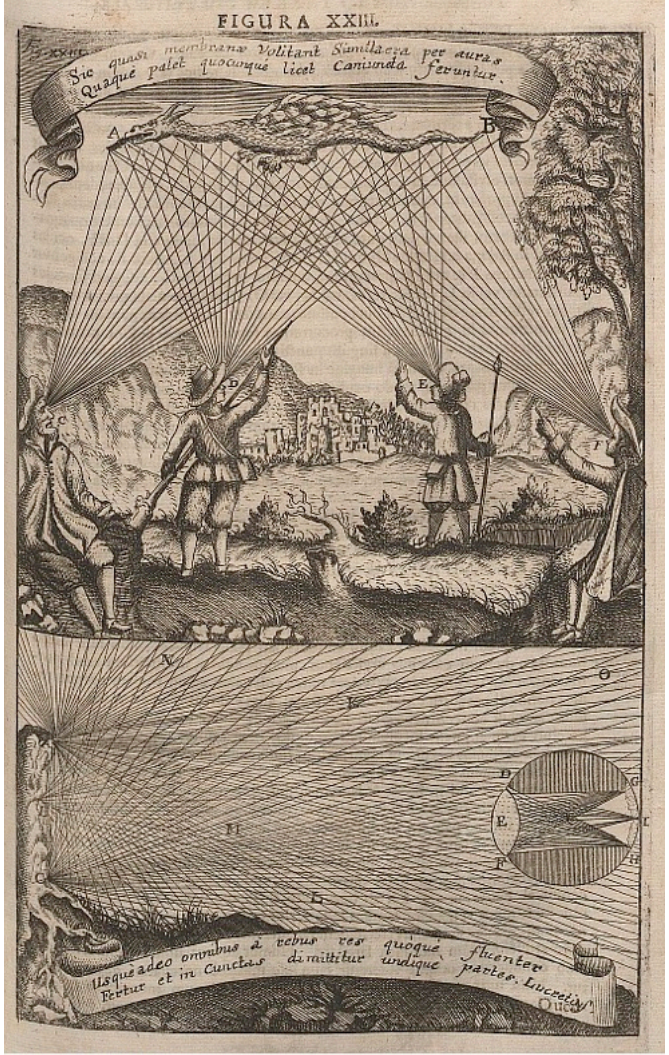
CS-503
Visual Intelligence: Machines and Minds

Amir Zamir

Lecture 1

École
polytechnique
fédérale
de Lausanne

Vision: understanding the world through light.

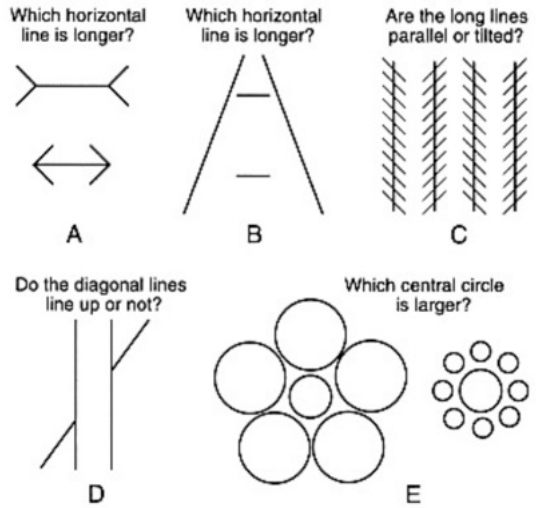


Quelle: Deutsche Fotothek

Vision: understanding the world through light.

This Course:

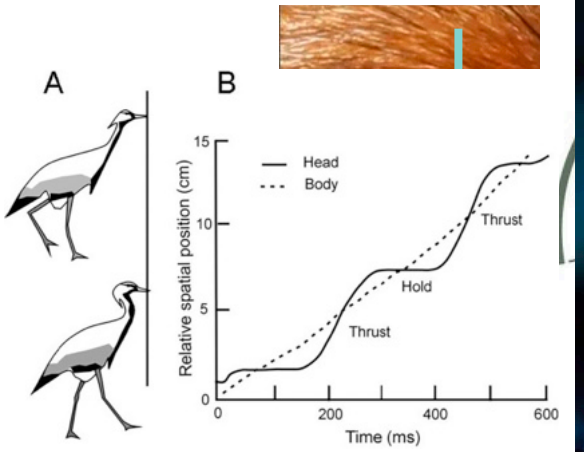
- **Basics**: biological vision, theories, etc.

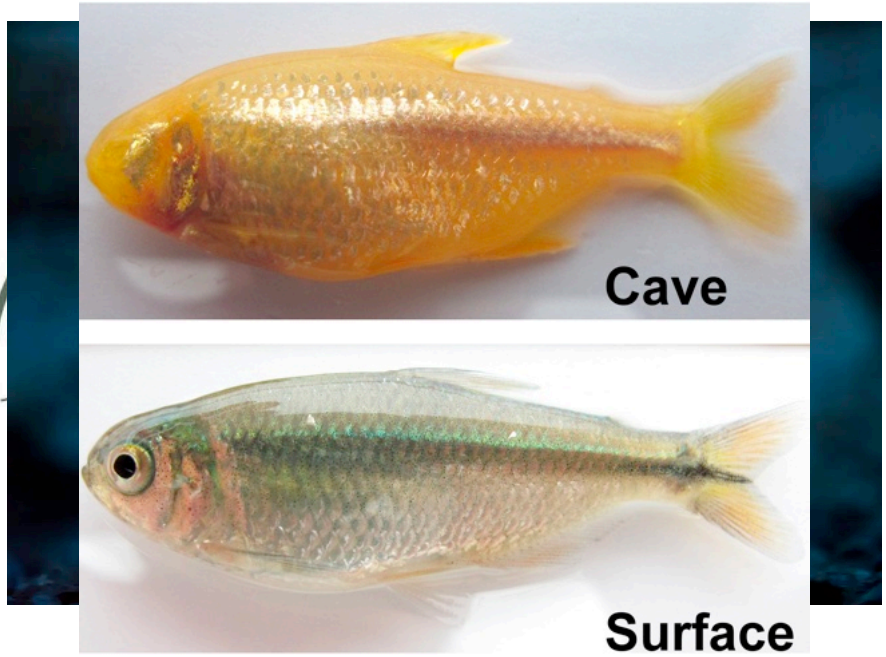


- **Foundation Models & Modern Tools**: visual and multimodal models, etc.

deepseek
Gemini
Qwen

- **Visual perception** is closely linked to agent's actions, body, and surrounding environment.

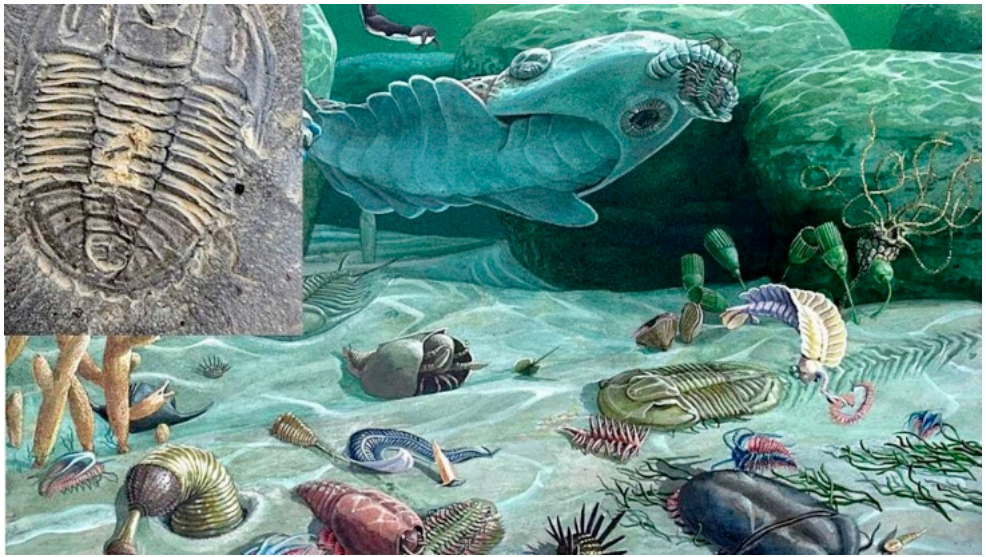




• M. Bank et al., 2015. M. Land, 2002

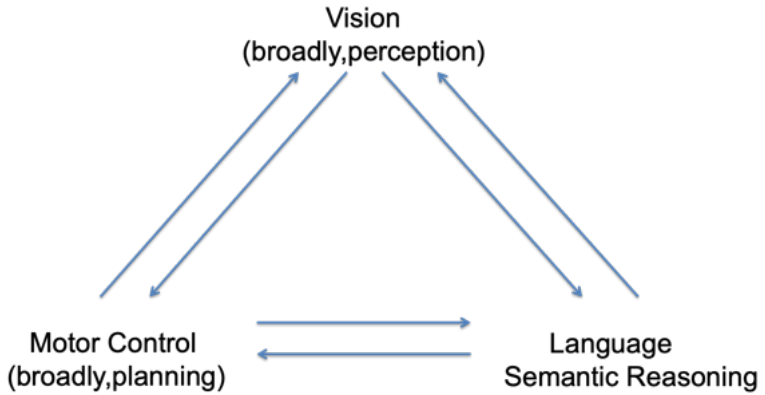
Why vision+action?

- Cambrian explosion: ~500 million years ago.
- Light switch theory, A. Parker.



Why vision+action?

- ~500 million years ago: light switch. Mostly **locomotion**.
- ~5 million years: bipedalism appeared. Freed hands for tool making/use/ **manipulation**.
- 50k-150k: modern homo sapiens. **Language** appears.

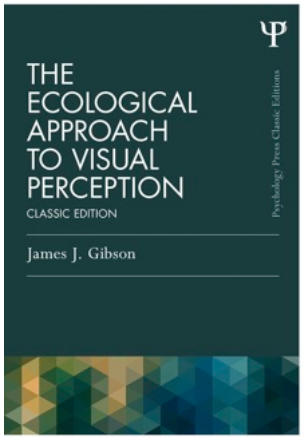


J. Malik

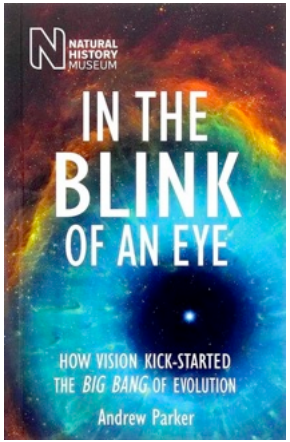
Why vision+action?

- Abundant evolutionary, biological, practical, and computational arguments.

The "ecological vision" argument



Light switch theory, A. Parker



A Critique of Pure Vision, Churchland et al.

A Critique of Pure Vision[1]

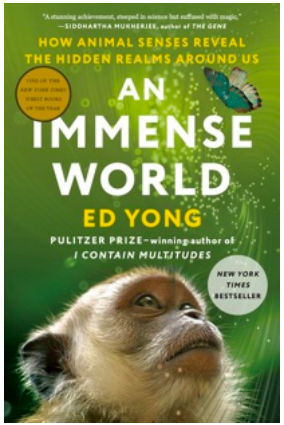
Patricia S. Churchland, V. S. Ramachandran, and Terrence J. Sejnowski

INTRODUCTION

Any domain of scientific research has its sustaining orthodoxy. That is, research on a problem, whether in astronomy, physics, or biology, is conducted against a backdrop of broadly shared assumptions. It is these assumptions that guide inquiry and provide the canon of what is reasonable—of what "makes sense." And it is these shared assumptions that constitute a framework for the interpretation of research results. Research on the problem of how we see is likewise sustained by broadly shared assumptions, where the current orthodoxy embraces the very general idea that the business of the visual system is to create a detailed replica of the visual world, and that it accomplishes its business via hierarchical organization and by operating essentially independently of other sensory modalities as

Animal Eye, M. Land

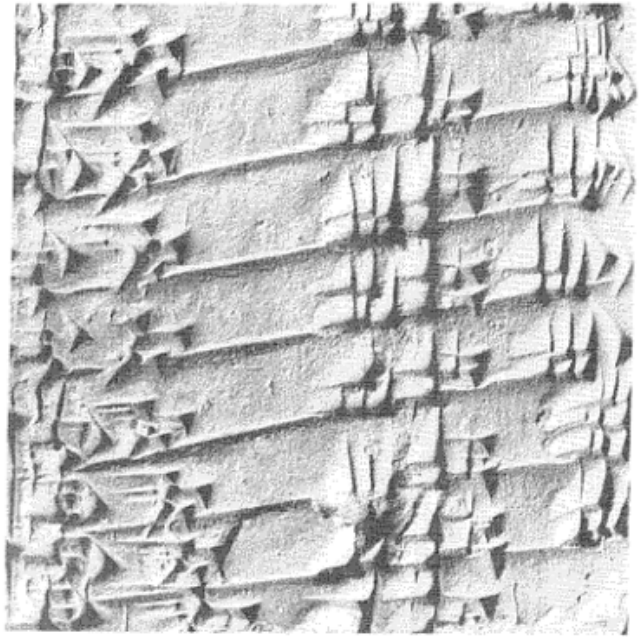
"Umwelt"



Why the real world/ecology?

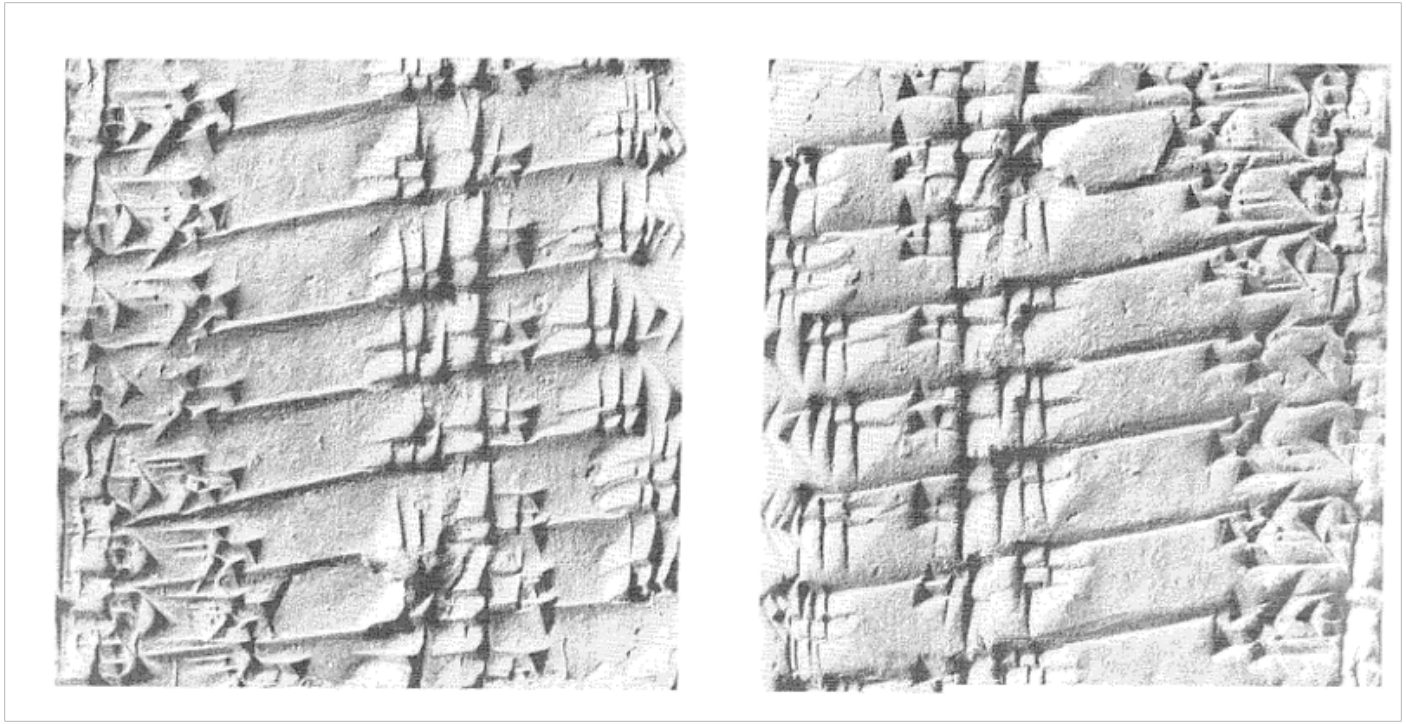
Why the real world/ecology?

Why the real world/ecology?

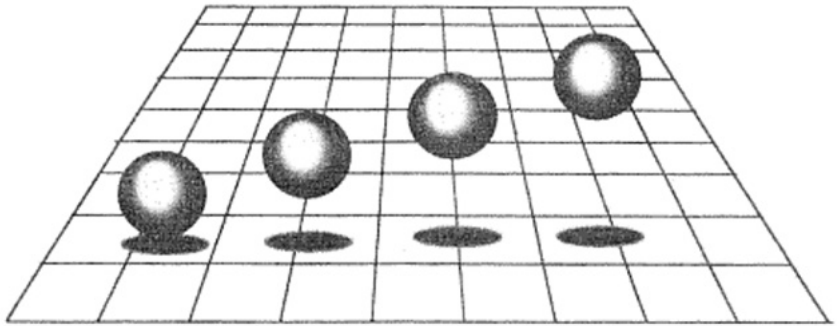


Why the real world/ecology?

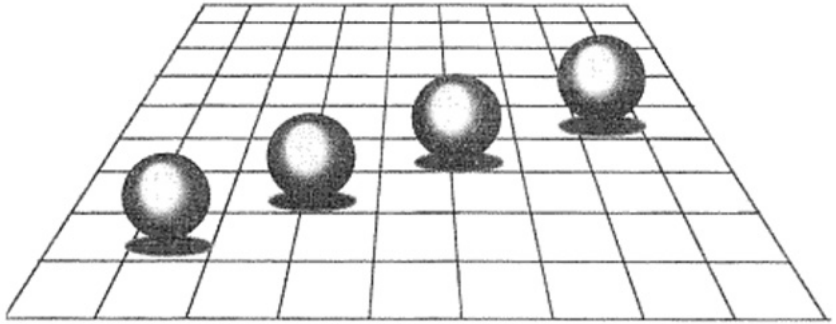
Why the real world/ecology?



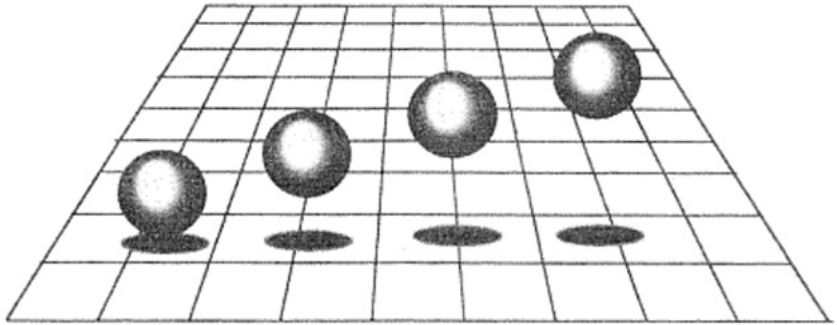
Why the real world/ecology?



Why the real world/ecology?

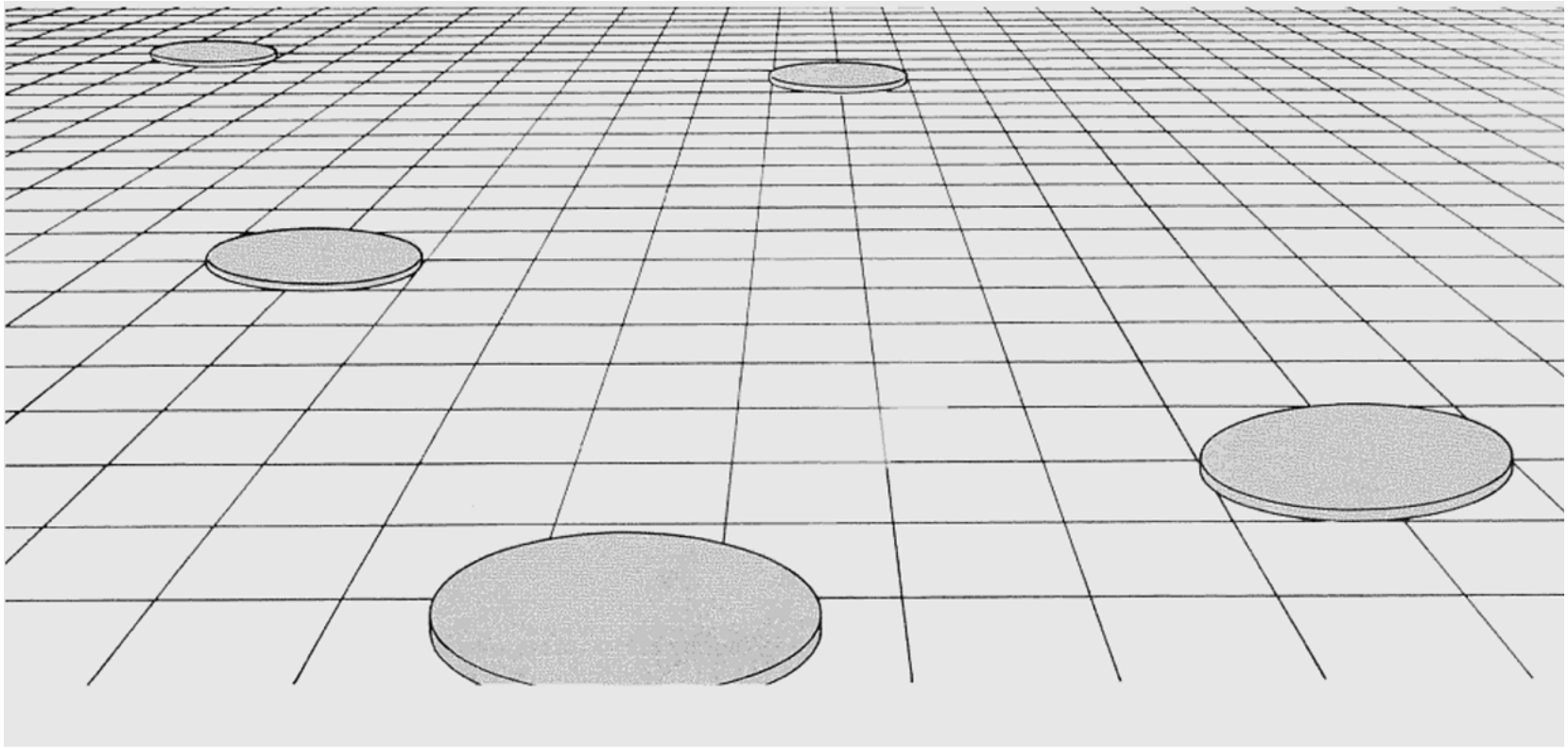


A



B

Why the real world/ecology?



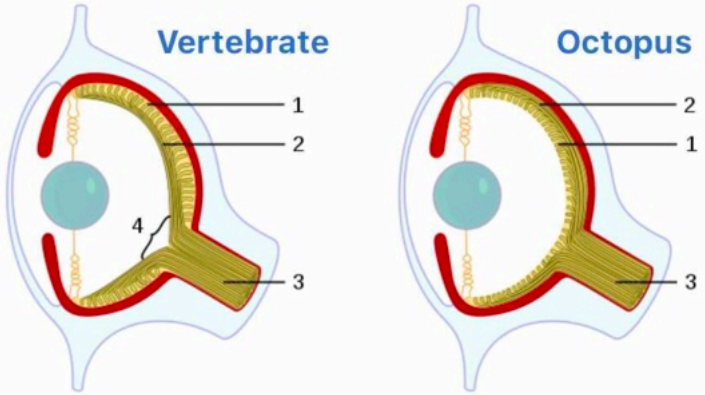
Why natural vision?

Why natural vision?

- Because most working examples of sophisticated perceptual intelligence are there!

But only looking at nature isn't enough.

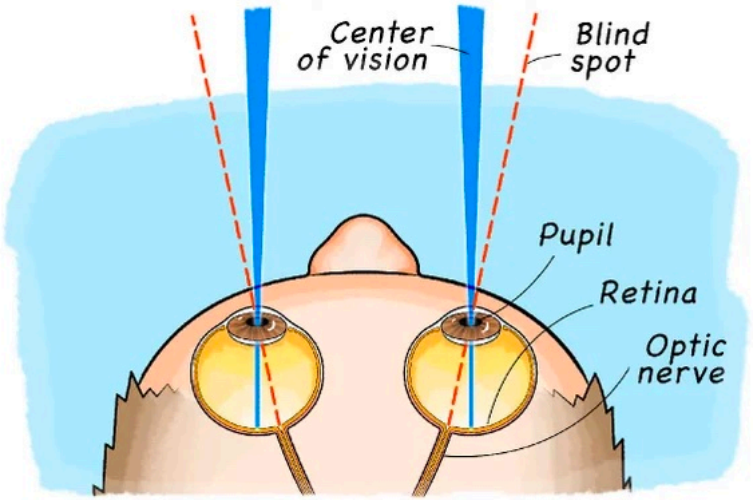
- Copying evolutionary bugs: blind spot (scotoma)



In vertebrate eyes, the nerve fibers route *before* the retina, blocking some light and creating a blind spot where the fibers pass through the retina and out of the eye. In octopus eyes, the nerve fibers route *behind* the retina, and do not block light or disrupt the retina. In the example, **4** denotes the vertebrate blind spot, which is notably absent in the octopus eye. In both images, **1** denotes the retina and **2** the nerve fibers, including the optic nerve (**3**).

But only looking at nature isn't enough.

- Copying evolutionary bugs: blind spot (scotoma)
  - Check it out: look straight ahead. Close one eye. extend your arm and thumb out. Move slowly from center to side. Your thumb will disappear around ~15 degree from center.



But only looking at nature isn't enough.

- Cover your left eye and look at the plus sign, move closer of further away to find where the circle disappears. (Try on your screen, not the projector, since you have to move your head a lot)
  - That location is where your optic nerve attaches to your eyeball and is where no visual information is processed due to the lack of rods and cones, otherwise known as the blind spot

But only looking at nature isn't enough.

- Wright Brothers
  - *"Learning the secret of flight from a bird was a good deal like learning the secret of magic from a magician. After you once know the trick and know what to look for you see things that you did not notice when you did not know exactly what to look for."*

ORVILLE WRIGHT
DAYTON, OHIO

December 27, 1941.

Mr. Horace Lytle, President,
The J. Horace Lytle Company,
Dayton, Ohio.

Dear Mr. Lytle:-

Your letter of November 26th was duly received, but having become buried among other papers, it has just come to my attention again.

I can not think of any part bird flight had in the development of human flight excepting as an inspiration. Although we intently watched birds fly in a hope of learning something from them I can not think of anything that was first learned in that way. After we had thought out certain principles, we then watched the bird to see whether it used the same principles. In a few cases we did detect the same thing in the bird's flight.

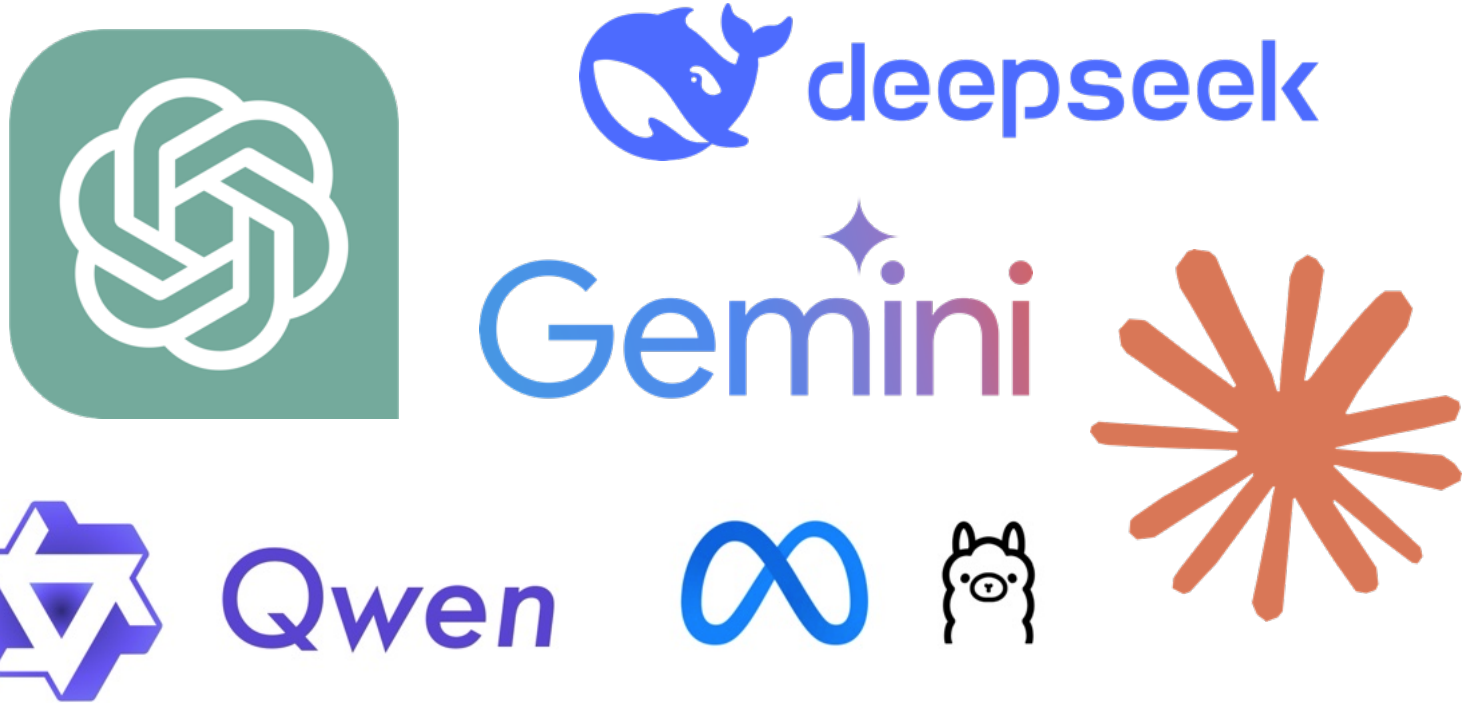
Learning the secret of flight from a bird was a good deal like learning the secret of magic from a magician. After you once know the trick and know what to look for you see things that you did not notice when you did not know exactly what to look for.

Sincerely yours,

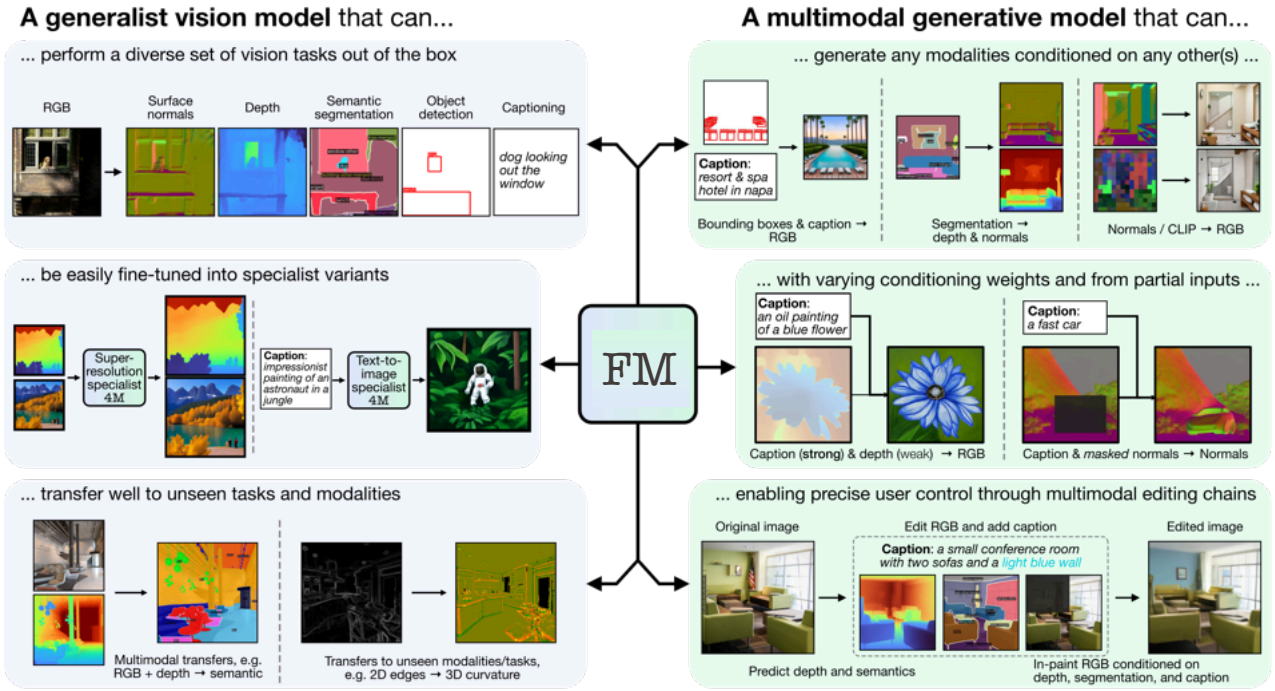
Orville Wright

But only looking at nature isn't enough.

2025: “Foundation Models”



Goal: develop a “Foundation Model” from scratch

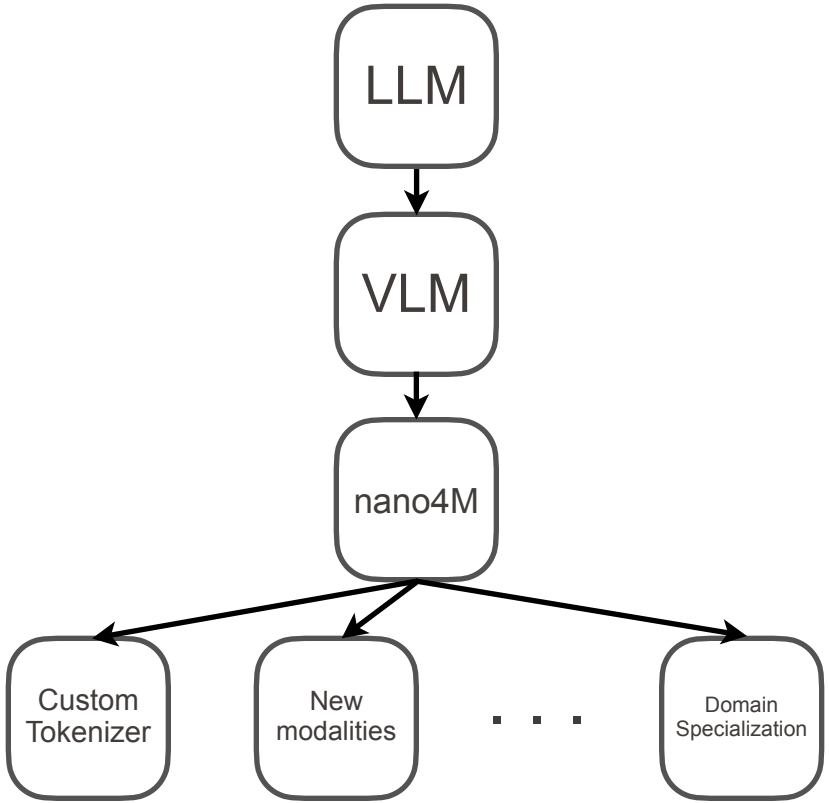


- develop and enhance a foundation (multimodal large vision-language) model end-to-end

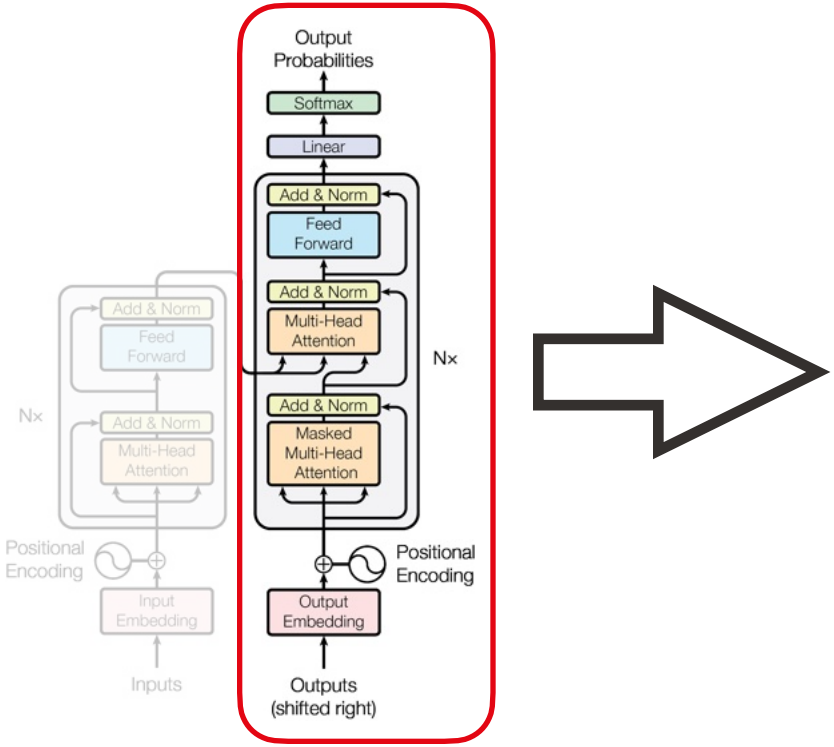
Goal: develop a “Foundation Model” from scratch

- **Basic Task**: develop a nano FM
  1. LLM
  2. Text-to-Image generation
  3. nano4M
- **Extensions**: propose your own, e.g.
  - Custom tokenizers
  - New modalities
  - Domain specialization
  - Reasoning
  - etc.

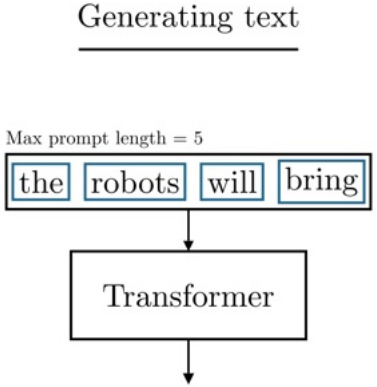
Project model development flow



1. Develop: LLM

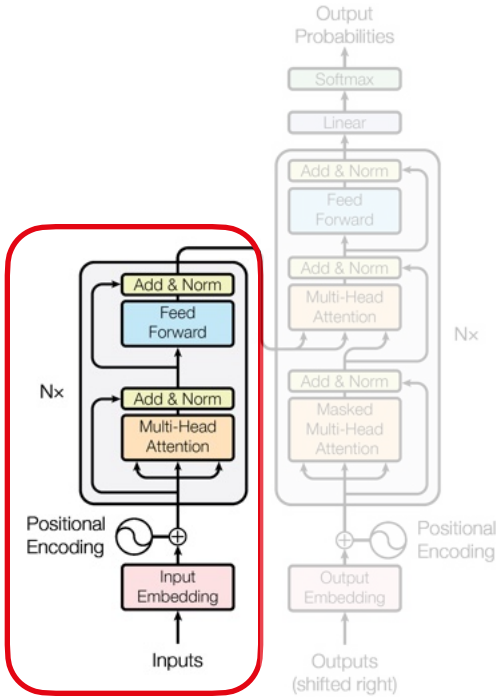


Transformer Decoder

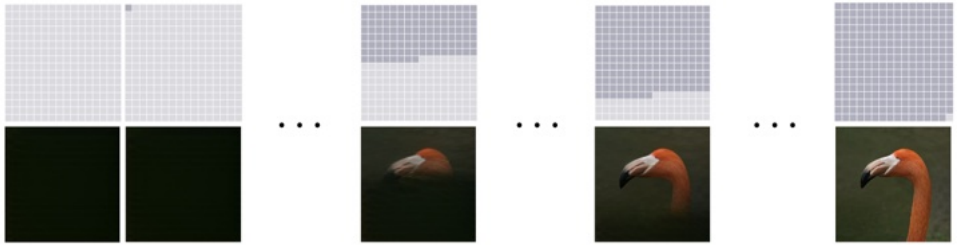


- From Attention Is All You Need and The Animated Transformer

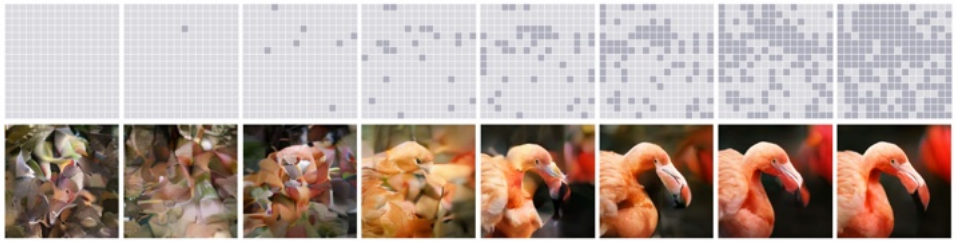
2. Develop: Text-to-Image Generator



Transformer Encoder



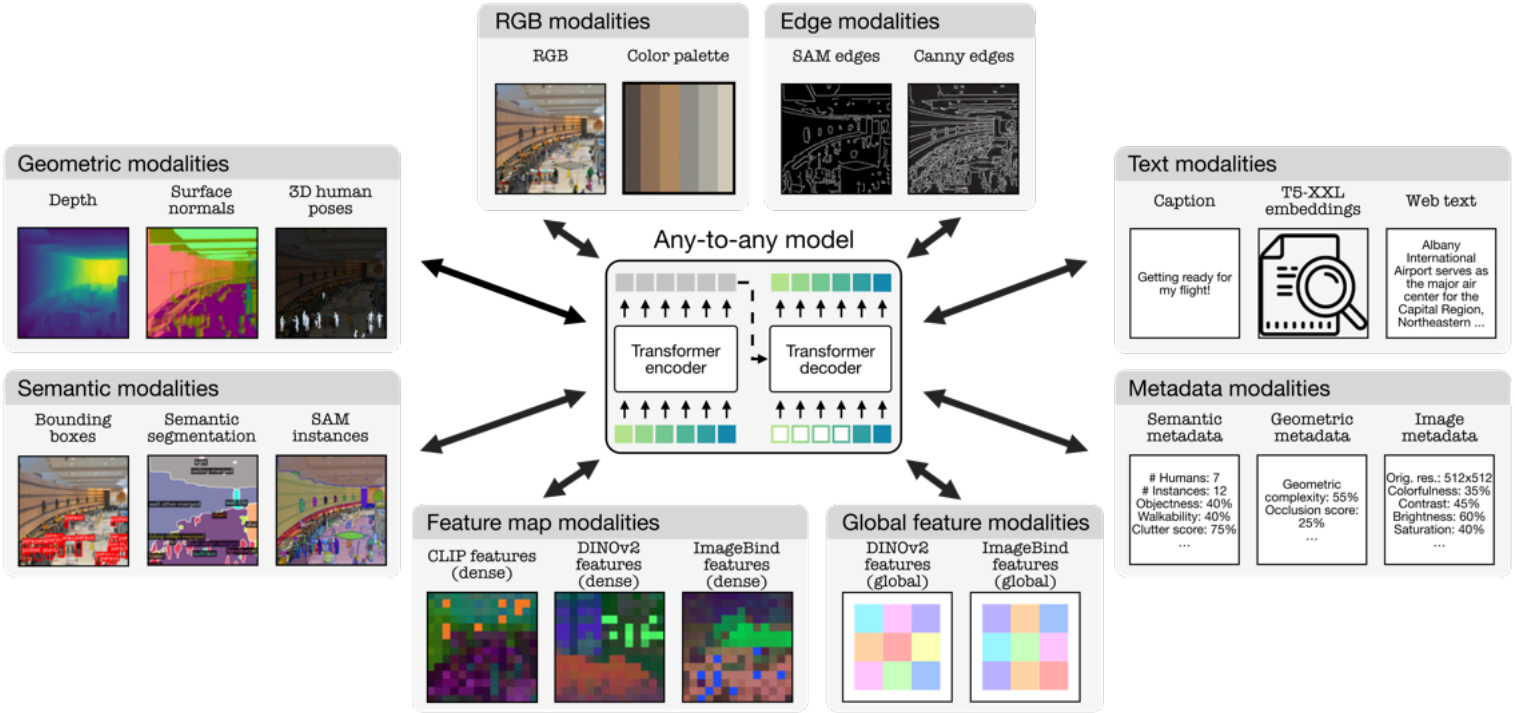
Autoregressive Generation



Masked Image Generation (with MaskGIT)

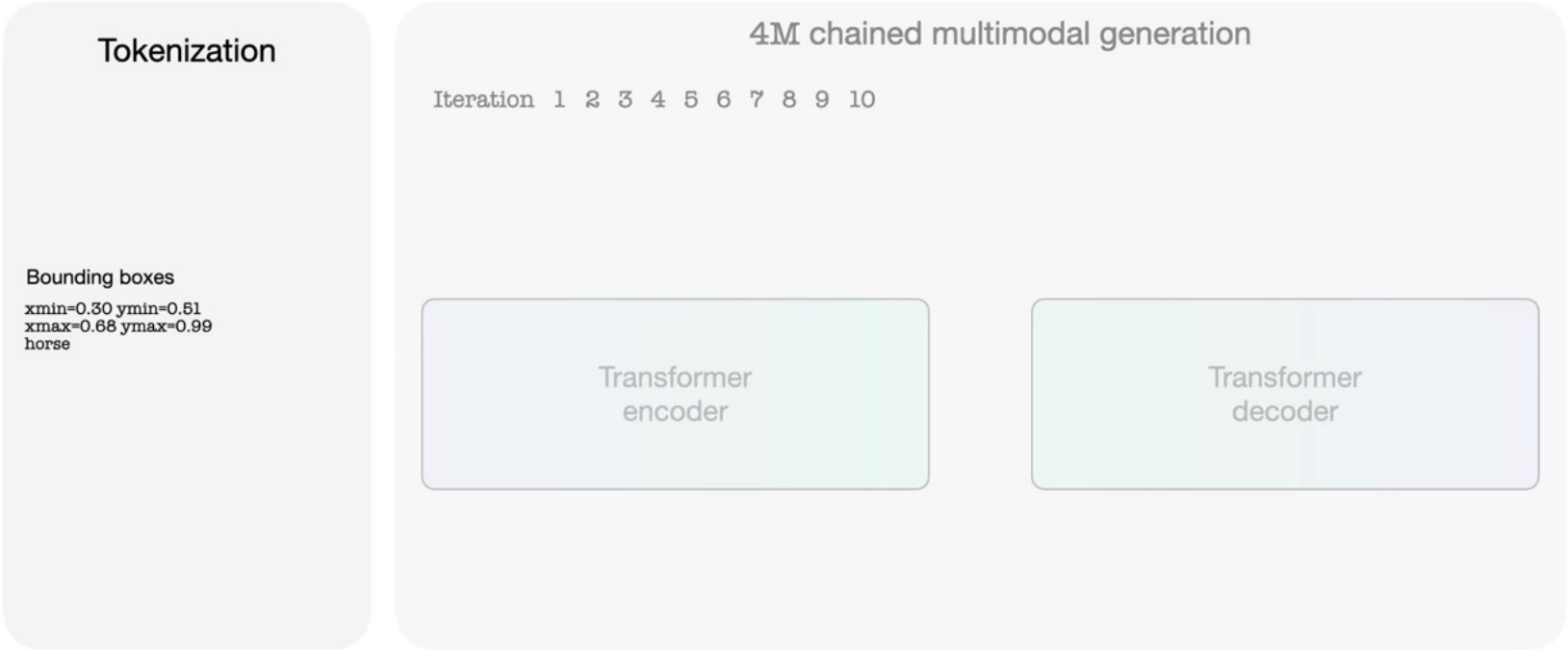
- From Attention Is All You Need and MaskGIT

3. Develop: nano4M



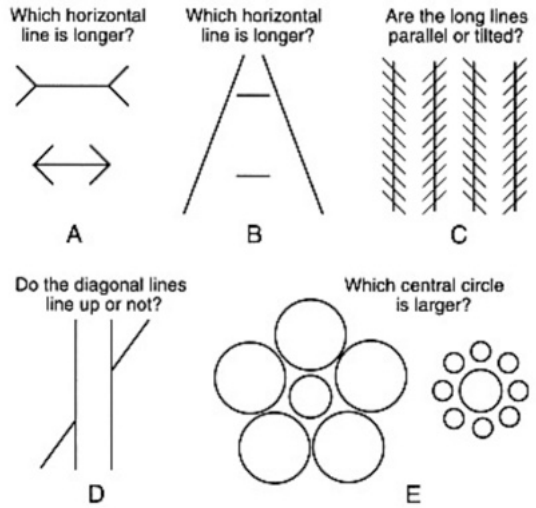
Multi-modal Encoder-Decoder

3. Develop: nano4M



Course Topics and Lectures

- **Basics**: biological vision, theories, etc.



- **Foundation Models**: visual and multimodal models.

deepseek
Gemini
Qwen

Course Topics and Lectures

- **Foundation/Concept Lectures**
  - e.g. theories of vision, sensory umwelt, perception-action loop, multimodality, etc.
- **SOTA/Tools Lectures**
  - e.g., Architectural/transformer advances, FMs, etc.
- **Domain Expert Lectures**
  - e.g., neuroscience, RL, etc.

- Credit for a fraction of visuals/slides: James J. Gibson, Michael Land, Bruno Olshausen, Jitendra Malik, Stella Yu, Alexei Efros, Dan-Eric Nilsson.

course logistics and policies

Teaching Team

- Instructor: Amir Zamir (amirzamir.com)
- Doctoral Teaching Assistants: (vimm-ta@groupes.epfl.ch)
  - Rishubh Singh (head TA)
  - Roman Bachmann
  - Zhitong Gao
- Group: https://vilab.epfl.ch/

Who is this course for?

- Anyone interested in open-ended learning and research in vision and perceptual AI.
  - In particular **active & multimodal** intelligent agents in the **real world**.
  - Emphasis on exploring and what is **not** possible today, than what it is.
- Part of the course is about defining **problems** than providing **solutions**.
  - A vast sea between the two. Eg:
    - Solution: we know how to train a DNN to predict 1000 predefined classed given an image
    - Problem: how to use DNNs to have an agent that interacts with an environment, continually learns a model of it, and performs in it.
  - Often there is no clear-cut solution.

Prerequisites (important)

- Machine Learning
  - E.g. Machine Learning (CS-433) or Introduction to Machine Learning (CS-233) or equivalent course on the basics of machine learning.
- Deep Learning
  - E.g. Deep Learning (EE-559) or Artificial Neural Networks (CS-456) or equivalent course on the basics of deep learning.
- For projects with an active agent component, Reinforcement Learning or control theory.
- Expertise at one of the common Deep Learning frameworks, e.g. Pytorch.
- Prior experience with Computer Vision (e.g. CS-442) or vision data recommended.

Grading

- **Project (60%)**
  - Project proposal (15%)
  - Project progress reports (15%)
  - Final project presentation (15%) and report (15%)
    - Presentation: the last week of the semester.
    - Report/presentation templates provided.
- **Homework (40%)**
  - 3-5 homework and notebook assignments.
- **Late Policy**: total of 3 late days with no penalty. 0 otherwise.

Projects

- Get inspired!
- Most of your learning will happen while doing the project and homework.
- Desired characteristics of projects:
  - Related to vision, perceptual agents, etc.
  - Cool and creative (in problem selection, solution formulation, implementation, demonstration)
  - Thought-through.
  - Well-executed.
  - Does not have to be hardware or software. Up to you.
  - You learn something from it.

Projects

- BristleBot
- Seemingly complex behavior can come out of simple pieces.
- An active agent doesn't have to look like Boston Dynamic's Atlas.

https://www.evilmadscientist.com/

Existing Vision/ Robotic/FM Challenges

- E.g.
  - CVPR Embodied AI Challenges

Analysis/debunking Projects

Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht* UC Berkeley
Rebecca Roelofs UC Berkeley
Ludwig Schmidt UC Berkeley
Vaishaal Shankar UC Berkeley

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.

A Metric Learning Reality Check

Kevin Musgrave[1], Serge Belongie[1], Ser-Nam Lim[2]

[1]Cornell Tech [2]Facebook AI

Abstract. Deep metric learning papers from the past four years have consistently claimed great advances in accuracy, often more than doubling the performance of decade-old methods. In this paper, we take a closer look at the field to see if this is actually true. We find flaws in the experimental setup of these papers, and propose a new way to evaluate metric learning algorithms. Finally, we present experimental results that show that the improvements over time have been marginal at best.

From Same Photo: Cheating on Visual Kinship Challenges*

Mitchell Dawson[1][0000−0002−6719−6584], Andrew Zisserman[1][0000−0002−8945−8573], and Christoffer Nellåker[2][0000−0002−2887−2068]

[1] Visual Geometry Group (VGG), Dept. of Engineering Science, University of Oxford
{mdawson,az}@robots.ox.ac.uk
[2] Nuffield Dept. of Women's & Reproductive Health, Big Data Institute, IBME, University of Oxford
christoffer.nellaker@bdi.ox.ac.uk

Abstract. With the propensity for deep learning models to learn unintended signals from data sets there is always the possibility that the network can "cheat" in order to solve a task. In the instance of data sets for visual kinship verification, one such unintended signal could be that the faces are cropped from the same photograph, since faces from the same photograph are more likely to be from the same family. In this paper we investigate the influence of this artefactual data inference in published data sets for kinship verification.

Past example

- **Hunting for Insights: Investigating Predator-Prey Dynamics through Simulated Vision and Reinforcement Learning**
- Arvind Menon, Lars C.P.M. Quaedvlieg, Somesh Mehra
- https://arvind6599.github.io/PredatorPreyWebsite/

Past example

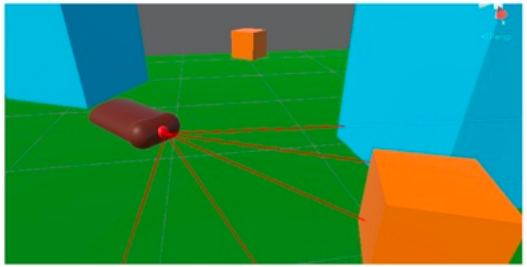
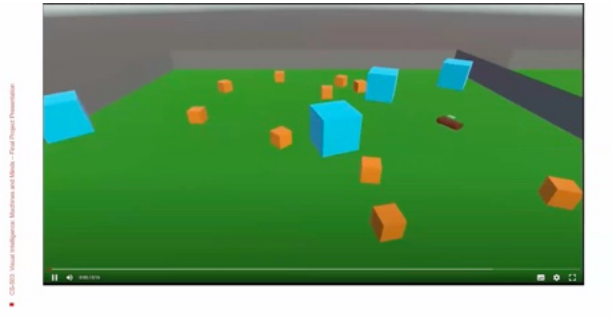


Figure 3. Visualization of the agent's vision with 5 eyes



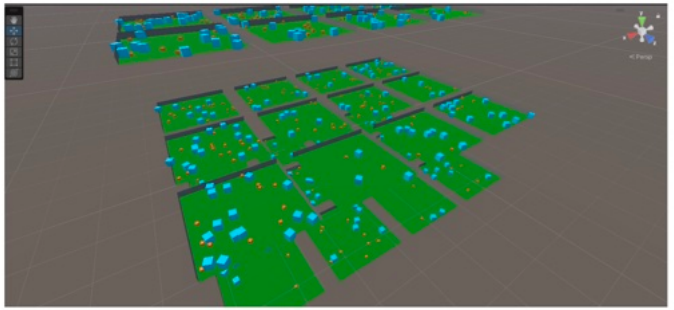


Figure 2. Multiple environments to speed up training

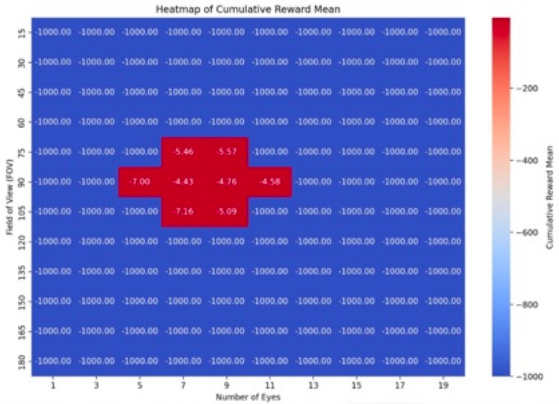
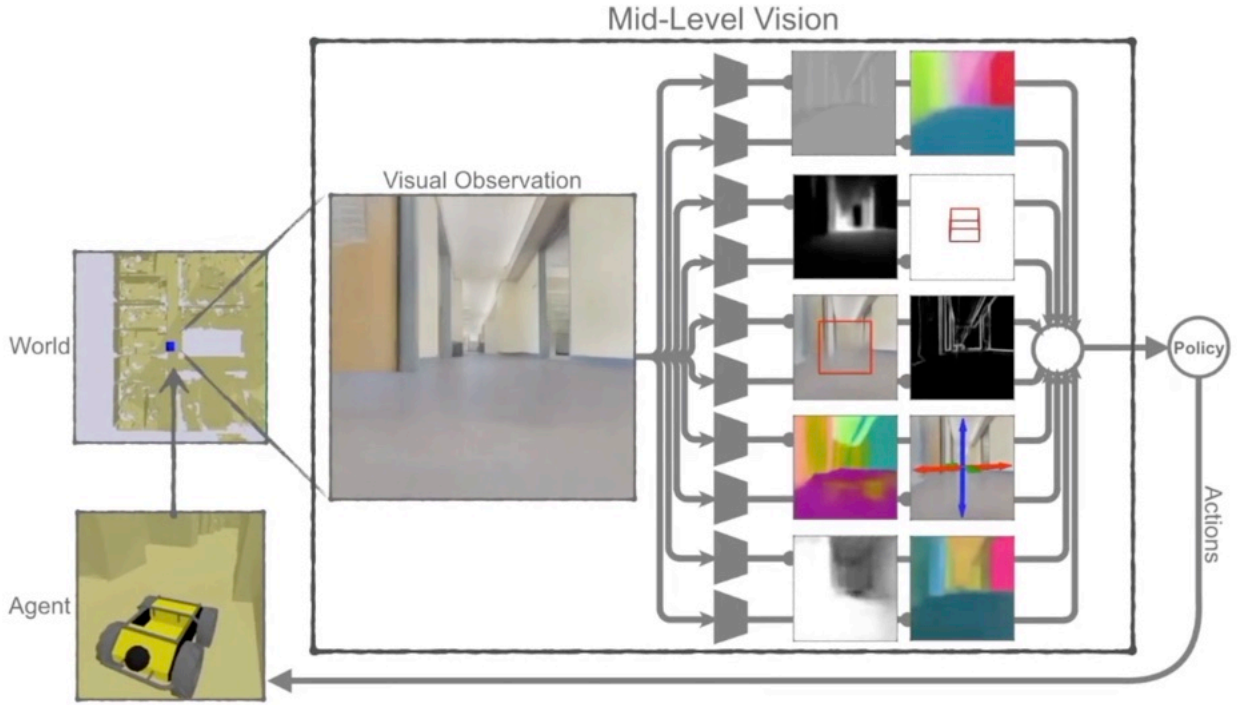


Figure 6. Exploration result starting with 9 eye and 90 fov, that converged on 7 eyes and 90 fov

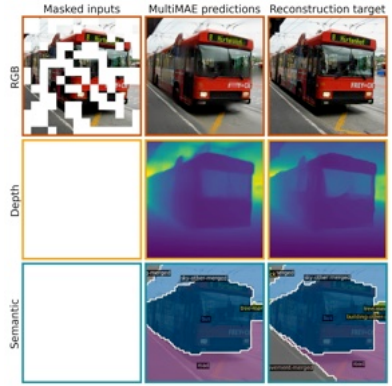
- **Vision Evolution: Natural Selection shaped by environment**
- Yassine Abdennadher, Ambroise Borbely, Matthieu Andre.

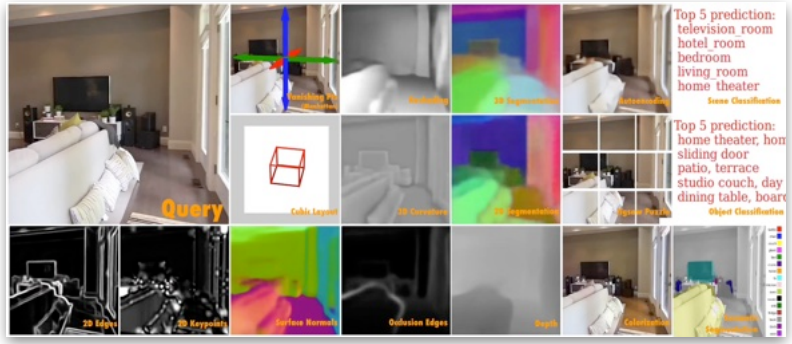
Past example

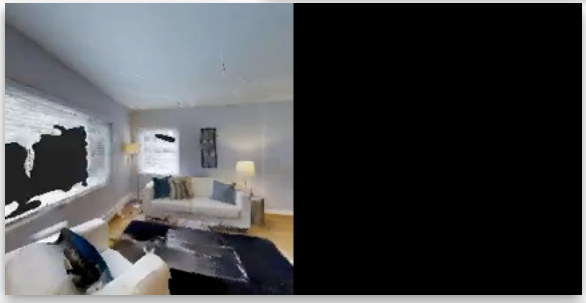


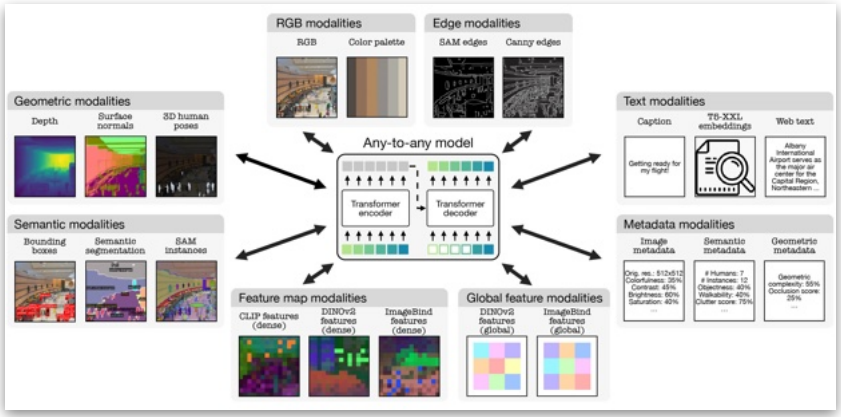
- Bradley Emi
- Mid-Level Vision, CoRL 2019, CoRL 2020. Sax et al.

Relevant Lab Projects https://vilab.epfl.ch/









Projects: Think about the unsolved

Projects

- More categorical examples:
  - Make a visual agent inspired by a simple biological organism.
  - Extend the FMs you develop.
  - Simulate the sensory “umwelt” of an animal.
  - Add a camera to any visually blind (or too heavily sensor loaded) system out there.
  - Stationary cameras: extract actionable intelligence. E.g. a camera in the corner of the living room, the passive surveillance cameras, etc.
  - Apply/analyze one of the concepts discussed in the class.
    - Take a concept from one of the books and apply it computationally.
  - Any of relevant CVPR/ICCV/ECCV/NeurIPS challenges.
  - Analysis study
- We will give you feedback on your proposals. Do seek feedback from TAs extensively.

Projects

- The goal is doing something cool and learn. We're here to help you. We'll be flexible with solid efforts.
- Use the teaching team.
- The top projects are of acceptable quality to top-tier AI/vision/ML conferences.
  - (CVPR/ICCV/ECCV/NeurIPS/ICML/ICLR/CoRL/RSS/ICRA)
- The Best Project will be selected and advertised. Awarded with conference registration.
- Well-distribute your time (continuous evaluation).
- **Teaming policy**: encouraged, but $1+1 \precnsim 2$.
- **Merging with an existing project**:
  - In principle, okay. But 1) No work by others 2) New work done. The report should clarify.

Compute Resources

- SCITAS
  - Anyone at EPFL can submit jobs via their scheduler (SLURM)
  - We reserved some GPUs for this course. More available if capacity unused.
- Google Cloud
- Azure
- Free tier (open to everyone)
  - Google Cloud: $300 for 90 days
  - Azure: $200 for 30 days

Misc.

- **Front loading lectures**: About 8 weeks of lectures 2x per week, then you're free to work on your projects full time.
  - The class hours turn into support sessions for troubleshooting and discussion with TAs.
- **Books and reading material** references on Moodle and the library (use them!)
  - Additional resources and reading will be occasionally placed on Moodle
  - **Have a sharing, fun, passionate culture**: find something related interesting? An article, video, book, meme? Just post it on the forum for everyone.
- **Amir's office hours**: Tuesdays 15:00-16:00 (by prior email appointment)
- **Feedback From:**
  - Anonymous. Any thoughts and suggestions welcome.
  - https://forms.gle/K7RKKafcEgfEJjhz7 (link on Moodle)


Equal opportunity
makes us brighter
EPFL


Diversity is
our strength
EPFL


Listen and dare
to speak up
?!
@:)
EPFL


Look out
for each other
EPFL


Kindness
brings serenity
EPFL


Together
we go further
EPFL

Mental health
Brilliant
ideas in a
peaceful mind

Enjoy the Course!

https://vilab.epfl.ch/