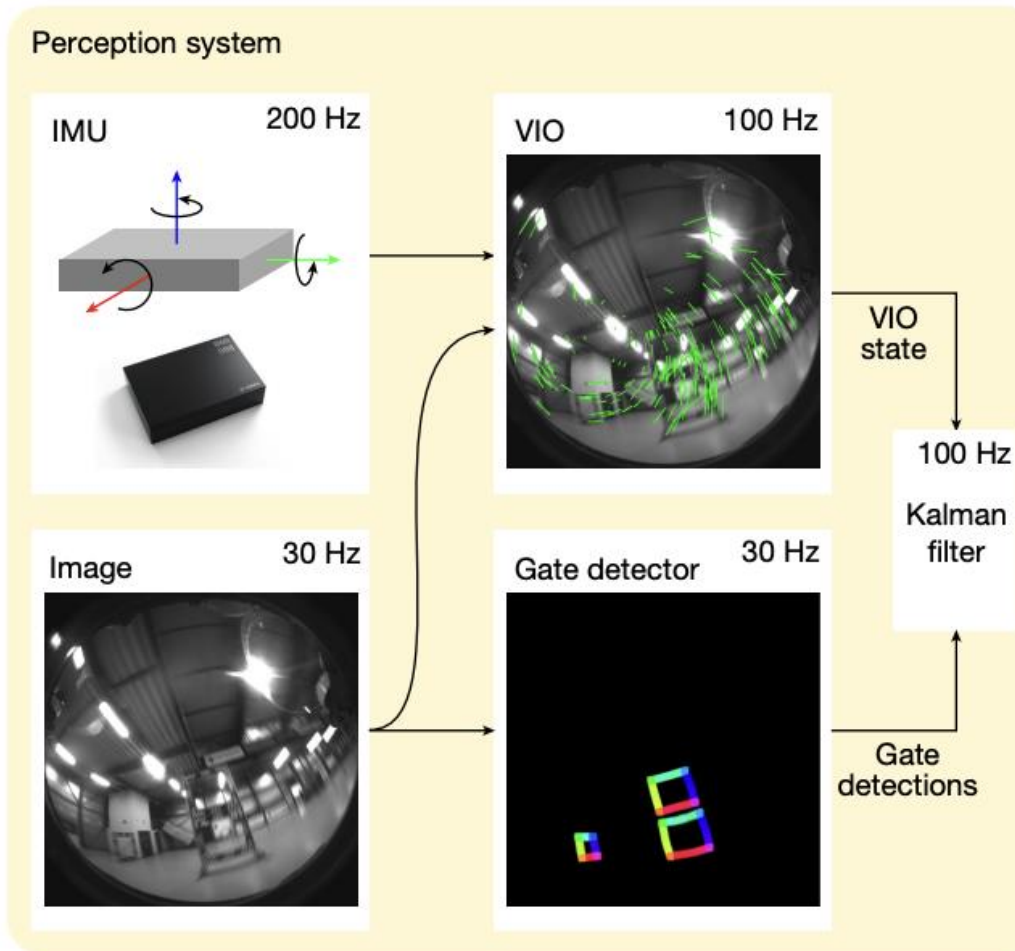


*Deep Drone Racing: From Simulation to the Real World Using Domain Randomization.* Loquercio et al.

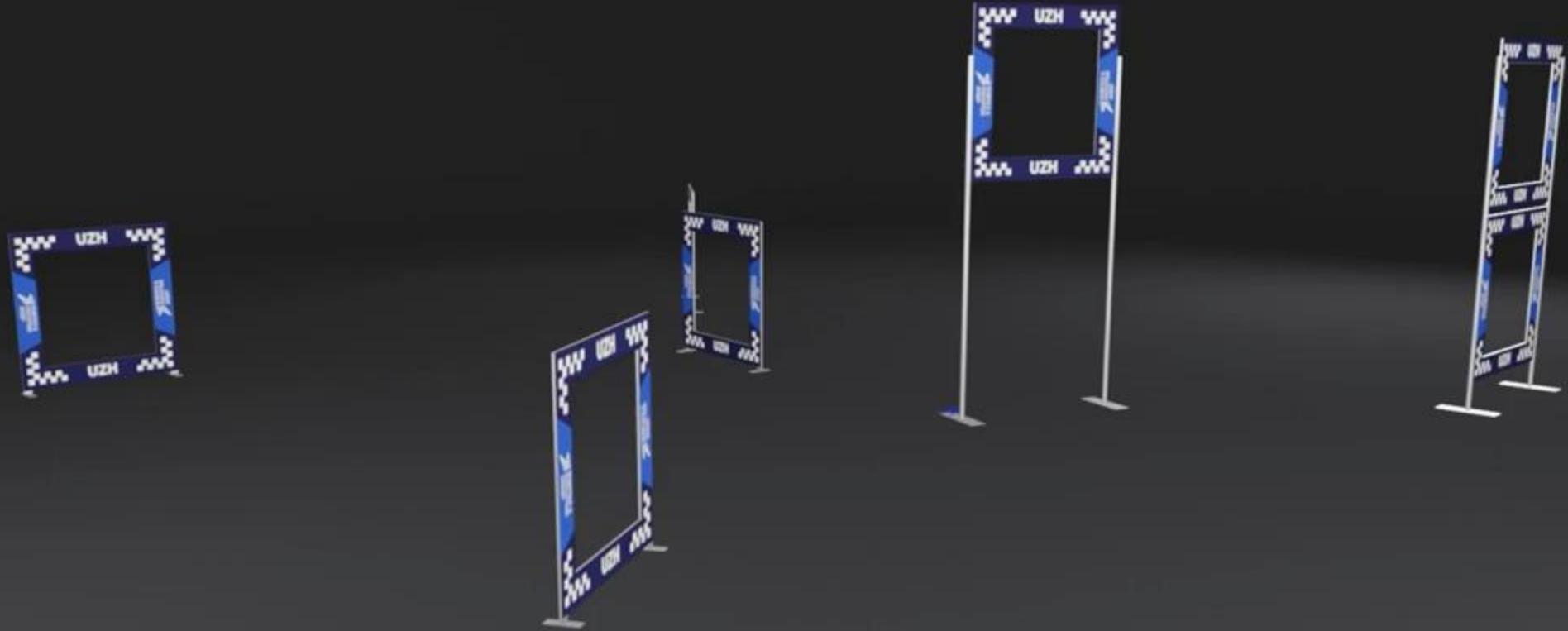
T-RO Best Paper Honorable Mention



# A Modular Approach

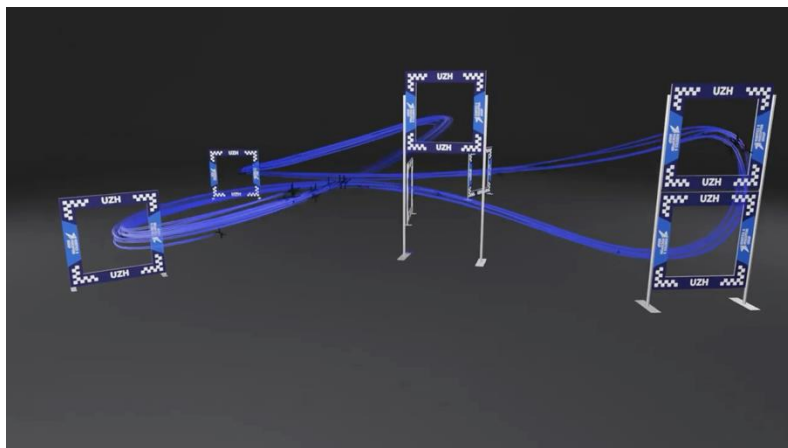


# Training

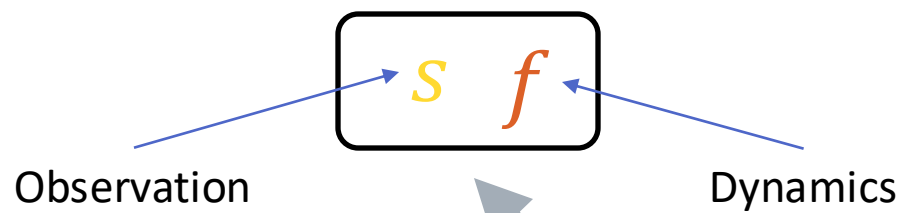
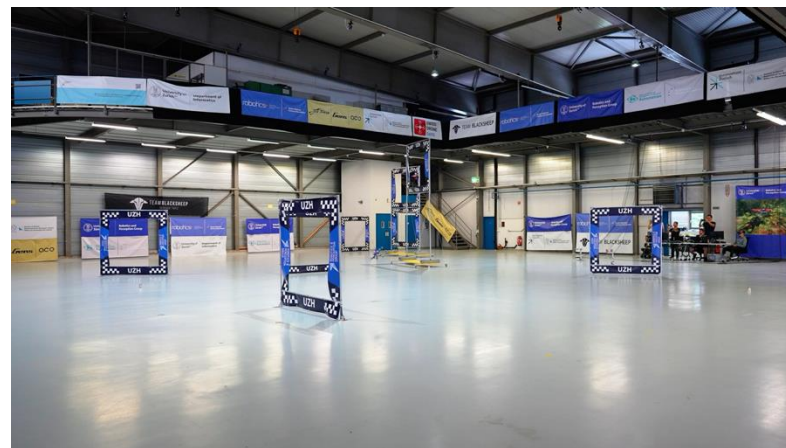




Simulation



Real World



ROBOTICS &  
PERCEPTION  
GROUP

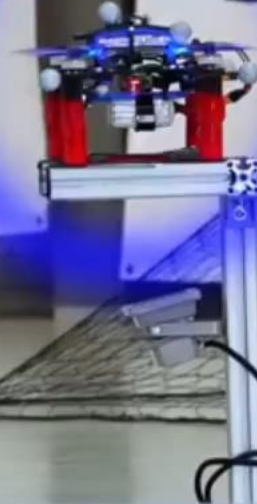


Universität  
Zürich <sup>UZH</sup>

Robotics and  
Perception Group



Canton of Zurich  
Department for Economic Affairs  
Office for Economy and Labour



robotics<sup>+</sup>

Swiss  
Cent  
in R

ROBOTICS  
PERCEPTION  
GROUP



TEA

UZH



# Making the comparison as fair as possible

- The same drone.
- Compensation for human perception latency at the start.

## **But**

- We use an onboard inertial measurement unit (IMU). But our camera updates only at 30Hz (120Hz for humans).
- We have lower latency (40ms vs ~200ms for humans). Unclear if that matters since the environment is predictable.

# Statistics of Racing against Professional Pilots

## Head-to-Head Racing Results

	Number of Races	Best Time-to-Finish	Wins	Losses	Win Ratio
A. Vanover vs. Swift	9	17.956 s	4	5	0.44
T. Bitmatta vs. Swift	7	18.746 s	3	4	0.43
M. Schaepper vs. Swift	9	21.160 s	3	6	0.33
Swift vs Human Pilots	25	<b>17.465 s</b>	15	10	<b>0.60</b>



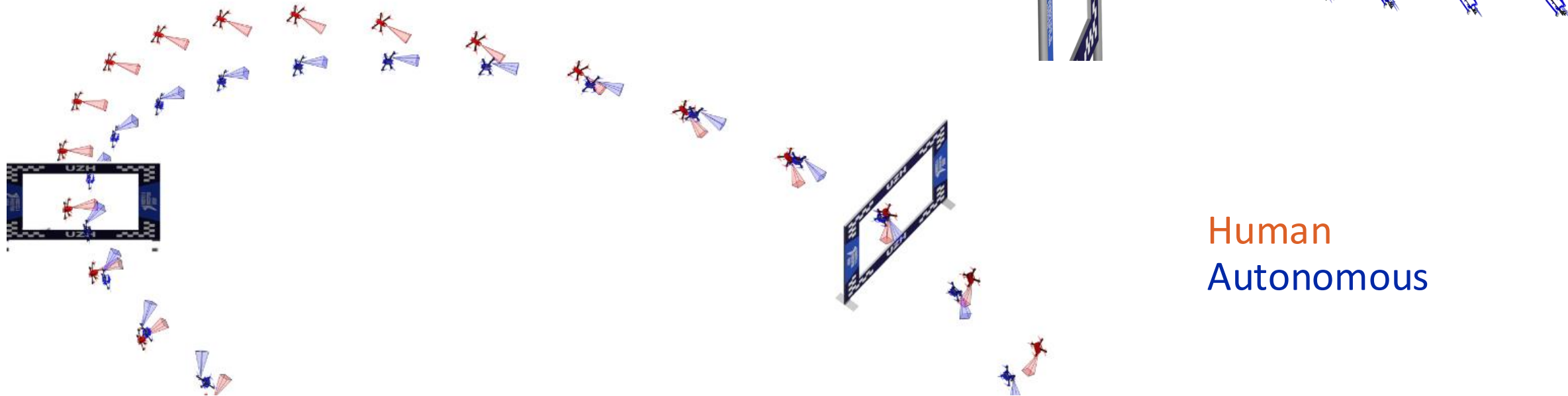
# Differences Human vs. Autonomous

The Autonomous Drone ...

... does not always fly faster

... is faster at the start

... takes a tighter path in difficult maneuvers



The international journal of science / 31 August 2023

# nature

## DRONE RACER

AI pilot beats human  
competitors in real-  
world championship



Vol. 620, No. 7976  
© 2023 Nature.com

Article | [Open access](#) | Published: 30 August 2023

## Champion-level drone racing using deep reinforcement learning

[Elia Kaufmann](#) , [Leonard Bauersfeld](#), [Antonio Loquercio](#), [Matthias Müller](#), [Vladlen Koltun](#) & [Davide Scaramuzza](#)

[Nature](#) **620**, 982–987 (2023) | [Cite this article](#)

# The Human Champions









Universität Zürich  
Robotics and Perception Group



# My Definition of Embodied Intelligence





# How to get there?

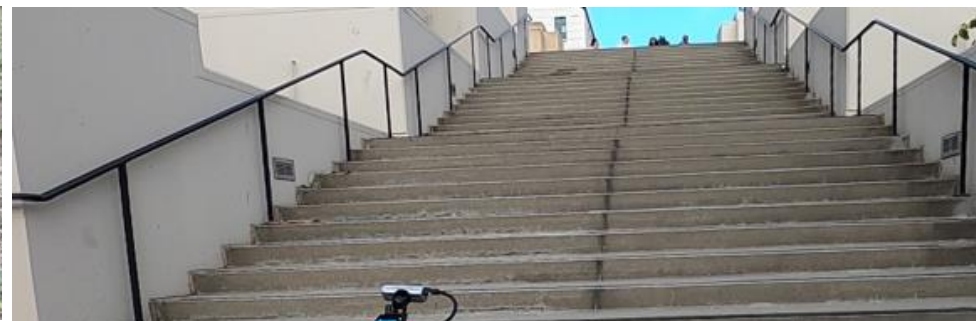
- “Collect a lot of teleoperation data”
- .
- .
- .
- .
- .
- “Tune costs/rewards”

# How to get there?

- “Collect a lot of teleoperation data.”
- .
- .
- “Learn to predict the world.” (akin to self-supervised learning)
- .
- .
- “Tune costs/rewards”

# Learning Visual Locomotion with Cross-Modal Supervision

Loquercio A., Kumar A., Malik J.





# Previous Work on Vision-Based Locomotion

## LEARNING VISION-GUIDED QUADRUPEDAL LOCOMOTION END-TO-END WITH CROSS-MODAL TRANSFORMERS

Ruihan Yang\* Minghao Zhang\*  
UC San Diego Tsinghua University

We propose to address quadrupedal learning (RL) with a Transformer-based information and high-dimensional comotion has made great advance randomization for training blind. Our key insight is that proprioceptive immediate reaction, whereas an agent can learn to proactively maneuver by anticipating changes in the environment. We introduce *LocoTransformer*, an end-to-end learning method in challenging simulated environments. We transfer our learned policy indoors and in the wild with unseen environments. Our method significantly improves over baseline performance, especially when training videos is at <https://richalya.com>

## Learning robust perceptive locomotion for quadrupedal robots in the wild

TAKAHIRO MIKI<sup>1,\*</sup>, JOONHO LEE<sup>1</sup>, JEMIN HWANG<sup>2</sup>, MARCO HUTTER<sup>1</sup>

<sup>1</sup>Robotic Systems Lab, ETH Zurich, Zurich, Switzerland

<sup>2</sup>Robotics and Artificial Intelligence Lab, KAIST, Daejeon, Korea

<sup>3</sup>Intelligent Systems Lab, Intel, Jackson, WY, USA

\* Corresponding author: tamiki@ethz.ch

Compiled January 20, 2022

Legged robots that can operate autonomously in the wild have many opportunities for exploration into under-explored environments. Efficient locomotion: perceiving the terrain before stepping on it, the gait ahead of time to maintain speed and locomotion has remained a grand challenge in robotics. One of the main reasons on which the robot cannot step – or are missing information – can degrade due to difficult lighting, dust, fog, etc. For this reason, the most robust and general solution is to use a camera. However, this severely limits locomotion speed, because the robot has to move accordingly. Here we present a robust and general method for perception for legged locomotion. We leverage an end-to-end learning method and exteroceptive input. The encoder is trained on multiple modalities without resorting to heuristics. The controller was tested in a variety of environments and seasons and completed an hour-long hike in the

## Legged Locomotion in Challenging Terrains using Egocentric Vision

Ananye Agarwal<sup>1</sup>, Ashish Kumar<sup>2</sup>, Jitendra Malik<sup>1,2</sup>, Deepak Pathak<sup>1</sup>

<sup>1</sup>Carnegie Mellon University, <sup>2</sup>UC Berkeley

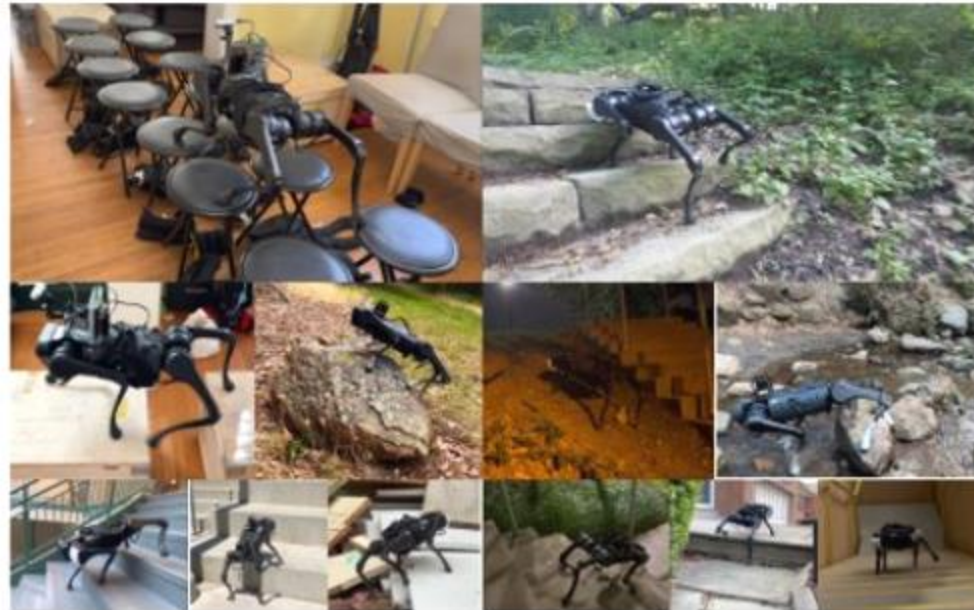
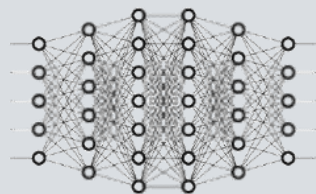
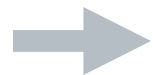


Figure 1: Our robot can traverse a variety of challenging terrain in indoor and outdoor environments, urban and natural settings during day and night using a single front-facing depth camera. The robot can traverse curbs, stairs and moderately rocky terrain. Despite being much smaller than other commonly used legged robots, it is



Neural Networks



**Actions**

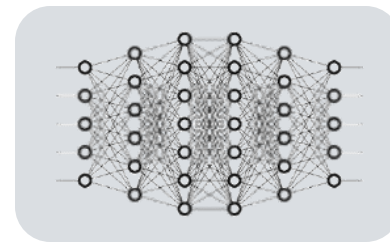








RGB Vision



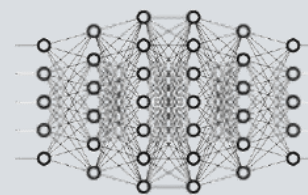
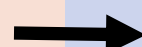
# Real World

# Simulation

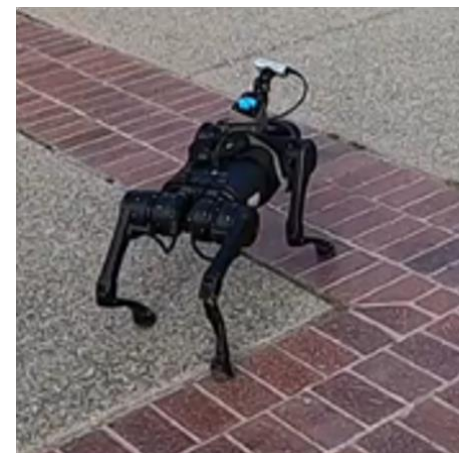
RGB Vision



Terrain  
Properties



Proprio-  
ception



Hwangbo et al., 2019  
Lee et al., 2020  
Kumar et al., 2020



RGB Vision

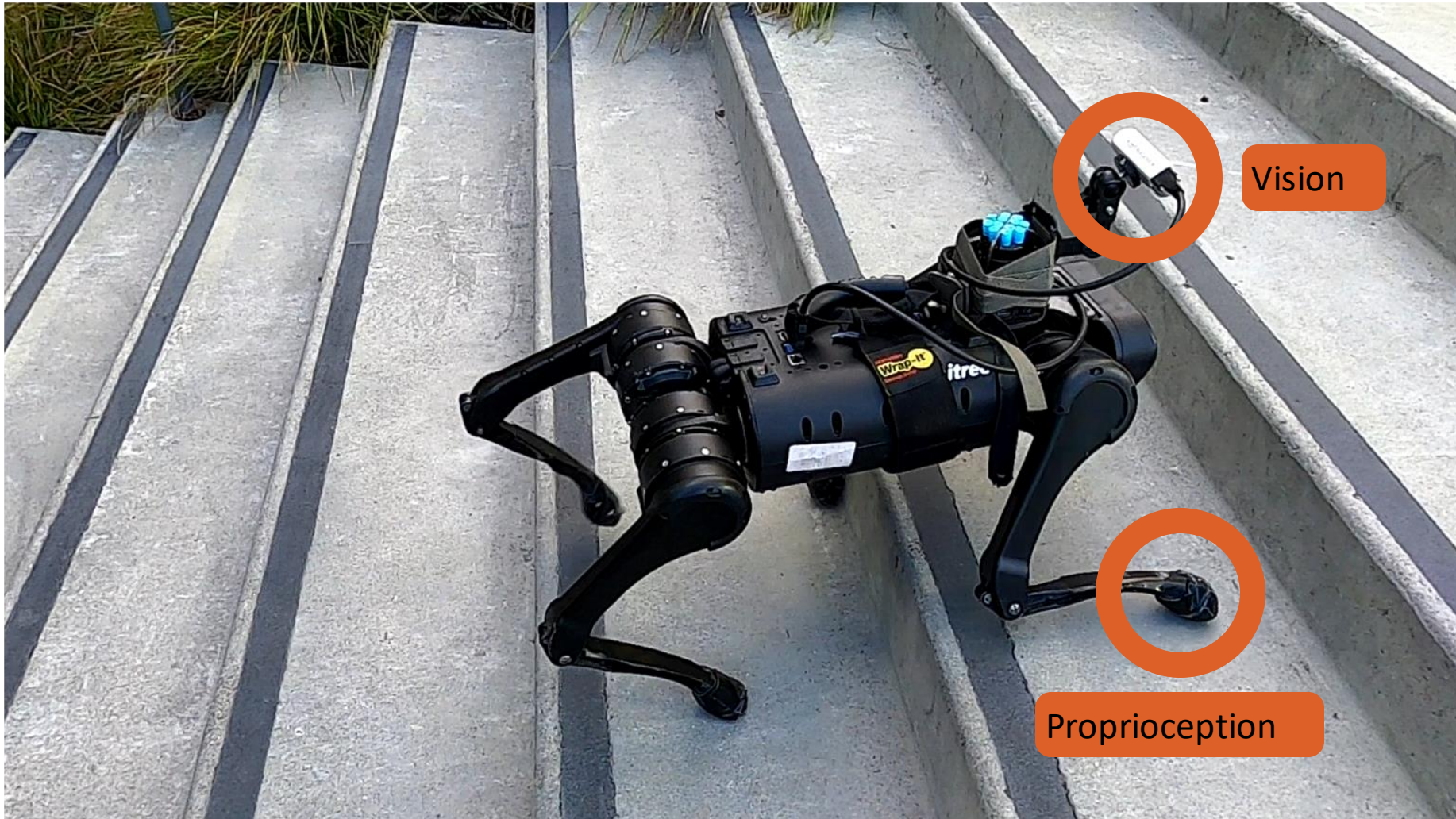


Terrain  
Properties

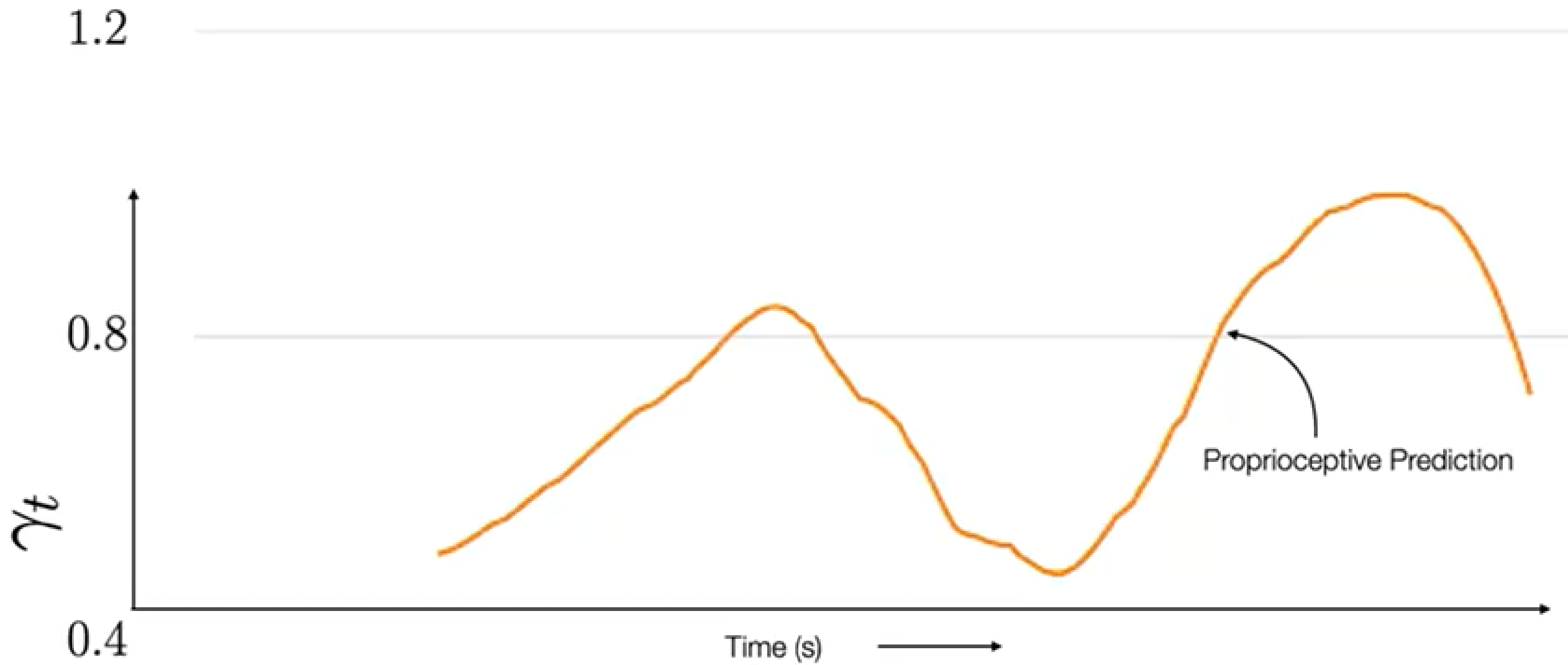
How do we train this estimator?

1. We can't use existing datasets
2. Humans can't provide annotations

# Proprioception to Estimate Terrain Properties



# Cross-Modal Supervision





Blind



# Vision-Based





# Day 1 (2X)









# Discrete Terrain





# Construction Zone







# Visual Plasticity

Before Adaptation



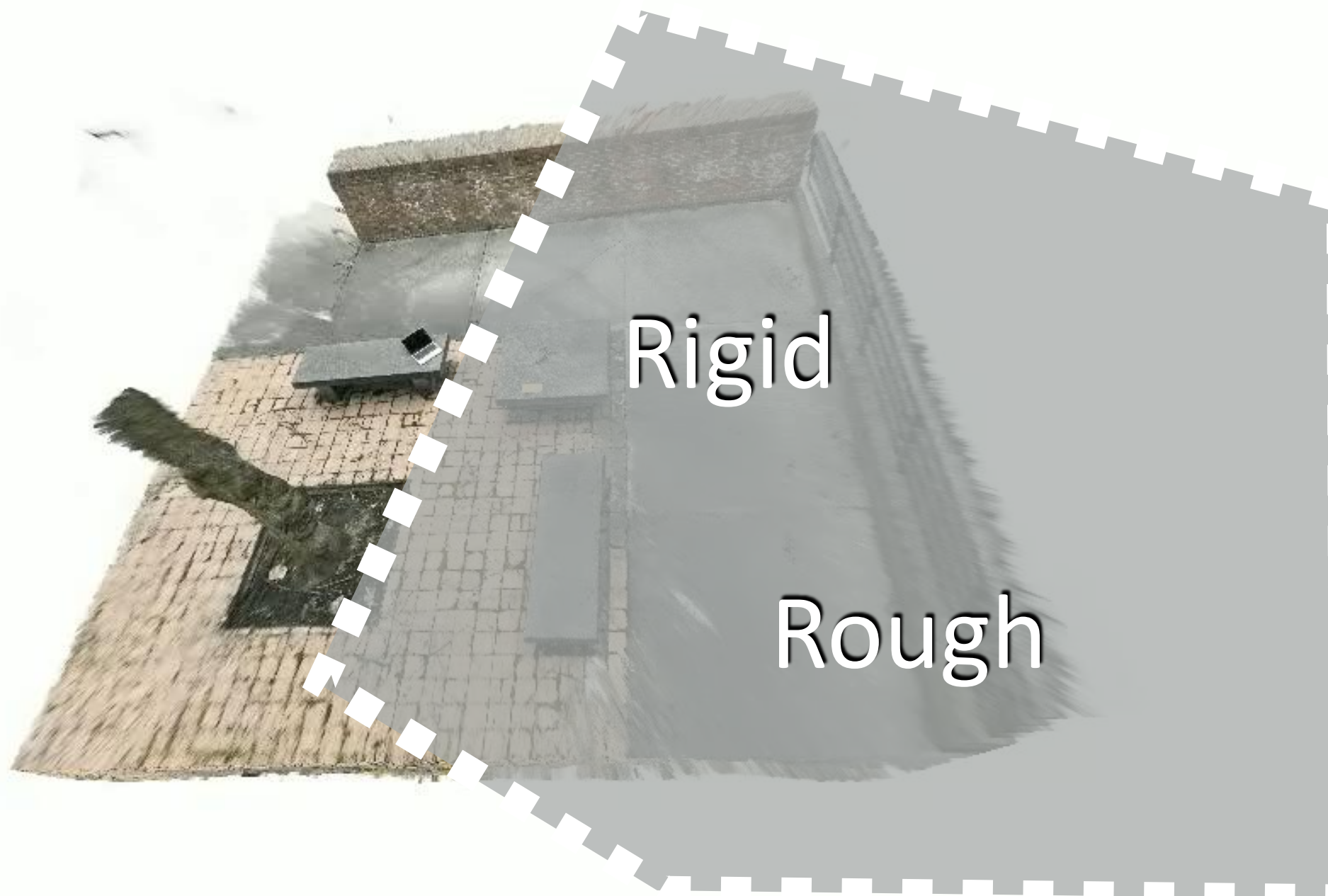
After 1min of data



# Takeaways

- Use a self-supervised loss (predict one sensor from the other) to recover from failures and/or adapt to novel conditions.
- Interaction is a tool to learn about the environment.





Rigid

Rough



Soft

Crumbly





# Hearing Hands:

## Generating Sounds from Physical Interactions in 3D Scenes



Yiming Dou



Wonseok Oh



Yuqing Luo



Antonio Loquercio

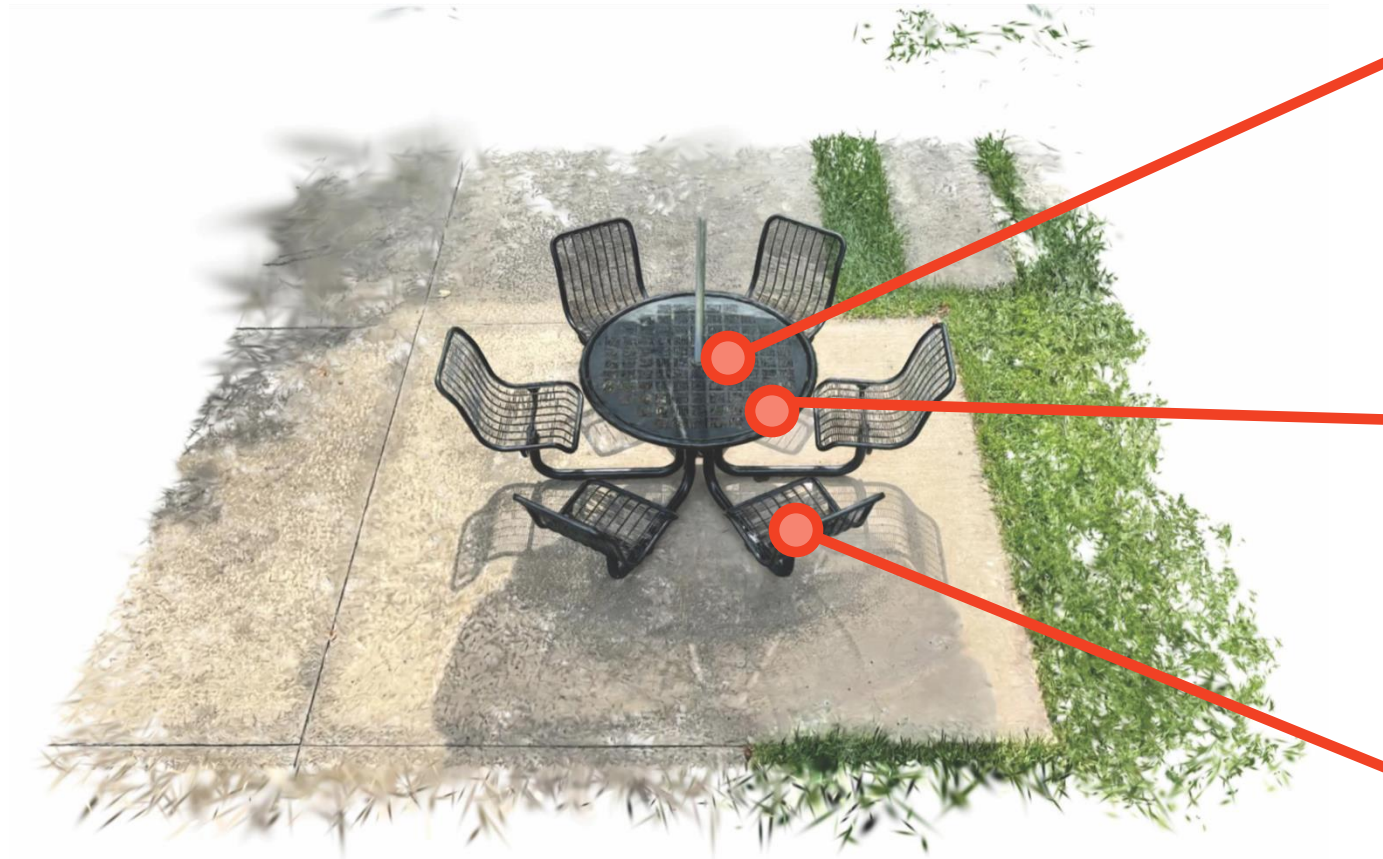


Andrew Owens



Poster #151, Fri 10:30-12:30  
(poster session 1)

# Predicting the sound of actions





# Predicting the Sound of Actions

- **Step 1:** Pick a location to interact with in a 3D scene



# Predicting the Sound of Actions

- **Step 1:** Pick a location to interact with in a 3D scene
- **Step 2:** Record the desired hand motion



# Predicting the Sound of Actions

- **Step 1:** Pick a location to interact with in a 3D scene
- **Step 2:** Record the desired hand motion
- **Step 3:** Generate synthetic interaction sound





# Predicting the Sound of Actions



# Predicting the Sound of Actions

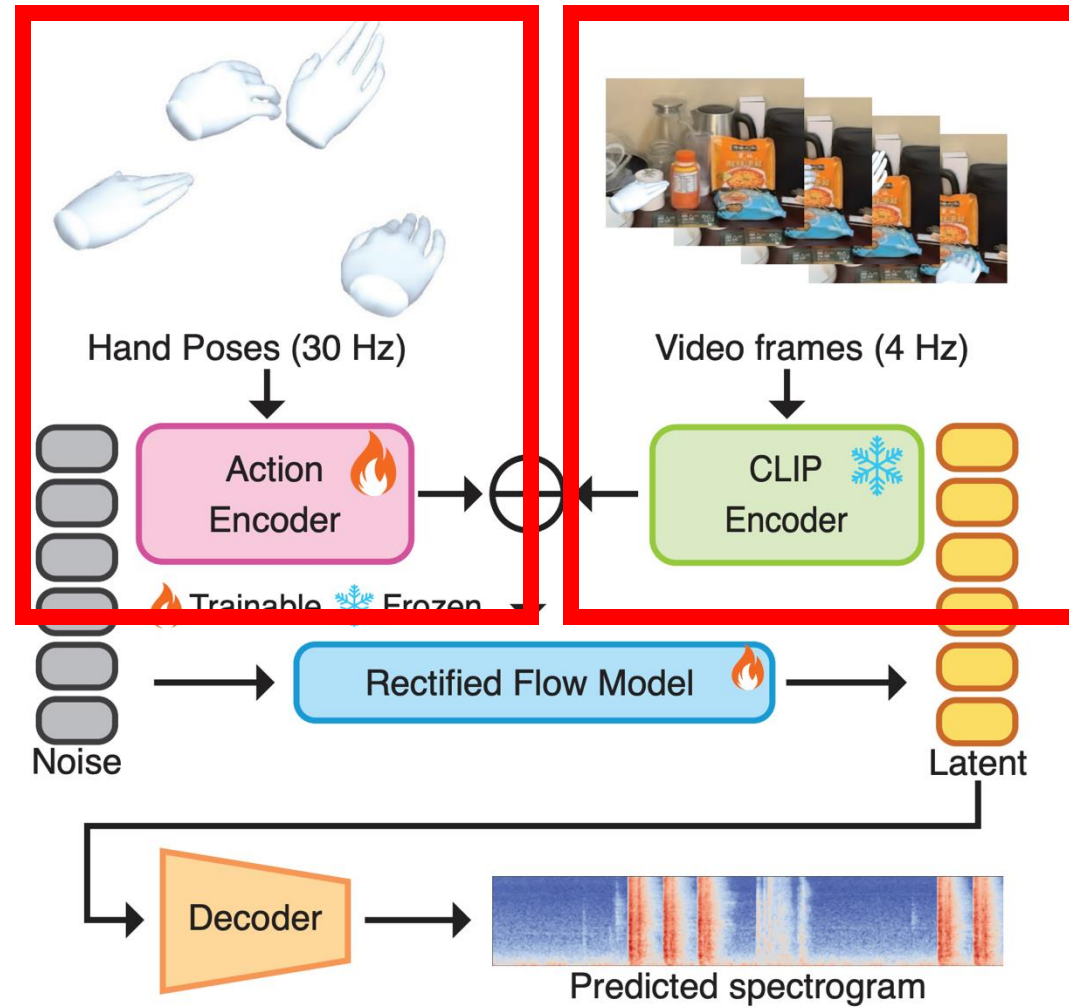


# Predicting the Sound of Actions





# Sound generation model



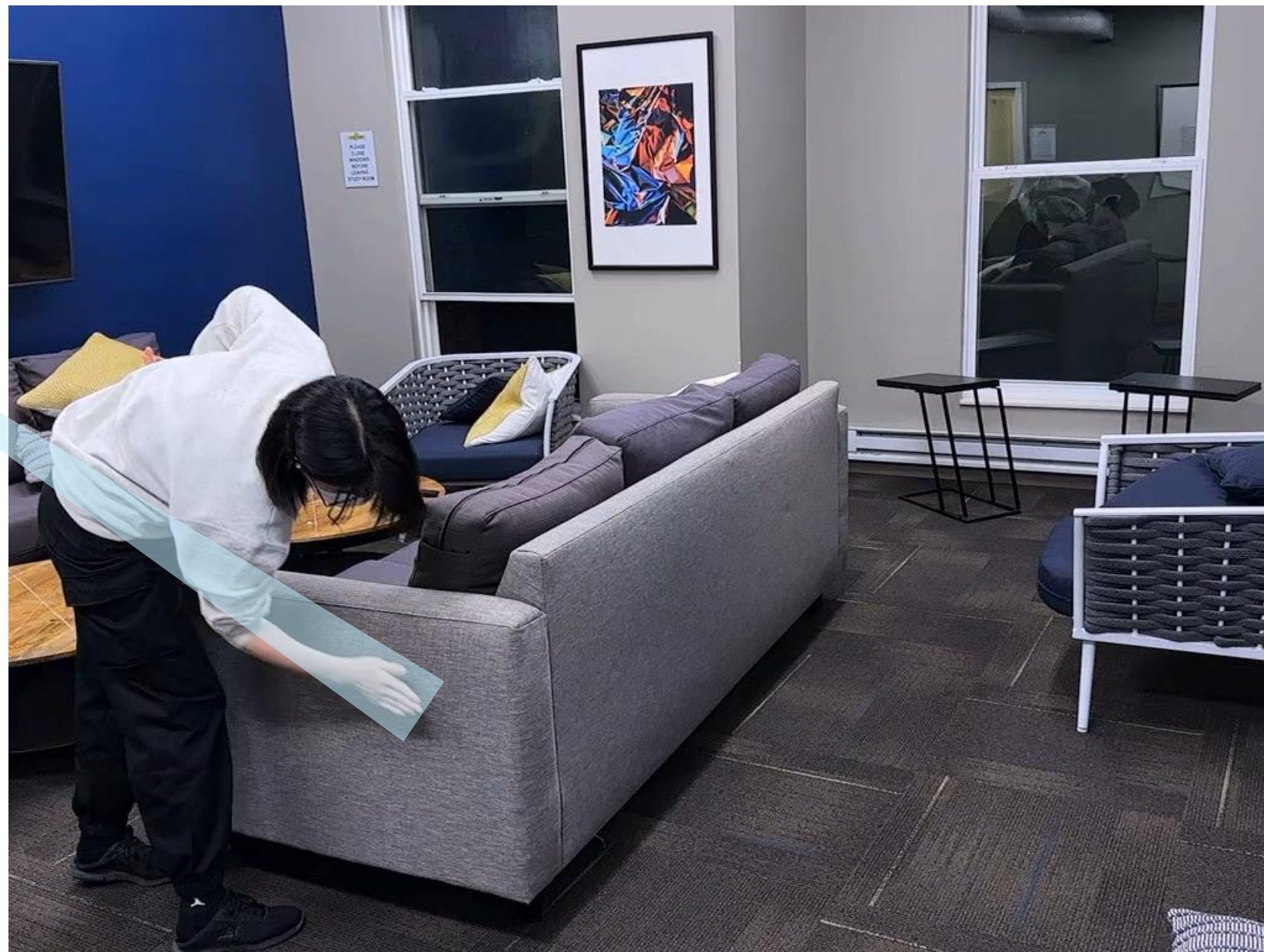
# A Dataset of Hand-Generated Sounds

Register to the  
existing reconstruction



# A Dataset of Hand-Generated Sounds

Register to the  
existing reconstruction





# A Dataset of Hand-Generated Sounds



# A Dataset of Hand-Generated Sounds

Original Video



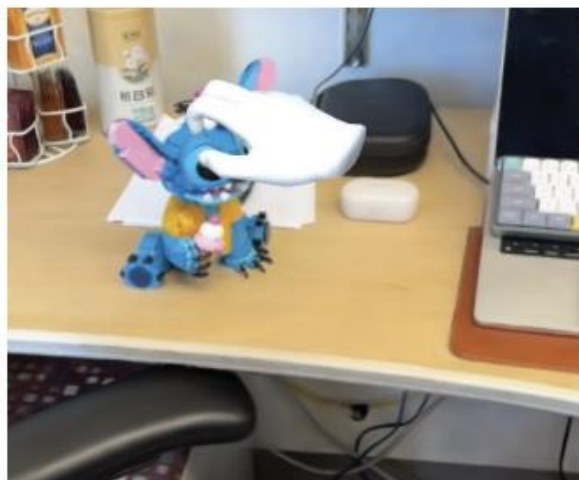
Rendered Video



Rendered Video (top-view)

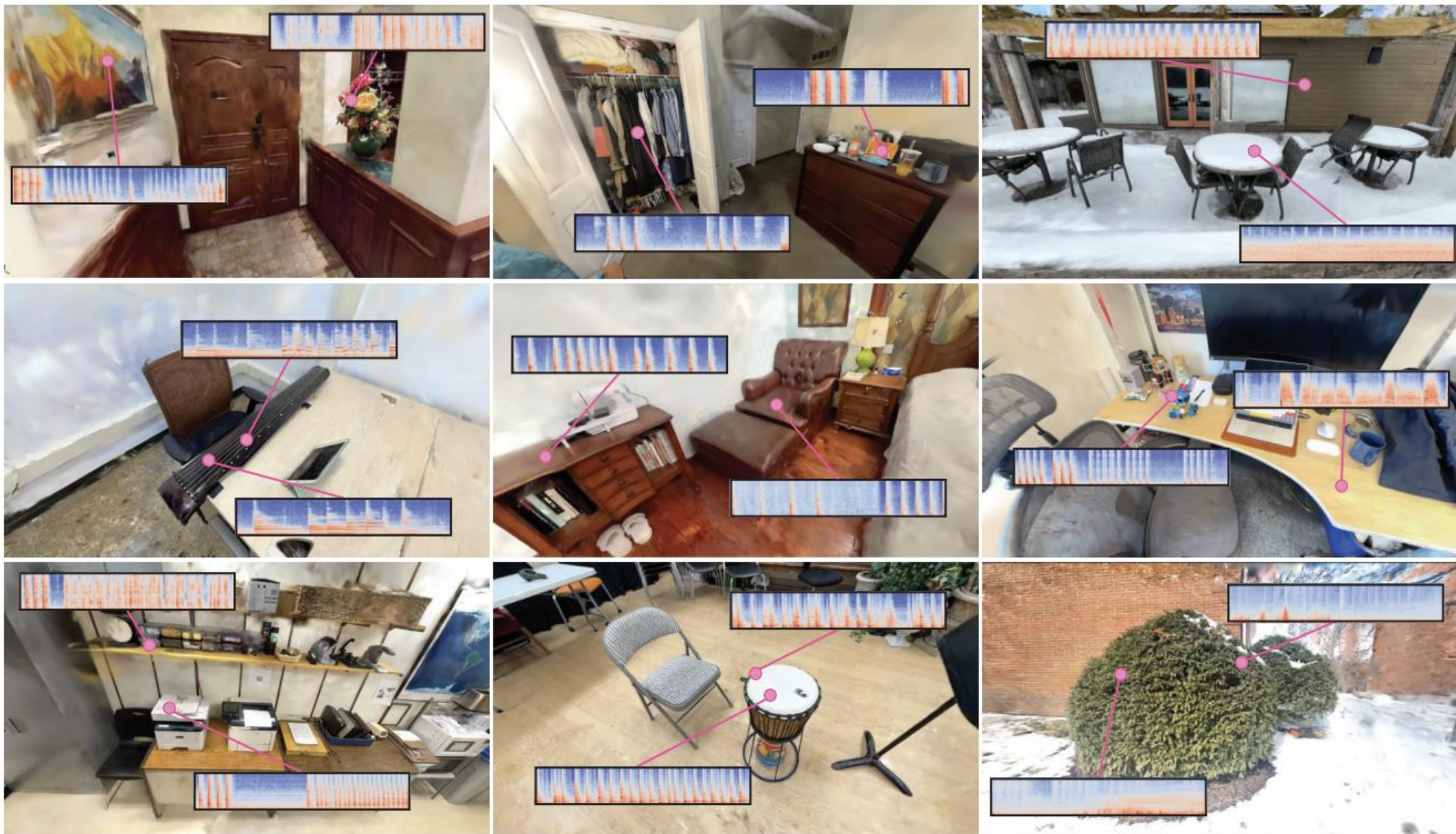


Rendered Video (side-view)





# A Dataset of Hand-Generated Sounds



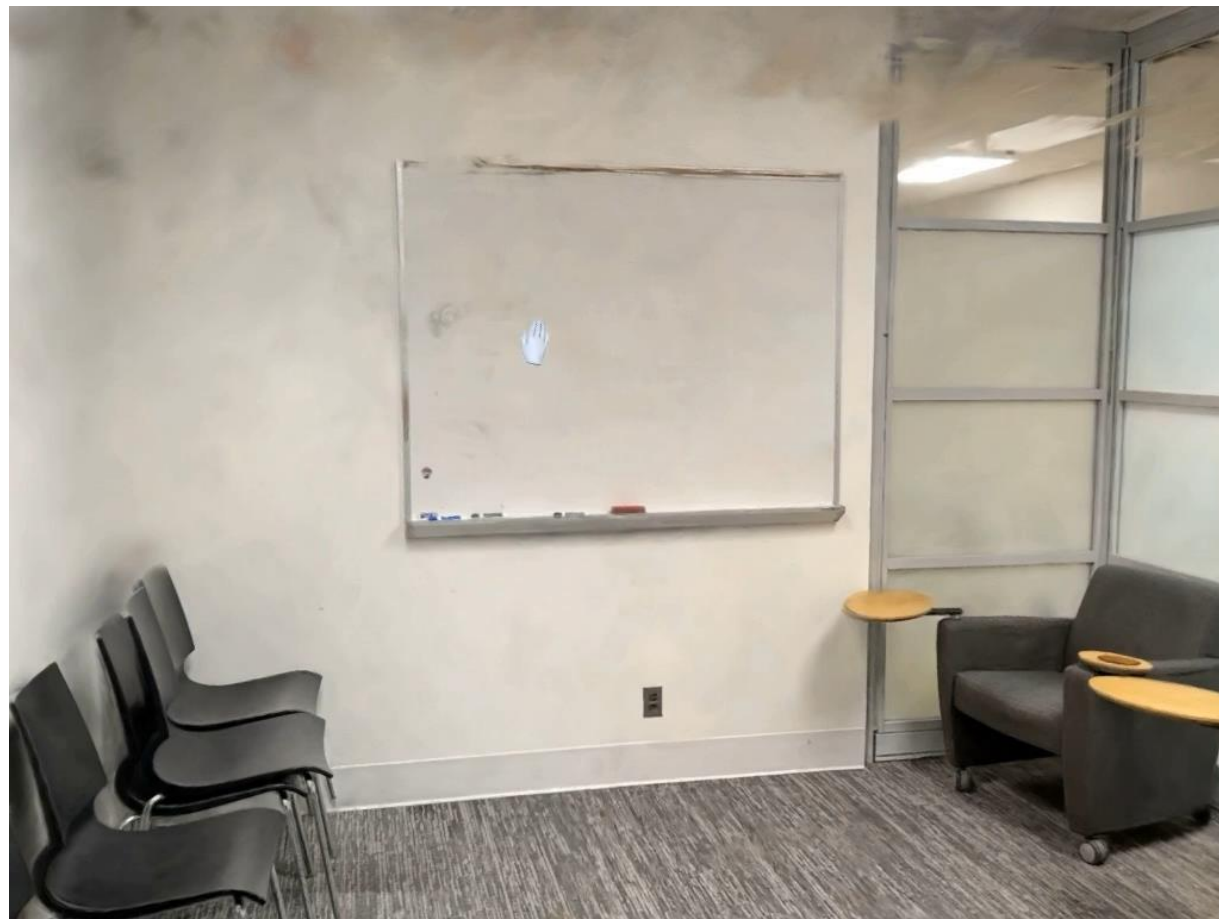


Let's Play a Game

# Which one is generated?



Real



Generated

# Which one is generated?



Real



Generated



# Which one is generated?

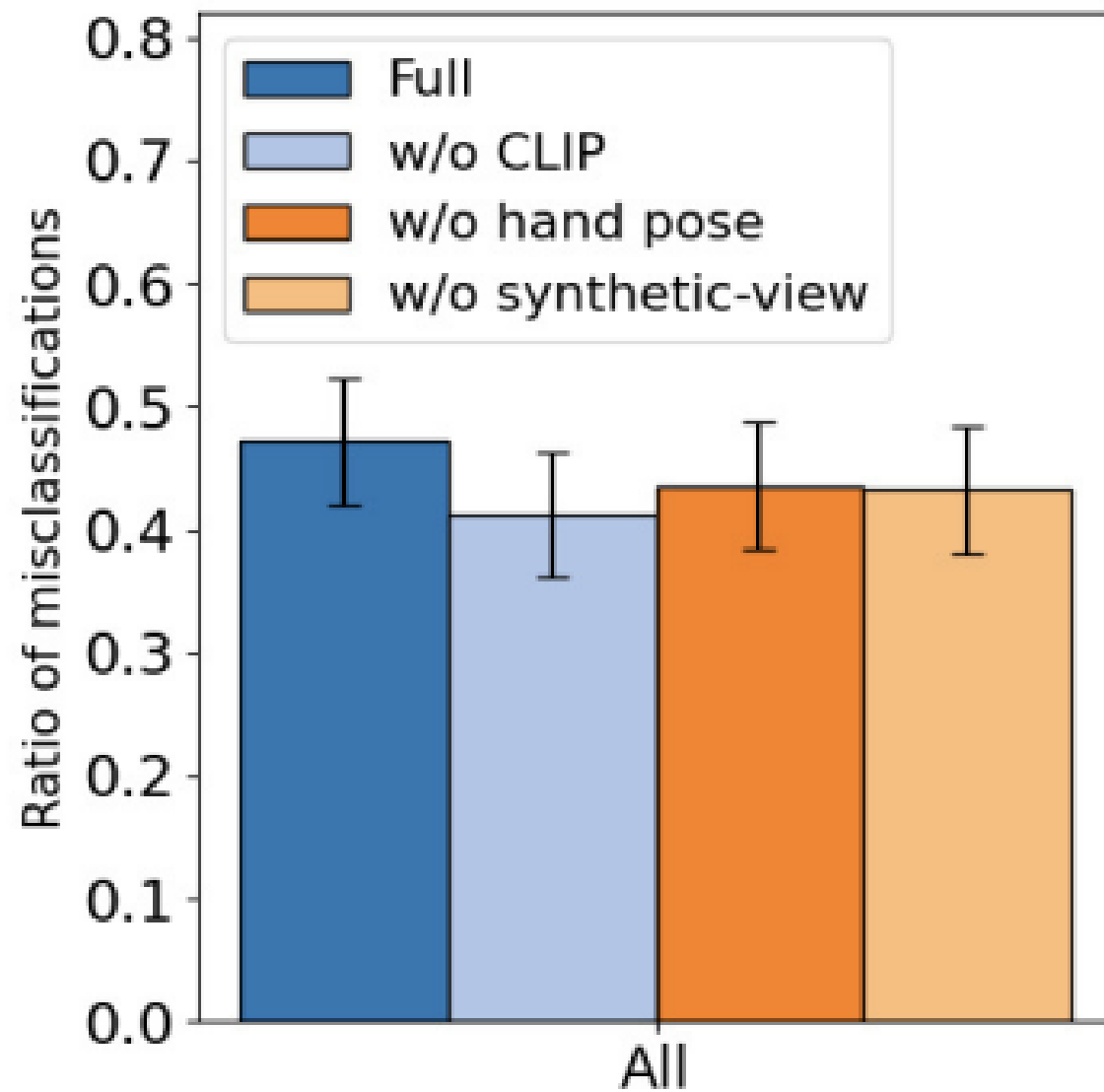


Real



Generated

# User Study



# Human Perception of Sound

## What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception

William W. Gaver  
*Rank Xerox EuroPARC*

Everyday listening is the experience of hearing events in the world rather than sounds per se. In this article, I take an ecological approach to everyday listening to overcome constraints on its study implied by more traditional approaches. In particular, I am concerned with developing a new framework for describing sound in terms of audible source attributes. An examination of the continuum of

Two types of sound perception:

1. Musical Listening
2. Everyday Listening





The Survival Bot

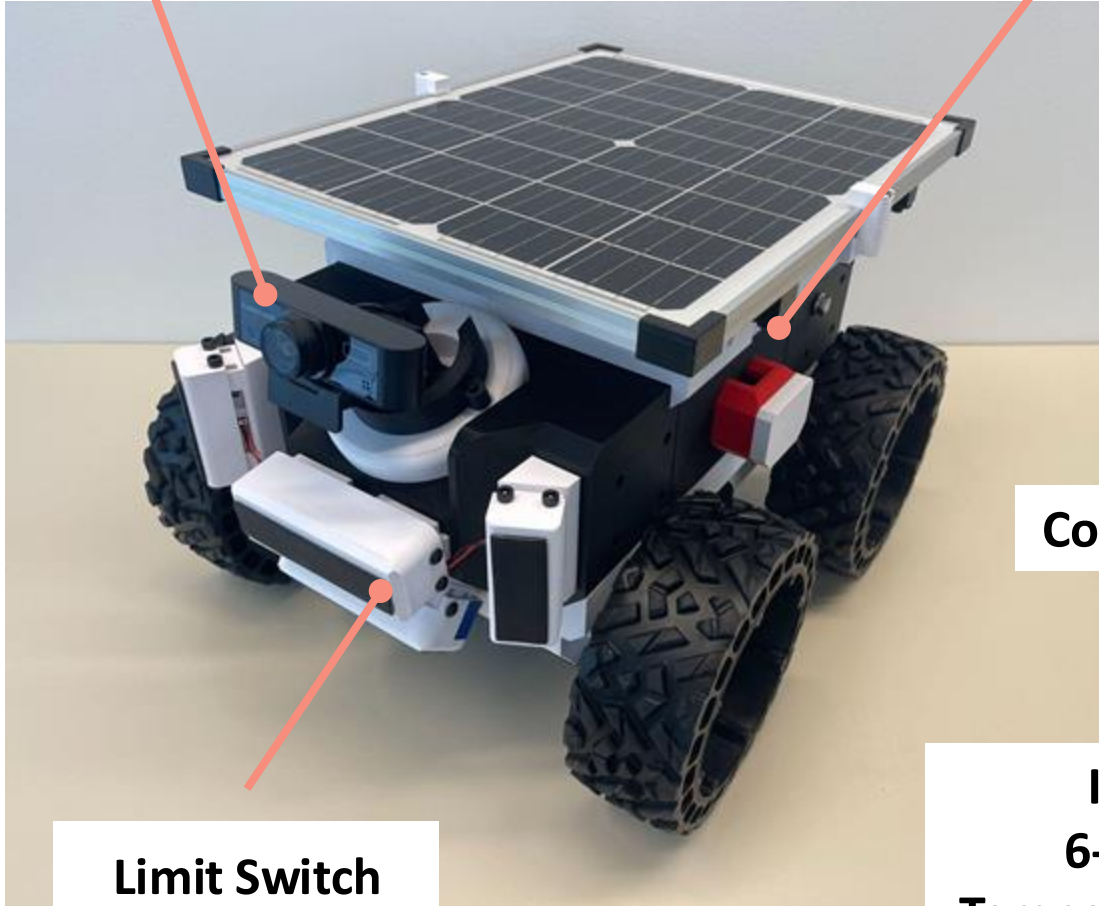


# A Diverse Array of Sensors

Camera

Microphones (x2)

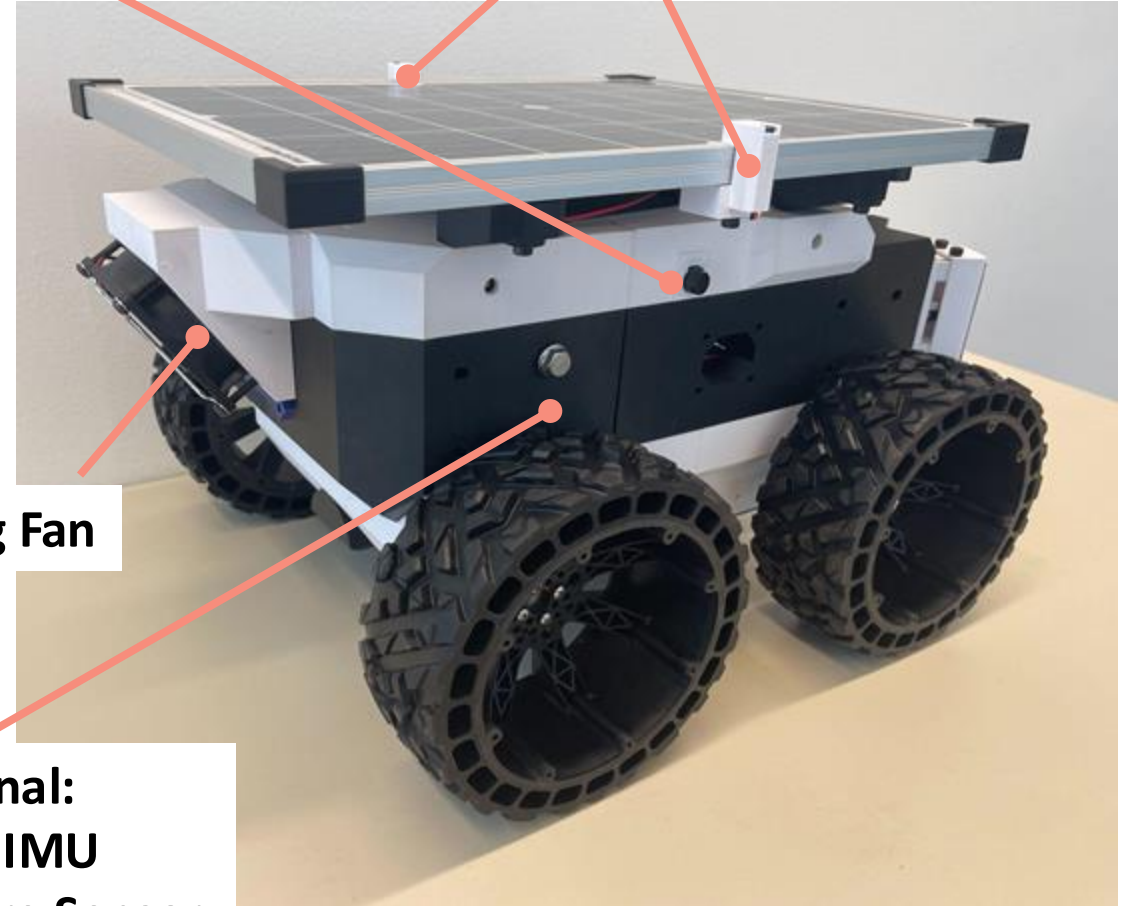
Photoresistors (x2)



Limit Switch  
Bumpers (x4)

Cooling Fan

Internal:  
6-axis IMU  
Temperature Sensor  
Humidity Sensor



# The Beauty of Real World





# Next Steps: Month-Long Learning



# Takeaways

- Embodied intelligence is the ability to deal with novelty, failure, and uncertainty.
- Interaction gives an agent the opportunity to learn about themselves and the environment.
- Get out of the lab!



Thank you!

