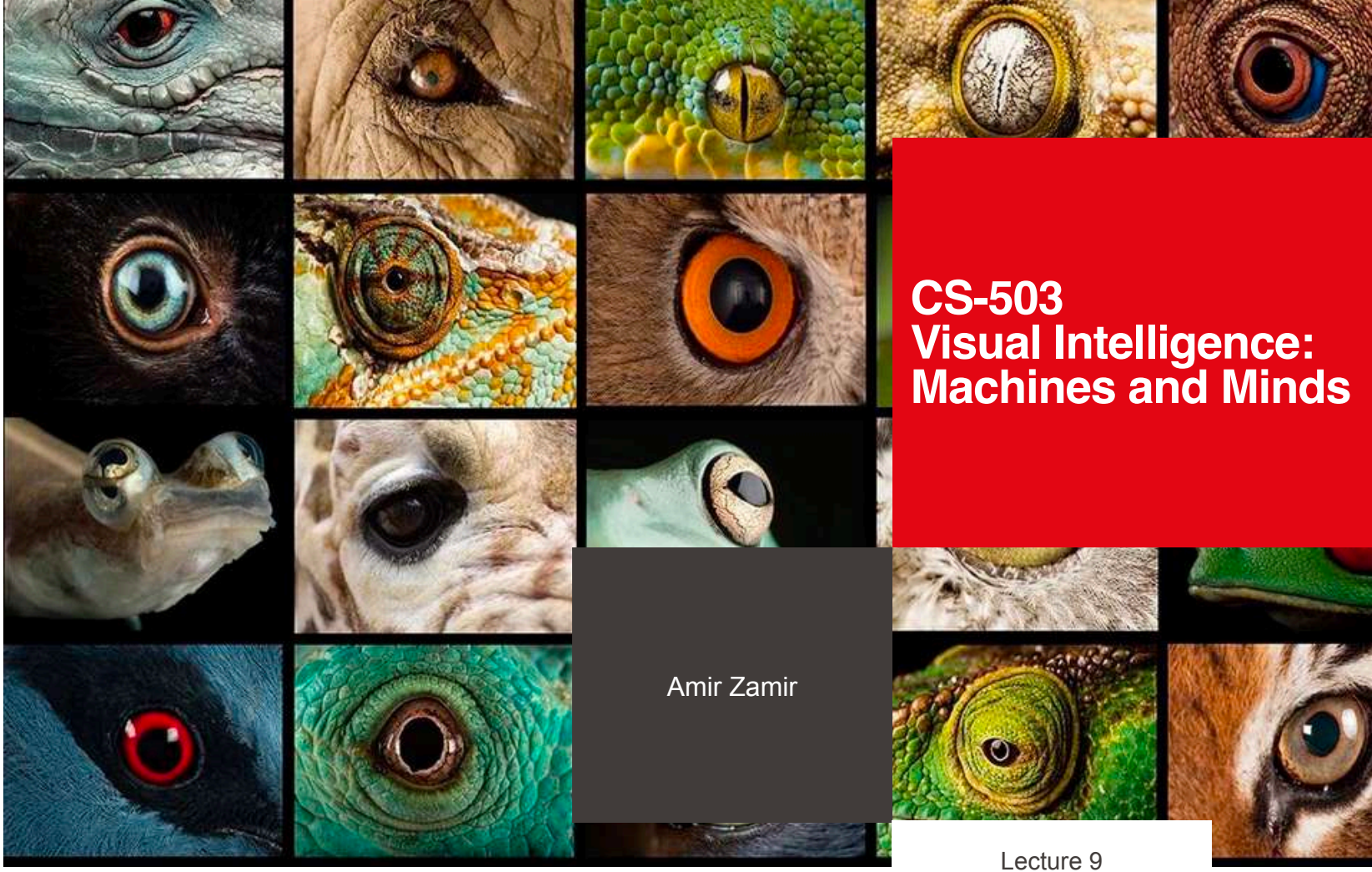


CS-503
Visual Intelligence: Machines and Minds

Amir Zamir

Lecture 9

École
polytechnique
fédérale
de Lausanne

Logistics

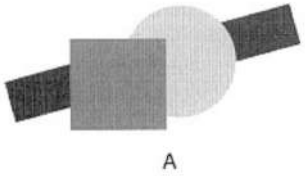
Logistics

5E	21.03	- Architectures notebook Q&A - Active agents preview
5	21.03	- lecture 5
	26.03	- Transformers notebook assignment due (released 15.03)
6E	28.03	- Active agents notebook Q&A
6	28.03	- lecture 6
~~7~~	~~3.04~~	~~No class (holiday week)~~
8E	11.04	- Project pitch and teaming (by students)
8	11.04	- ~~lecture 8~~ - Project Q&A
9E	18.04	- Project proposals Q&A - Cover up Lecture
9	18.04	- lecture 9
	18.04	- Active agents notebook assignment due (released 28.04)
	19.04	- Project proposals due
10E	25.04	- Presentation on Robustness of vision models - Project troubleshooting - Proposal revision Q&A
10	25.04	- lecture 10
	26.04	- Project proposals due, when revision is needed.
11E	02.05	- Presentation on synthetic training data via generative models - Project troubleshooting - Robustness Notebook Q&A
11	02.05	- lecture 11
~~12~~	~~09.05~~	~~No class (holiday)~~
13E	16.05	- Project troubleshooting
13	16.05	- lecture 13
	10.05	- Project progress report due
	12.05	- Robustness notebook assignment due
14E	23.05	- Project troubleshooting
14	23.05	- lecture 14
	24.05	- Moodle homework due (release 19.05)
15E	30.05	- Final project presentation
15	30.05	- Final project presentation
	31.05	- Project report due

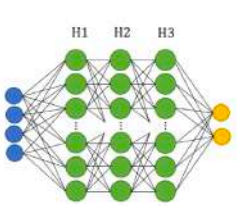
Deliverable due. Lecture. Exercise session. Holiday.

Recap

Perception as modeling the environment

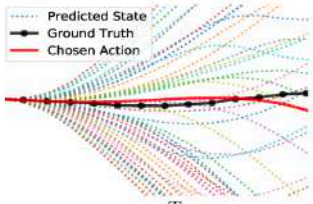


- How and why?
  - The evolutionary utility of vision toward survival and reproduction, in the environment.
- *The observer is constructing a **model** of what environment situation might have produced the observed pattern of sensory stimulation*



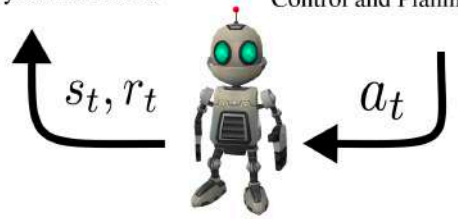
$s_{t+1} = f_\theta(s_t, a_t)$

Dynamics Model

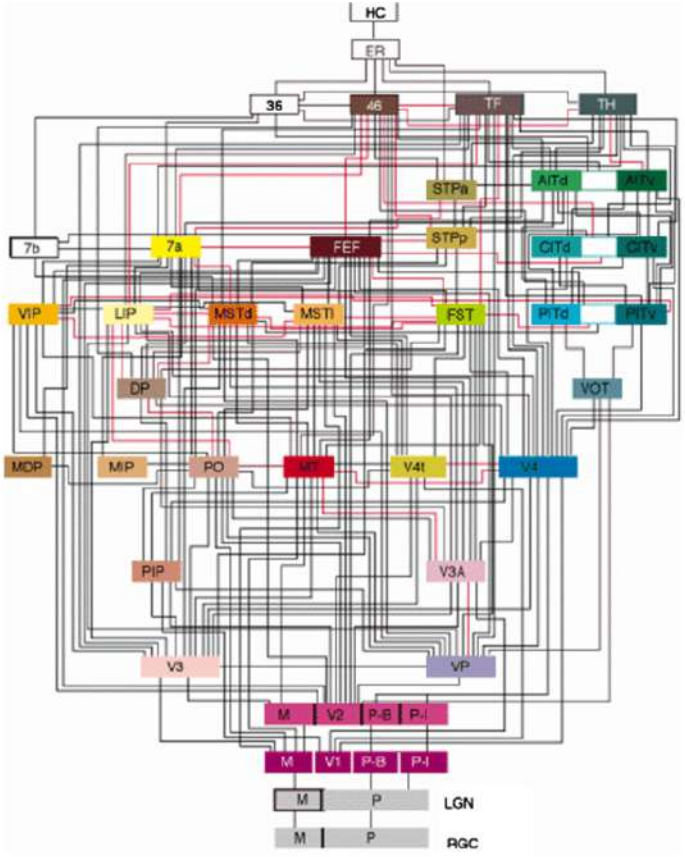


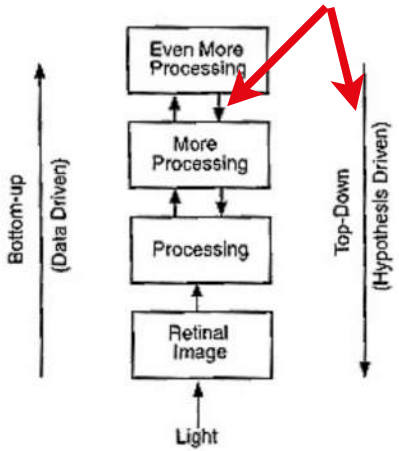
$\arg\max_{u(t)} \sum_{t=0}^{T} r(s_t, a_t)$

Control and Planning



Feedback

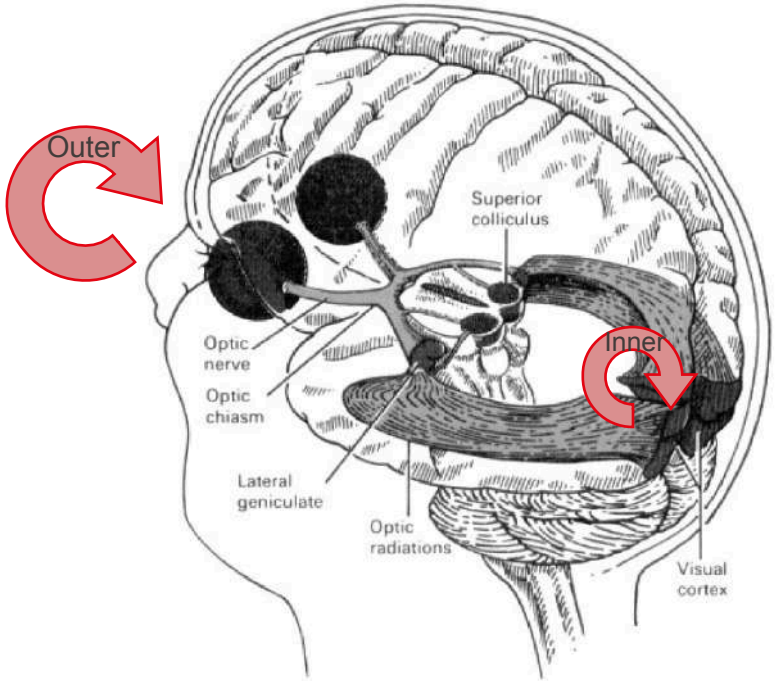




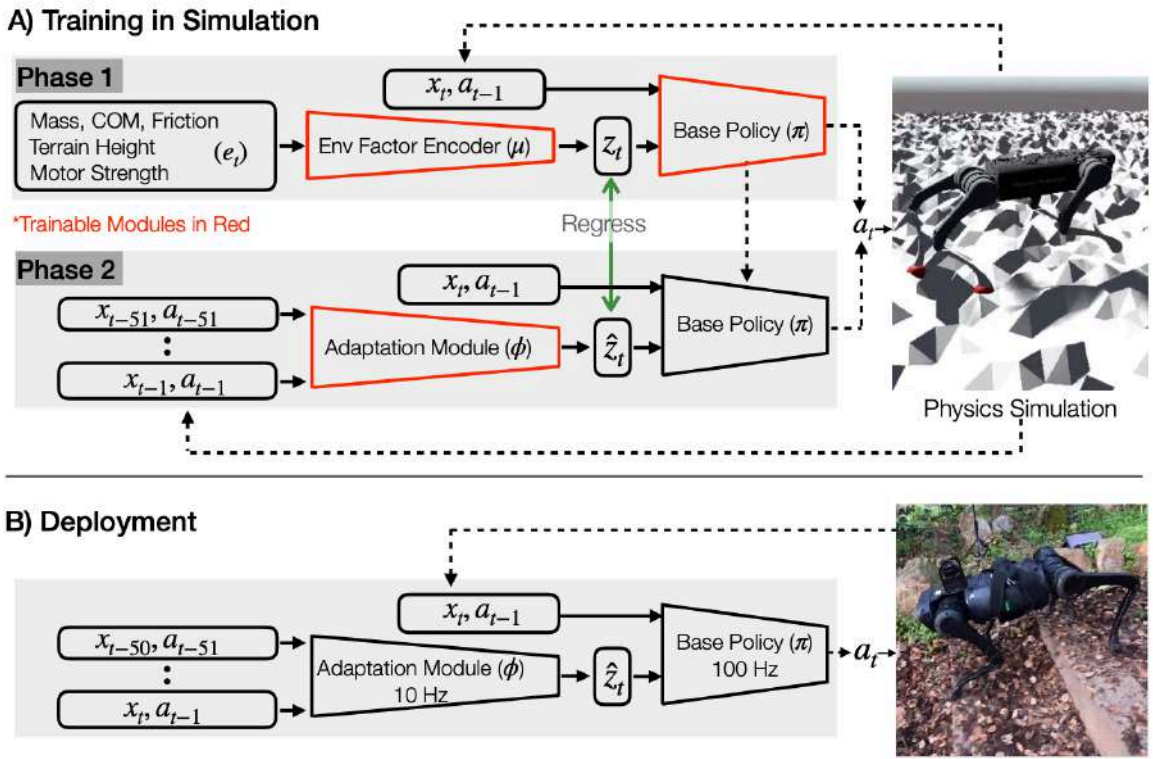
- Felleman & Van Essen (1991)

Feedback

- Inner loop
  - top-down processing without external feedback from the world.
  - e.g. IEF (iterative error feedback, 2016), Attention, Feedback Networks (2017), diffusion.
- Outer loop
  - with external feedback from the world
  - e.g. RMA (2021), RNA (2023), Most vision-action loop (e.g. Mid-level 2019), “Test-Time Training” (2020)
- (All of the above are test-time feedback)

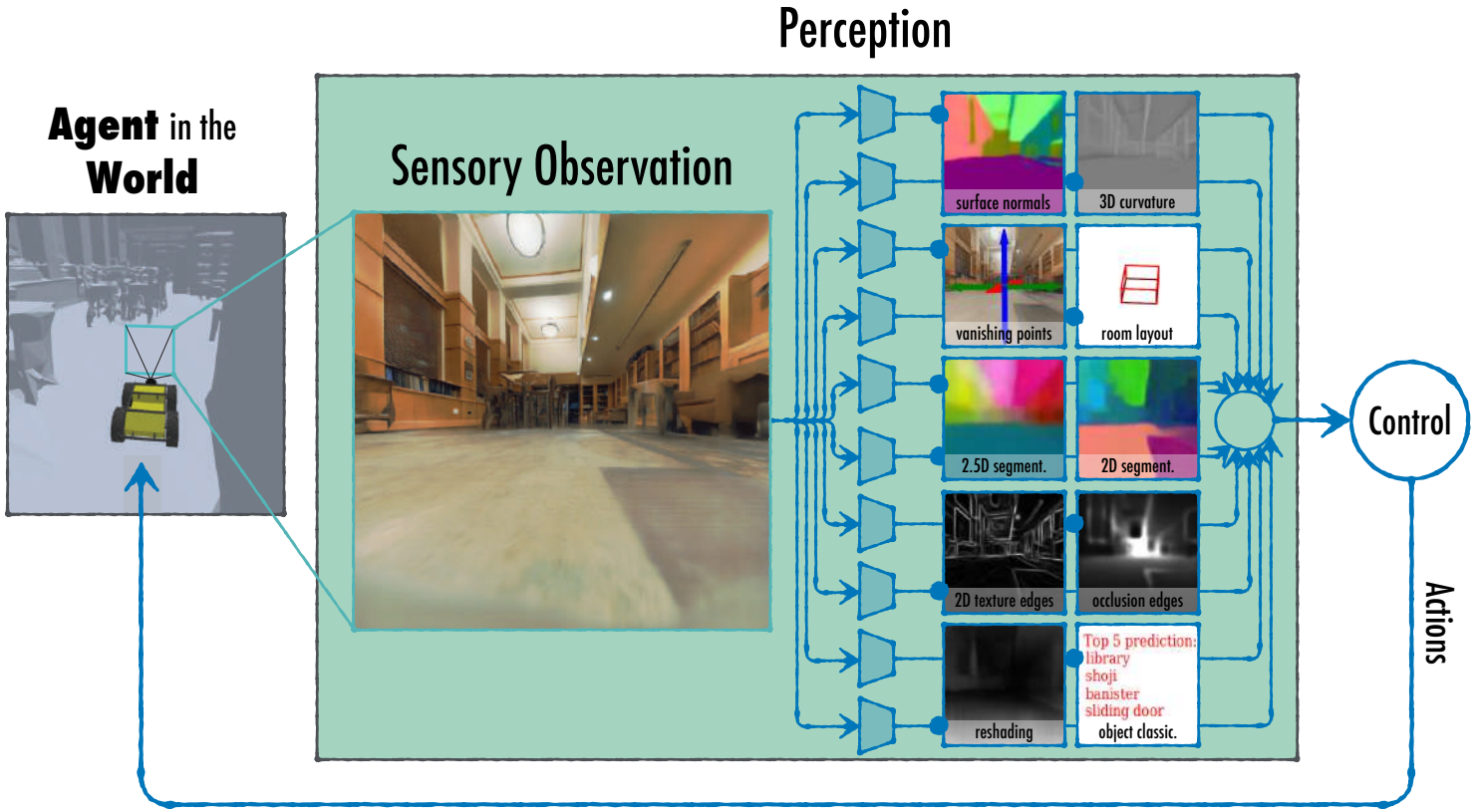


Outer loop Feedback



RMA: Rapid Motor Adaptation for Legged Robots, Kumar et al., RSS 2021

Vision In-the-loop




Frequency
Rank order of items

Gibson Environment

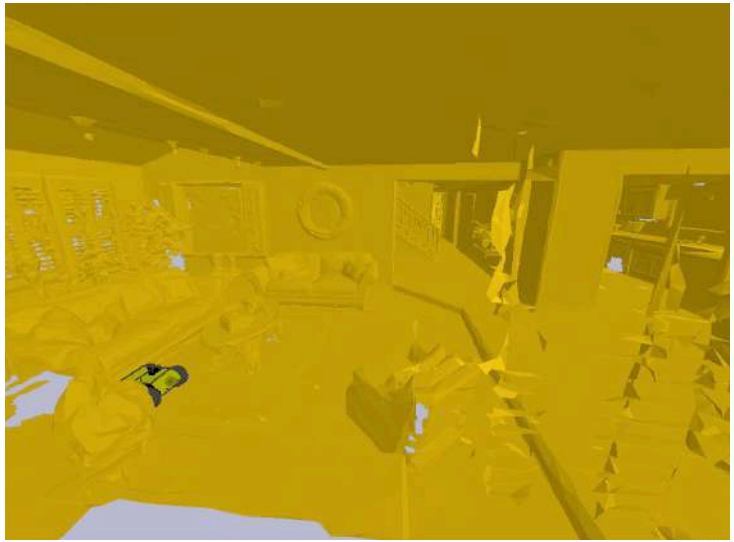
Large Real Space

Active Agent

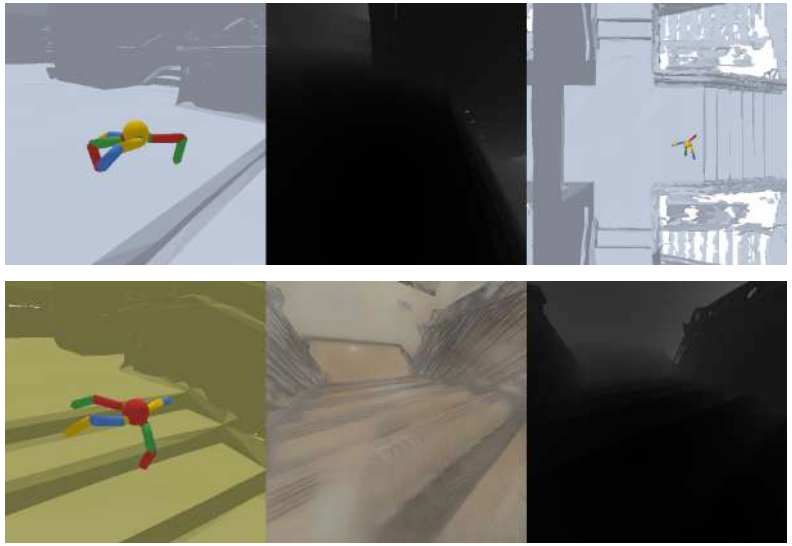
RGB Frame Stream

Additional Modalities

Sample perceptual agents trained in Gibson (using Reinforcement Learning)

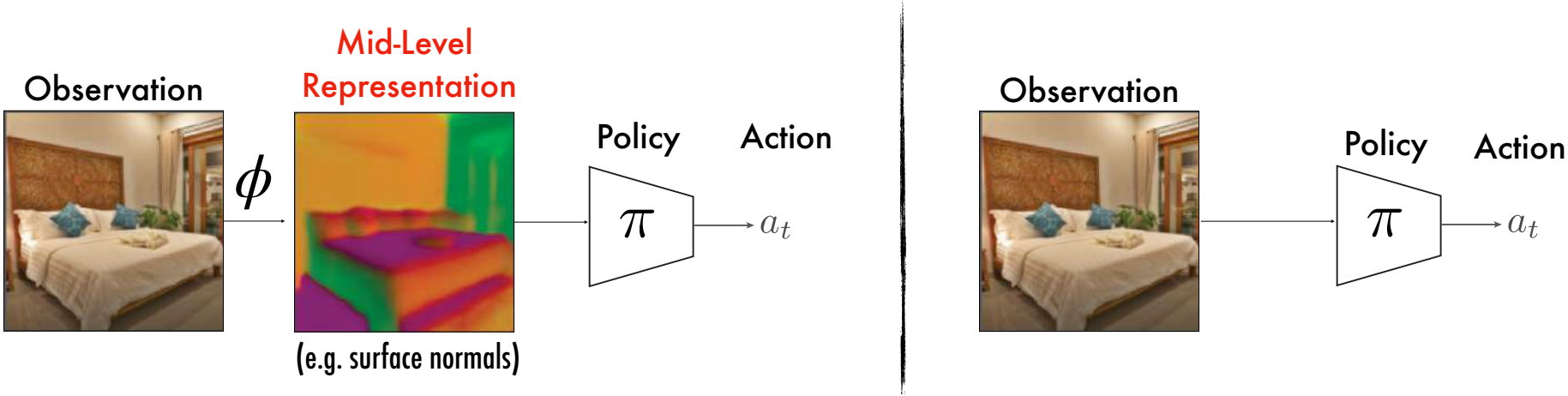


Local planning ("go to the target")



Stair climb

Setup



Learning with Perceptual Priors Vs "Tabula Rasa" (scratch) Learning

Tested hypothesis 1: Does mid-level vision **accelerate learning**?

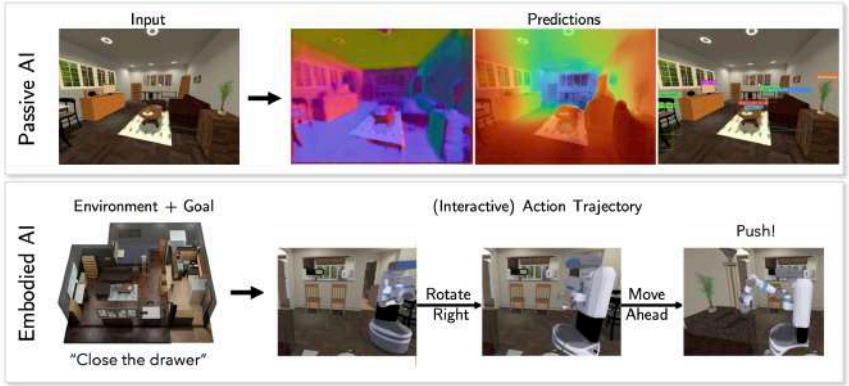
Tested hypothesis 2: Can mid-level features **generalize** better to unseen spaces?

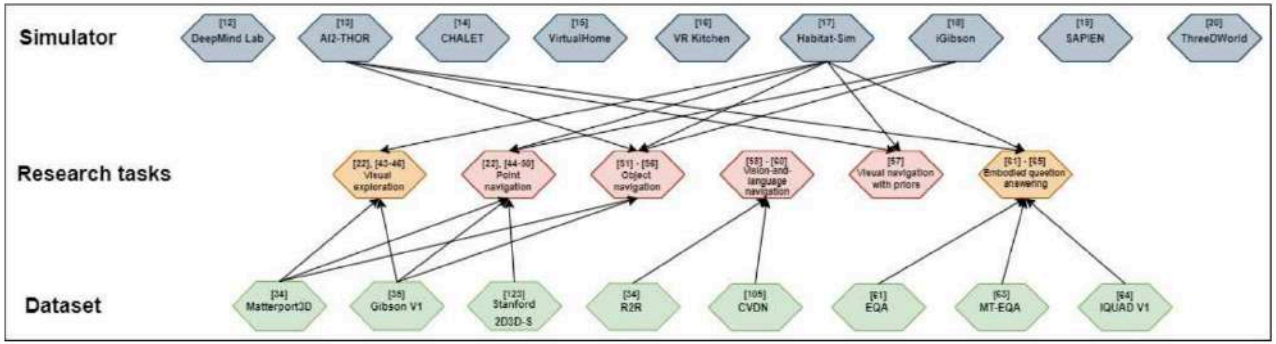
- "Mid-Level Visual Representations Improve Generalization and Sample Complexity for Learning Visuomotor Policies". Sax, Emi, Zamir, Guibas, Savarese, Malik. Arxiv 2018. CoRL 2019.
- "Robust Policies via Mid-Level Visual Representations: An Experimental Study in Manipulation and Navigation". Chen, Sax, Pinto, Lewis, Armeni, Savarese, Zamir, Malik. CoRL 2020

Standardized embodied vision efforts (as of ’24/’25)

Common Tasks (2023)

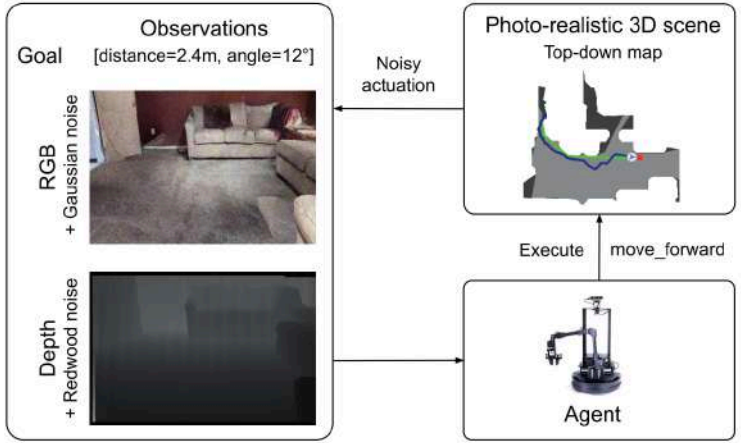
- (1) visual navigation
- (2) rearrangement
- (3) embodied vision-and-language





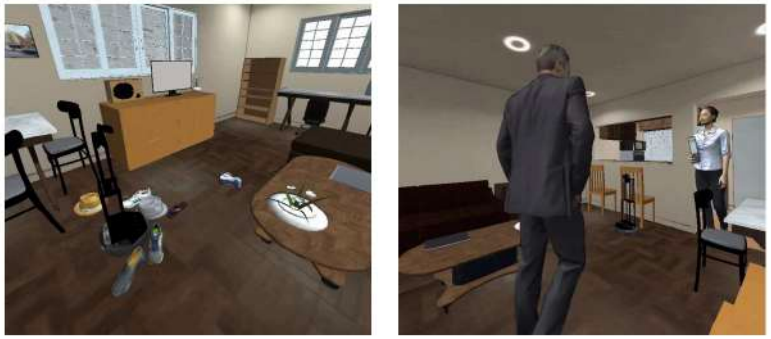
Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

PointNav

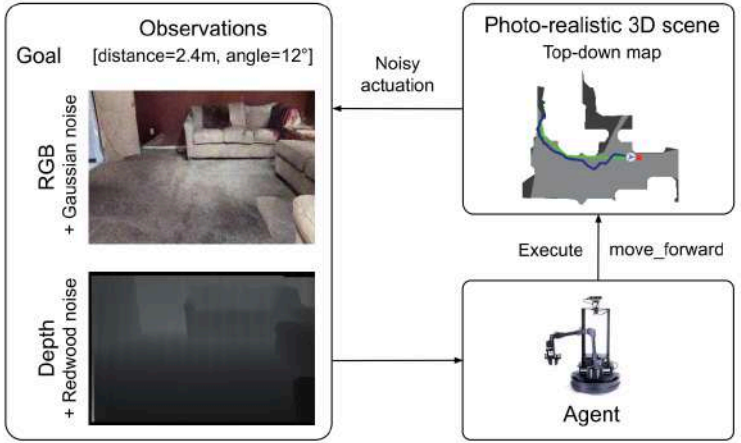


Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

PointNav++

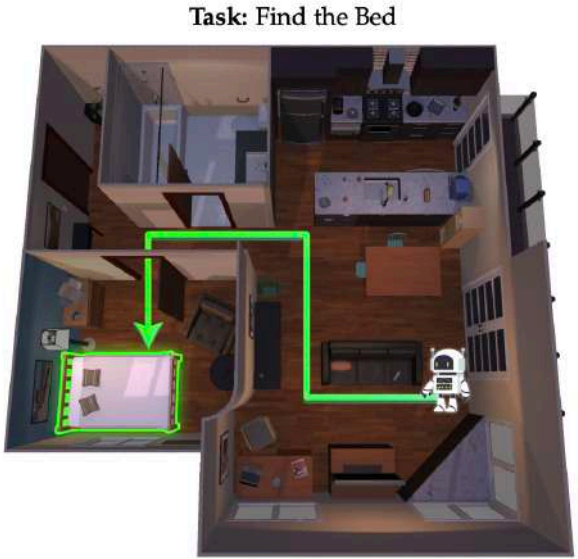


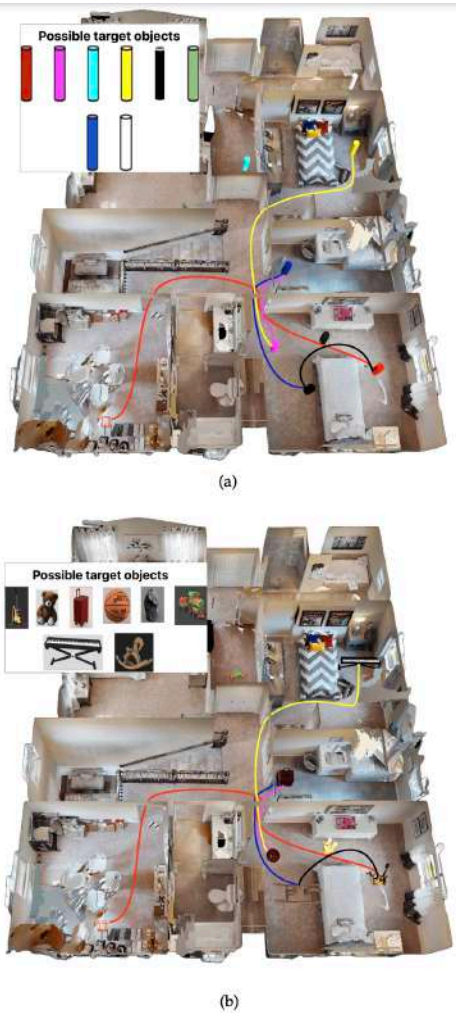
(a) Interactive Navigation (b) Social Navigation



Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

ObjectNav





Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

Nav++

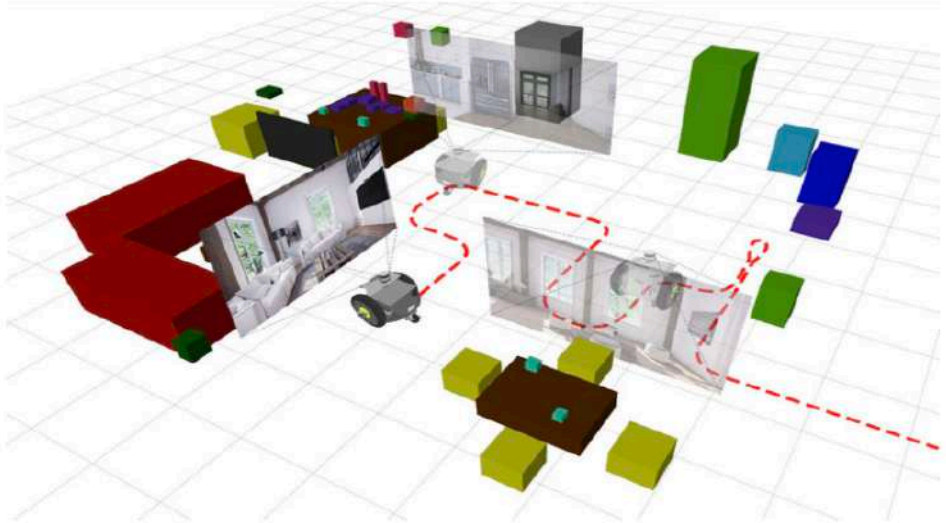
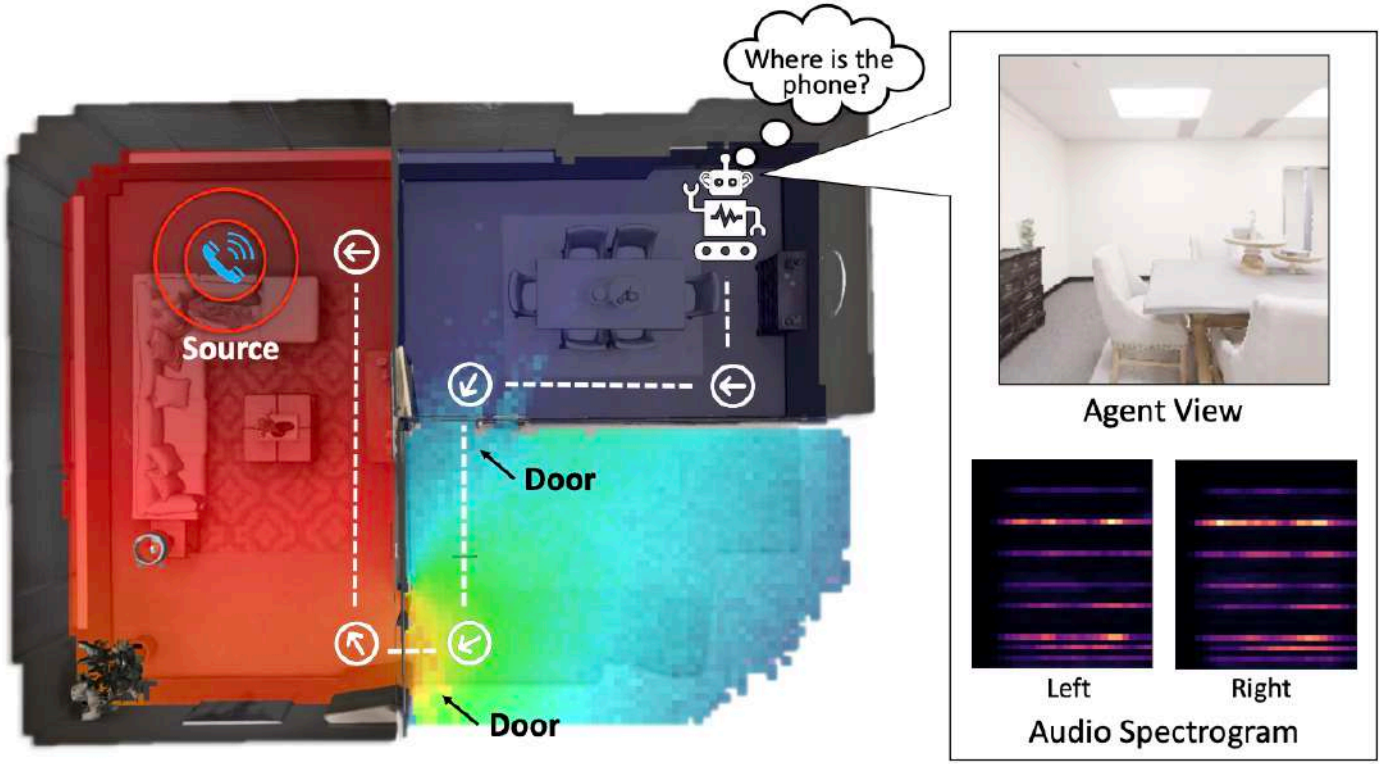


Figure 7. In the *RVSU Semantic SLAM* task, an autonomous agent explores environment to create a semantic 3D cuboid map of objects.

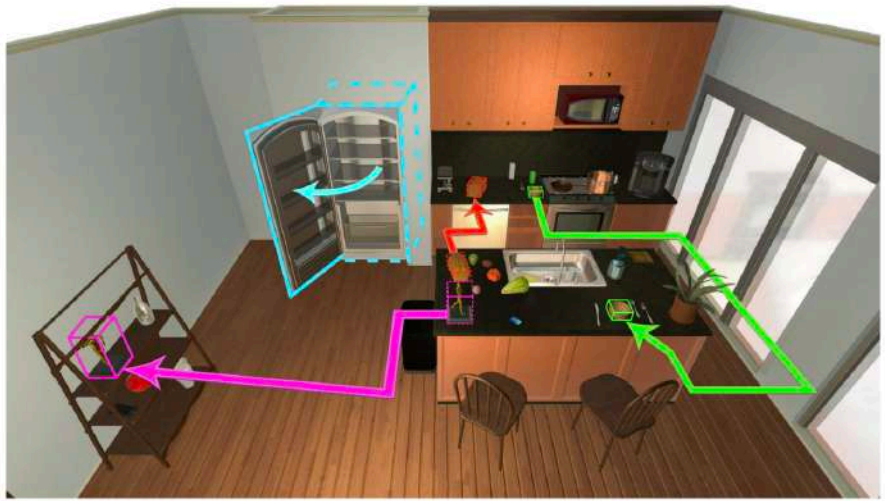
Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

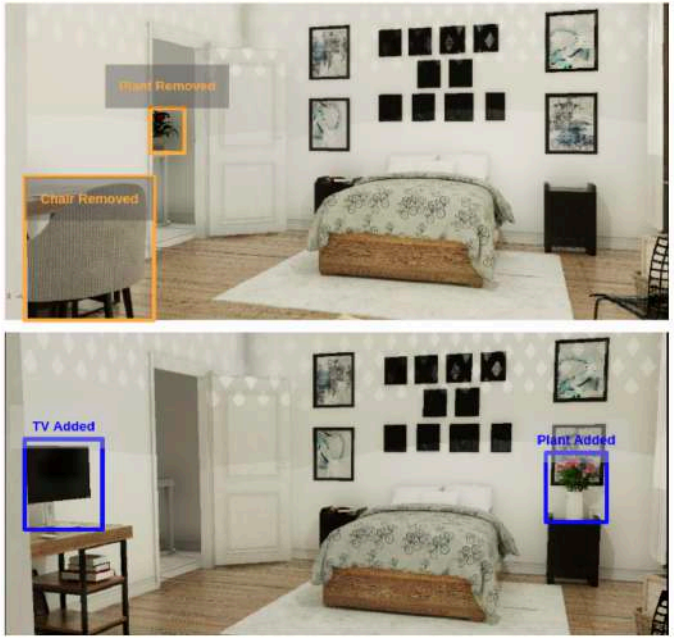
Nav++ (multimodal)



Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

Rearrange





Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

Embodied vision-and-language

You are in a bedroom. Turn around to the left until you see a door leading out into a hallway, go through it. Hang a right and walk between the island and the couch on your left. When you are between the second and third chairs for the island stop.

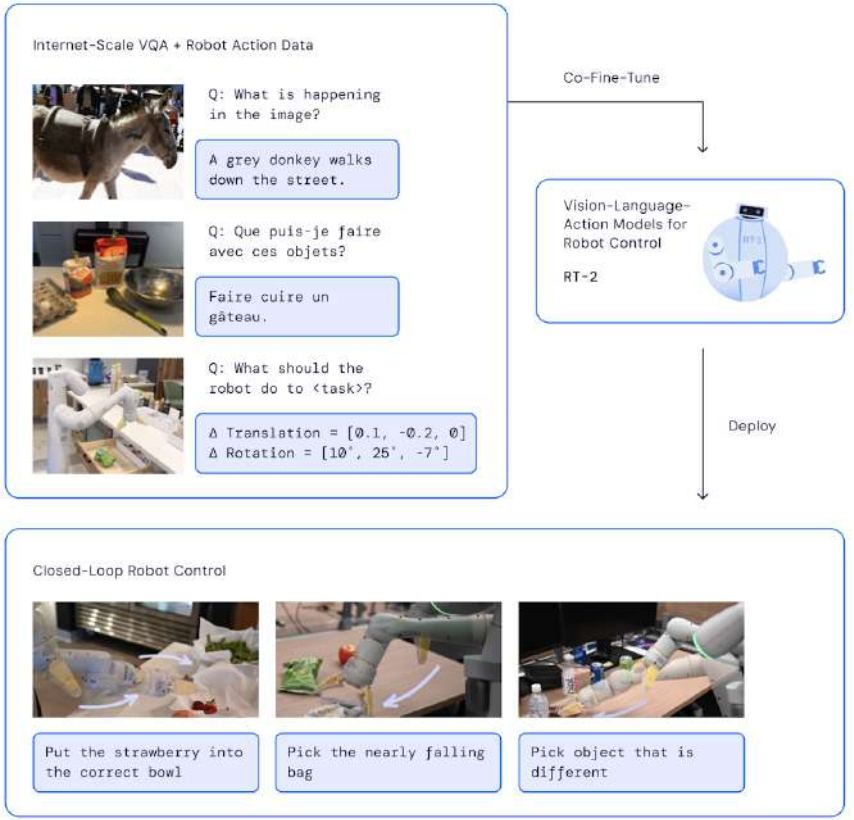
Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

Empirical lessons

		Best End-to-end			Best Modular		
Challenge	**Simulator**	**Method**	**Success**	**Rank**	**Method**	**Success**	**Rank**
ObjectNav	Habitat	Habitat-Web	60	2	Stretch	60	1
Audio-Visual Navigation	SoundSpaces	Freiburg Sound	73	2	colab_buaa	78	1
Multi-ON	Habitat	-	-	-	exp_map	39	1
Navigation Instruction Following	VLN-RxR	CMA Baseline	13.93	10	Reborn	45.82	1
Interactive Instruction Following	AI2-THOR	APM	15.43	14	EPA	36.07	1
Rearrangement	AI2-THOR	ResNet18 + ANM	0.5	6	TIDEE	28.94	1

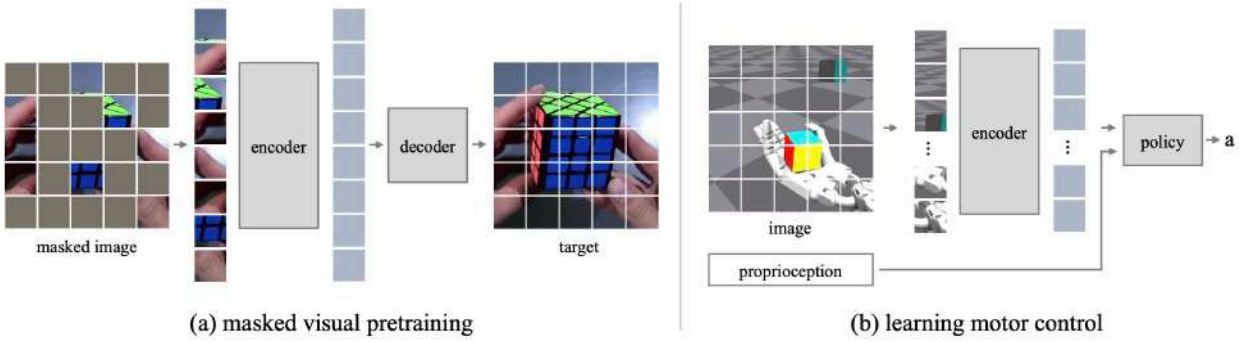
Retrospectives on the Embodied AI Workshop, Deitke et al., 2022

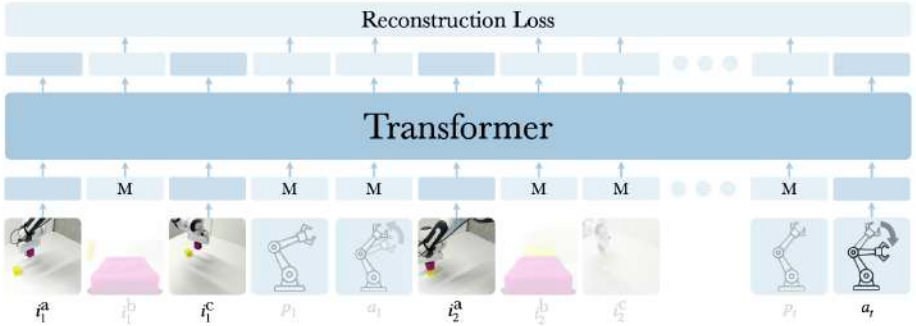
LLMs in robotics pipelines



- RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, Google, 2023.

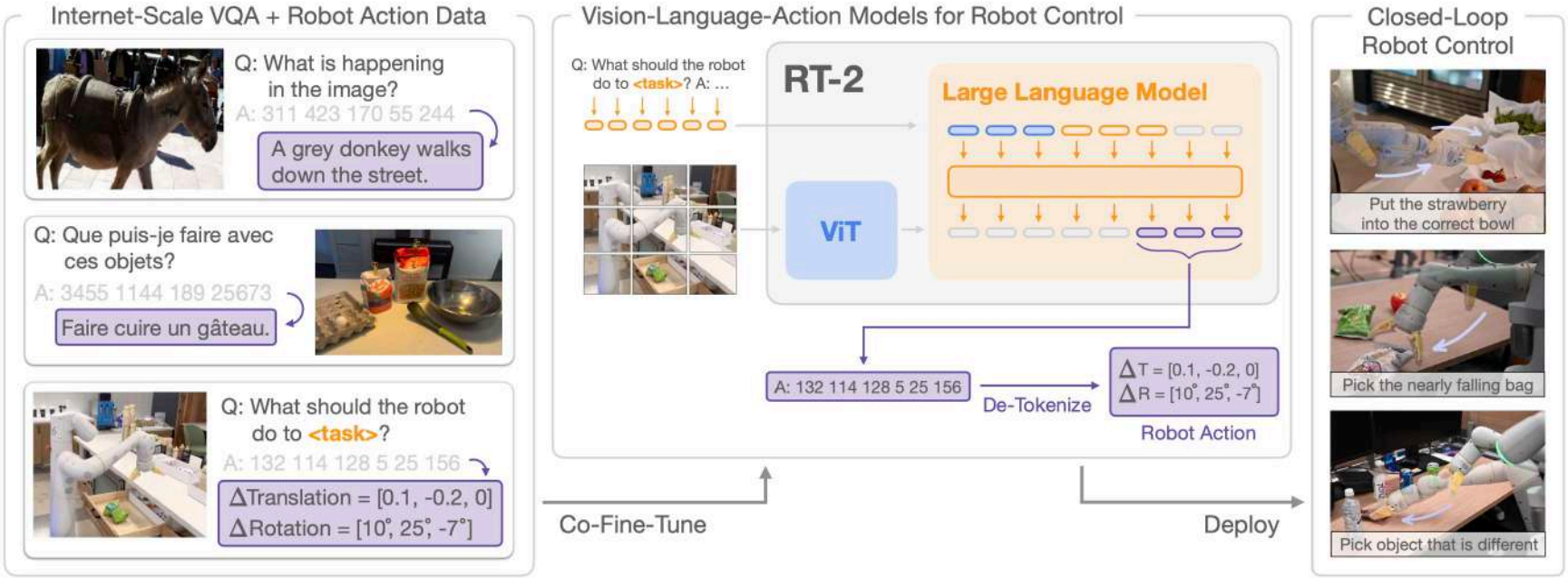
Multi-modal learning → Motor Control





Robot Learning with Sensorimotor Pre-training, Radosavovic, Shi, Fu, Goldberg, Darrell, Malik. 2023
Real-World Robot Learning with Masked Visual Pre-training, Radosavovic, Xiao, James, Abbeel, Malik, Darrell. CoRL 2022
Masked Visual Pre-training for Motor Control, Xiao, Radosavovic, Darrell, Malik. ArXiv 2022
MultiMAE: Multi-Modal Multi-Task Masked Autoencoders, Bachmann, Mizrahi, Atanov, Zamir. ECCV 2022

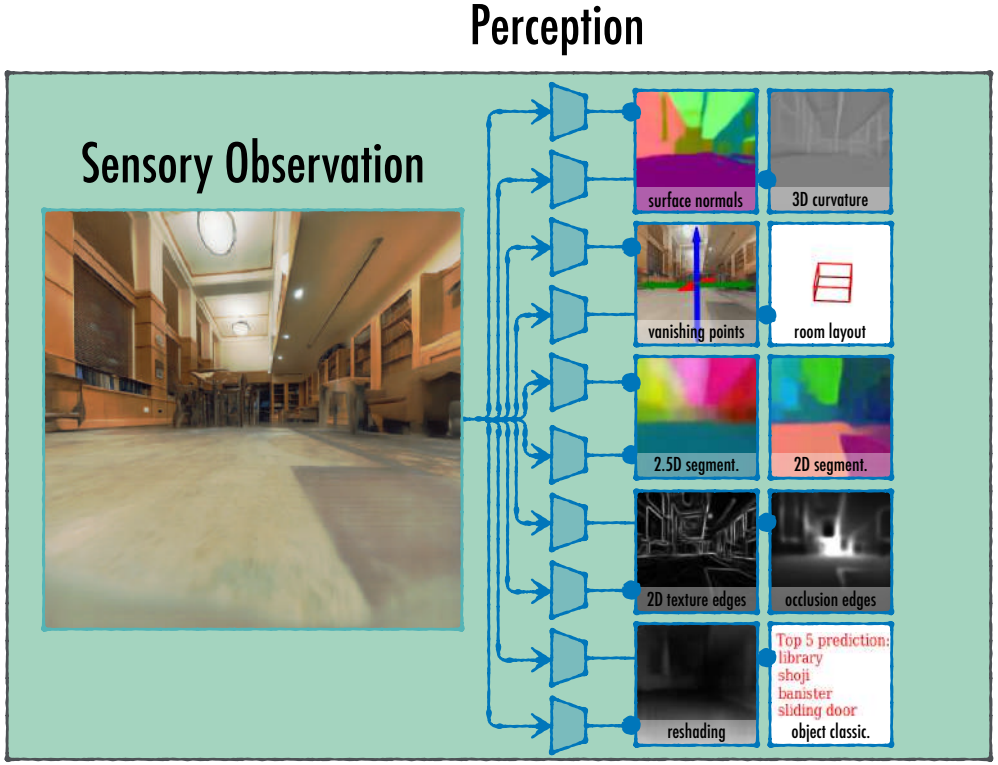
Multi-modal learning → Motor Control



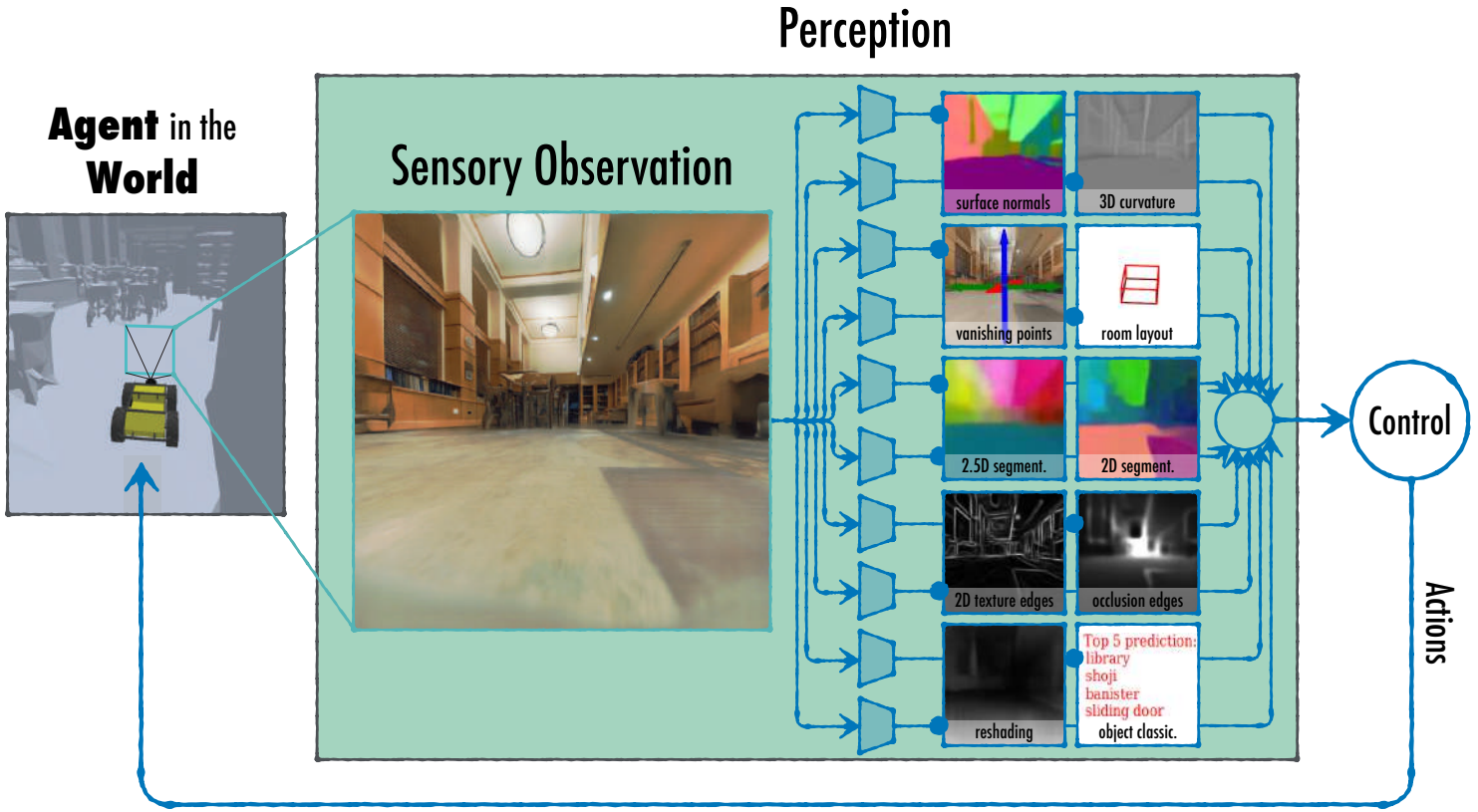
RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, 2023.
PaLM-E: An Embodied Multimodal Language Model, 2023.

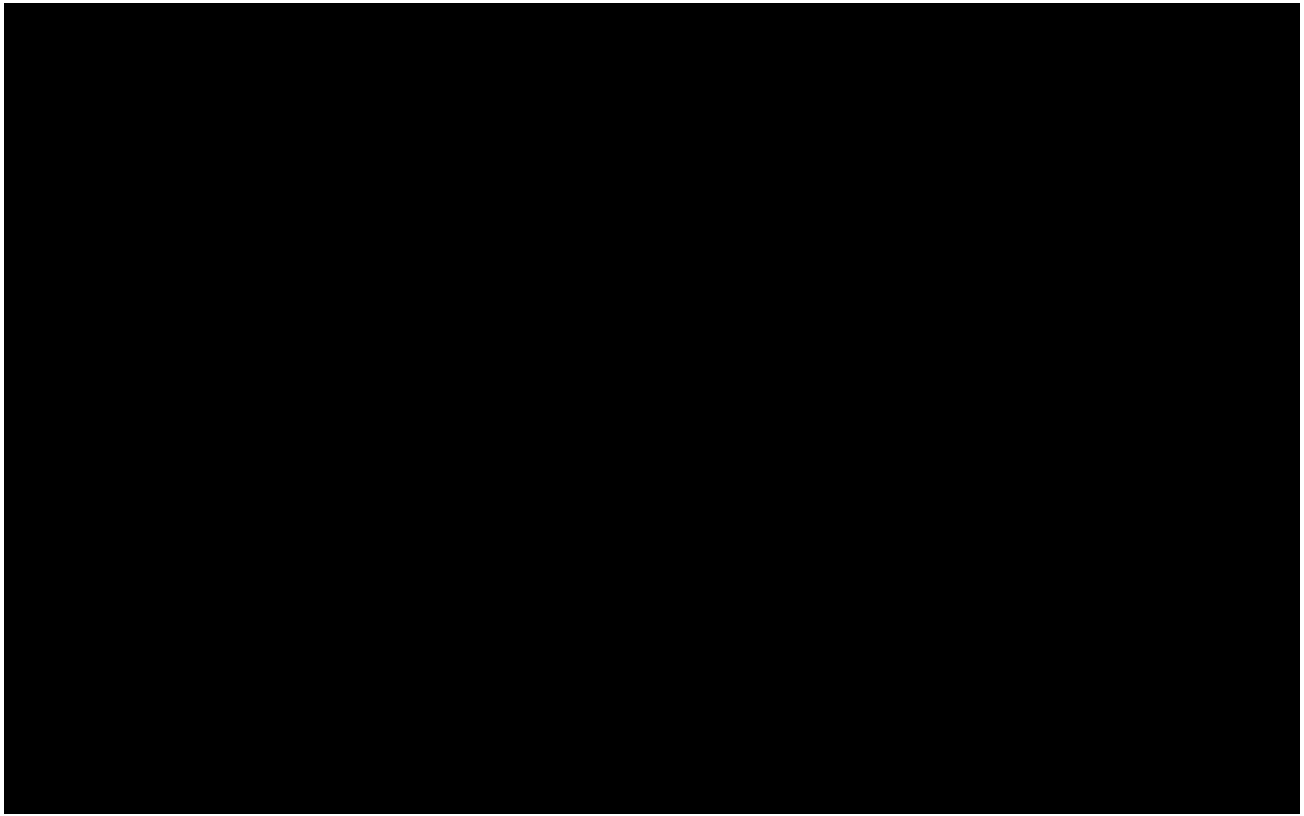
Sensorimotor Contingency

Vision In-the-loop



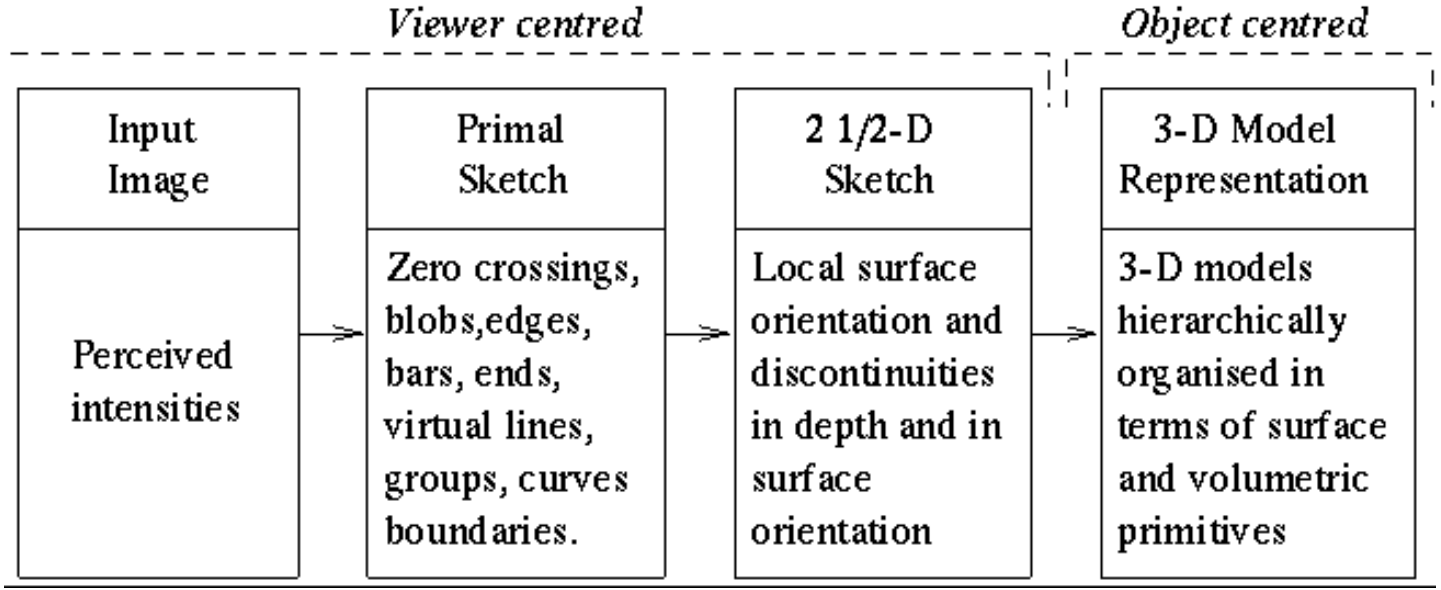
Vision In-the-loop





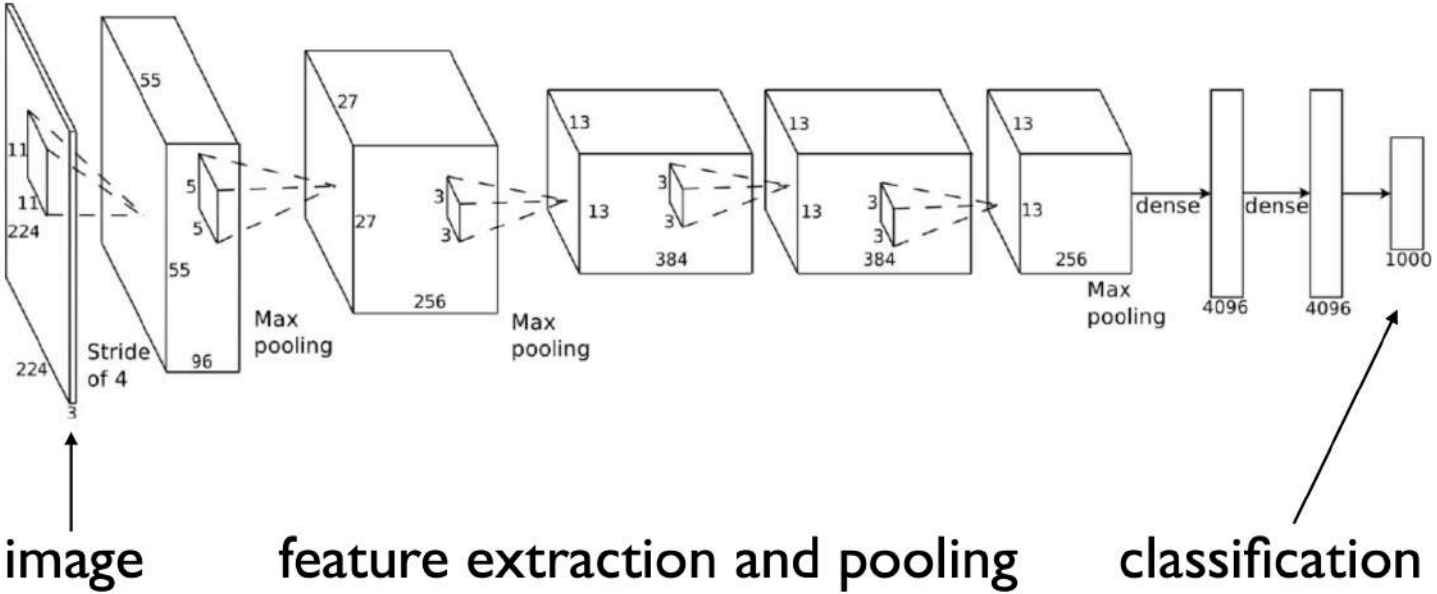
Bower lab, Caltech. B Olshausen.

The approach of David Marr



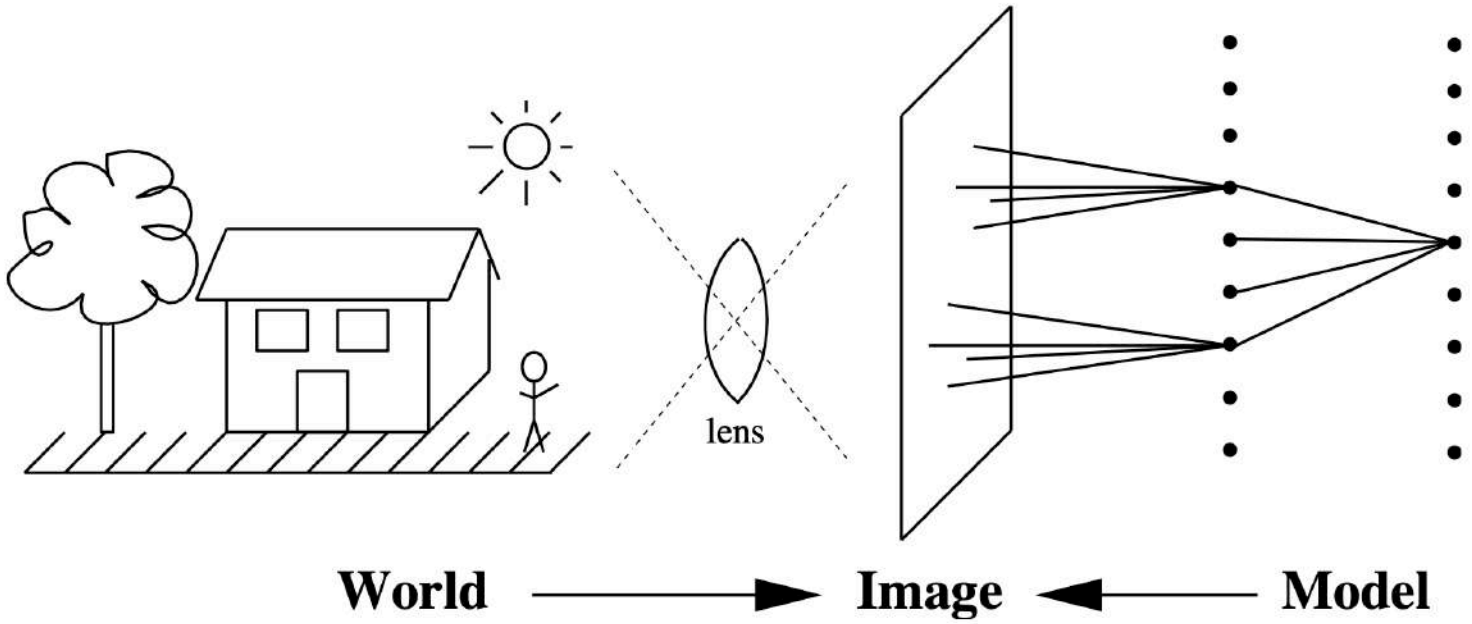
B. Olshausen

'AlexNet'
(Krizhevsky, Sutskever & Hinton 2012)

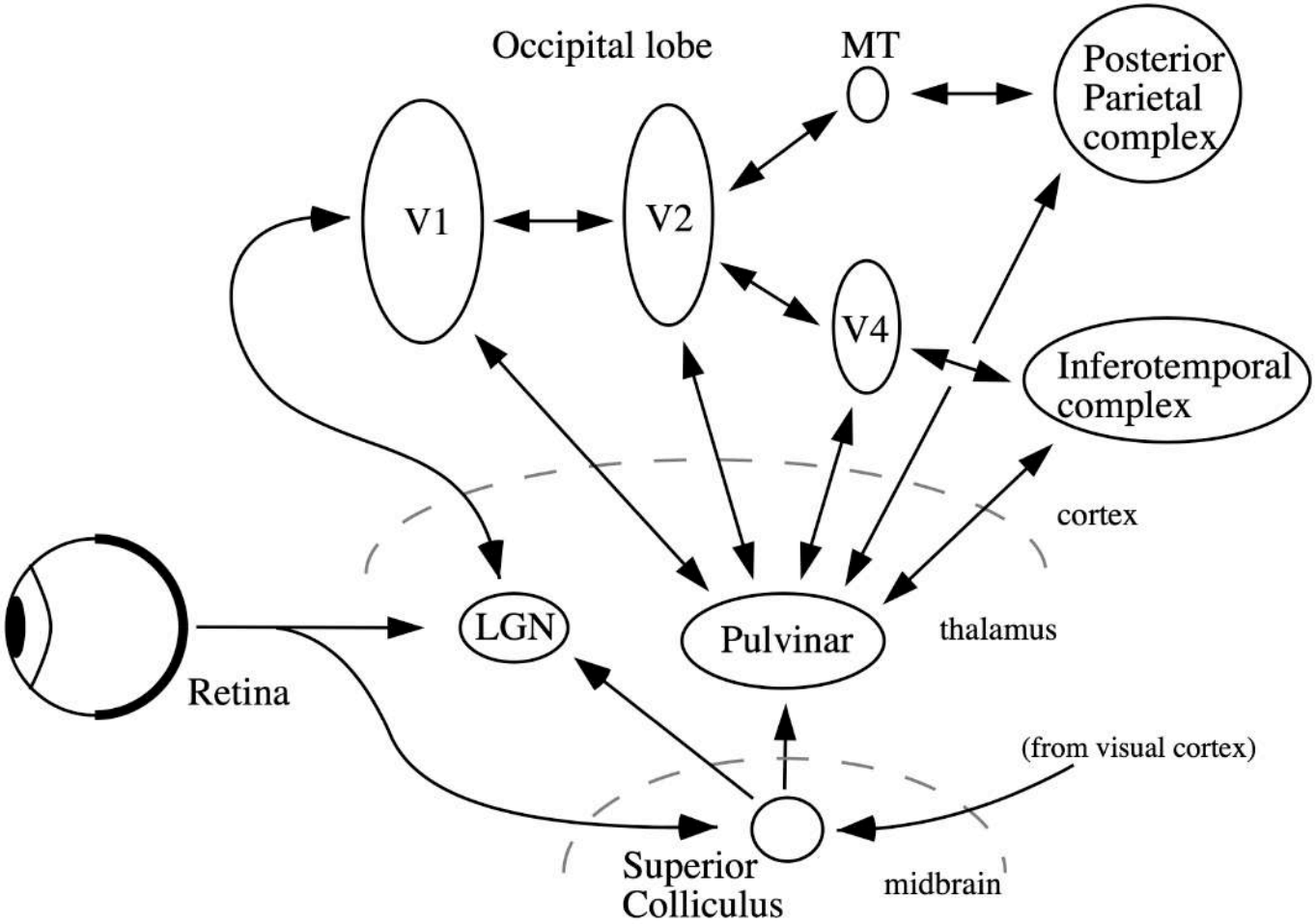


B. Olshausen

Vision as inference

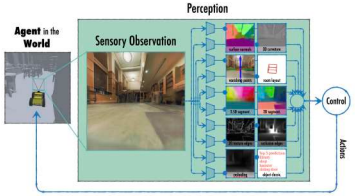


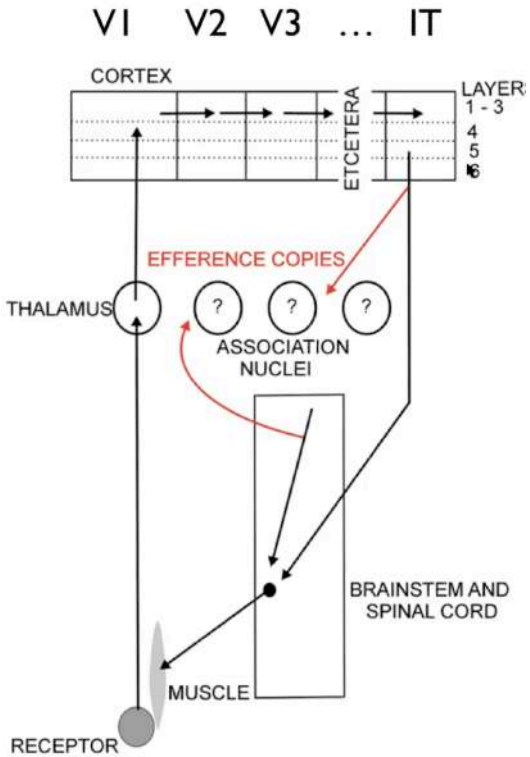
B. Olshausen

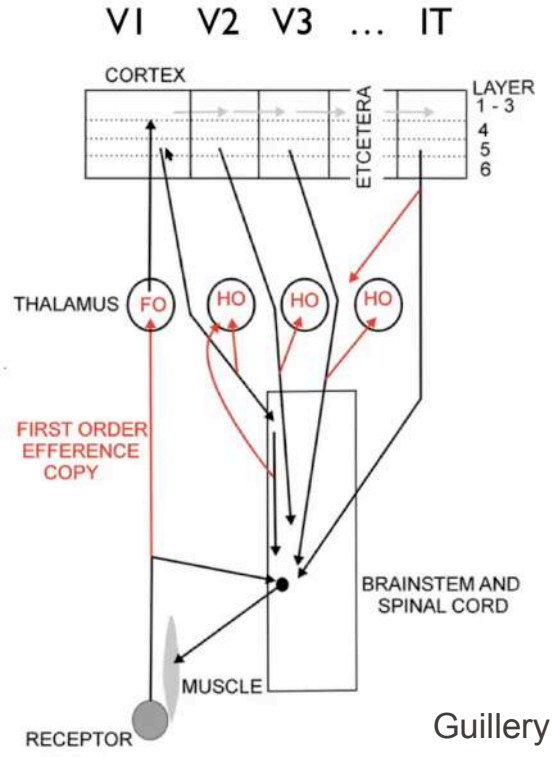

Occipital lobe
MT
Posterior Parietal complex
V1
V2
V4
Inferotemporal complex
cortex
LGN
Pulvinar
thalamus
Retina
(from visual cortex)
Superior Colliculus
midbrain

B. Olshausen

Anatomically tighter connection between vision & action



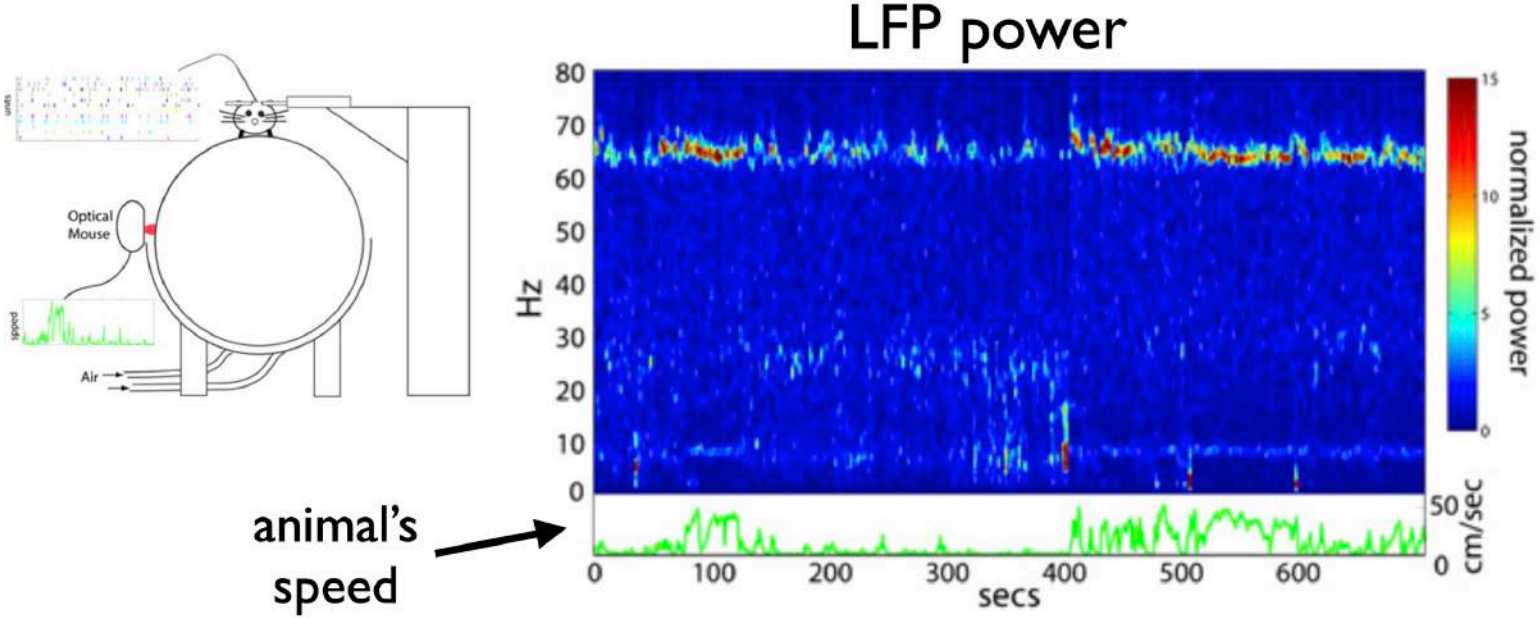




B. Olshausen
Guillery & Sherman, 2011

Anatomically tighter connection between vision & action

- Activity in V1 notably increases during locomotion



Neil & Stryker, 2010

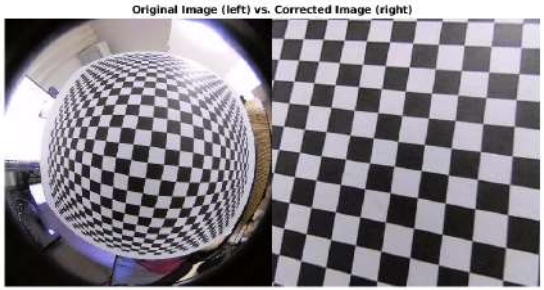
Sensorimotor Contingency

Imagine a team of engineers operating a remote-controlled underwater vessel exploring the remains of the Titanic, and imagine a villainous aquatic monster that has interfered with the control cable by mixing up the connections to and from the underwater cameras, sonar equipment, robot arms, actuators, and sensors. What appears on the many screens, lights, and dials, no longer makes any sense, and the actuators no longer have their usual functions. What can the engineers do to save the situation?

O'Regan & Noe (2001)

Sensorimotor Contingency

(Cross) Calibration



Imagine a team of engineers operating a remote-controlled underwater vessel exploring the remains of the Titanic, and imagine a villainous aquatic monster that has interfered with the control cable by mixing up the connections to and from the underwater cameras, sonar equipment, robot arms, actuators, and sensors. What appears on the many screens, lights, and dials, no longer makes any sense, and the actuators no longer have their usual functions. What can the engineers do to save the situation? By observing the *structure of the changes* on the control panel that occur when they press various buttons and levers, the engineers should be able to deduce which buttons control which kind of motion of the vehicle, and which lights correspond to information deriving from the sensors mounted outside the vessel, which indicators correspond to sensors on the vessel's tentacles, and so on.

O'Regan & Noe (2001)

Sensorimotor Contingency

2:55

Erismann & Kohler 1931.
Stratton 1897.

Sensorimotor Contingency

A sensorimotor account of vision and visual consciousness

J. Kevin O'Regan
Laboratoire de Psychologie Expérimentale, Centre National de Recherche Scientifique, Université René Descartes, 92774 Boulogne Billancourt, France
oregan@ext.jussieu.fr http://nivea.psycho.univ-paris5.fr

Alva Noë
Department of Philosophy, University of California at Santa Cruz, Santa Cruz, CA 95064
anoe@cats.ucsc.edu http://www2.ucsc.edu/people/anoe/

visual experience does not arise because an internal representation of the world is activated in some brain area.

...

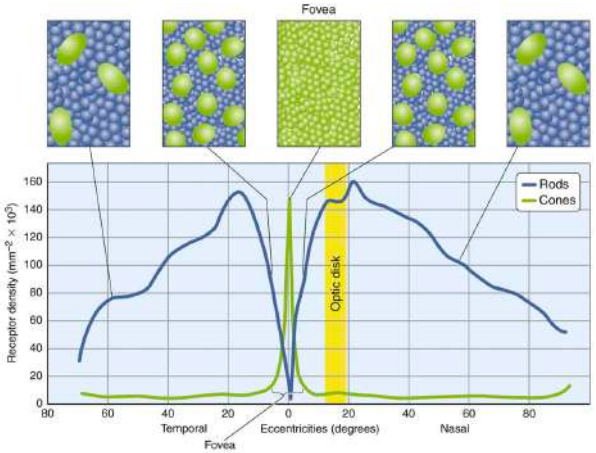
Indeed, there is no "re"-presentation of the world inside the brain:

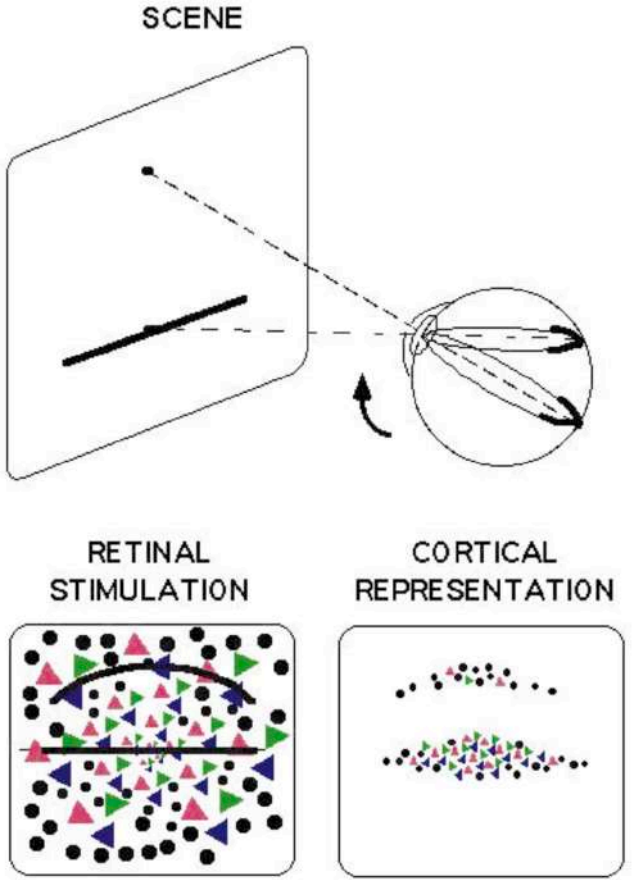
...

The experience of seeing occurs when the outside world is being probed according to the visual mode.

B. Olshausen
O'Regan & Noe (2001)

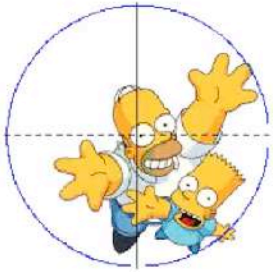
Sensorimotor Contingency

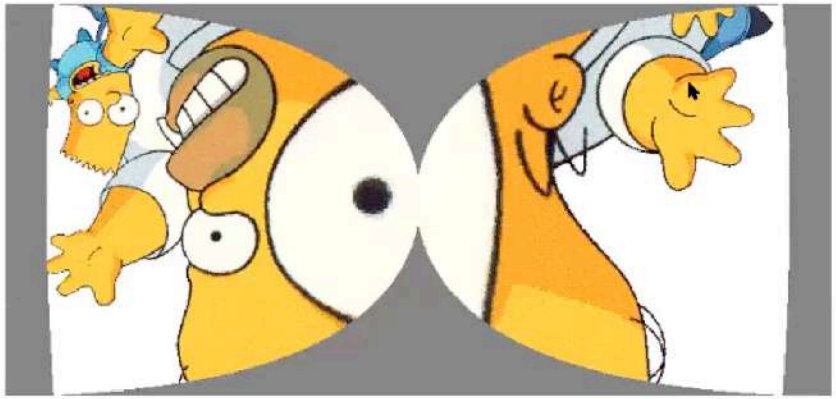




O'Regan & Noe (2001)

V1 representation during eye movement





courtesy of Arash Fazl

B. Olshausen

2020-11-17 16:51:21

Sensorimotor Contingency

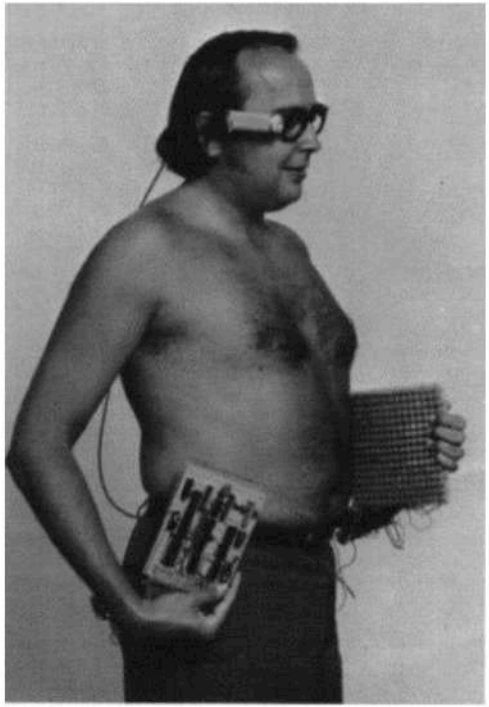
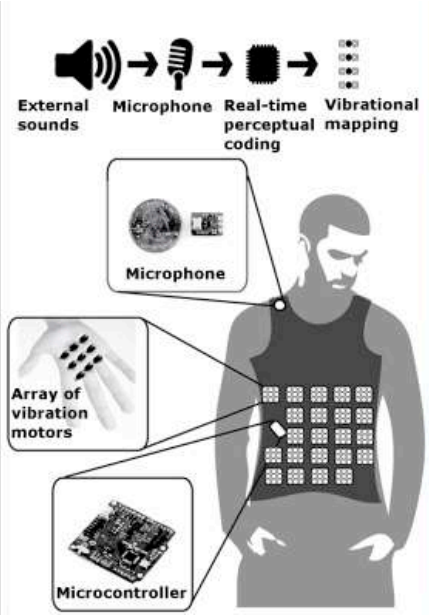
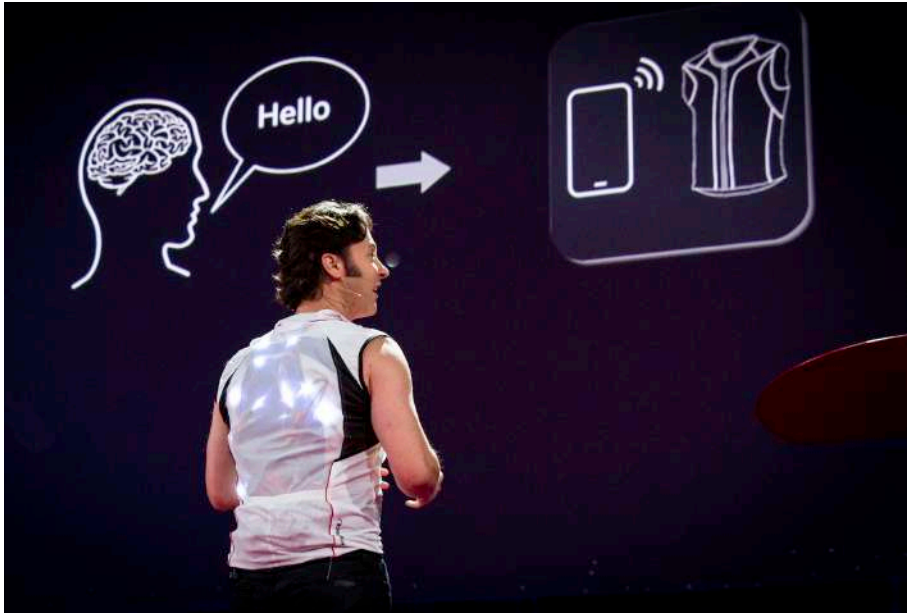


Figure 7. A blind subject with a "Tactile Visual Substitution system" (TVSS). A TV camera (mounted on spectacle frames) sends signals through electronic circuitry (displayed in right hand) to an array of small vibrators (left hand) which is strapped against the subject's skin. The pattern of tactile sitmulation corresponds roughly to a greatly enlarged visual image. (Photograph courtesy of P. Bach-y-Rita). From Morgan (1977).

O'Regan & Noe (2001)

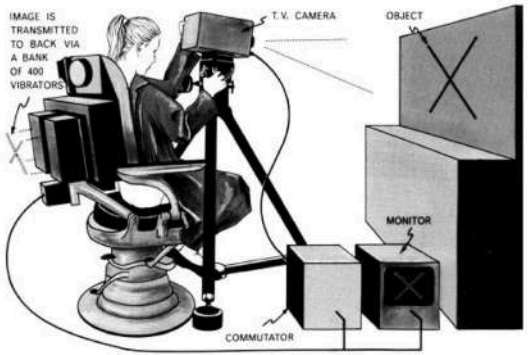
Sensorimotor Contingency



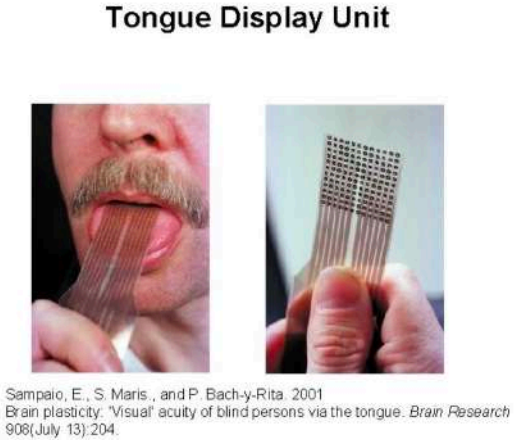


David Eagleman

Sensorimotor Contingency



Bach-y-Rita et al., Vision substitution by tactile image projection, Nature (1969)



Sampaio, E., S. Maris , and P. Bach-y-Rita. 2001
Brain plasticity: 'Visual' acuity of blind persons via the tongue. *Brain Research* 908(July 13):204.

David Ha 2022. Erismann & Kohler 1931. Stratton 1897.
Paul Bach-y-rita (1934-2006) (the father of sensory substitution.

Behavior-based Robotics

- Intelligence Without Representation, Rodney Brooks, 1987.

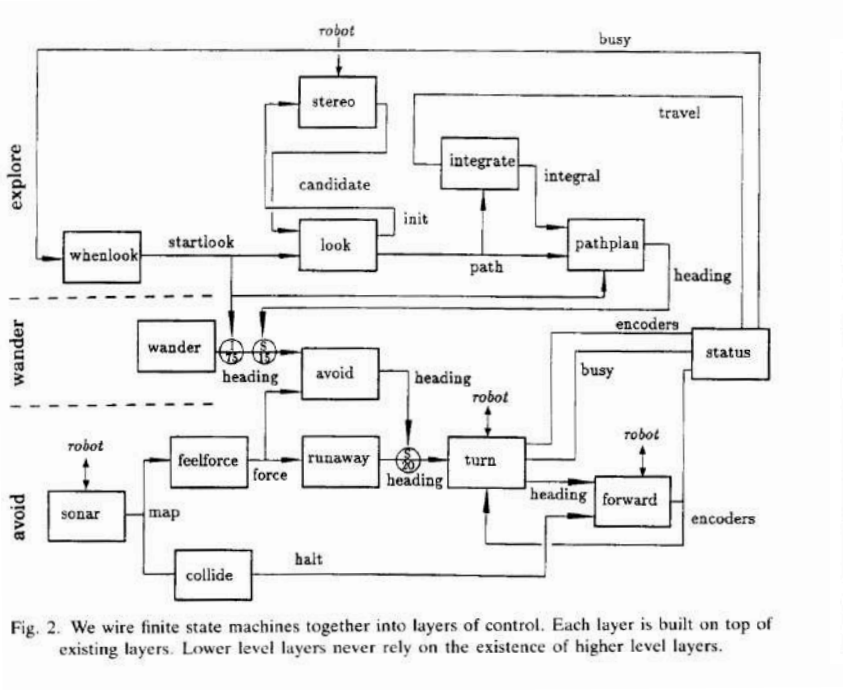
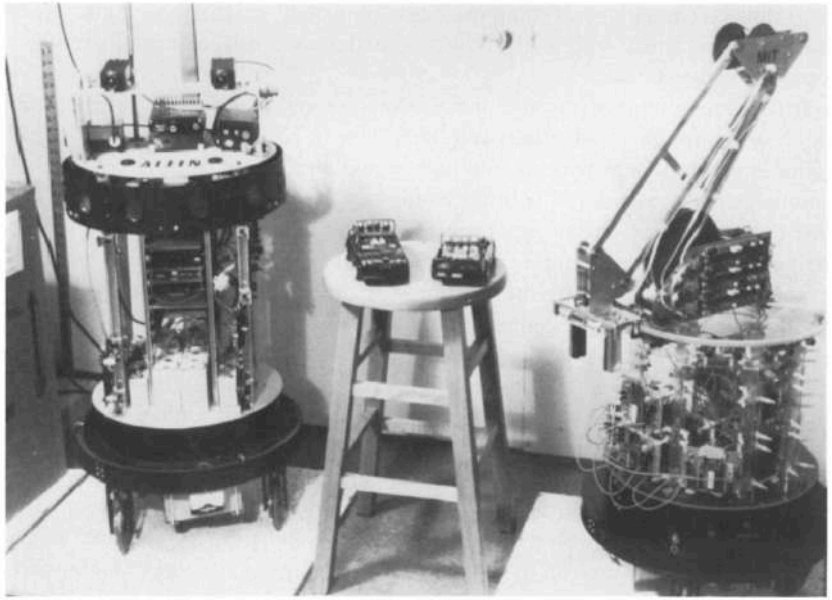


Fig. 2. We wire finite state machines together into layers of control. Each layer is built on top of existing layers. Lower level layers never rely on the existence of higher level layers.



Sensorimotor Contingency

Allegory of the Cave, Plato

Sensorimotor Contingency

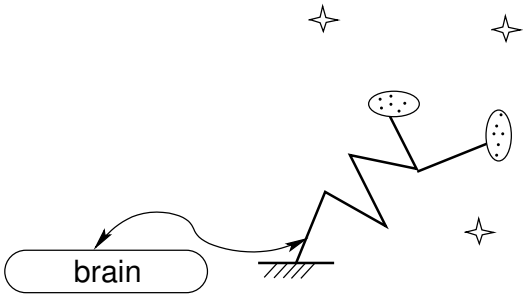
LETTER Communicated by Anthony Bell

Is There Something Out There? Inferring Space from Sensorimotor Dependencies

D. Philipona
david.philipona@polytechnique.org
Sony CSL, 75005 Paris, France

J.K. O'Regan
oregan@ext.jussieu.fr
Laboratoire de Psychologie Expérimentale, CNRS, Université René Descartes, 92774 Boulogne-Billancourt Cedex, France

J.-P. Nadal
Jean-Pierre.Nadal@lps.ens.fr
Laboratoire de Physique Statistique, Ecole Normale Supériure, 75231 Paris Cedex 05, France



$$S = \psi(M, E).$$

B. Olshausen

Sensorimotor Contingency

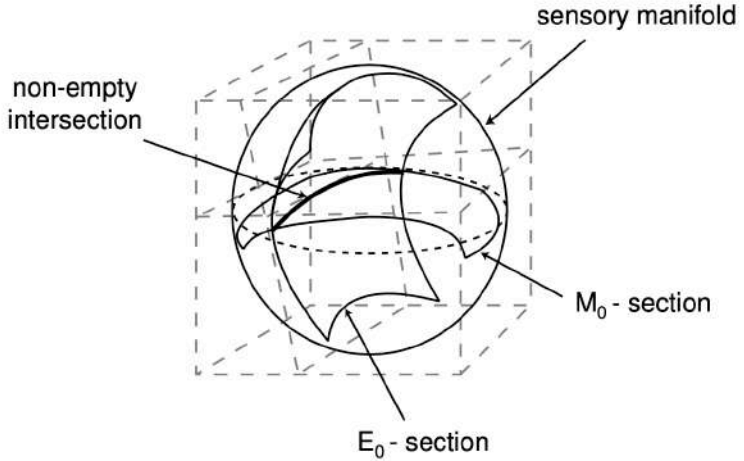


Figure 2: The sensory manifold in the neighborhood of S_0, the E_0 and M_0-sections (see text). These two manifolds are transverse, and their intersection is the manifold of the sensory inputs accessible through either motion of the exteroceptive body or motion of the environment.

$$S = \psi(M, E)$$

$$dS = \frac{\partial \psi}{\partial M}|_{(M_0,E_0)} \cdot dM + \frac{\partial \psi}{\partial E}|_{(M_0,E_0)} \cdot dE$$

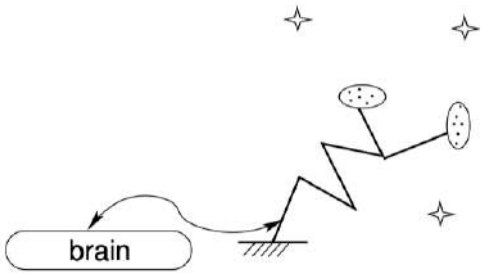
$$\{dS\} = \{dS\}_{dM=0} + \{dS\}_{dE=0}$$

$$\dim\{dS\}_{dM=0} + \{dS\}_{dE=0} = \dim\{dS\}_{dM=0} + \dim\{dS\}_{dE=0} - \dim\{dS\}_{dM=0} \cap \{dS\}_{dE=0}$$

Philipona et al. 2003

Sensorimotor Contingency

- The learned degrees of freedom



Characteristics	Organism 1	Organism 2	Organism 3
Dimensions of motor commands	40	100	100
Dimensions of exteroceptive inputs	40	80	80
Number of eyes	2	4	4
Diaphragms	None	Reflex	Controlled
Number of lights	3	5	5
Light luminance	Fixed	Variable	Variable
Dimensions found for body (p)	12	24	28
Dimensions found for environment (e)	9	20	20
Dimensions found for both (b)	15	38	41
Deduced dimension of rigid group (d)	6	6	7

Philipona et al. 2003

Sensorimotor Contingency

- But:
  - Some things are easier (upside down goggles). Some things are harder (luminance reversal).

In: Carpenter G, Grossberg S (Eds): Neural networks for vision and image processing. Bradford Books, MIT Press, 1992.

VISUAL ADAPTATION TO A NEGATIVE, BRIGHTNESS-REVERSED WORLD: SOME PRELIMINARY OBSERVATIONS

Stuart Anstis, Dept of Psychology, UC San Diego, La Jolla CA 92093

Abstract

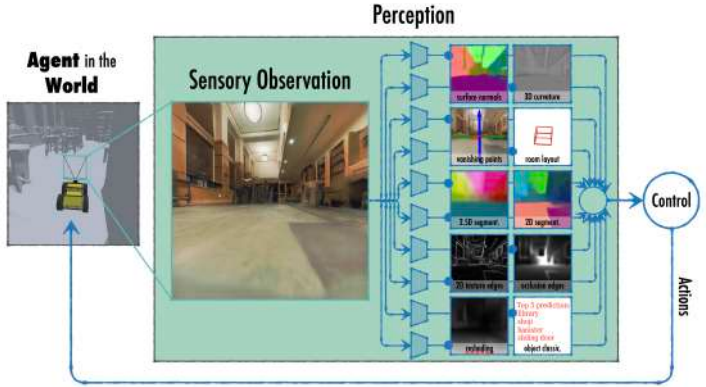
There have been many studies of visual adaptation to spatial rearrangements, starting with Stratton's (1897) classic studies on adaptation to an upside-down world. These have been reviewed by Rock (1966) and Howard (1982). Luminance information is crucial to such visual tasks as extracting shape from shading and recognising faces. If a picture of bumps and hollows is turned upside down, or reversed in brightness, the perceived depth reverses (Ramachandran 1988). Cavanagh and Leclerc (1989) have shown that shadows are treated as such only if they are darker than unshadowed regions. Extraction of depth by shape from shading seems to be an early process which precedes perceptual grouping (Ramachandran 1988a, b) and pop-out in visual search tasks (Enns 1990). Is shape from shading affected by perceptual experience? Hershenberg (1971*?) showed that chicks reared with grains lit from below preferred to peck at photographs of grains lit from below versus lit from above. If humans adapt to reversed luminance, will they "unlearn" that light comes from above, or that light is brighter than darkness?

Summary

- What is “correct”?
  - What appears to matter is the closed connection with downstream utility/actipn ⇒ things get (constantly) calibrated vs. being hard-wired to be the “correct” way
  - **Engineering implication**: close the connection with downstream utility of vision (and learning continually) vs. hard engineering a known configuration
    - (not the current “continual learning”)
    - Inductive biases still matter.

Summary

- (Again) vision should be viewed +action.
- The architectural connection between vision and action may be denser than what we think.
- Calibration, sensorimotor contingency, and representations
- Some structure and selective relearning is still in play.



Questions?

https://vilab.epfl.ch/