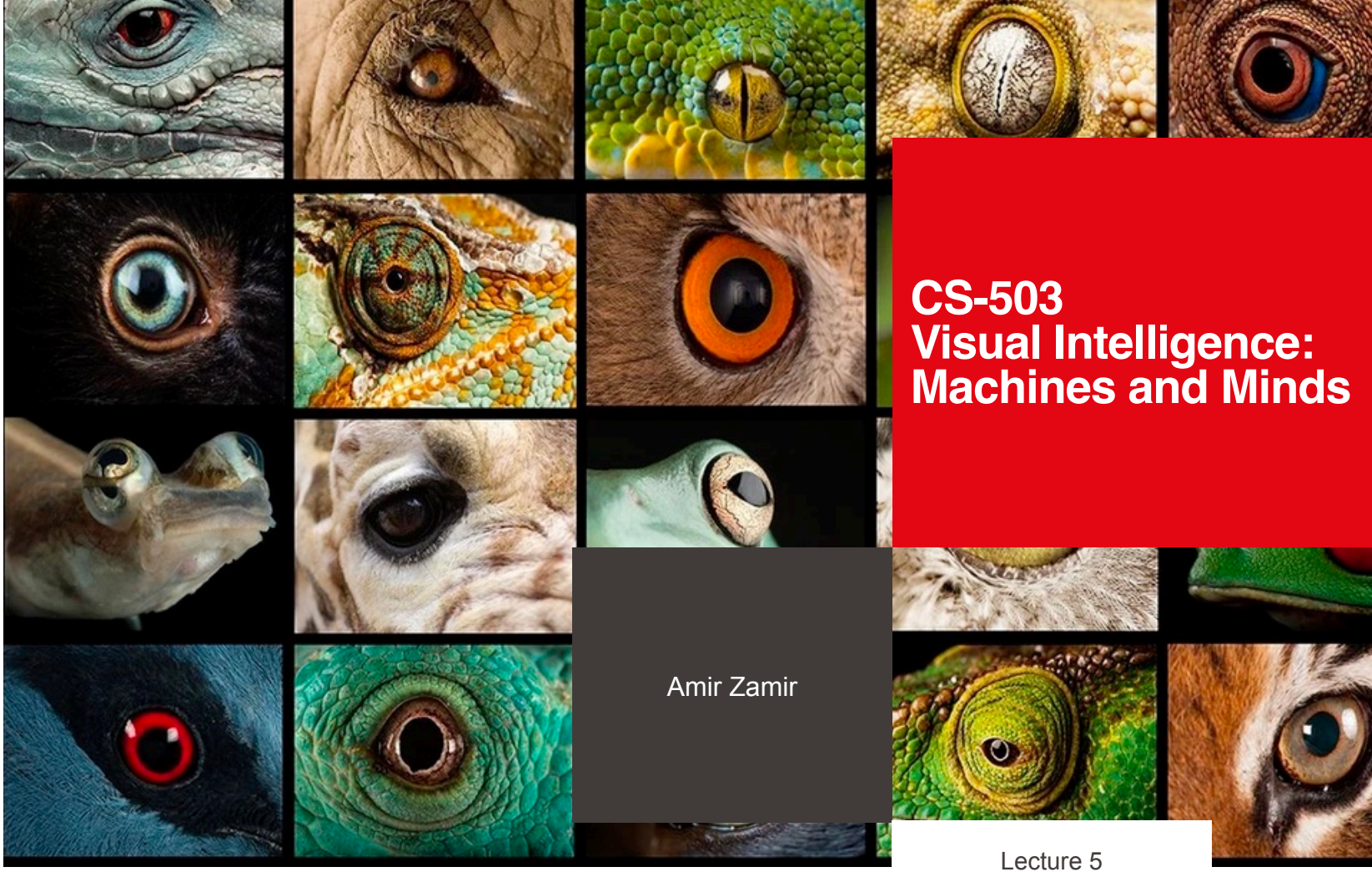


# CS-503
# Visual Intelligence: Machines and Minds

Amir Zamir

Lecture 5

École polytechnique fédérale de Lausanne

# Logistics

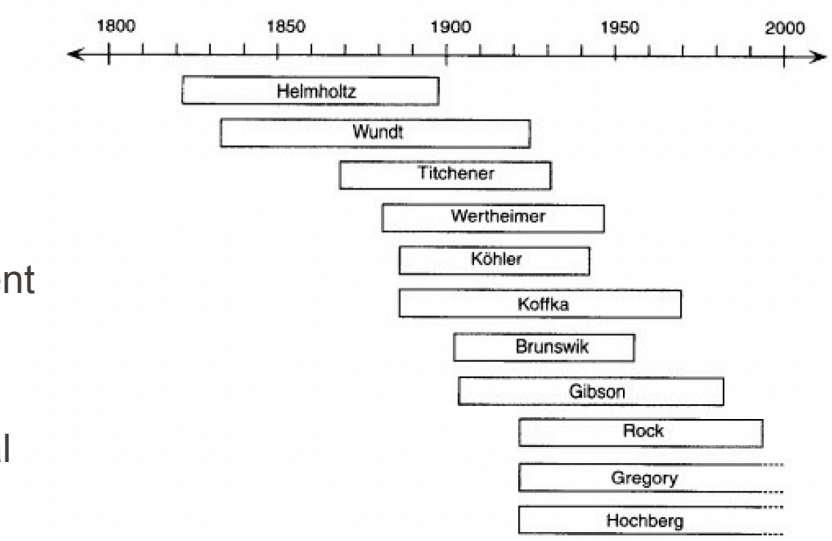
# Logistics

- First assignment notebook due 11/03/2025 23:59 CET.

| Week Num. | Date | Item |
|---|---|---|
| 1 | 20.02 | - lecture 1 |
| 2a | 25.02 | - lecture 2 |
| 2b | 27.02 | - lecture 3 |
| 3a | 04.03 | - lecture 4 |
| 3b | 06.03 | - lecture 5 |
| 4a | 11.03 | - lecture 6 (+ Q&A) |
| | 11.03 | - Transformers notebook assignment due |
| 4b | 13.03 | - lecture 7 |
| 5a | 18.03 | - lecture 8 |
| 5b | 20.03 | - lecture 9 |
| 6a | 25.03 | - lecture 10 |
| 6b | 27.03 | - lecture 11 (+ Q&A) |
| | 01.04 | - Active agents notebook assignment due |
| 7a | 01.04 | - lecture 12 |
| 7b | 03.04 | - lecture 13 |
| 8a | 08.04 | - lecture 14 |
| 8b | 10.04 | - lecture 15 (+ Matchmaking session) |
| | 13.04 | - Project proposals due |
| | 15.04 | - all subsequent sessions from 15.04 onwards are for Q&A |
| | 18.04 | - Project proposals due, when revision is needed. |
| | 22.04 | - MidSem break - No classes |
| | 25.04 | - MidSem break - No classes |
| | 29.04 | - Foundation Models assignment due |
| | 01.05 | - lecture 16 |
| | 09.05 | - Project progress report due |
| | 13.05 | - Robustness assignment due (extra credit) |
| | 20.05 | - Moodle homework due |
| | 26.05 | - Final project presentation video due |
| | 27.05 | - Final project presentation Part I |
| | 29.05 | - Final project presentation Part II |
| | 30.05 | - Project report due |

# Recap

# Vision theories




- What's a "theory"? An integrated/consistent set of statements/hypotheses about underlying principles of something.
  - That not only organizes and explains known facts (eg existing experimental results), but also makes predictions about new ones.

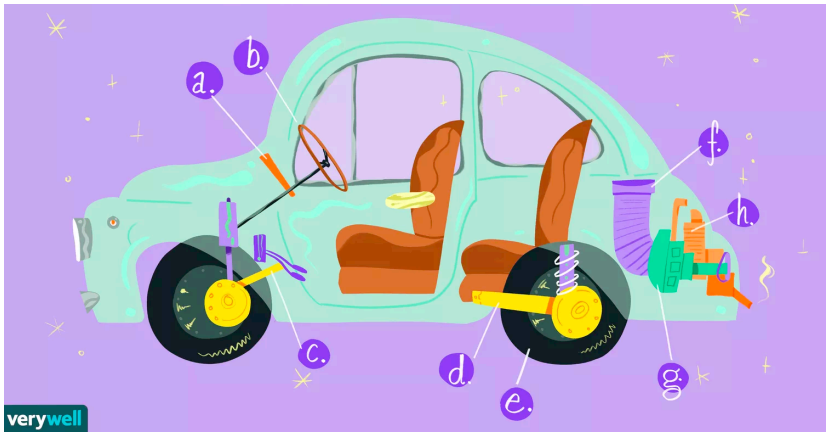
| THEORY | NATIVISM vs. EMPIRICISM | ATOMISM vs. HOLISM | ORGANISM vs. ENVIRONMENT | PRINCIPAL ANALOGY | METHOD |
|---|---|---|---|---|---|
| Structuralism | Empiricism | Atomism | Organism | Chemistry | Trained Introspection |
| Gestaltism | Nativism | Holism | Organism | Physical Field Theory | Naive Introspection |
| Ecological Optics | Nativism | Holism | Environment | Mechanical Resonance | Stimulus Analysis |

Stephen Palmer

# Structuralism



- Progressive “concatenation" of “sensory atoms”

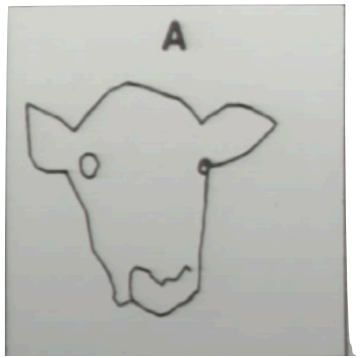


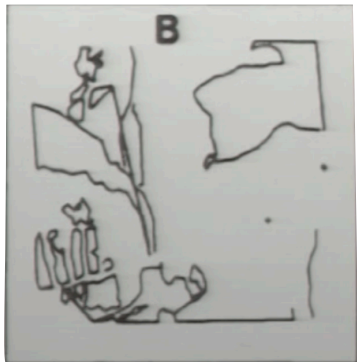

Wilhelm Wundt

# Gestaltism



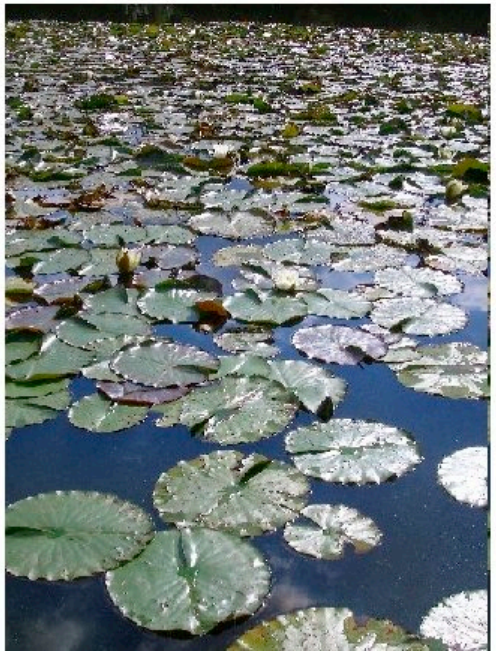
- Whole is more than the sum of parts.

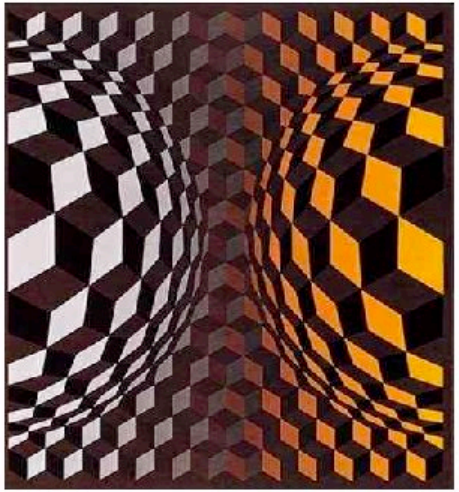



# Ecological Approach

- Gibson: “Ask not what's inside your head, but what your head's inside” (Mace, 1977)

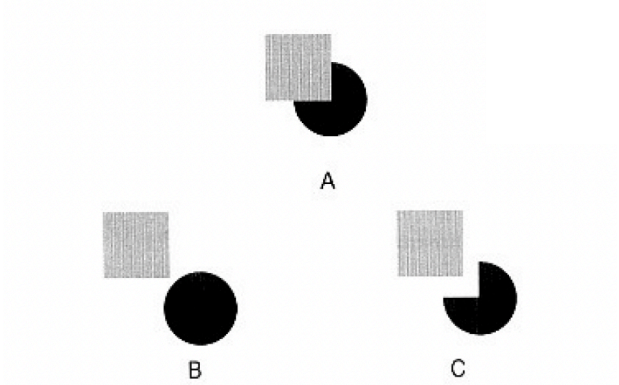


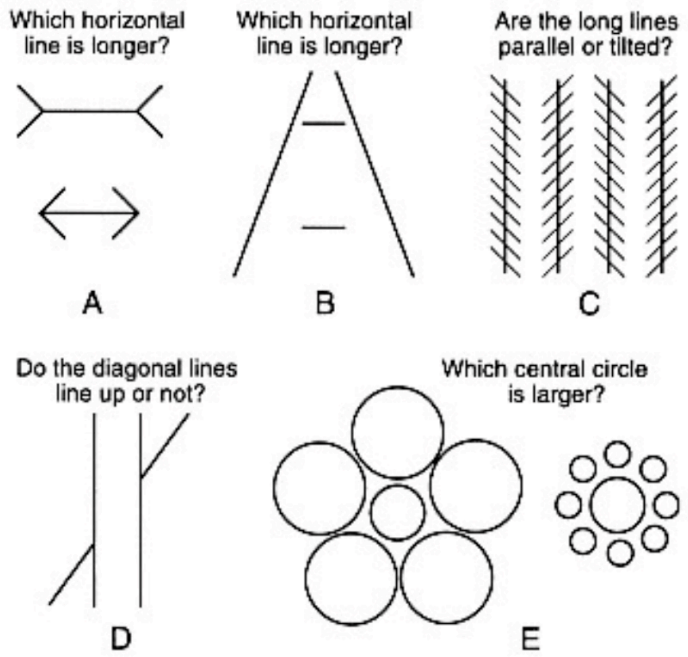


# Constructivism



- Vision:
  - an indeterminate inverse problem from retinal images.
  - a “reconstruction” of the reality.
- Something besides the retinal image is needed.
- Likelihood Principle



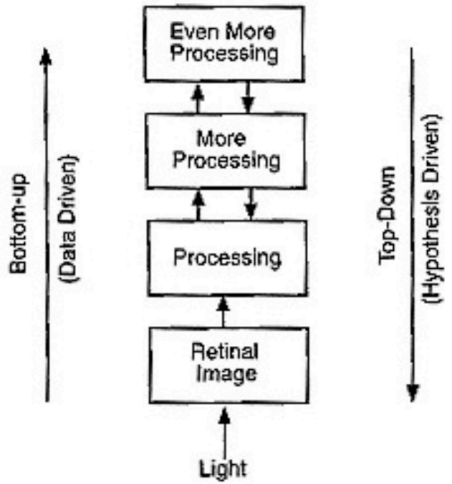



# Feedback



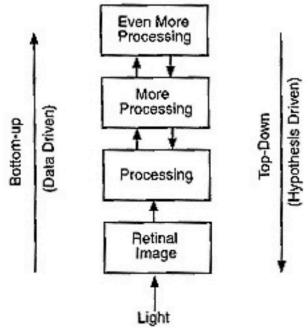
- Vision appears to be more than bottom-up association.

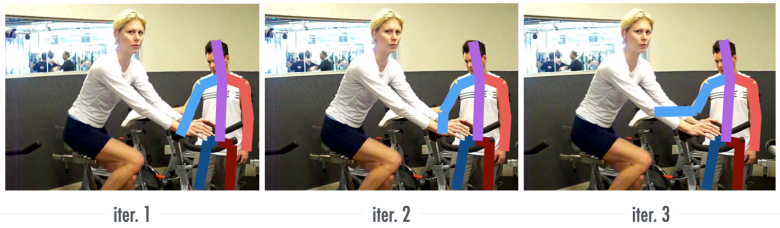
Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.



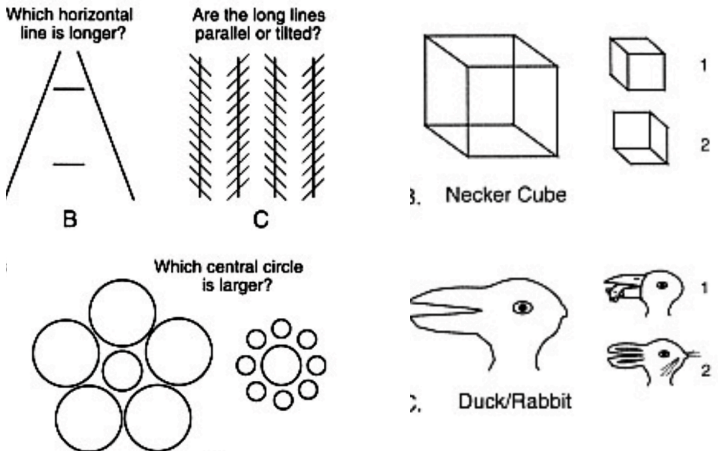
# Feedback

- Feedback:
  - 1) conditional processing
  - 2) hypothesis/expectation driven processing of lower representation.
- Vision appears to be more than bottom-up association.



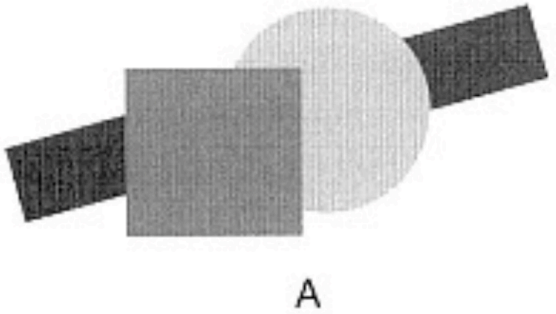

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.



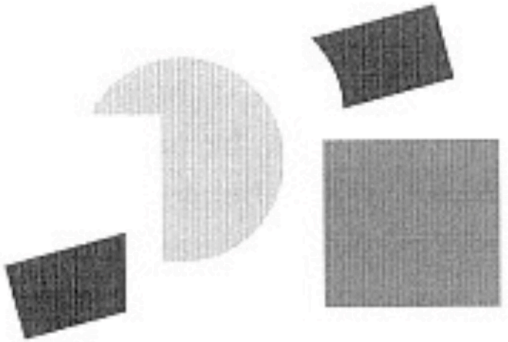
- Vision: A *useful* *reconstruction* of the world in a bottom-up and *top-down* way.

# Perception as modeling the environment

- *The observer is constructing a **model** of what environment situation might have produced the observed pattern of sensory stimulation*



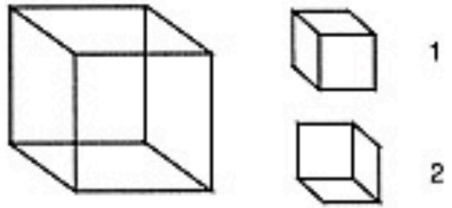
A



B

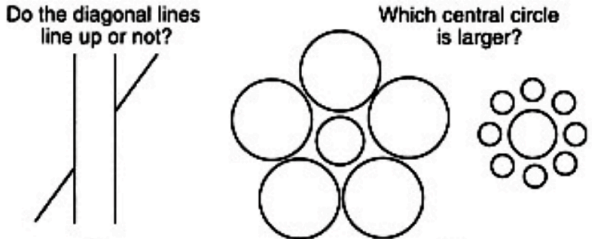
# Perception as modeling the environment

- *The observer is constructing a **model** of what environment situation might have produced the observed pattern of sensory stimulation*
- **Visual illusions**: the model is sometimes inaccurate.
- **Ambiguous figures**: the model is sometimes not unique.




Necker Cube


Duck/Rabbit



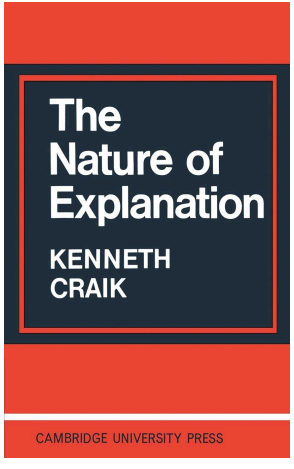
# Why/how does this all matter to us?

# Why/how does this all matter to us?

- Making a model: why?
- Visual “Completion”
- Look around
- Evolutionary benefit (e.g. energy)
- Planning
- “Completion”
  - ~prediction
  - ~in-painting/filling
  - ~model (eg “statistical model”)
  - ~assumption

# Mental Models

Kenneth Craik (1943)

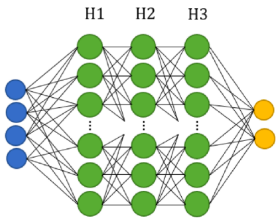



- *"If the organism carries a "small-scale model" of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilise the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it."*
- Model:
  - something that parallels a reality.
  - enables prediction, planning, and counterfactual reasoning/ imagination.
    - (vs reactive)
  - keeps the relevant aspects and simplifies others.

# Models

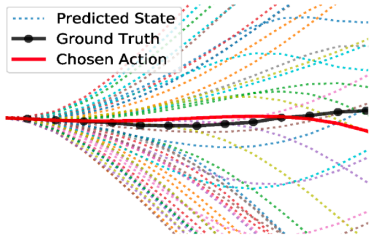


- Model-Based Reinforcement Learning



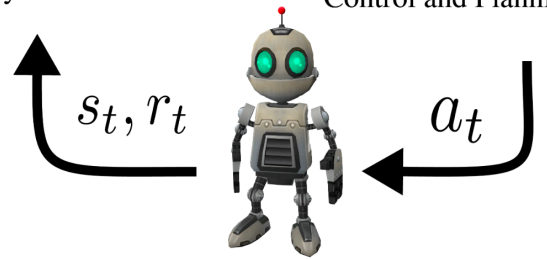

$$s_{t+1} = f_\theta(s_t, a_t)$$

Dynamics Model




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

Control and Planning




While improving:

1. Agent acts in environment
2. Learn model of dynamics

$$p_\theta = \arg\max_\theta \sum_{i=1}^{N} \log\, p_\theta(s_{t+1}|s_t, a_t)$$

3. Plan actions to maximize reward

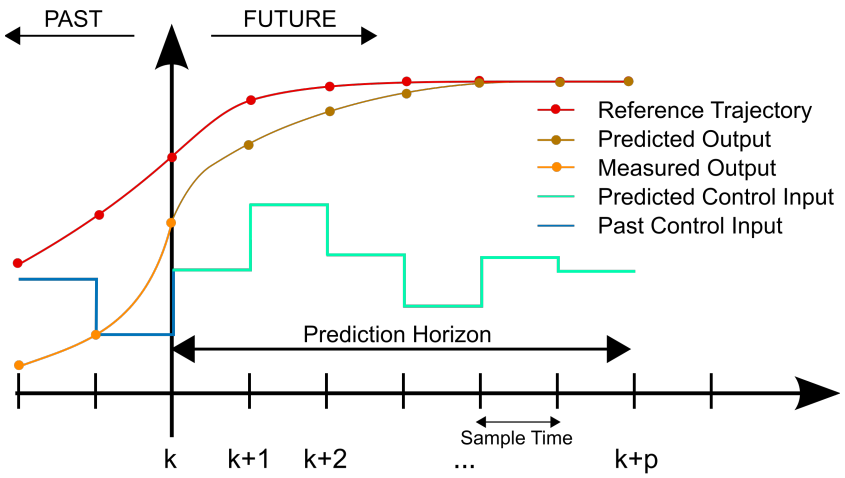
$$a^* = \arg\max_a \sum_{t=0}^{T} \gamma^t r(s_t, a_t)$$

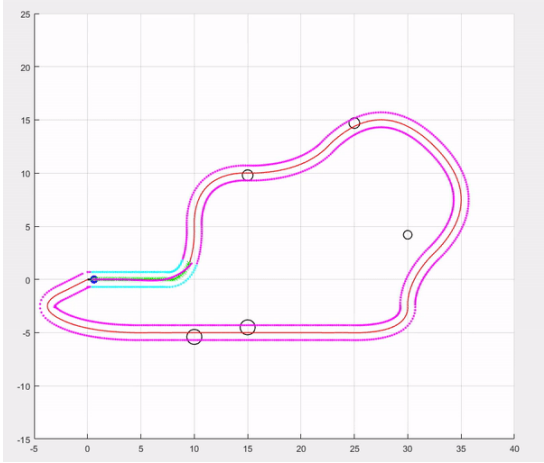
$$s.t.\ s_{t+1} \sim p_\theta(s_{t+1}|s_t, a_t)$$

# Models



- Model Predictive Control (MPC)
  - Act for the current time while using a model to plan accounting for longer future.



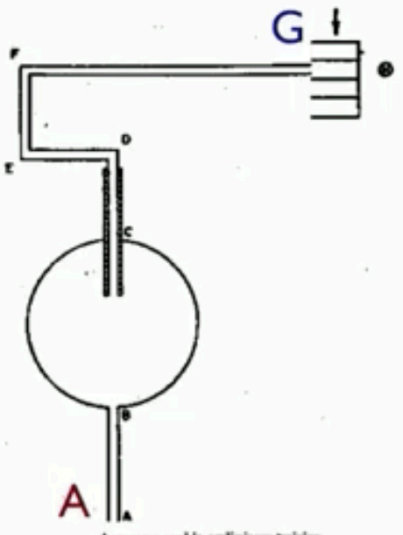
# Cognitive Maps (1948)

Zamir

## COGNITIVE MAPS IN RATS AND MEN [1]

BY EDWARD C. TOLMAN

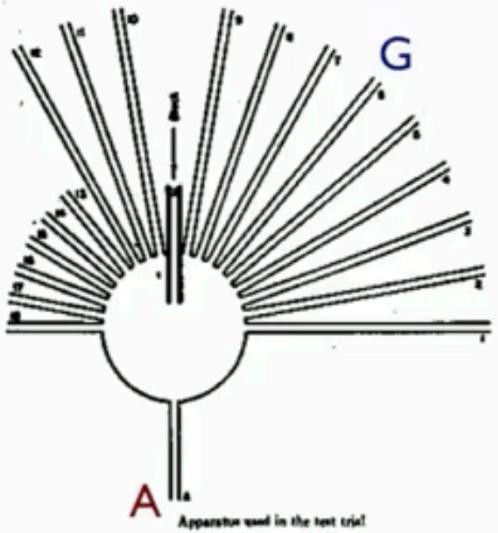
*University of California*




Apparatus used in preliminary training

FIG. 15

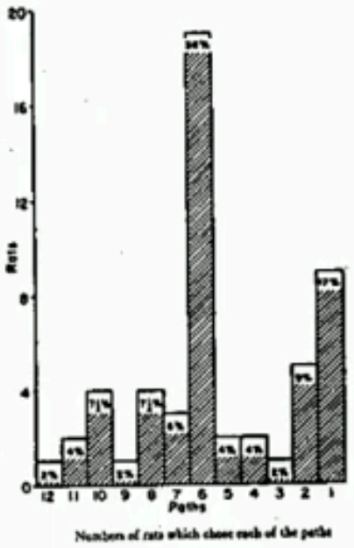
(From E. C. Tolman, B. F. Ritchie and D. Kalish, Studies in spatial learning. I. Orientation and the short-cut. *J. exp. Psychol.*, 1946, 36, p. 16.)




Apparatus used in the test trial

FIG. 16

(From E. C. Tolman, B. F. Ritchie and D. Kalish, Studies in spatial learning. I. Orientation and short-cut. *J. exp. Psychol.*, 1946, 36, p. 17.)

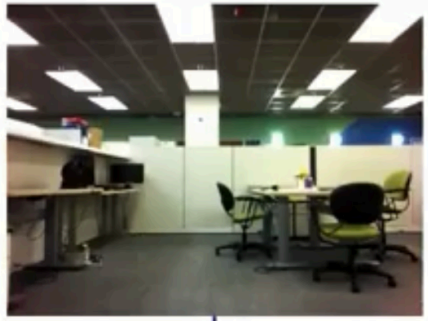



Numbers of rats which chose each of the paths

FIG. 17

(From E. C. Tolman, B. F. Ritchie and D. Kalish, Studies in spatial learning. I. Orientation and the short-cut. *J. exp. Psychol.*, 1946, 36, p. 19.)

# Computational Cognitive Maps

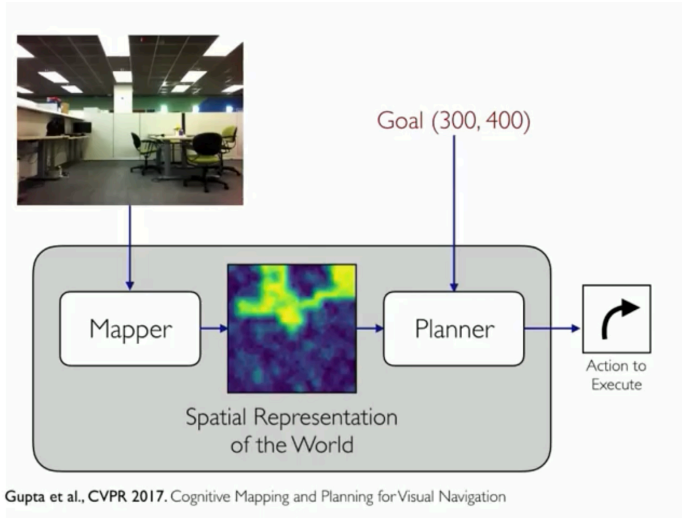




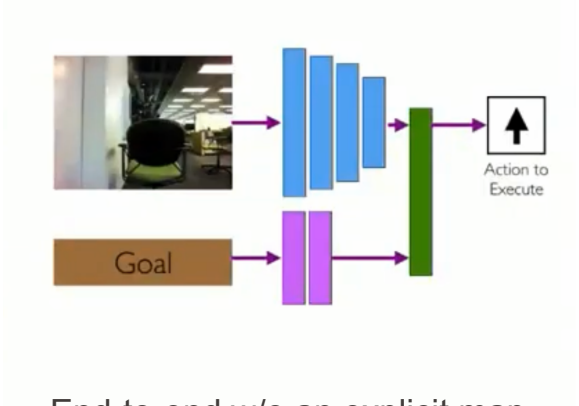

**Gupta et al., CVPR 2017.** Cognitive Mapping and Planning for Visual Navigation

# Computational Cognitive Maps

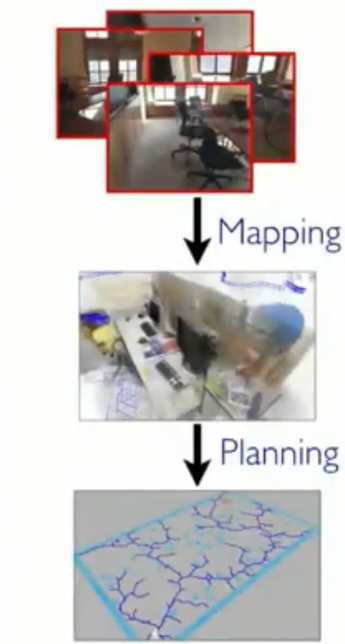





End-to-end w/ a map




End-to-end w/o an explicit map




Not end-to-end
(e.g. Classical SLAM)

# Latent Learning
## Tolman & Honzik 1930

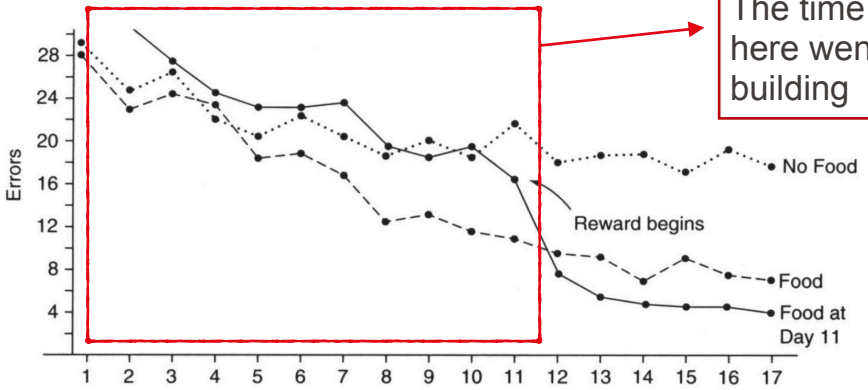


**Group 1**: Rewarded: Day 1 – 17: Every time they got to end, given food (i.e. reinforced).

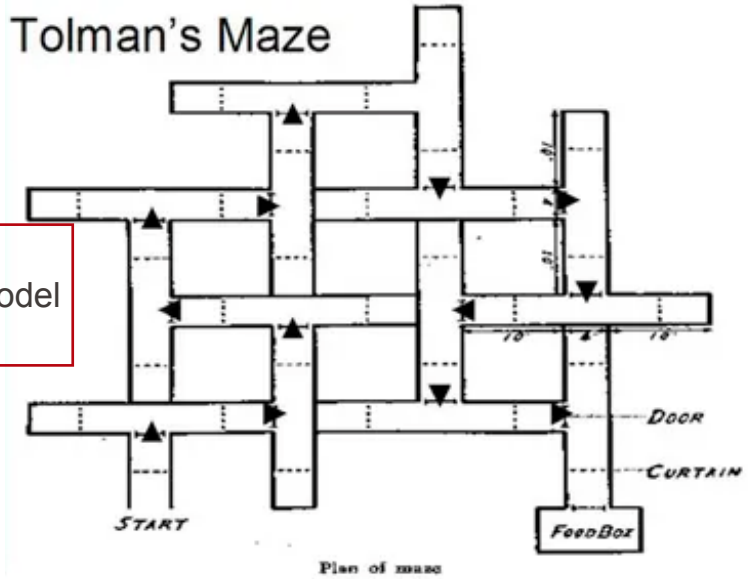
**Group 2**: Delayed Reward

- Day 1 - 10: Every time they got to end, taken out.
- Day 11 -17: Every time they got to end, given food (i.e. reinforced).

**Group 3**: No reward: Day 1 – 17: Every time they got to end, taken out.



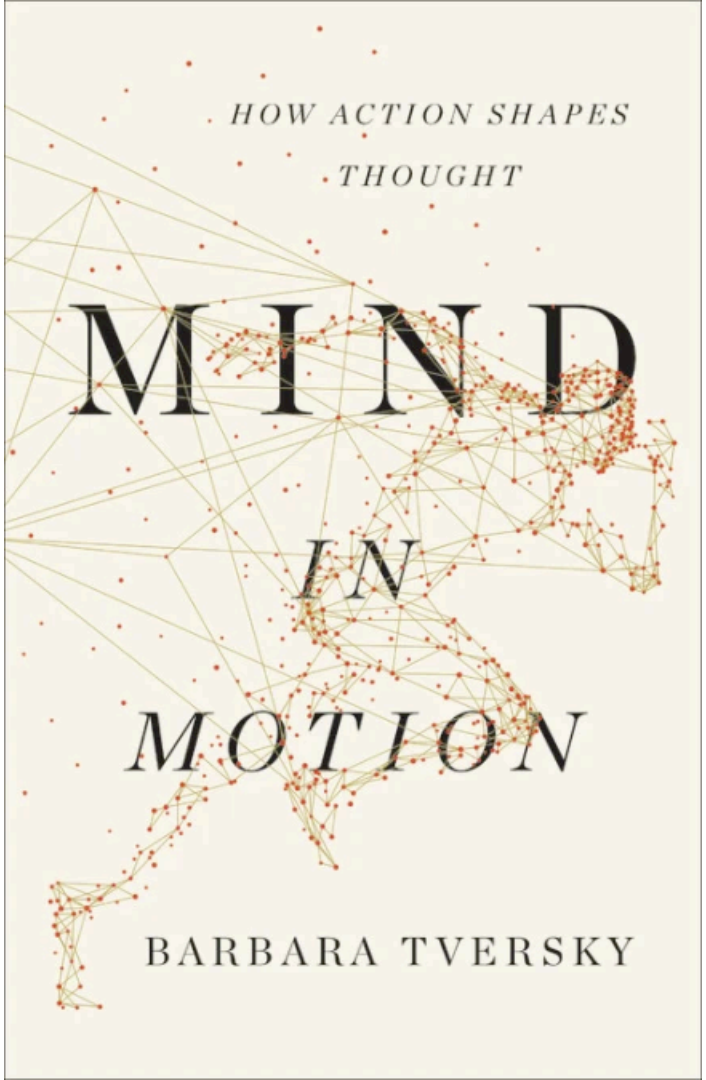

The time spent here went to model building



# Quick segue: “Spatial Thinking”

- A map is a special case of spatial organization of information.

Barbara Tversky



# Spatial Thinking

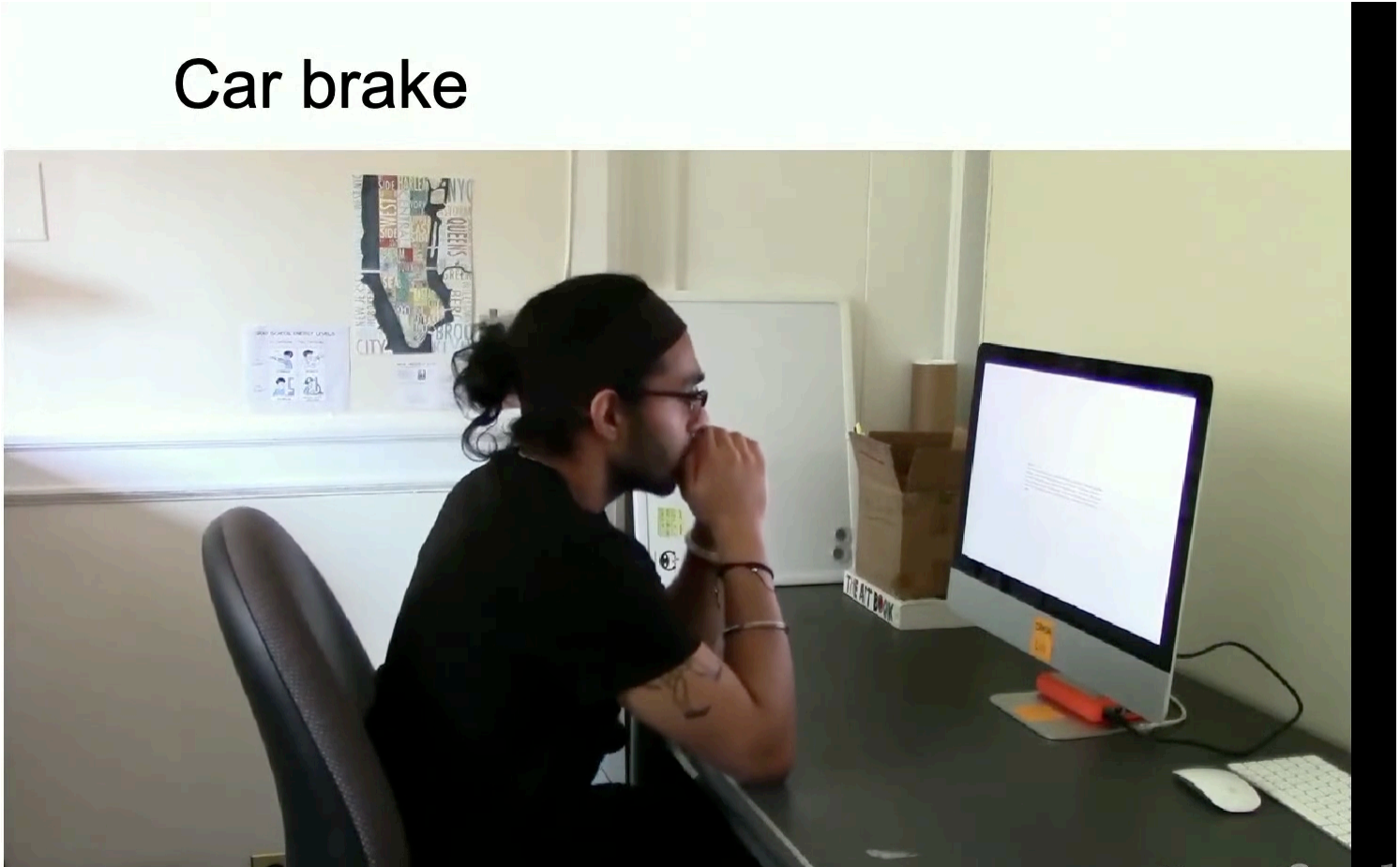


Car Brake

From the brake fluid reservoir, brake fluid enters and travels sideways and down the tube. As the brake fluid accumulates at the bottom of the tube, pressure is exerted on the small pistons inside the wheel cylinders. This causes the pistons to push outward toward the brake drum. The outward movement of the shoes causes friction along the inside of the brake drum, slowing the rotation of the wheel.

Barbara Tversky

# Spatial Thinking



# Spatial Thinking



Small environment, survey description

Etna is a charming town nestled in an attractive valley, entered on River Highway. River Highway runs east-west at the southern edge of the town of Etna. Toward the eastern border, River Highway intersects with Mountain Rd, which runs north of it. At the northwest corner of the intersection is a gas station. North of the gas station, Mountain Road will intersect with Maple Ave, which runs west.

Barbara Tversky

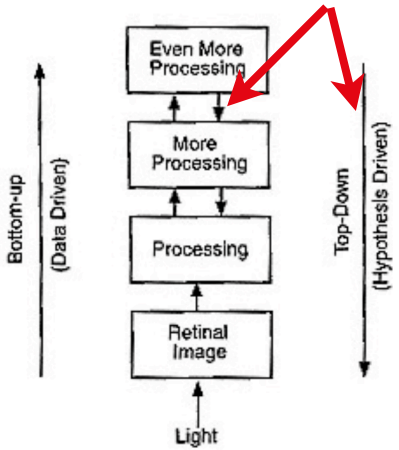
# Spatial Thinking



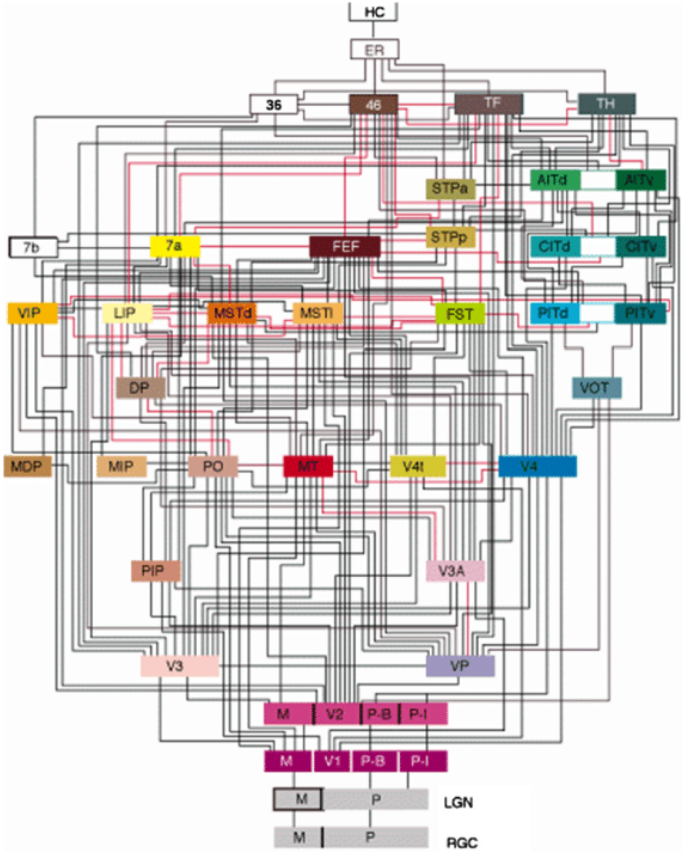


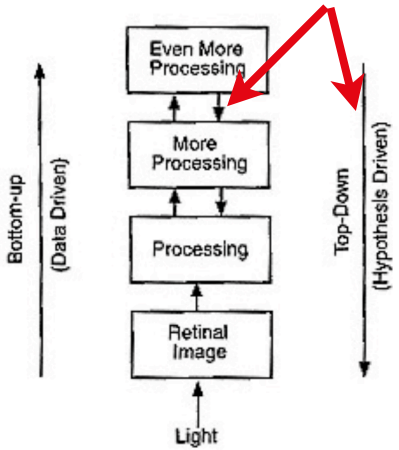
# Feedback

# Feedback


Even More Processing
More Processing
Processing
Retinal Image
Light
Bottom-up
(Data Driven)
Top-Down
(Hypothesis Driven)

# Feedback



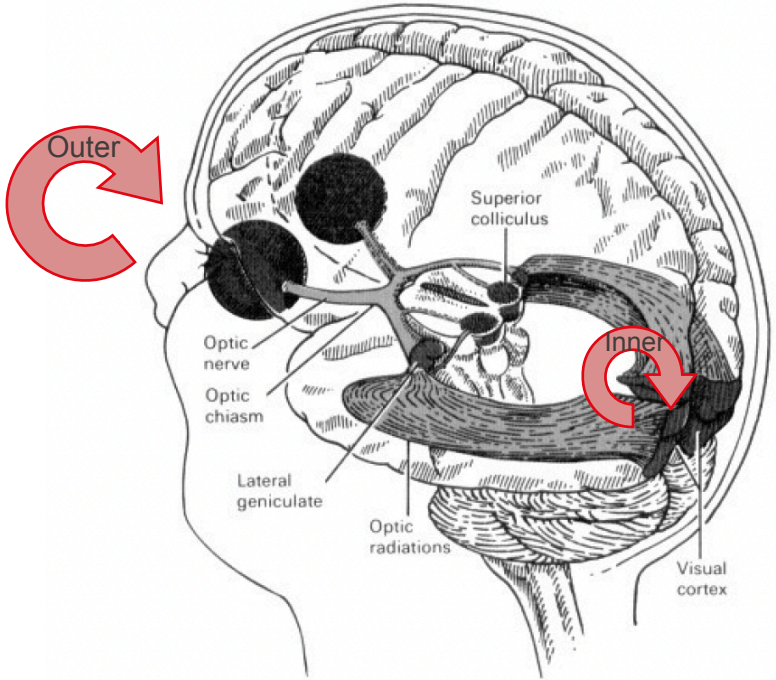




- Felleman & Van Essen (1991)

# Feedback

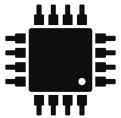


- Inner loop
  - top-down processing without external feedback from the world.
  - e.g. IEF (iterative error feedback, 2016), Attention, Feedback Networks (2017), diffusion.



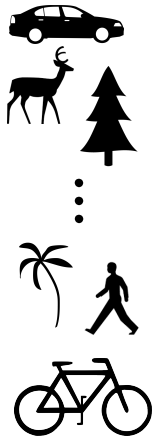
# *Prediction on a Budget*

*Time*



*Resources*

~120 kmh



1 second
33 meters

# Prediction on a Budget

Time

Resources

A tradeoff

~120 kmh

✗ foliage

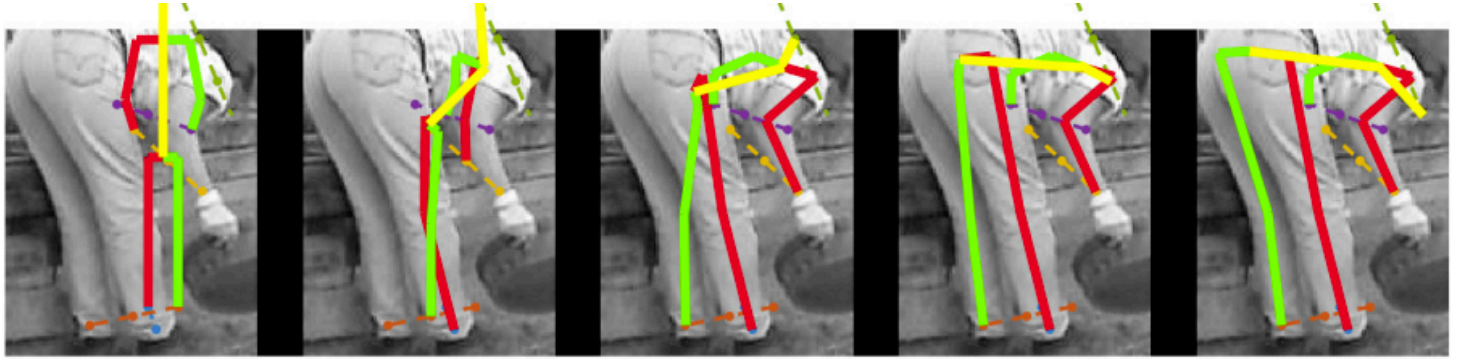
tandem bike

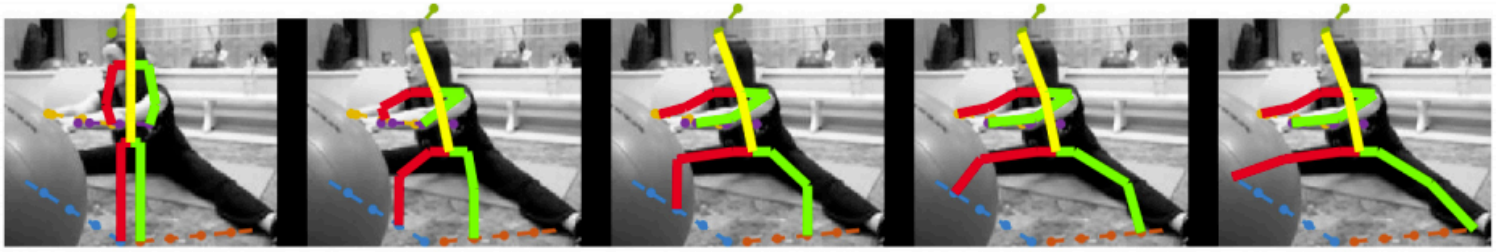
✓ wheeled vehicle

road bike

flat bar bike

# Feedback



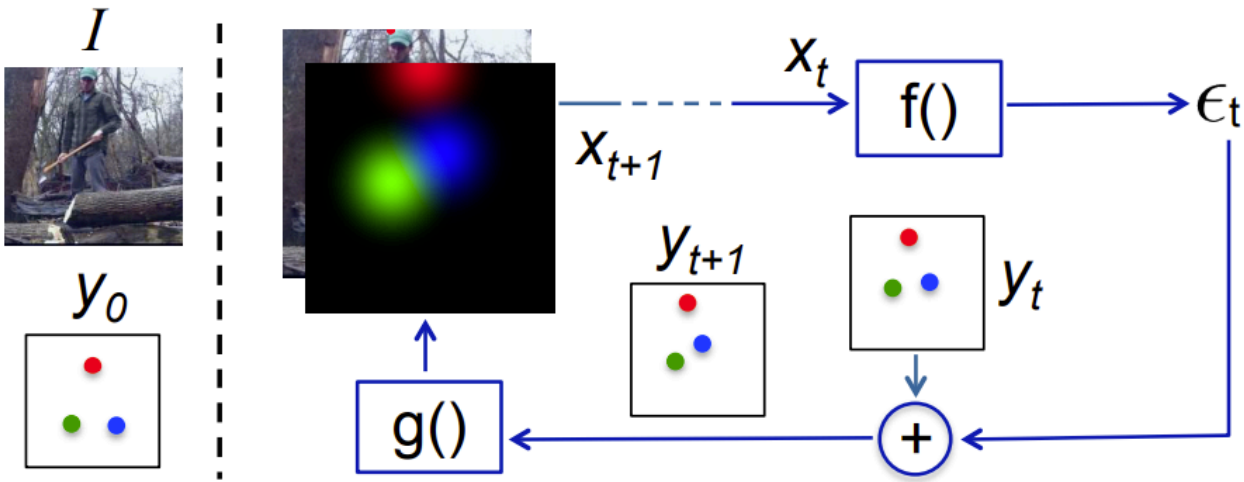


| | Head | Shoulder | Elbow | Wrist | Hip | Knee | Ankle | UBody | FBody |
|---|---|---|---|---|---|---|---|---|---|
| Iterative Error Feedback (IEF) | 95.2 | 91.8 | 80.8 | 71.5 | 82.3 | 73.7 | 66.4 | 81.4 | 81.0 |
| Direct Prediction | 92.9 | 89.4 | 74.1 | 61.7 | 79.3 | 64.0 | 53.3 | 75.1 | 74.8 |
| Iterative Direct Prediction | 91.9 | 88.5 | 73.3 | 59.9 | 77.5 | 61.2 | 51.8 | 74.0 | 73.4 |

Human Pose Estimation with Iterative Error Feedback, J Carreira, P Agrawal, K Fragkiadaki, J Malik, CVPR 2016

# Feedback



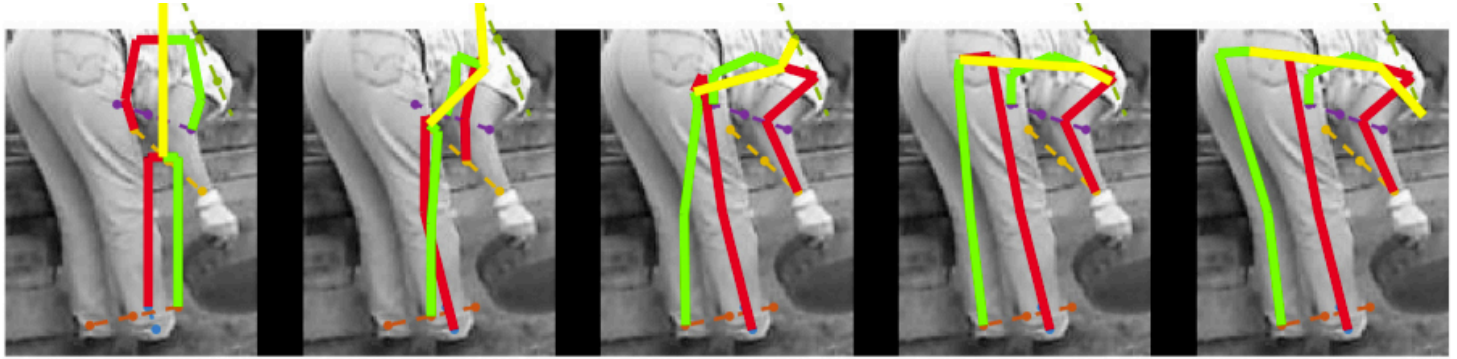


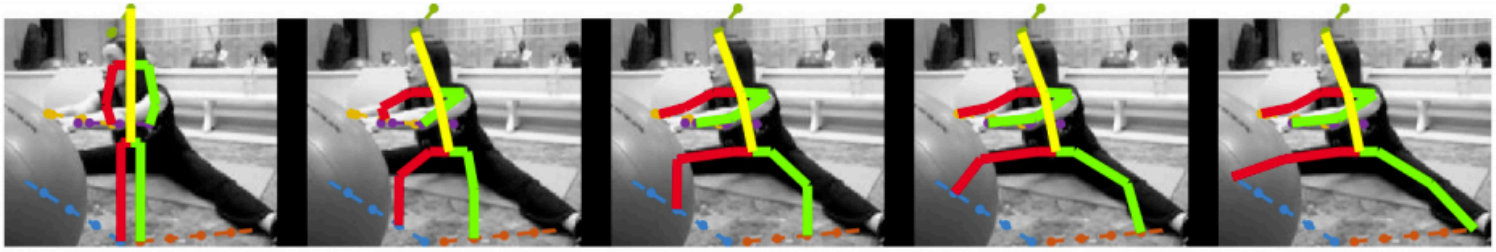

- $\epsilon_t = f(x_t)$
- $y_{t+1} = y_t + \epsilon_t$
- $x_{t+1} = I \oplus g(y_{t+1})$

$$\min_{\Theta_f, \Theta_g} \sum_{t=1}^{T} h(\epsilon_t, e(y, y_t))$$

Human Pose Estimation with Iterative Error Feedback, J Carreira, P Agrawal, K Fragkiadaki, J Malik, CVPR 2016

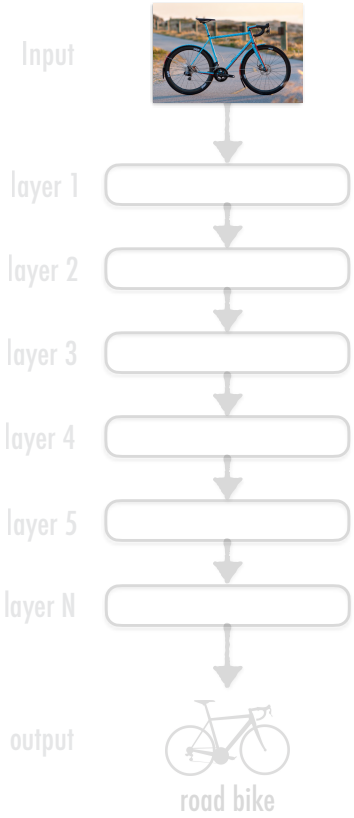
# Feedback





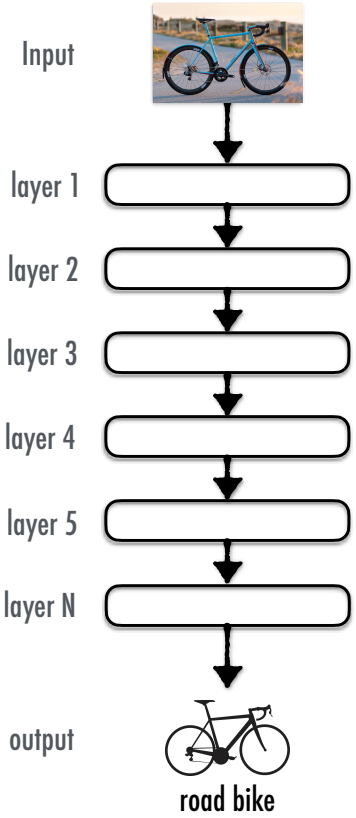
| | Head | Shoulder | Elbow | Wrist | Hip | Knee | Ankle | UBody | FBody |
|---|---|---|---|---|---|---|---|---|---|
| Iterative Error Feedback (IEF) | 95.2 | 91.8 | 80.8 | 71.5 | 82.3 | 73.7 | 66.4 | 81.4 | 81.0 |
| Direct Prediction | 92.9 | 89.4 | 74.1 | 61.7 | 79.3 | 64.0 | 53.3 | 75.1 | 74.8 |
| Iterative Direct Prediction | 91.9 | 88.5 | 73.3 | 59.9 | 77.5 | 61.2 | 51.8 | 74.0 | 73.4 |

Human Pose Estimation with Iterative Error Feedback, J Carreira, P Agrawal, K Fragkiadaki, J Malik, CVPR 2016



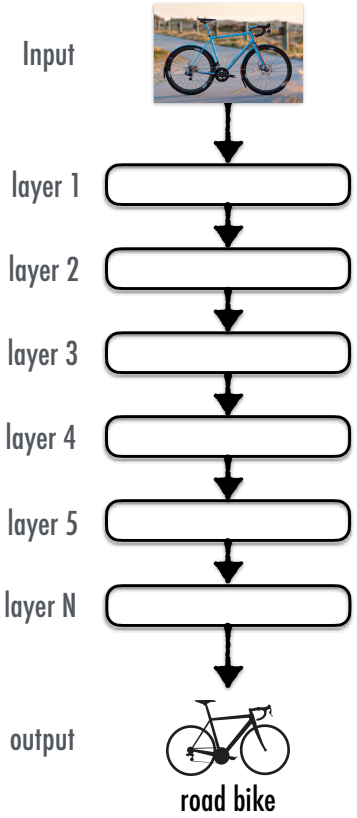

Feedforward model.

Feedback Networks, CVPR 2017



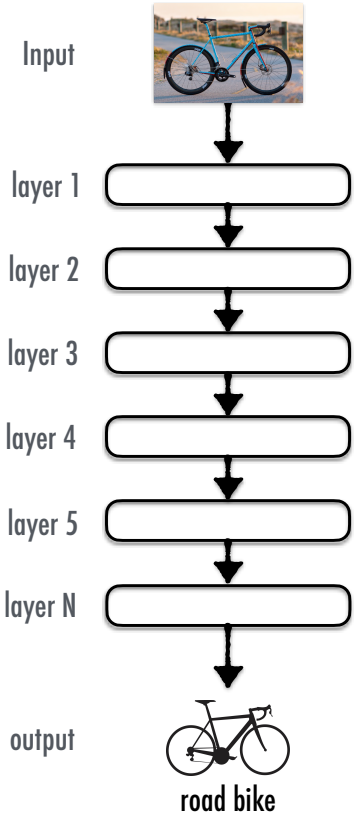

Feedforward model.

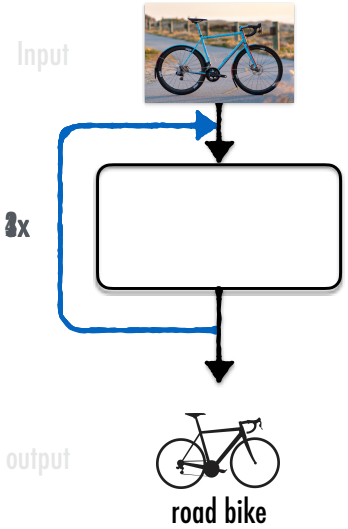
Feedback Networks, CVPR 2017




*Feedforward model.*

Feedback Networks, CVPR 2017

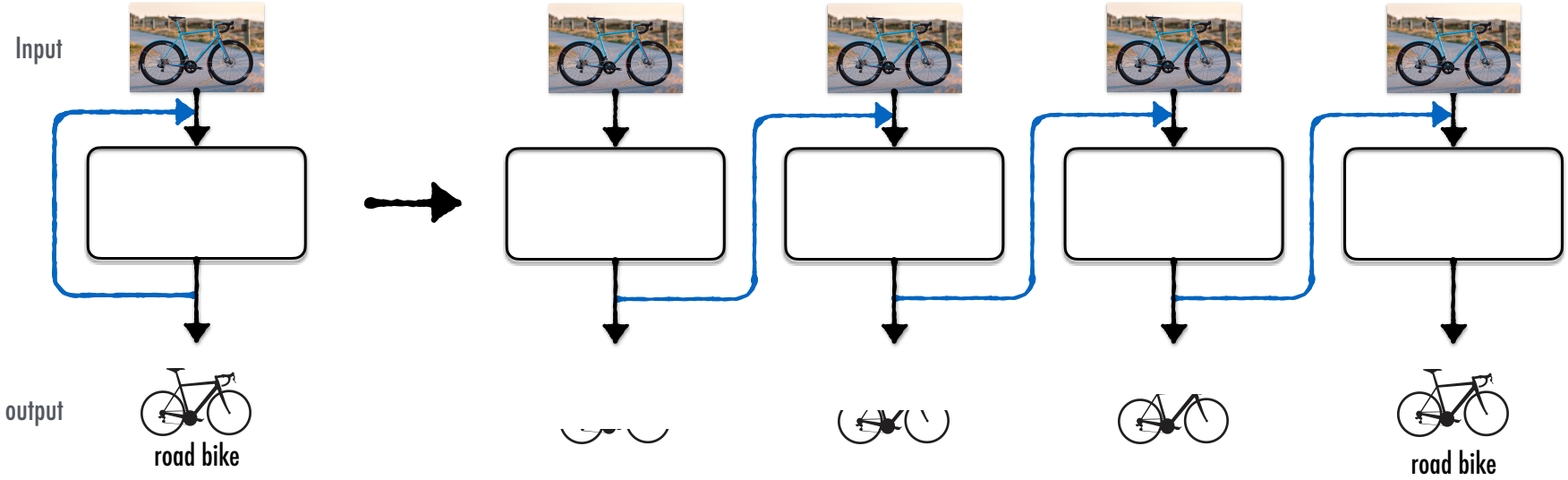


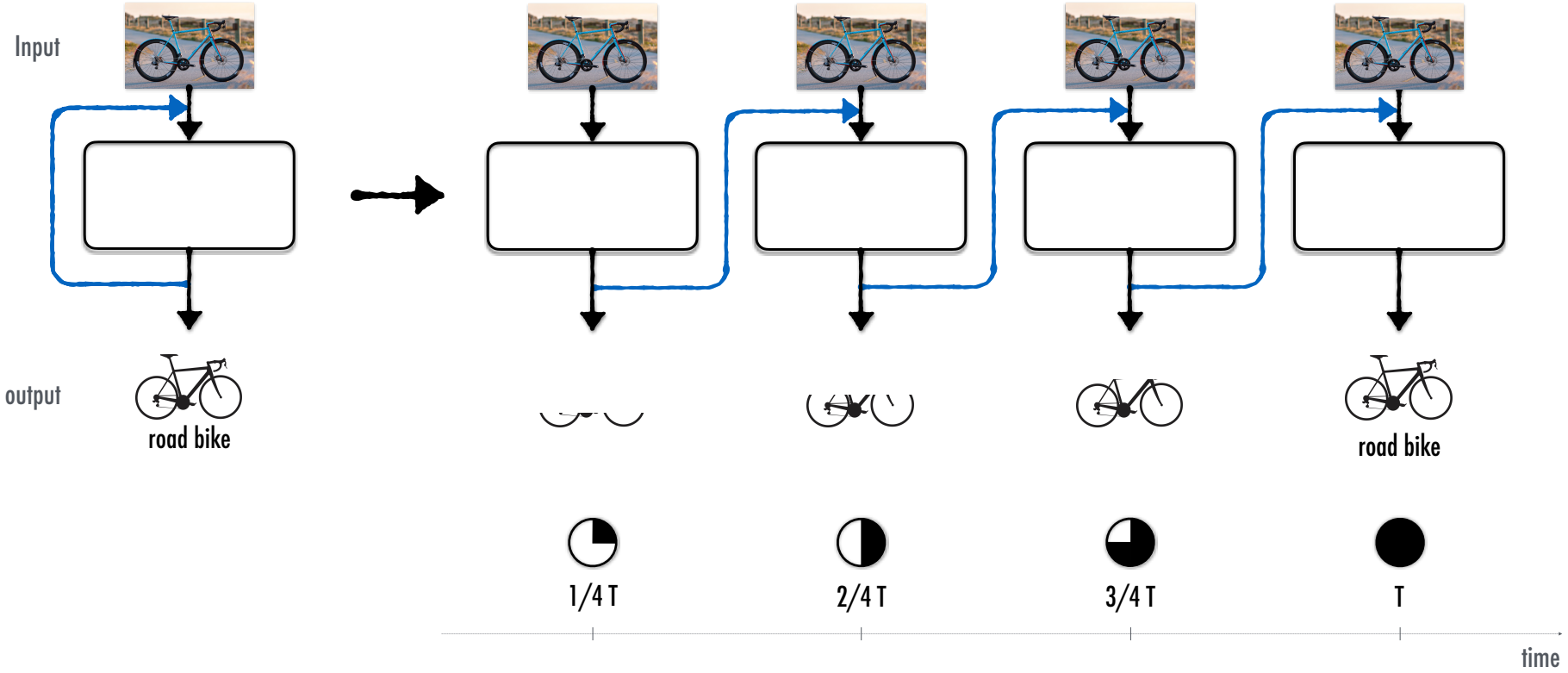

Feedforward model.

Feedback model.

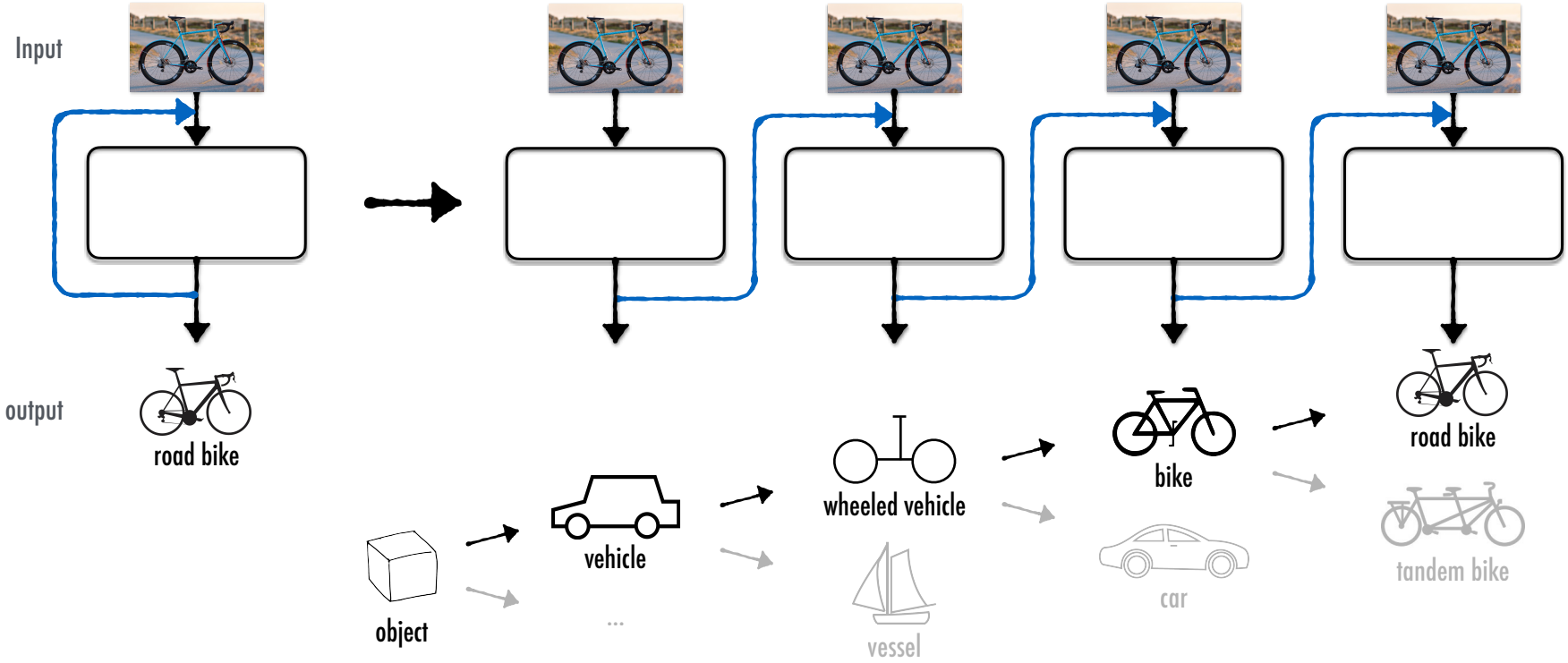
Feedback Networks, CVPR 2017




*Feedback model unrolled.*

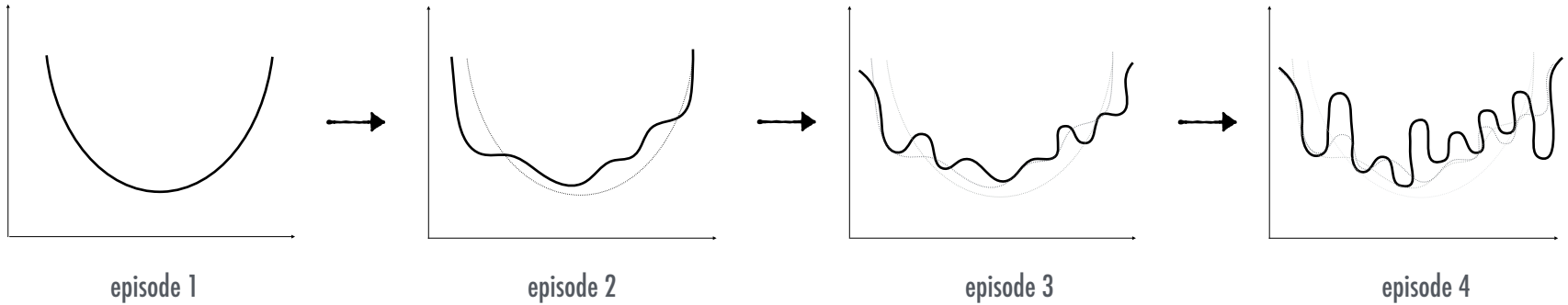



# *Advantage I: Early Prediction*



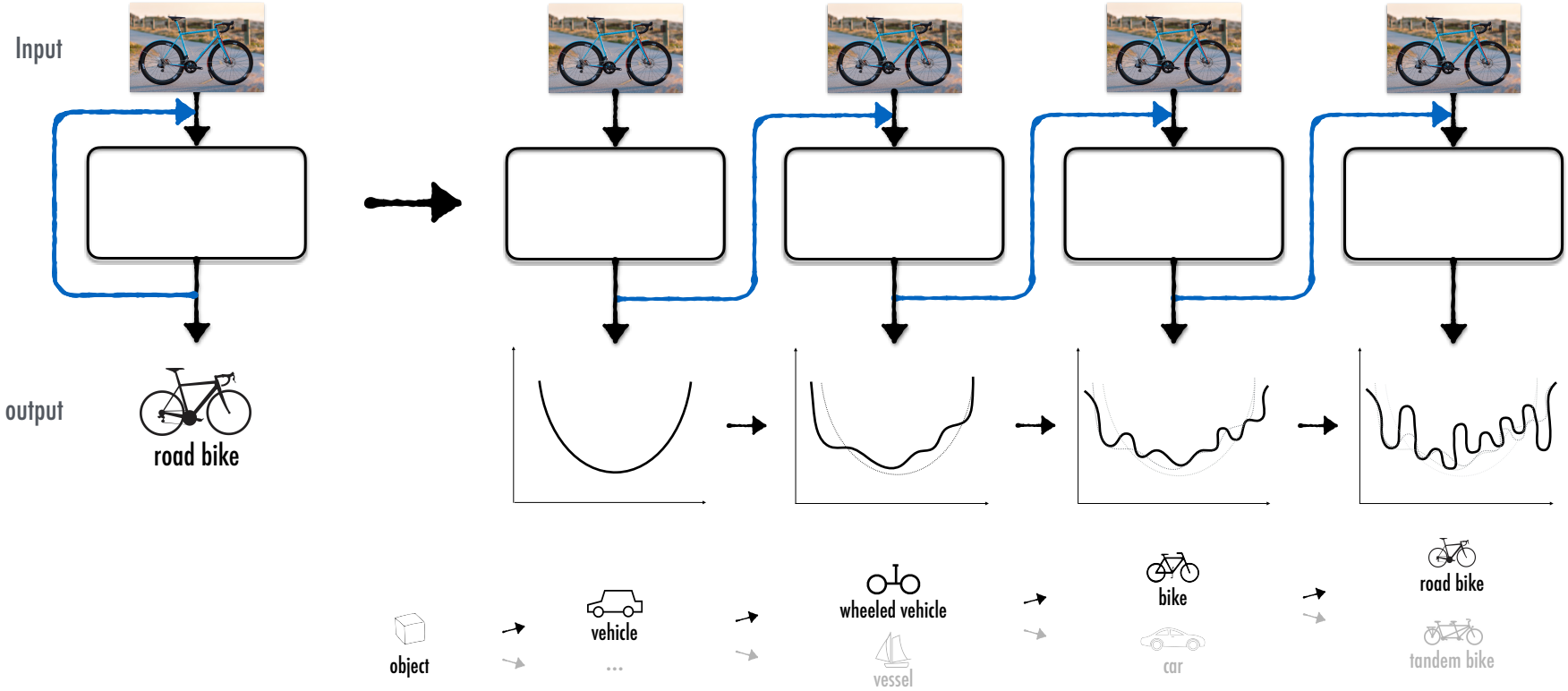

# *Advantage II: Taxonomic Prediction*

Objective function

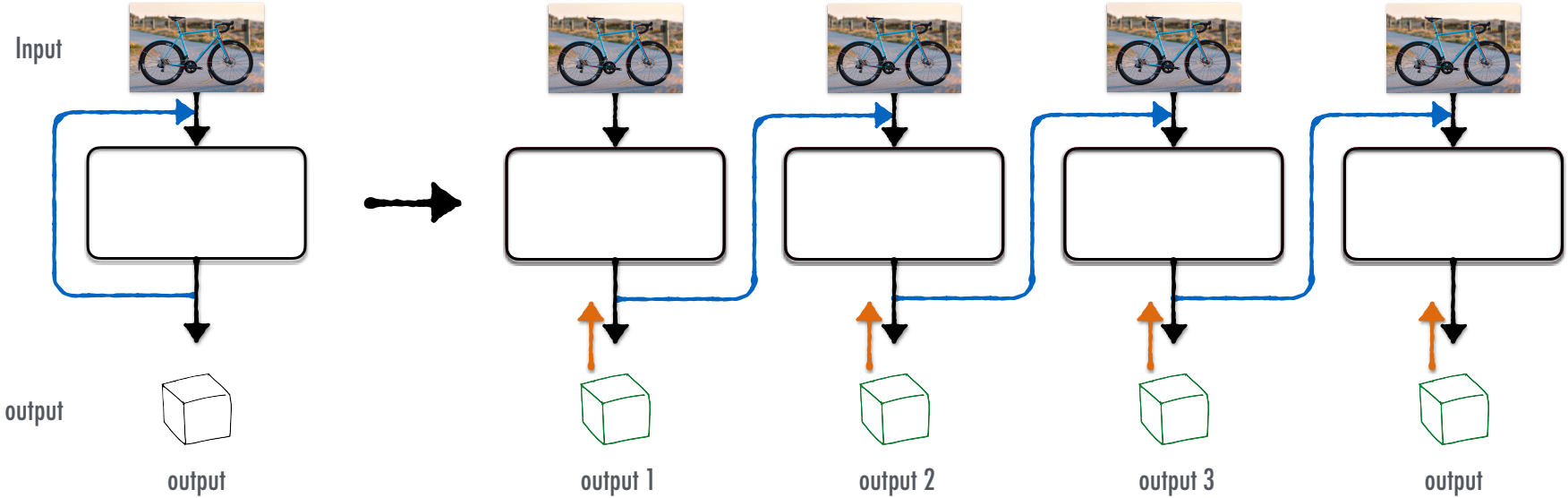



# *Episodic Curriculum Learning*

*Bengio et al., 2009*



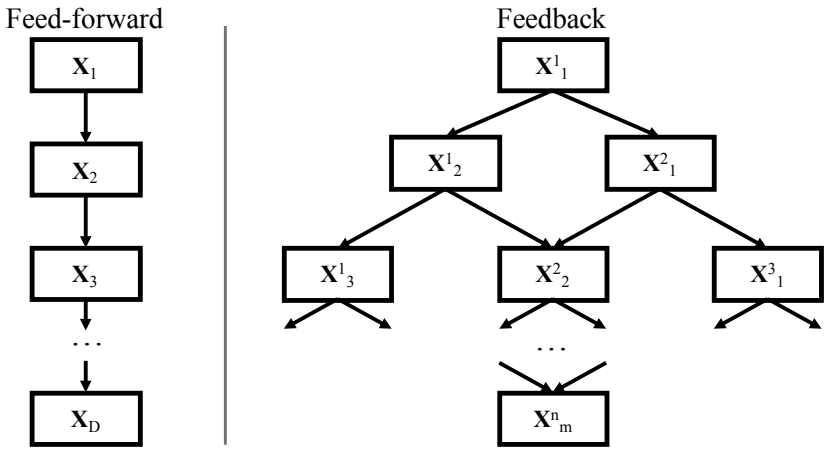

# *Advantage III: Episodic Curriculum Learning*

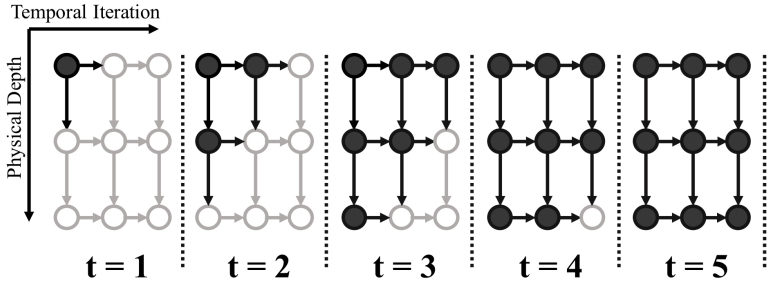



*Feedback requirements:*

1. *Recurrence ⇒ Recurrent Neural Networks (we used ConvLSTM)*
2. *Notion of output at each iteration ⇒ connected loss & back propagation per iteration*

# *The computation graph*

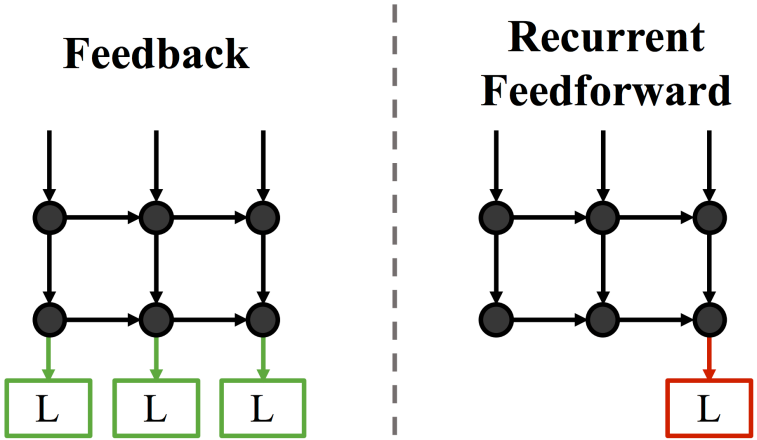



$$d_{ff} = mn - 1 > m + n - 1 = d_{fb}$$

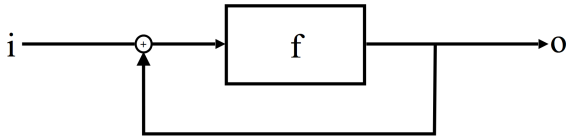
# *computational advantages*

Feedback Networks, CVPR 2017

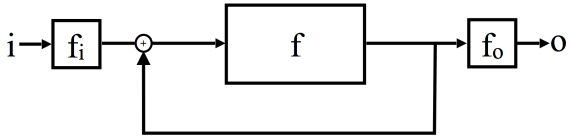



*Feedback vs Recurrent Feedforward*

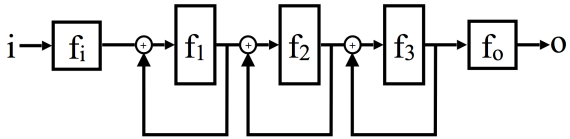
Feedback Networks, CVPR 2017




(a) Feedback via observed state



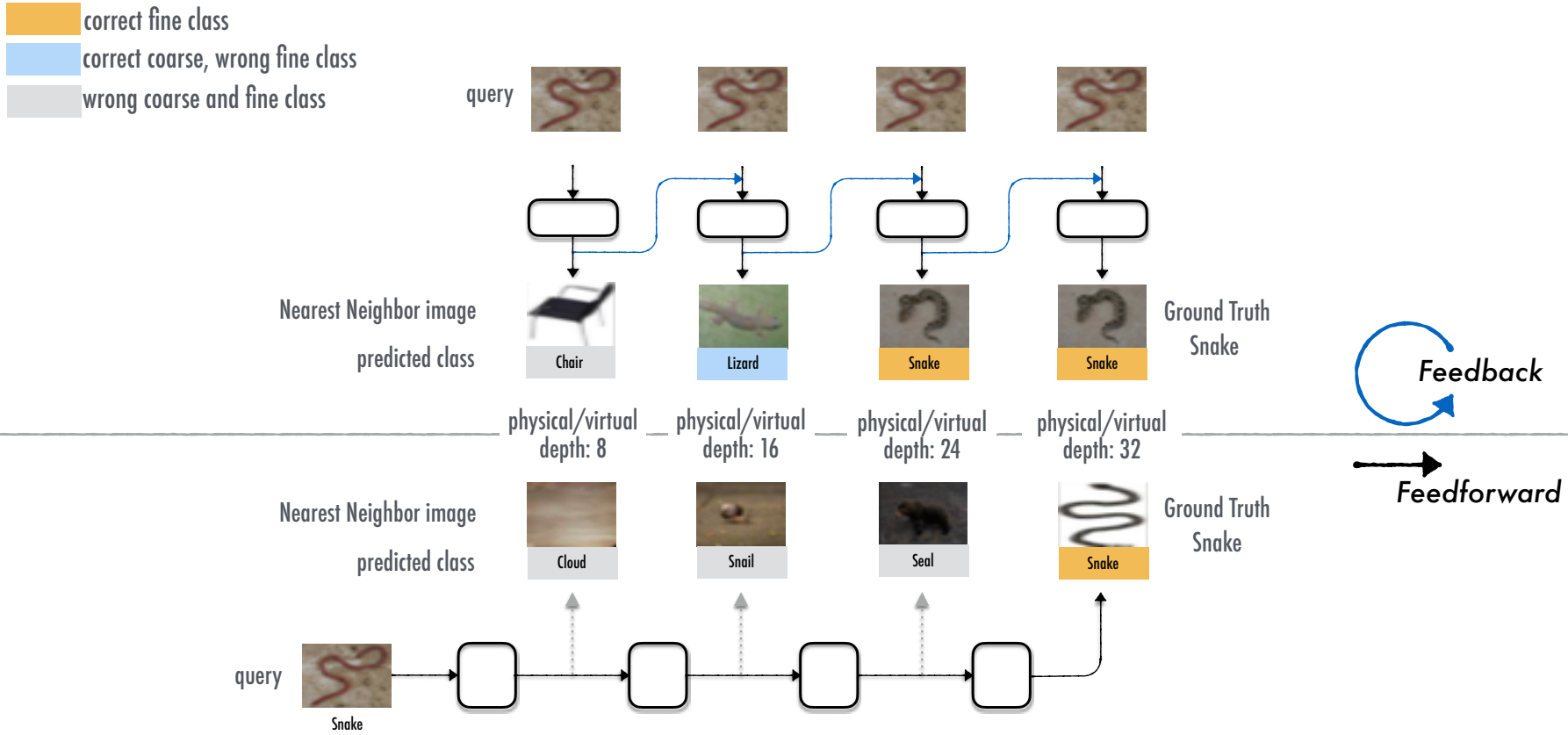

(b) Feedback via hidden state
(Non-distributed. Stack-All)




(c) Feedback via hidden state
(Distributed. Stack-3)

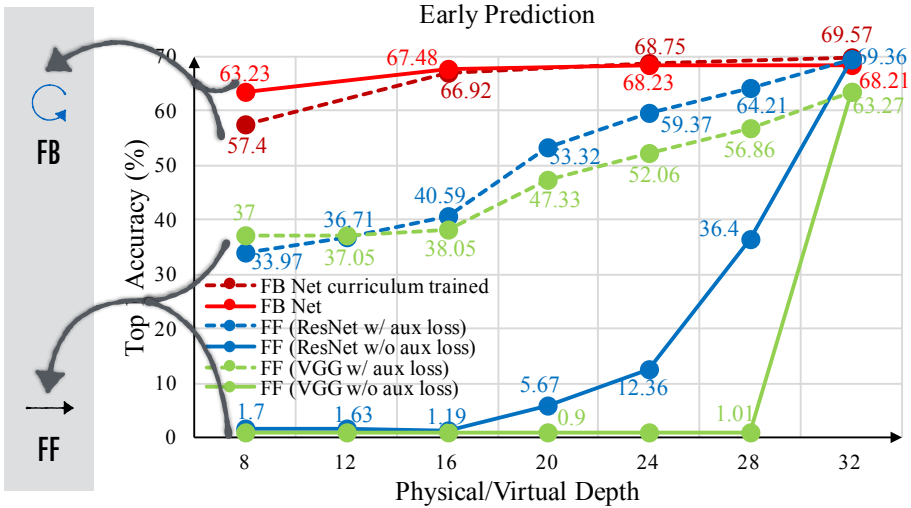
*Feedback in latent space*

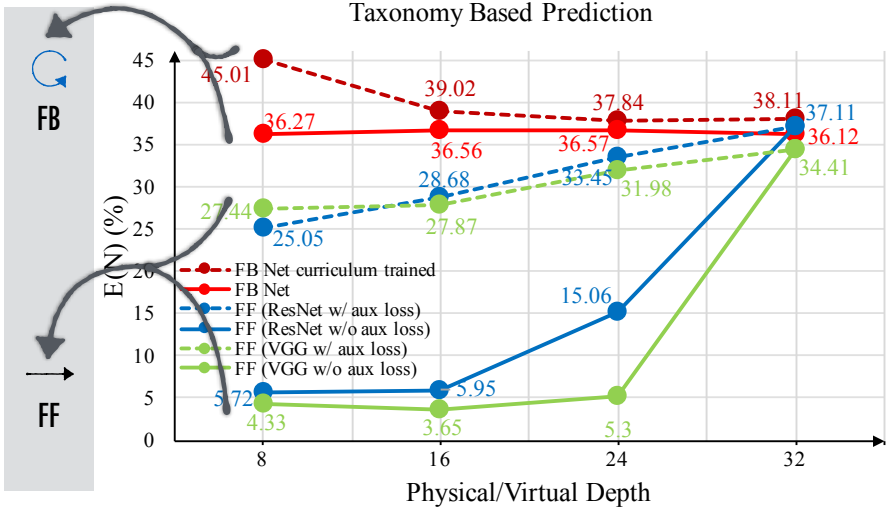
Feedback Networks, CVPR 2017



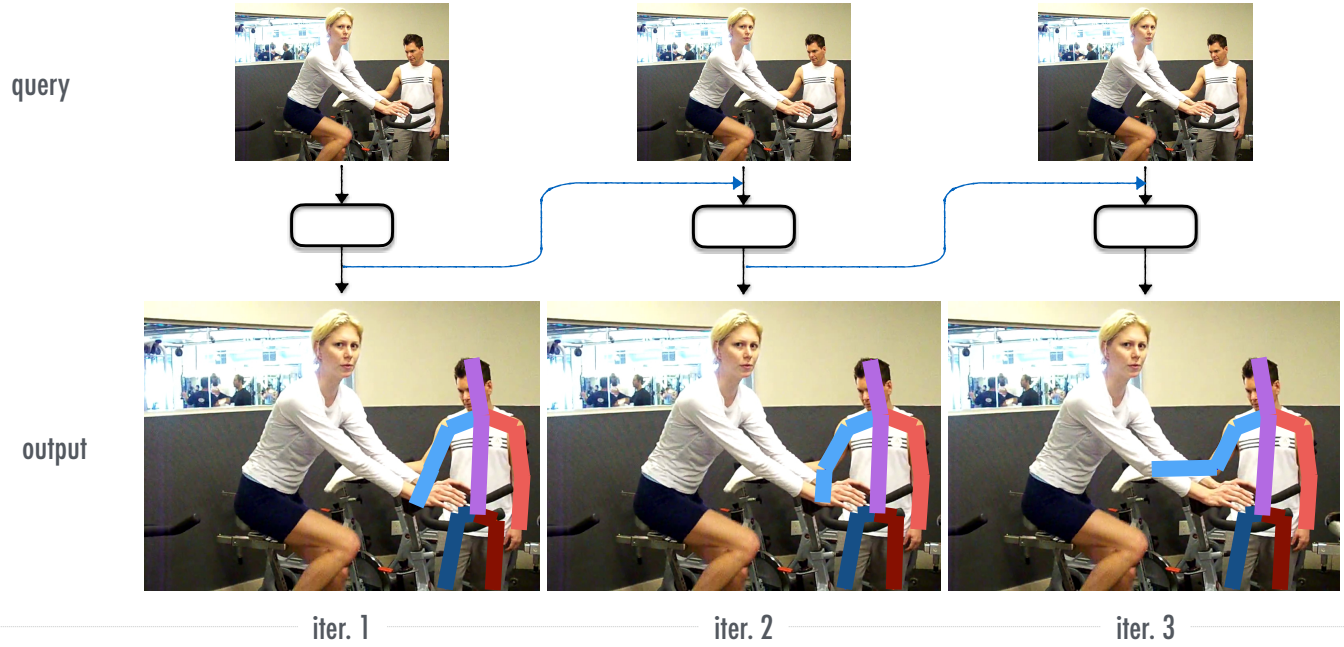

# *Qualitative results on CIFAR100 test set*

# Early Prediction & Taxonomic Prediction



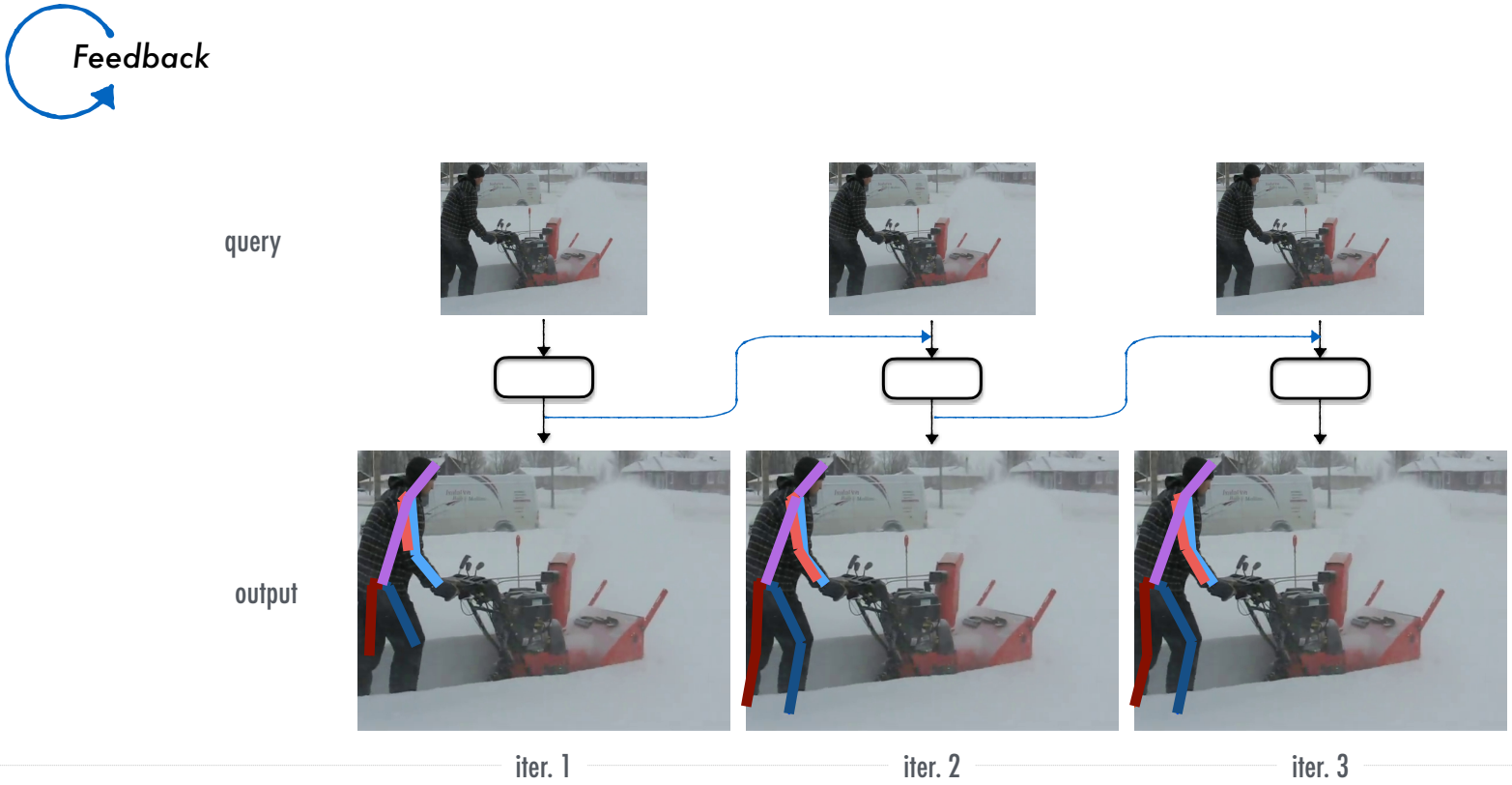




(details in the main paper)


# *Qualitative results on MPII human pose test set*

Feedback Networks, CVPR 2017




# *Qualitative results on MPII human pose test set*

Feedback Networks, CVPR 2017

# Questions?

https://vilab.epfl.ch/