Embodied Vision

Introduction and Logistic

Tsung-Wei Ke

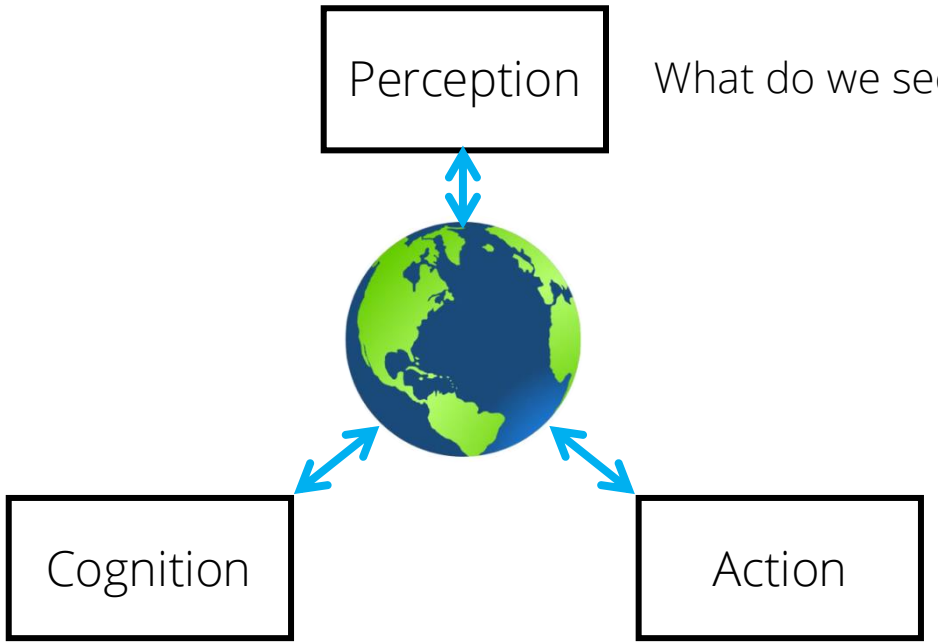
Spring 2026

Welcome!

Course website: https://ntu-ev.github.io/

About Me

Research interests:

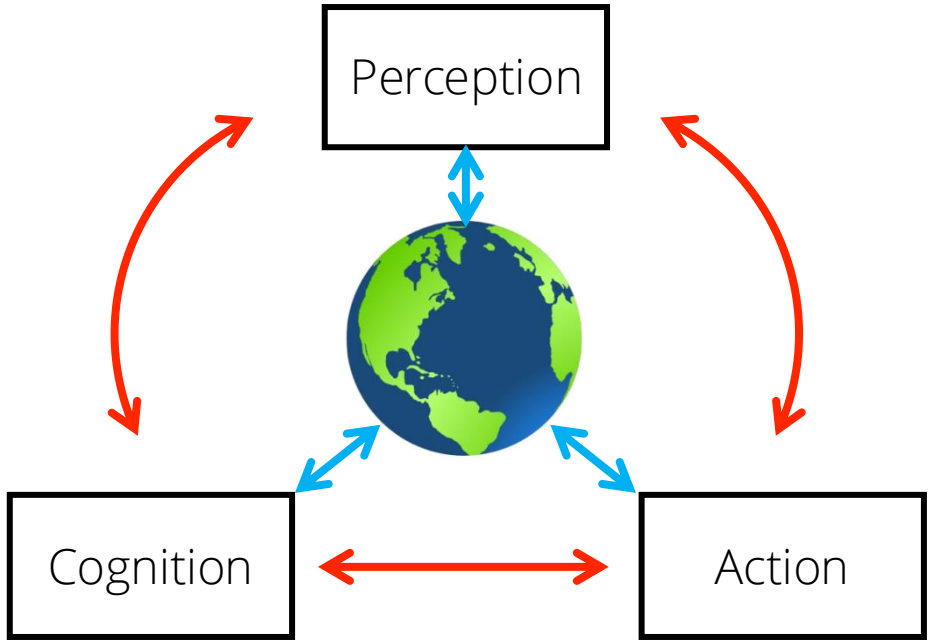


Tsung-Wei Ke 柯宗瑋

Embodied AI Lab
具形人工智慧實驗室

About Me

Research interests: the trio of perception, cognition and action in the real world



Tsung-Wei Ke 柯宗瑋

Embodied AI Lab
具形人工智慧實驗室

About TA: Jia-Wei

Jia-Wei Liao 廖家緯

NTU CSIE PhD Candidate

Research Interests:
- Generative AI in Vision
- Mathematical Modeling
- Data Science

Room: CSIE building 505

About TA: Shih-Hsin

Shih-Hsin Fang **方世昕**

NTU CSIE MS

Research Interest:
- Robotics
- Reinforcement learning

Room: CSIE building 544

About TA: Hung-Kai

Hung-Kai Chung 鍾閎凱

NTU EE MS

Research Interest:
- Video / Image generation
- Vision language action model

Room: MK-514

About TA: Po-Yi

Po-Yi Wu 吳柏毅

NTU CSIE Undergraduate

Research Interest:

- Robotic Learning
- Real-World Robotics
- Physics-Informed Training

Room: CSIE building 544

About TA: Hsin-Wei

Hsin-Wei Chen **陳忻韋**

NTU CSIE Undergraduate

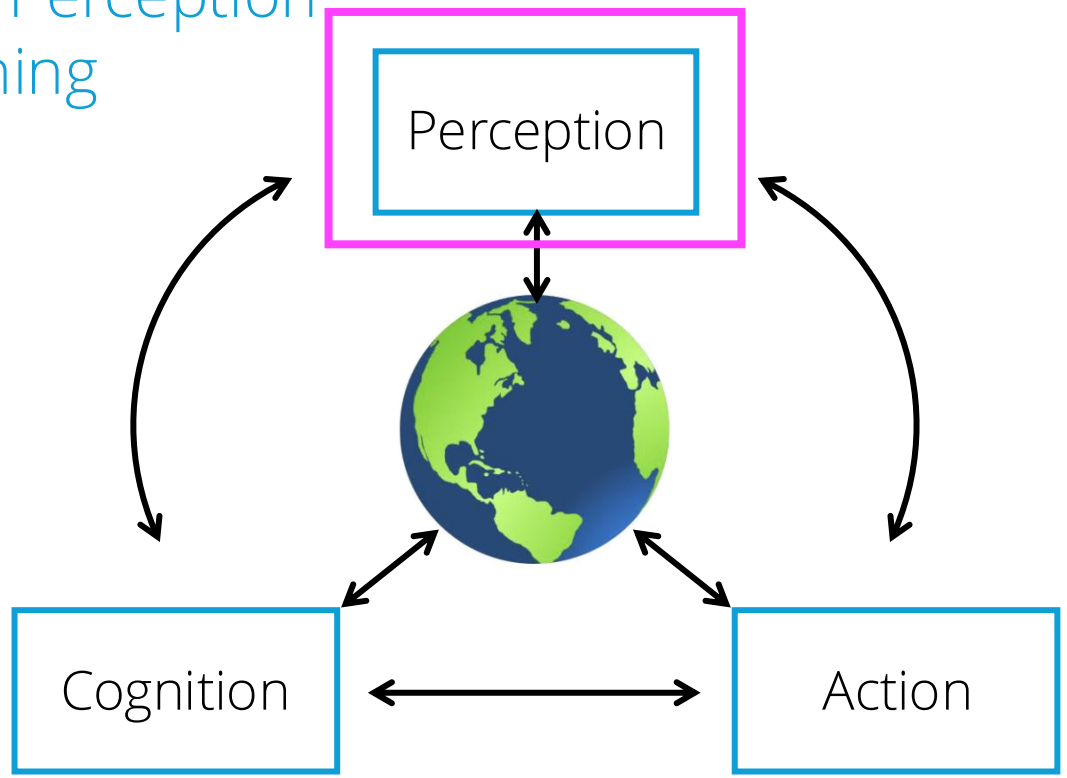
Research Interest:
- reinforcement learning
- human motion generation
- fluid simulations

Room: CSIE building 544

About This Class

Covered in Robot Perception and Learning

Covered in Embodied Vision!



Administrivia

Enrollment

- I am not enrolled but I would like to enroll in this class. What should I do?
  - If you're interested in enrolling, please fill out and submit this form. We'll send you the registration id.

https://forms.gle/ZS64KrqJmoeLzxK67

- I am interested in embodied AI / Vision research but I am not enrolled in this class. What should I do?
  - We'll release the slides and learning resources online.
  - Email me/TAs or come to my office hours. We're happy to chat!

- I'm not yet sure if this course is for me…
  - Hopefully, this lecture can help you decide

Prerequisite Knowledge

- Students should have a solid understanding of the following areas:
  1. Machine learning: stochastic gradient descent, loss function, optimization, neural network
  2. Calculus: differential, integral
  3. Linear algebra: matrices, vectors, norms, scalar/vector products, orthogonality, singular value decomposition…
  4. Probability: expectation, independence, Baye's Theorem…
  5. Linux system: setting up the required environment, being familiar with bash scripts
  6. Python programming: creating python projects, importing required packages, visualizing your results
  7. Pytorch programming: creating NNs, setting up training & evaluation pipeline

Prerequisite Hardware

- Grading in this course heavily depends on the coding assignments and final project. You should prepare for your own Linux machine with GPUs for the assignments and the final project. Otherwise, you can try cloud platforms for the access of GPUs.

- Common cloud platforms include:
  - Colab
  - Runpod
  - Google Cloud
  - Amazon AWS

Course Objectives

- The goal of this class is to guide you through “what are the essential components for embodied perception” that facilitates various downstream tasks, such as VR, AR and robotics. Going through each component thoroughly is not our goal.

- Upon completion of the course students should know:
  - How to model the world in 4D
  - How to model the world physically and dynamically
  - How to incorporate multi-sensory inputs
  - What are the essential components in action-centric perception

- Upon completion of the course students should be able to:
  - Write down the formulation of physic simulation
  - Implement a physic- and action-aware visual system

Evaluation

- Assignments: 30%
  - 3 assignments, 10% each
  - Individual submissions
  - Throughout the whole semester, students should submit their assignments before the deadline. **The score of the assignment will be multiplied by 0.9 for each additional day of delay. No grace day is granted**. The submission deadline is based on Taiwan's time zone.

- Final Projects: 70%
  - Teams of 3-4 members
  - Project proposal presentation (10%), two in-person meeting with TW (2 * 10%), and poster session (40%)
  - **In-person participation of these presentations are required**
  - Poster presentation will be graded by the instructor, TA and students

$$score = \frac{1}{N+2} score_{tw} + \frac{1}{N+2} \sum_{i}^{N} score_{TA,i} + \frac{1}{N+2} \frac{1}{|Students|} \sum_{i \in Students} score_i$$

Policy for Assignments

- You are encouraged to discuss with others, but **do not share your codes with them**! If you wrote the same code as others, you may waive the penalty by refactoring your code in-person within limited time, or otherwise, you'll get 15% total grade penalty (for each assignment).

- Please list your collaborator in the appendix of each assignment

- You are allowed to use AIs **at your own risk**. You are responsible for refactoring the code snippets generated by AIs. You'll get the penalty as long as your submitted codes are the same as others.

Policy for the Final Project

- You can form a team of 3-4 members. If you really want to work alone, come and chat with us.
- Project proposal presentation: You will prepare for a 5-minute oral presentation, describing the topic, experimental setup, todos and expected contribution of each member for the final project.
- Two in-person meeting with TW: You will discuss with me about your progress and the current challenges. I may be able to provide some help!
- You will prepare a poster to present your final project, describing the overall ideas, methodology and results. Performance and experimental results are not the key to attract attention! Tell a good story to impress your audience.

Tentative Schedule

Date	Course	Announcements	Student Presentation
2/23	Introduction		
3/2	2D Image Formation		
3/9	3D Representations		
3/16	3D Scene Reconstruction		
3/23	Visual Modeling in Time	HW1 Release	
3/30	Generative Modeling		
4/6	N/A		
4/13	Rigid Body Motion, Articulated Object Modeling, Articulated Object Generation	HW1 due	
4/20	Time Integration, Particle-based System, Position-Based Dynamics, Mass-Spring System in 3D	Deadline of Course Withdraw HW2 Release	Final Project Proposal
4/27	The Finite Element Method, Rigid-Body Modeling, Contact Modeling		
5/4	Material Point Method, Learning to Model Physics (Neural ODE, Lagrangian Network)		

Tentative Schedule

Date	Course	Announcements	Student Presentation
5/11	Learning to Model Physics (PINN, Neural Operator, Fourier Neural Operator), Interactive Simulation	HW3 Release HW2 due	Meet with TW on 5/13
5/18	Multi-sensory Perception (Guest lecture)		
5/25	SLAM, Perceptive Action Decision		
6/1	Visual language model for embodied perception	HW3 due	Meet with TW on 6/3
6/8	N/A (Final exams)		
6/15	N/A	N/A	Final Project Poster

Policy for the Final Project

Date	Announcements
4/20	Final project proposal
6/15	Poster session (9am-12pm)

- If you don't know what to work on, come join the office hours, we're happy to chat with you
- **Think big** and **own** your final project! Don't just consider it as one of the assignments in a class. If you do it well, the final project could end up an awesome paper in a top-tier conference.
- If you want more advice on your final project, come join the office hours or email or message us to arrange for meetings

Policy for course withdrawal

Date	Announcements
4/20	Deadline of Course Withdrawal

- Please be aware of the course withdrawal deadline!
- Any application of course withdrawal after 4/20 **will not** be accepted!
- Please be responsible for your teammates. Piggybackers will be penalized.

- If most of your teammates withdraw, you'd have two options:
  1. Keep working alone. We know it'll be tough, but you'll enjoy at the end of the day.
  2. Team up with others. We'll help you to join other teams.

Policy for in-person meeting with me

Date	Announcements
5/13	Chat with me
6/3	Chat with me

- We offer two in-person meetings to chat with Tsung-Wei about your final project (or anything else). We'd like to provide more guidance on your final project.
- Tell me about the issues / challenges of your final project. I may be helpful to address them.
- We will provide multiple time slots on these two days. First come, first served.

Communication

- Lecture:
  - Please don't hesitate to raise questions if you have any, I may not know the answers but I'll try my best
- Office hours:
  - Great opportunity to discuss project/research idea, confusion about the lecture, debugging issue, or any difficulty you have
  - Tsung-Wei's OH: 13:00-14:00 pm every Monday
- Email:
  - Tsung-Wei's email: twke@csie.ntu.edu.tw
  - TA's email: ntu-ev-2026@googlegroups.com
- NTU COOL:
  - Feel free to post if you have any problem/issue. The whole community will try to help you.

Disclaimer – New(ish) and Evolving Course

- I'm also learning these contents with you, and there will likely be bugs in material and lecture. Apology in advance! Please don't hesitate to point them out.
- Many contents are borrowed/adapted from other professors' lecture.
- Some math equations / notations might be wrong and confusing. Just point them out if you are confused.

Questions?

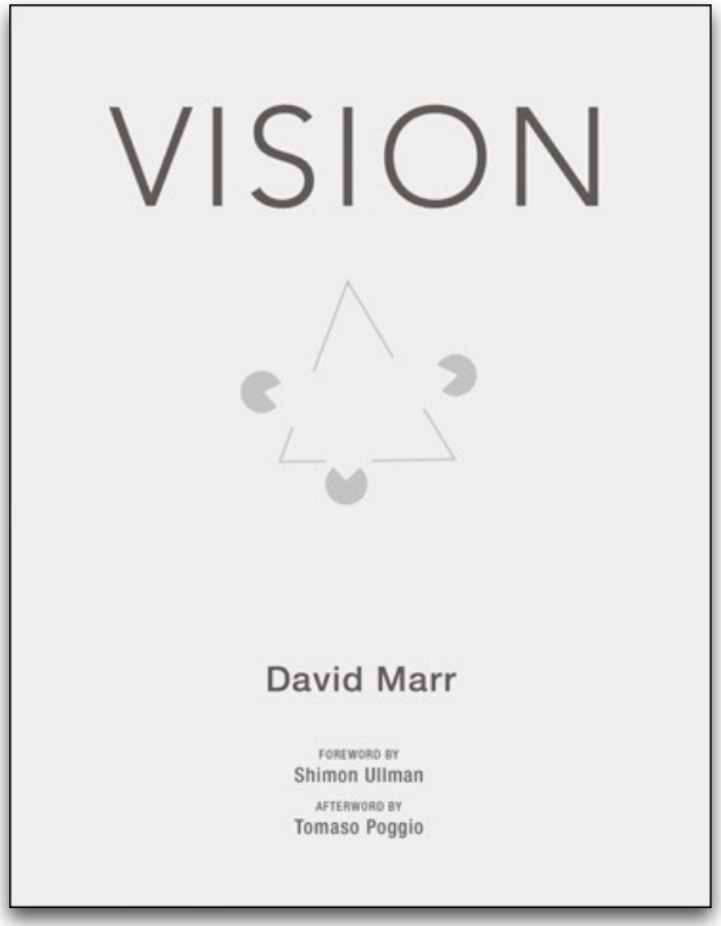
Why do we need this class?
Is this class different from Computer Vision?

The Goal of Computer Vision is to Enable Machines to See the World

To see

"What does it mean, to see? The plain man's answer (and Aristotle's, too). would be, to know what is where by looking."

To discover from images what is present in the world, where things are, what actions are taking place, to predict and anticipate events in the world.



Slide credit B. Freeman and P. Isola

What Does It Mean to See the World…

"I stand at the window and see a house, trees, sky. Theoretically I might say there were 327 brightnesses and nuances of colour. Do I have "327"? No. I have sky, house, and trees."

— Max Wertheimer, 1923

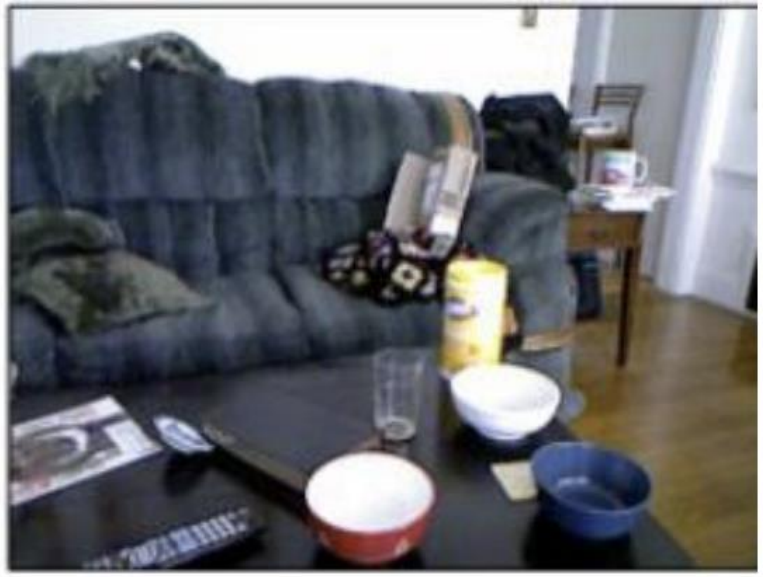
Slide credit P. Isola

What Does It Mean to See the World…

To see

"What does it mean, to see? The plain man's answer (and Aristotle's, too). would be, to know what is where by looking."

To discover from images what is present in the world, where things are, what actions are taking place, to predict and anticipate events in the world.



Slide credit A. Efros

Seeing the World Has Three Aspects

…as measurement

Goals: **Objective** (depth, distance, etc)

Represented by: meters, angles, 3D meshes, etc.

Related fields: mathematics, optics, physics, etc.

…as understanding

Goals: **Subjective** (objects, parts, affordances)

Represented by: words, human annotations, etc.

Related fields: statistics, learning, psychology, epistomology, etc.

… as generation

Goals: Objective (inpainting, dynamics, generation)

Represented by: pixels, D meshes, 3D point clouds, etc

Slides adapted from A. Efros

The Entire Research Field Starts in 1966 as an One-week Summer Project…

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group — July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

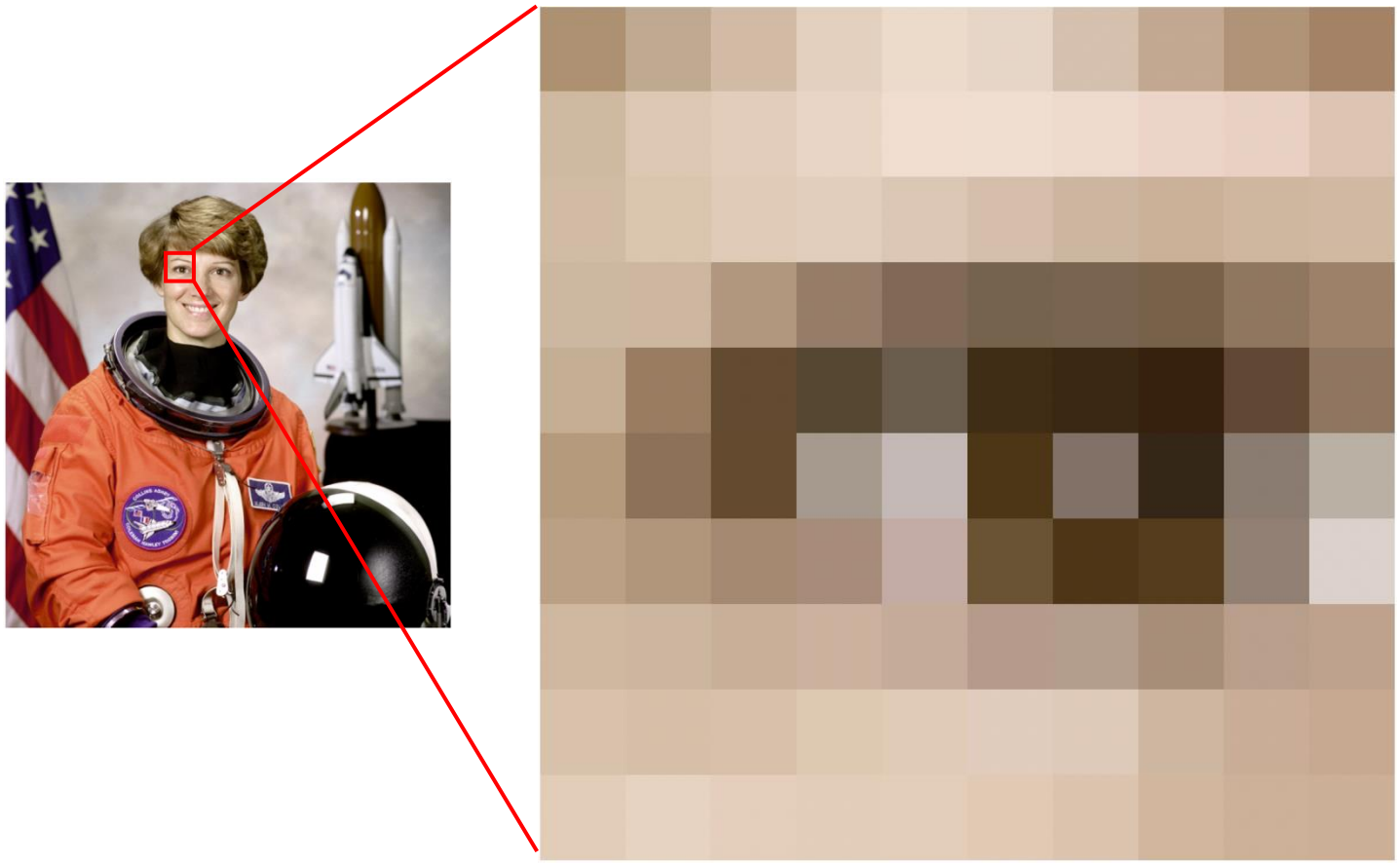
Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

Slide credit P. Isola

Vision is Hard...

What human see



What machines get

179	211	214	209	201	187	192	212	221	231
197	225	223	210	159	147	182	209	219	235
215	230	228	183	103	105	170	206	221	233
232	238	230	155	89	169	173	209	226	231
240	245	224	133	109	198	200	202	229	232
234	244	218	122	66	80	109	187	229	230
218	243	209	124	62	133	82	183	227	224
200	241	207	124	58	55	88	172	211	215
182	237	212	148	100	141	149	190	208	210
170	225	213	163	148	188	221	196	206	210

Challenges: Many Nuisance Parameters

Illumination

Object pose

Clutter

Occlusions

Intra-class appearance

Viewpoint

Slide credit K. Graumen

Challenges: Intra-Class Variation






Slide credit L. Fei-Fei, R. Fergus, and A. Torralba

Challenges: Ambiguity in Inverse Graphics (2D-to-3D)

Depth ambiguity from single-view images

Structure ambiguity from single-view images

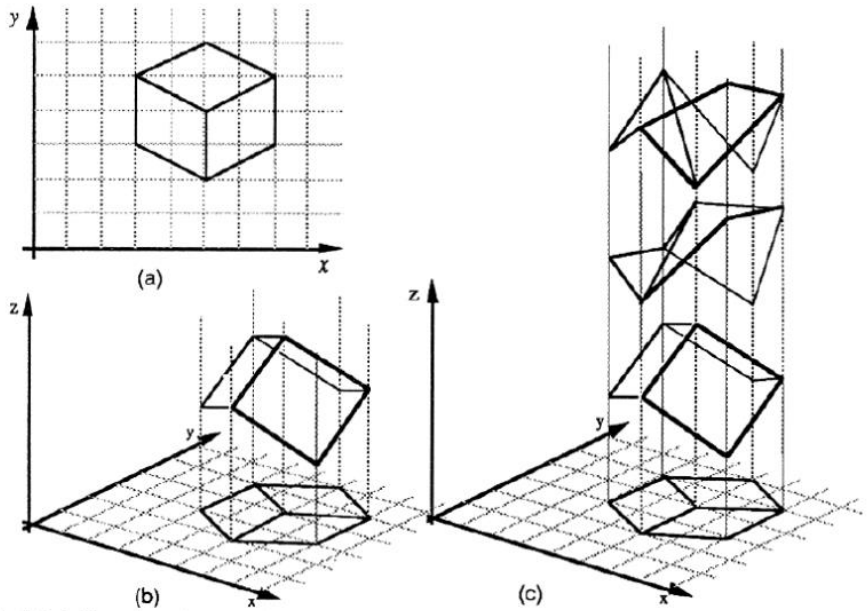
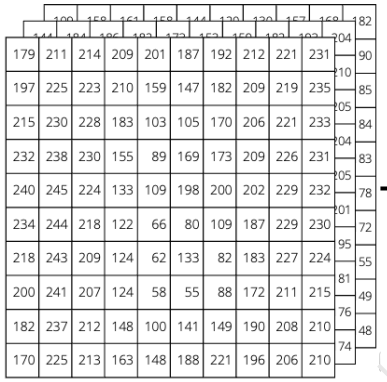


Figure 1. (a) A line drawing provides information only about the x, y coordinates of points lying along the object contours. (b) The human visual system is usually able to reconstruct an object in three dimensions given only a single 2D projection (c) Any planar line-drawing is geometrically consistent with infinitely many 3D structures.

Sinha & Adelson 93

Image credit A. Efros and L Lazebnik

A Generic Formulation of Computer Vision Models

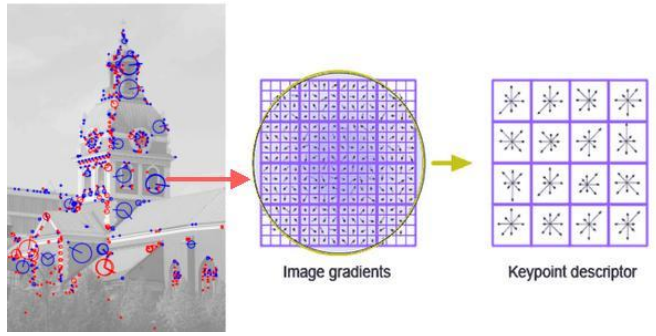


feature + classifier → Labels

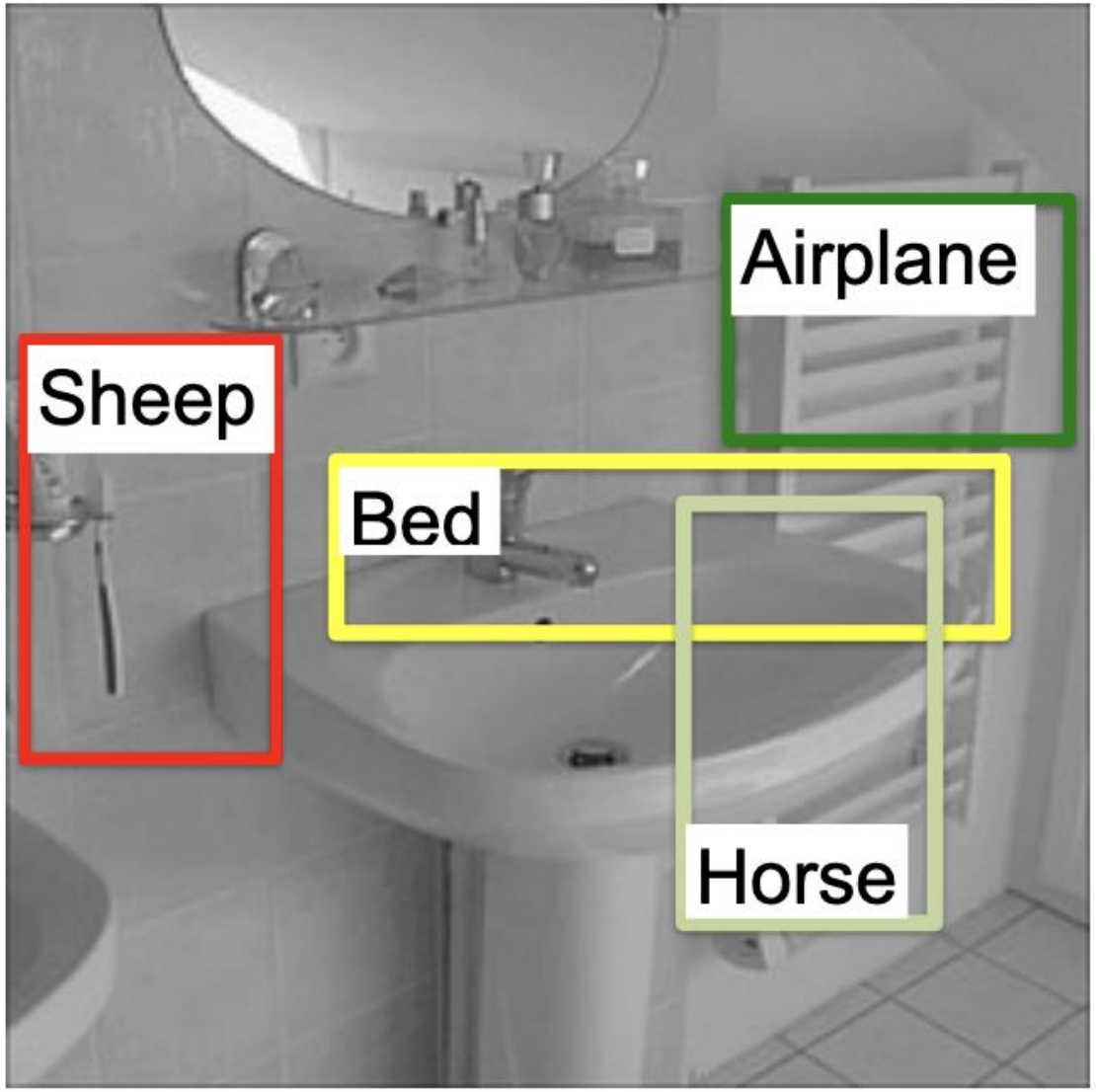
All we need is a better feature extractor and classifier

Histogram of Oriented Gradients

Scale-invariant Feature Transform



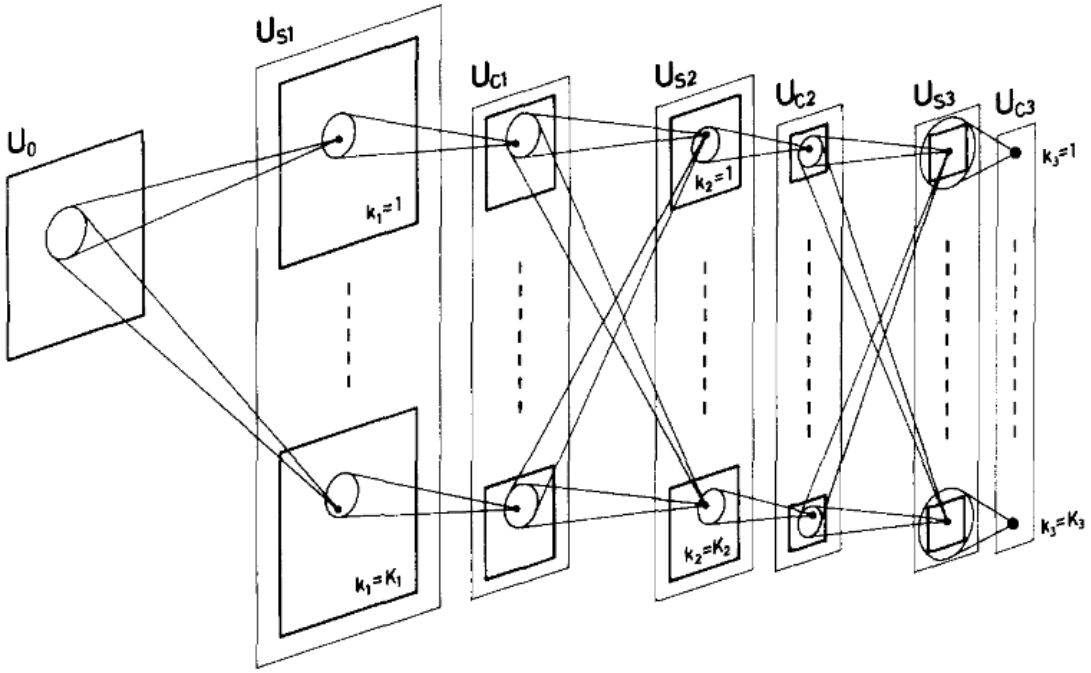
Things were Not Working Before 2012…



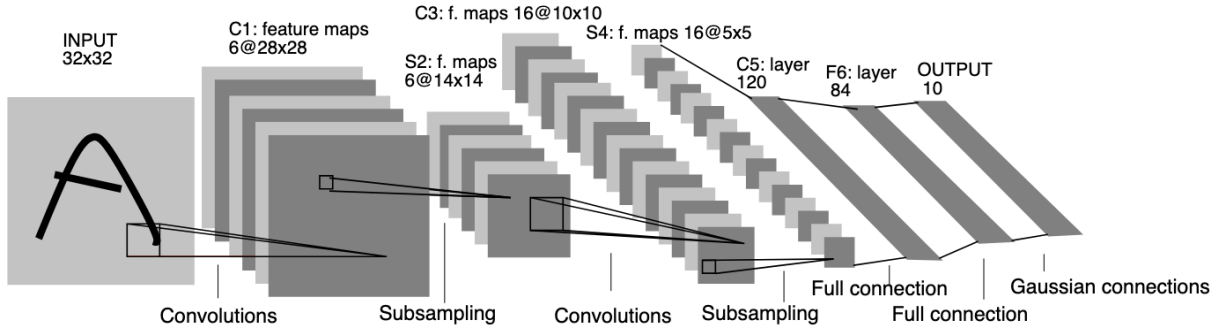
Slide credit B. Freeman and P. Isola

Shouldn't CNN Work? In Fact, the Idea was Introduced 40 Years Ago, but It didn't Work....

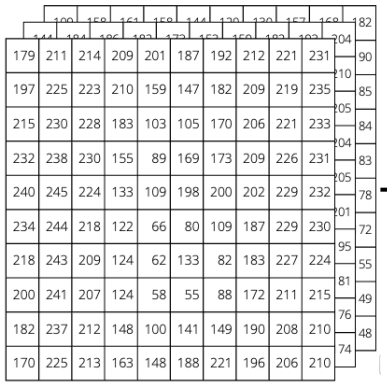
Neocognitron proposed by K. Fukushima (1980)



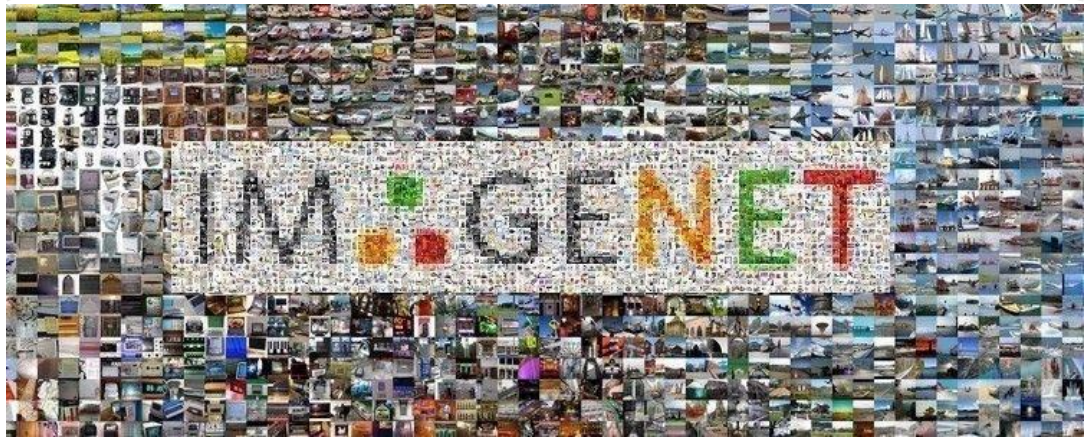
Lenet proposed by Y. Lecun (1998)



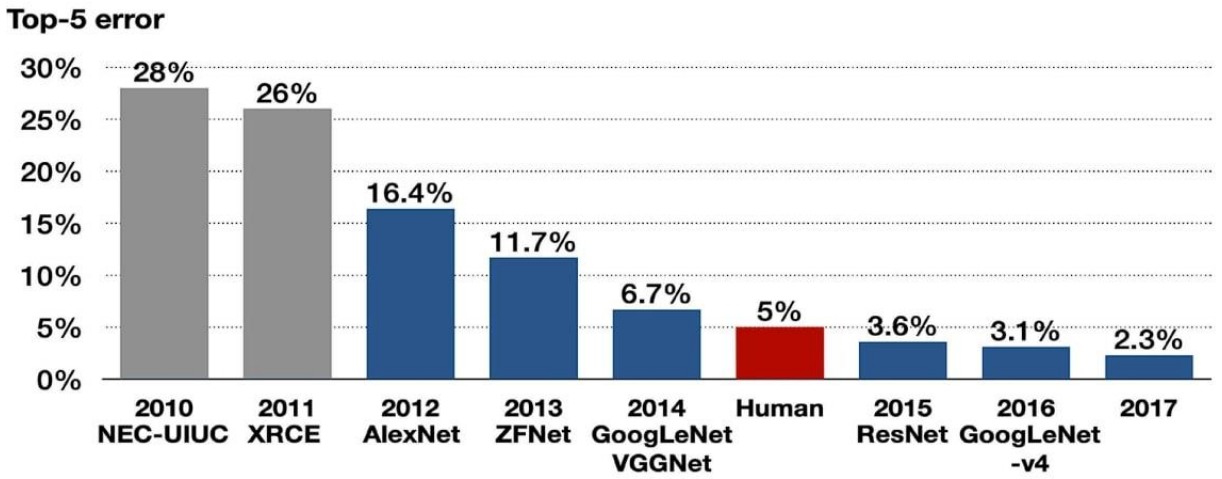
It Turns Out **Data** is the Key



CNN + classifier → Labels

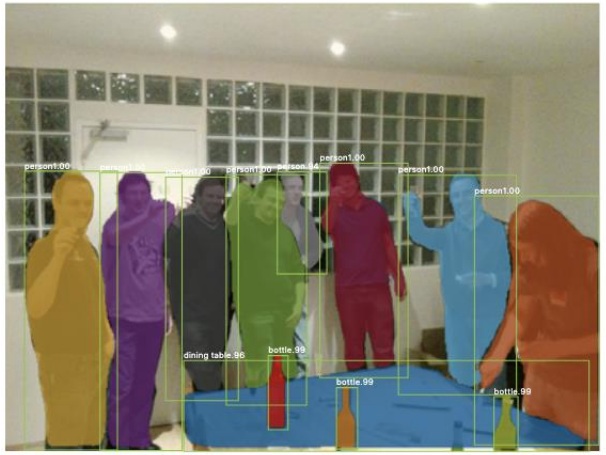


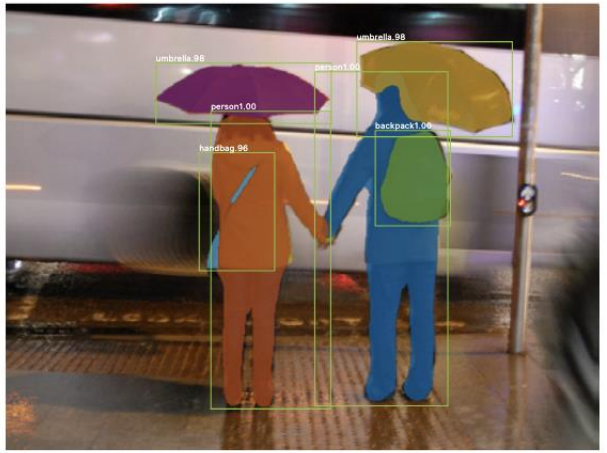
1,000,000 images with 1,000,000 labels

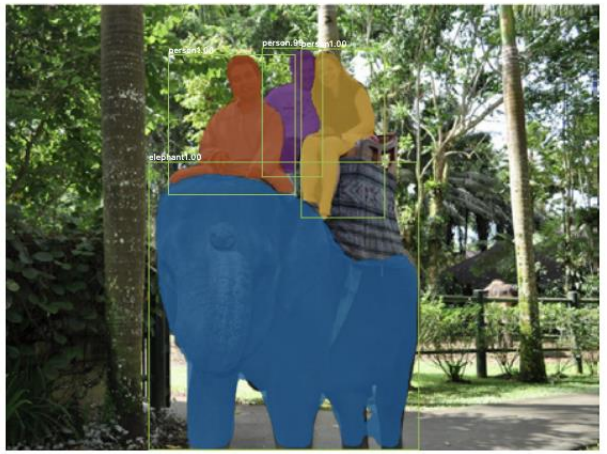


Computer Vision is Solved with the “Data is All You Need” Paradigm

Microsoft COCO: Common Objects in Context.

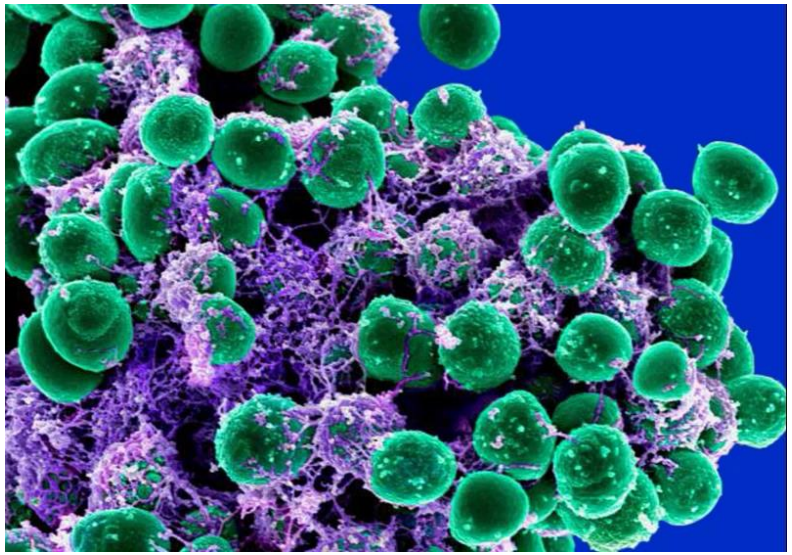






Mask R-CNN. He et al.

Computer Vision is Solved with the “Data is All You Need” Paradigm

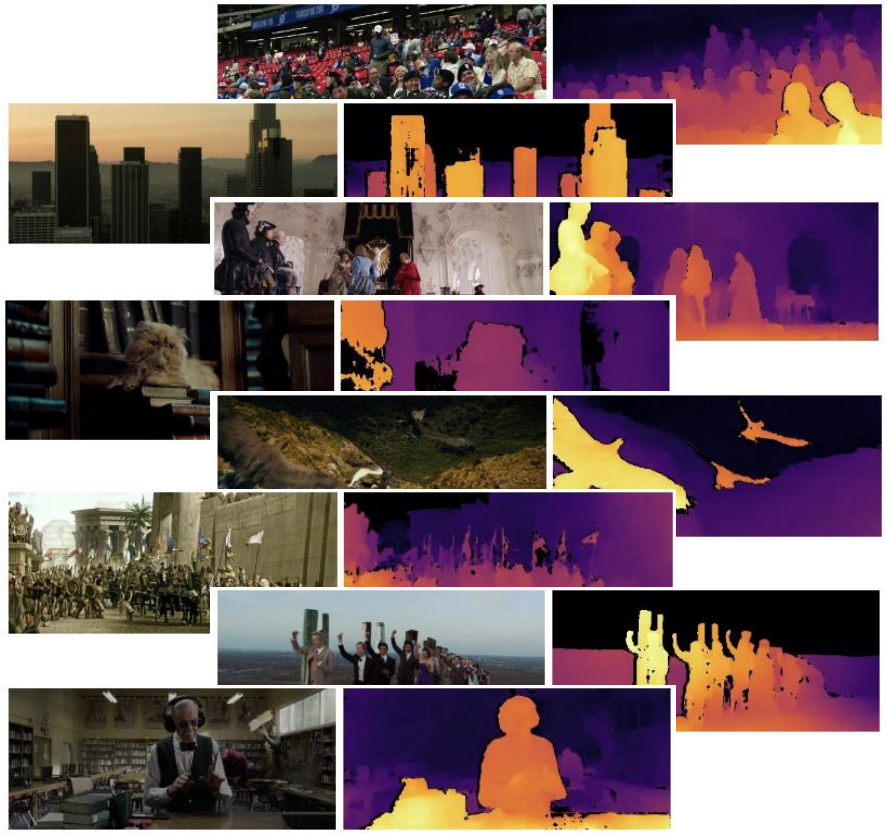


Segment Anything Model. Kirilov et al.

Computer Vision is Solved with the “Data is All You Need” Paradigm

SAM 2: Segment Anything in Images and Videos. Ravi et al.

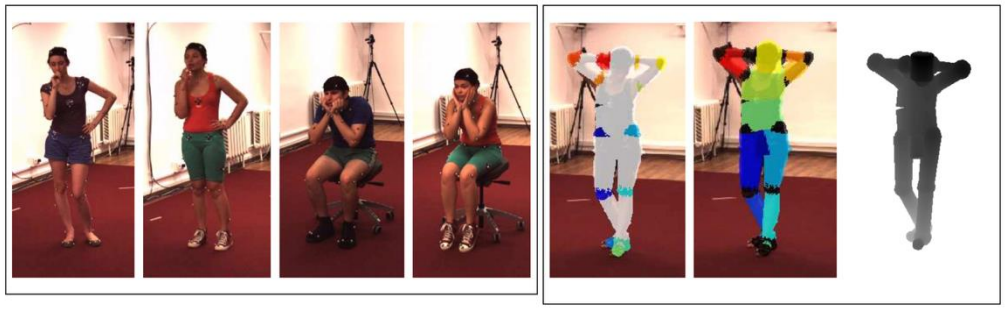
Computer Vision is Solved with the “Data is All You Need” Paradigm



3D Movies Dataset

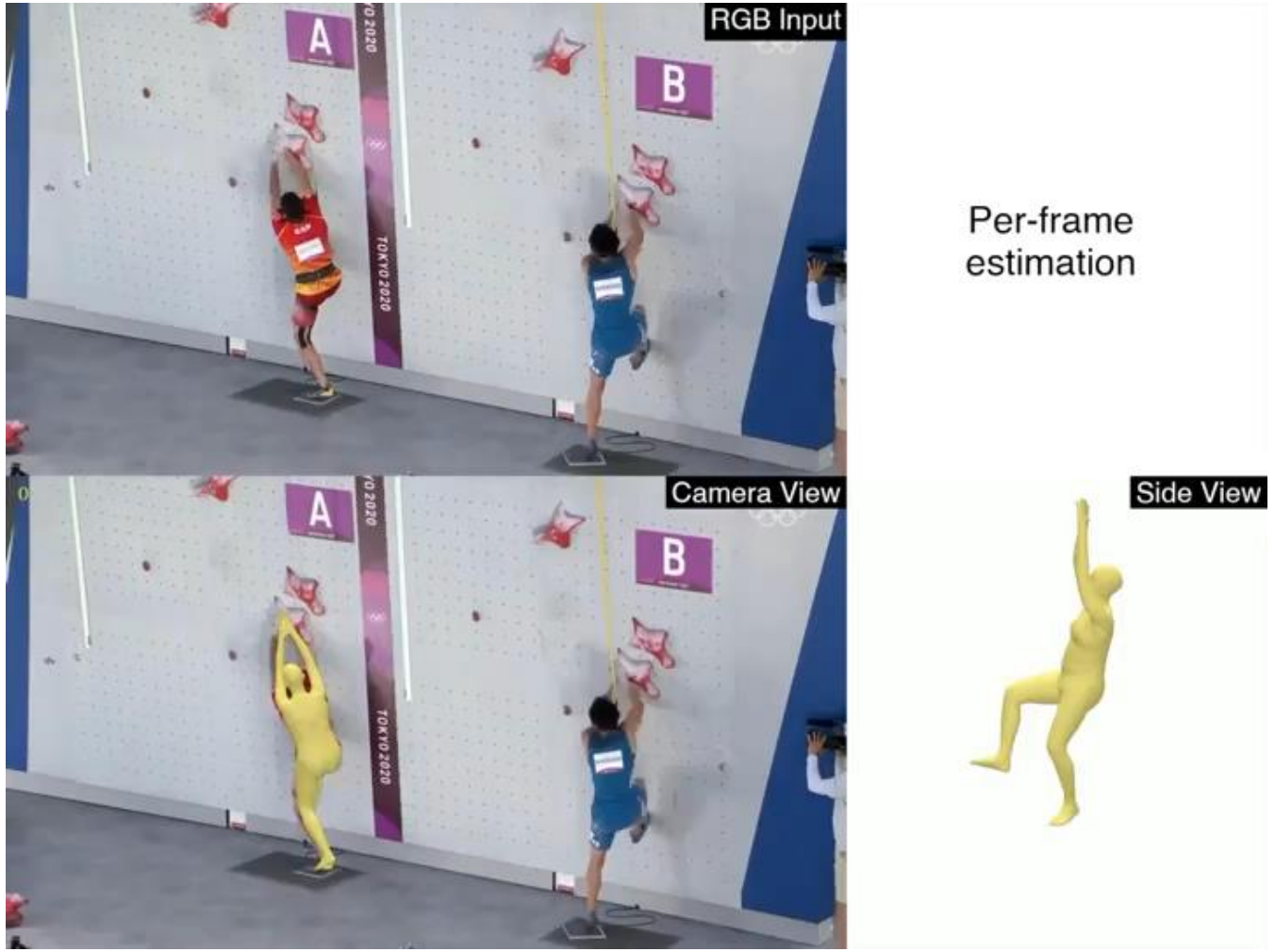
Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. Rantfl et al.

Computer Vision is Solved with the “Data is All You Need” Paradigm

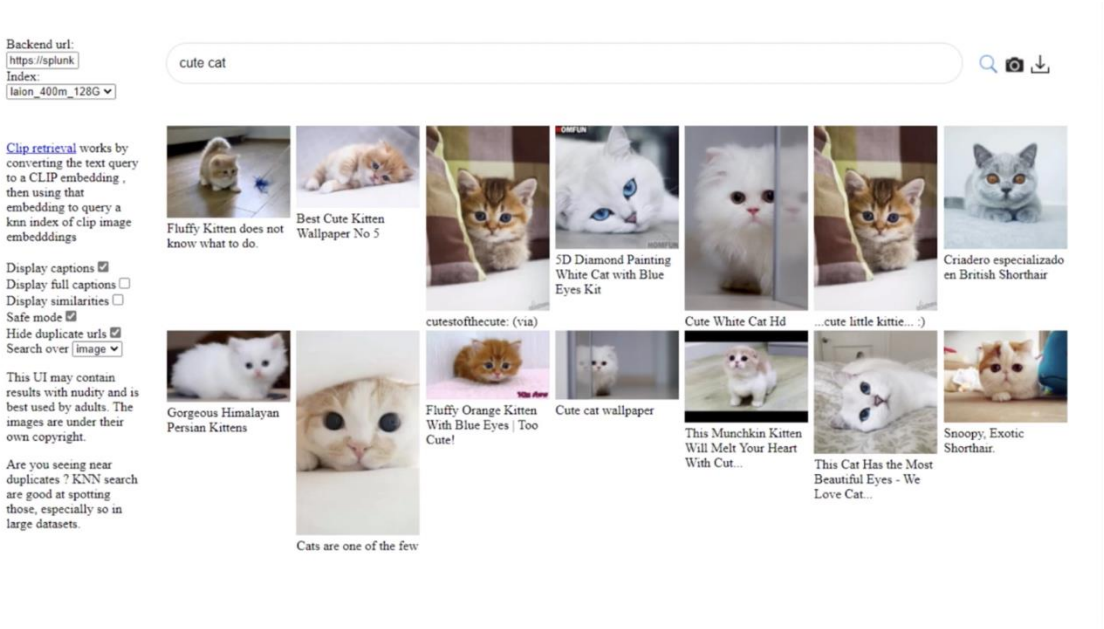


3.6M Human

MPII Human Pose Dataset



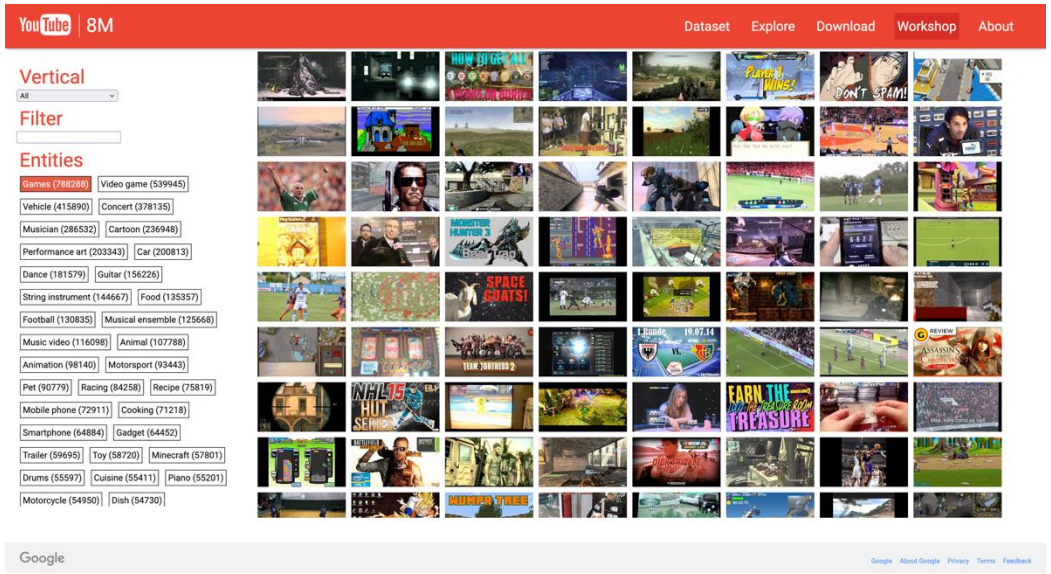
Computer Vision is Solved with the “Data is All You Need” Paradigm



Text-paired Images on the Internet

Stable Diffusion

Computer Vision is Solved with the “Data is All You Need” Paradigm



Videos on the Internet

SORA by OpenAI

Computer Vision is Solved with the “Data is All You Need” Paradigm

Objaverse-XL

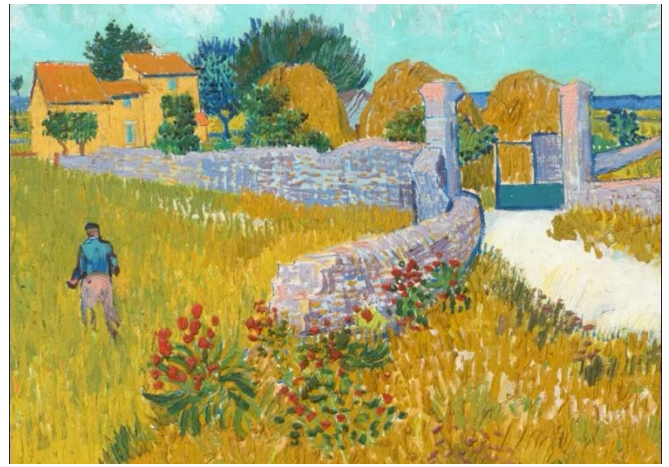
SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion. Voleti et al. ECCV 2024.

It Seems that We Have Solved the Computer Vision Task

This image features a small Chihuahua dog lounging on a bright pink couch. **The dog is wearing a soft pink bathrobe and oversized pink sunglasses, giving it a relaxed and stylish look. It is holding a remote control in one paw and has a bowl of popcorn in front of it, as if getting ready for a cozy movie night.** The dog's relaxed posture and accessories make the scene humorous and adorable.

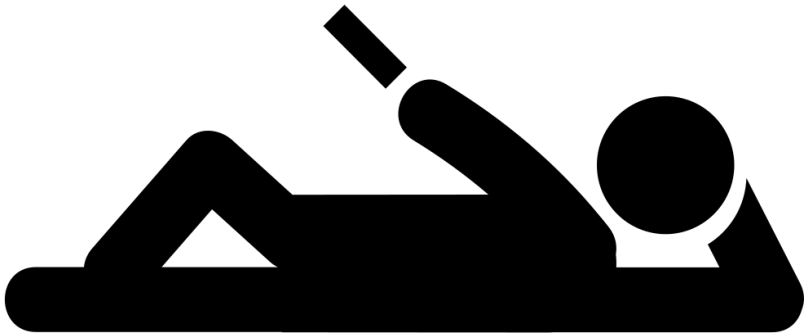
Image source: https://images.app.goo.gl/XeWqM1bvosZNQQyc9

Let's Take a Closer Look at the Tasks We've Solved



This image features a small Chihuahua dog lounging on a bright pink couch. **The dog is wearing a soft pink bathrobe and oversized pink sunglasses, giving it a relaxed and stylish look. It is holding a remote control in one paw and has a bowl of popcorn in front of it, as if getting ready for a cozy movie night.** The dog's relaxed posture and accessories make the scene humorous and adorable.

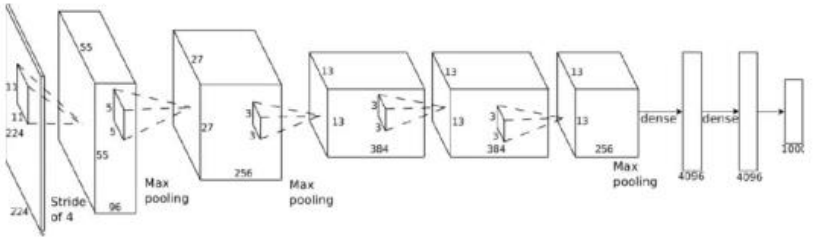
Vision is Over-Simplified: Passively Parsing the Information Received from the World



Icon source: https://www.pngegg.com/en/png-zhyar

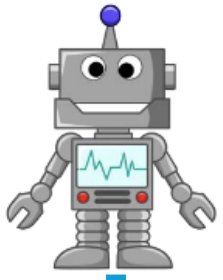
Passive labeling

o_t



Tiger

y_t



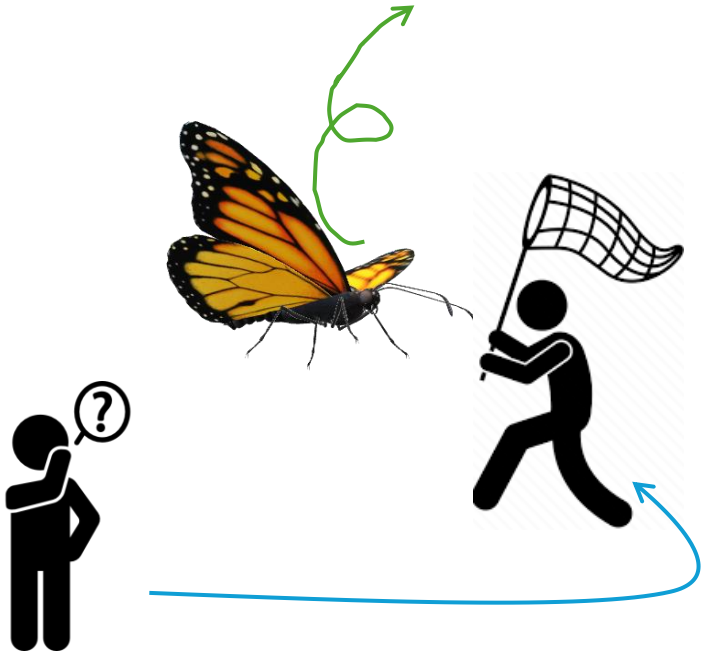
Observations o_{t+1}

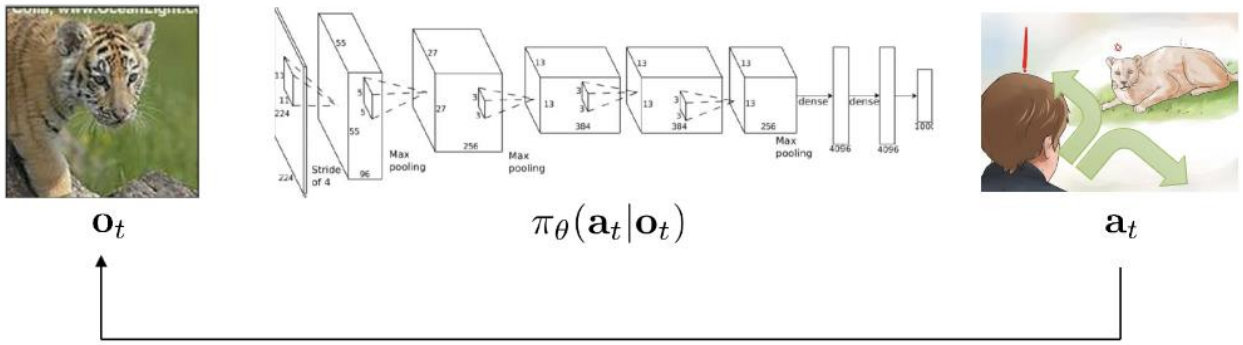
Observations o_t

Label y_t

Image source S. Levine

Humans are Embodied Agents: We Act, We Perceive and We Predict





$\mathbf{s}_t$ – state
$\mathbf{o}_t$ – observation
$\mathbf{a}_t$ – action

$\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ – policy
$\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ – policy (fully observed)

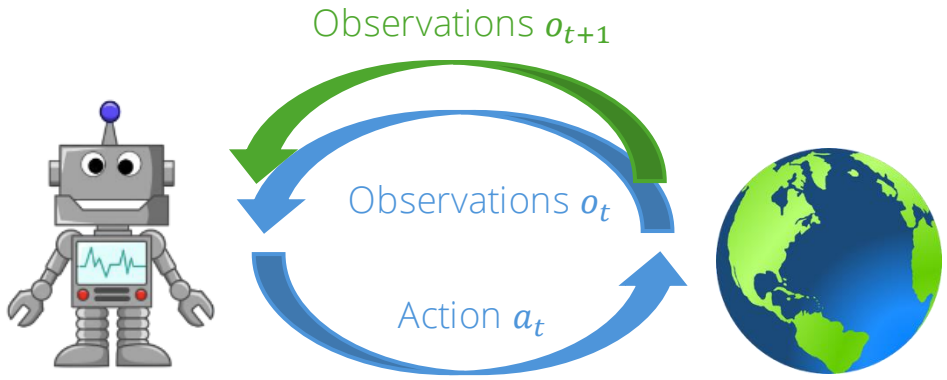
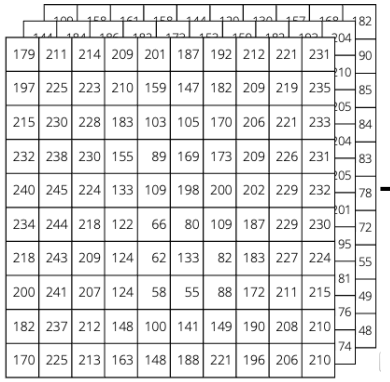
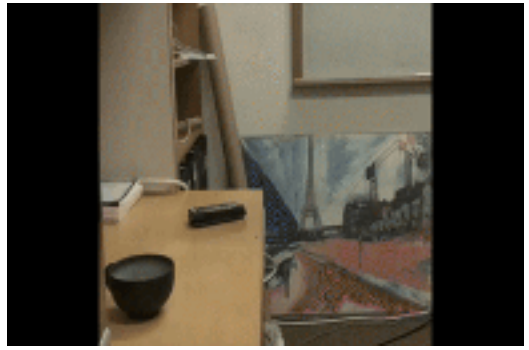


Image source S. Levine

An Example of a Passive Generative Model



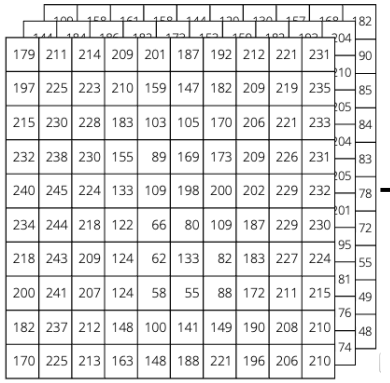
179	211	214	209	201	187	192	212	221	231
197	225	223	210	159	147	182	209	219	235
215	230	228	183	103	105	170	206	221	233
232	238	230	155	89	169	173	209	226	231
240	245	224	133	109	198	200	202	229	232
234	244	218	122	66	80	109	187	229	230
218	243	209	124	62	133	82	183	227	224
200	241	207	124	58	55	88	172	211	215
182	237	212	148	100	141	149	190	208	210
170	225	213	163	148	188	221	196	206	210



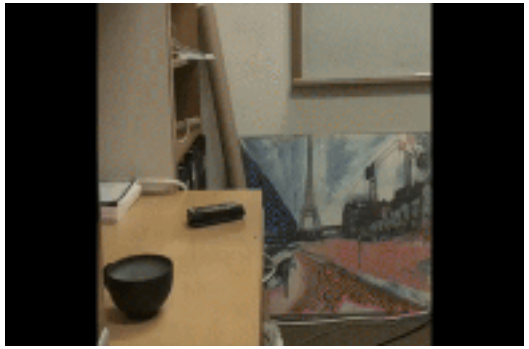
Video generative model

An Example of an Embodied Generative Model

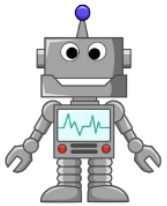
Laws of Physics
Environmental parameters
Object shape
Object 3D location (4D motion): $o_1, o_2, \ldots, o_{T-1}$
Object physical properties



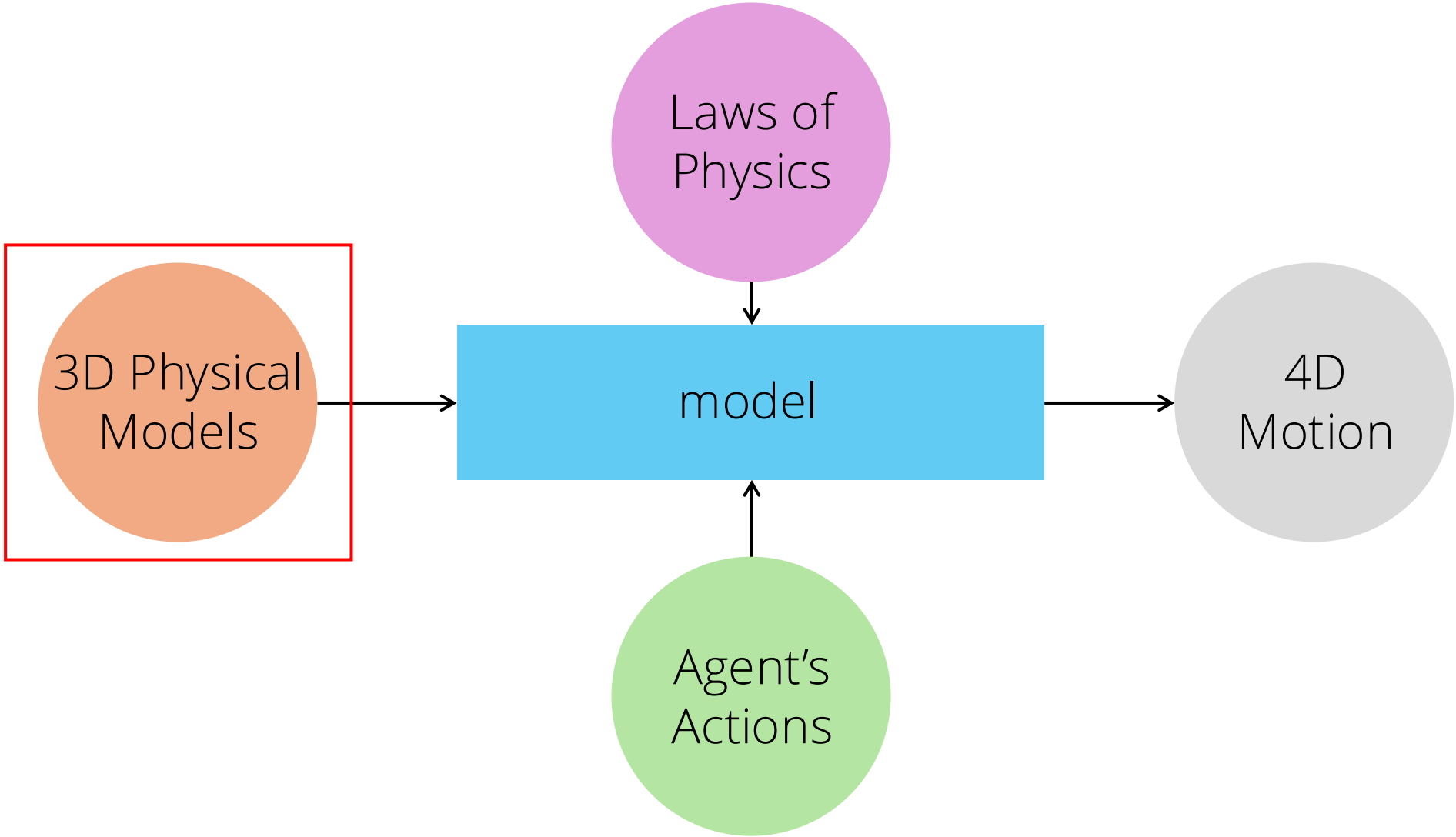
Video generative model



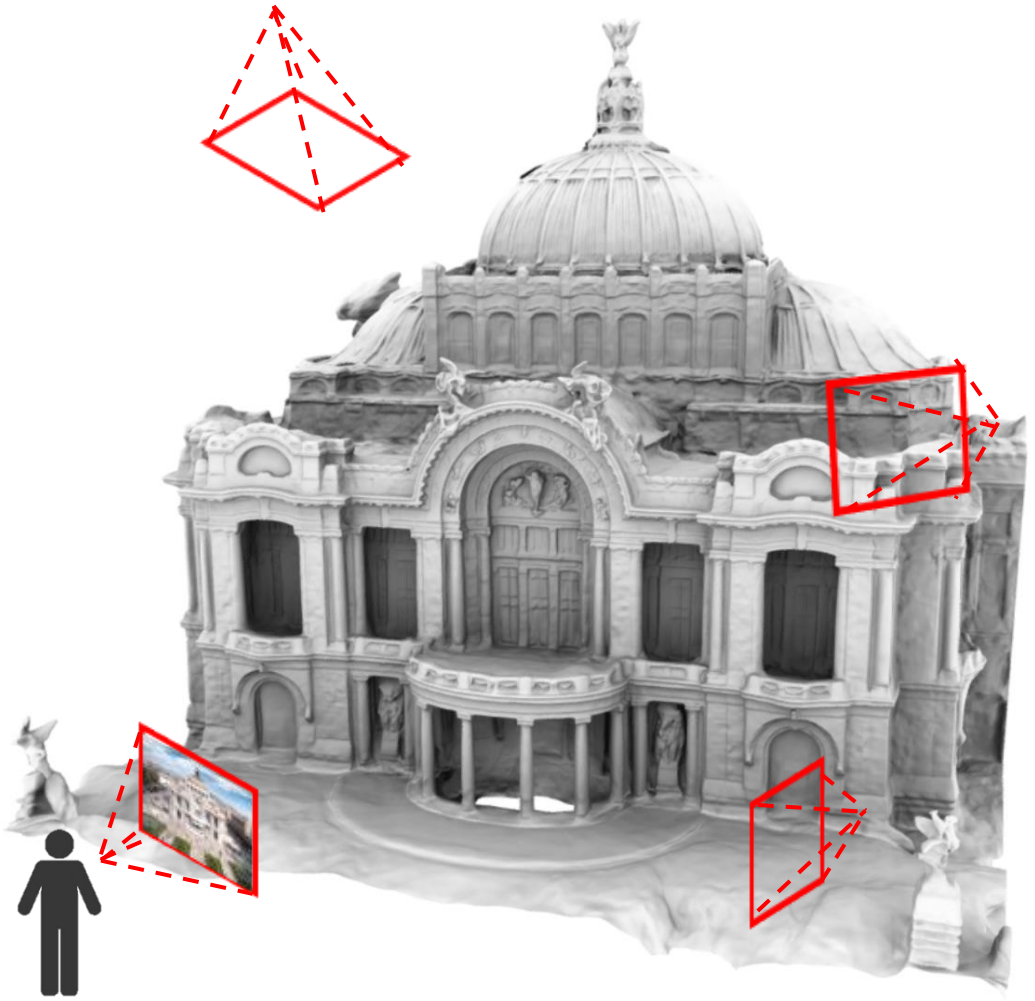
Action: $a_1, a_2, \ldots, a_{T-1}, a_T$
3D Location (4D motion): $l_1, l_2, \ldots, l_{T-1}$
Shape
Physical properties
…



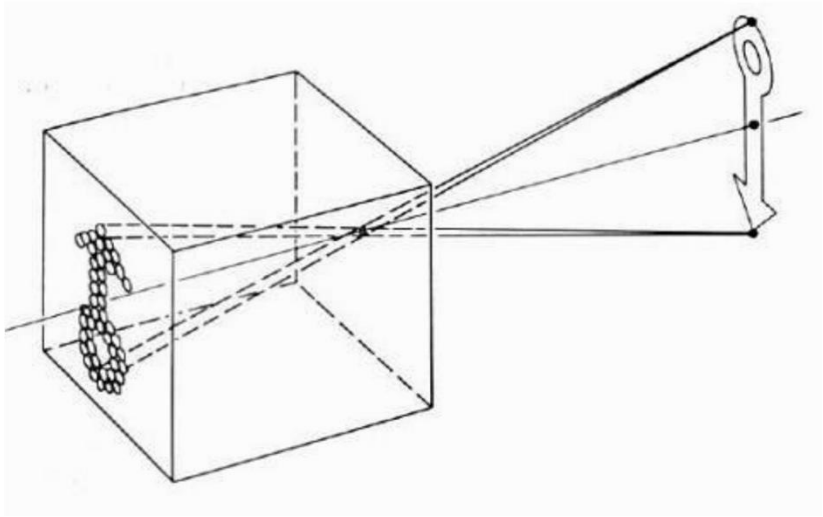
An Example of an Embodied Generative Model



What are the Essential Components in Embodied Vision? Reconstructing 3D Models from 2D Visual Observations



2D Image Formation from the 3D World



Pinhole camera model

A simplified model of photometric image formation

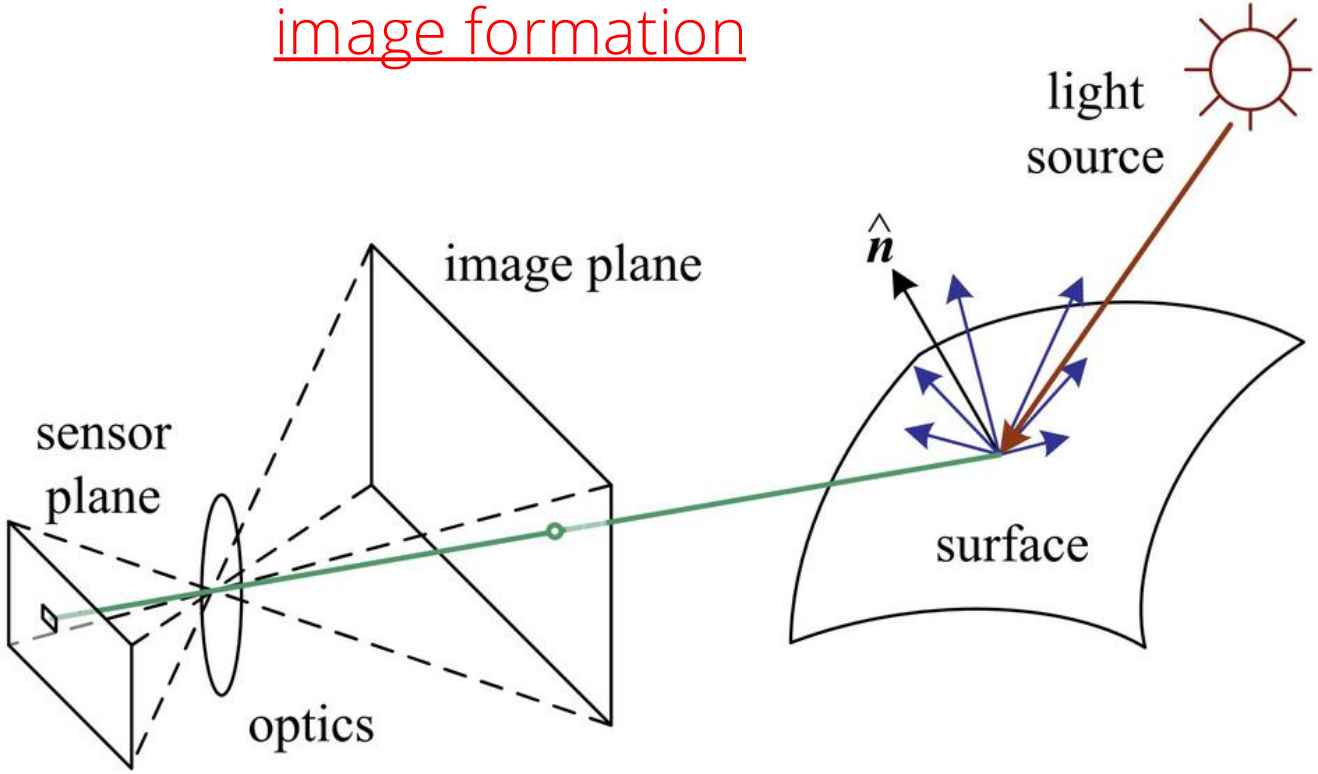


Image source S. Seitz text book

3D Model Reconstruction from 2D Images

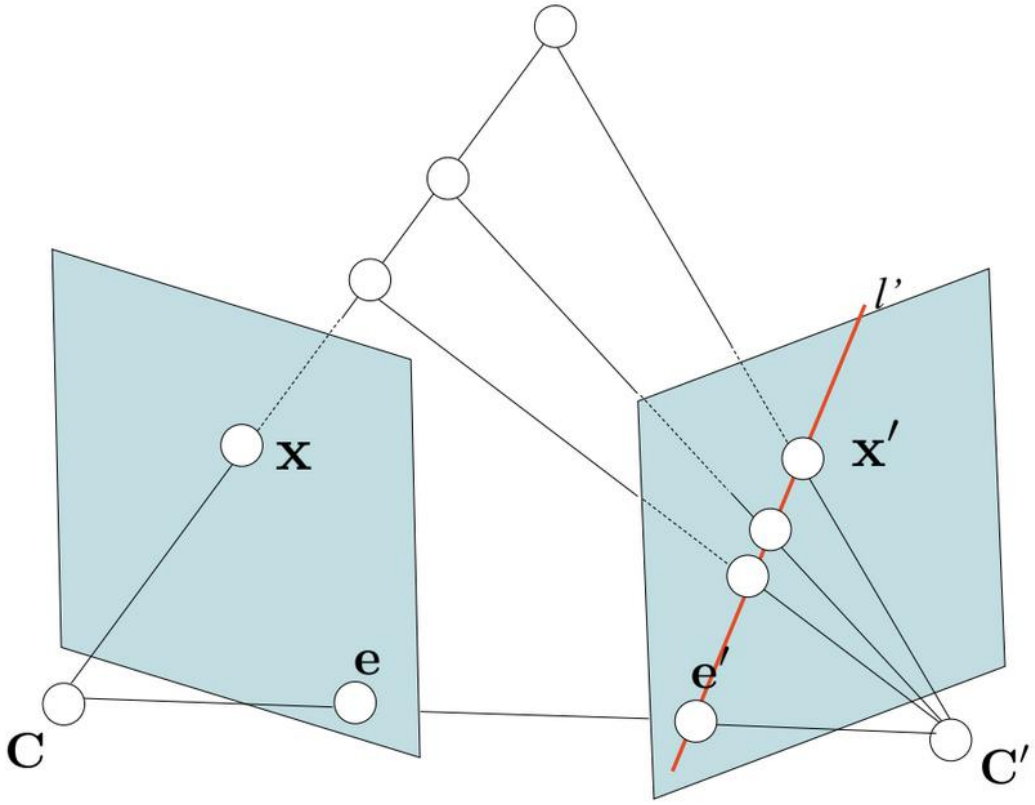


Image source M. Hebert

Image source https://3dvision.princeton.edu/courses/SFMedu/
Image credit S. Tulsiani

3D Model Reconstruction from 2D Images

Building Rome in a Day. Agarwal et al.

Representations of 3D Models

Depth map

Image source Paul Bourke

Pointcloud

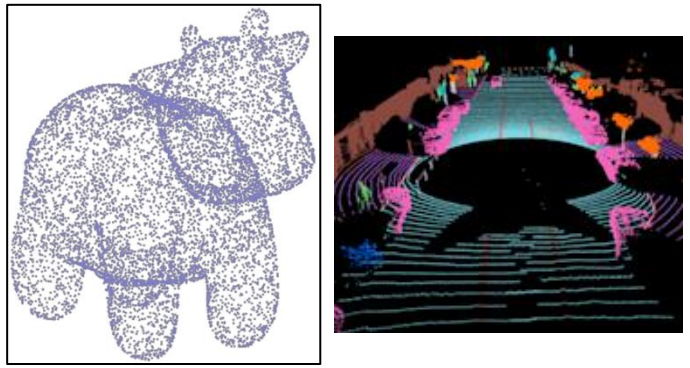


Image source S. Tulsiani Image source Waymo

Mesh

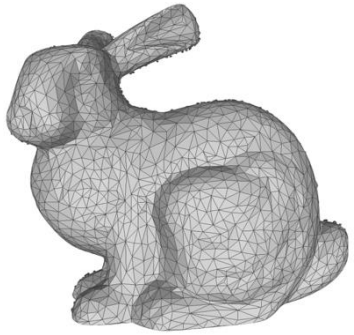
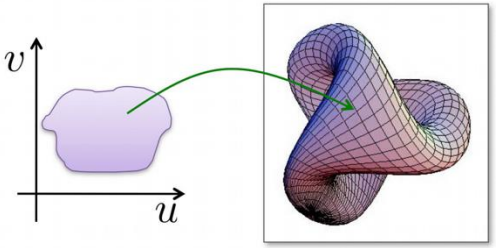


Image source S. Tulsiani

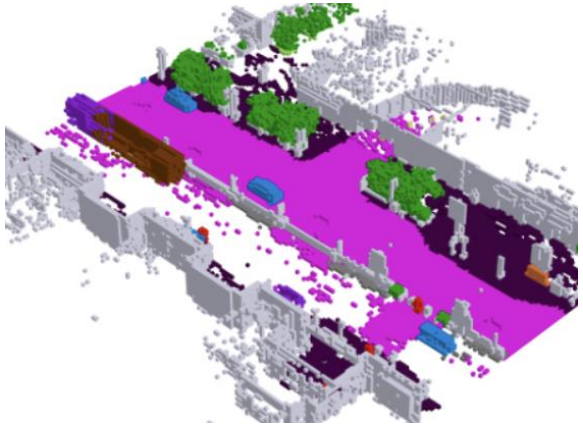
Parametric Surface



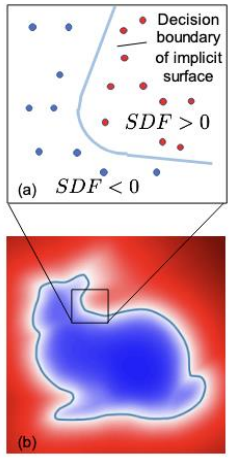
$$f(\mathbf{u}) = \mathbf{p} \in \mathbb{R}^3;\ \mathbf{u} \in \mathcal{M}$$

Image source S. Hao

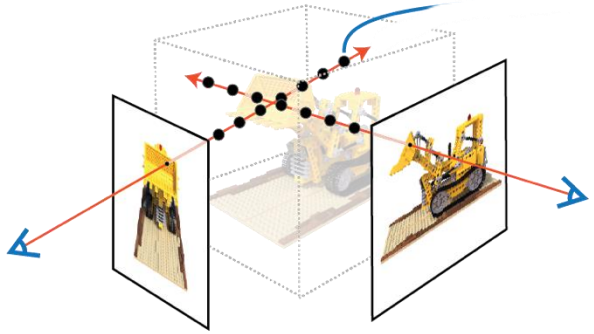
Voxels



Signed Distance Function

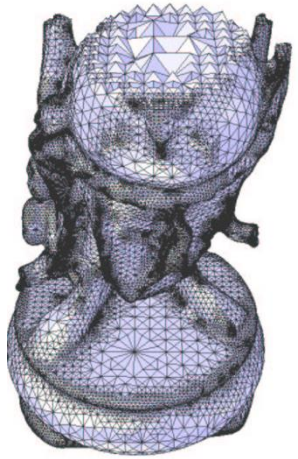


NERF

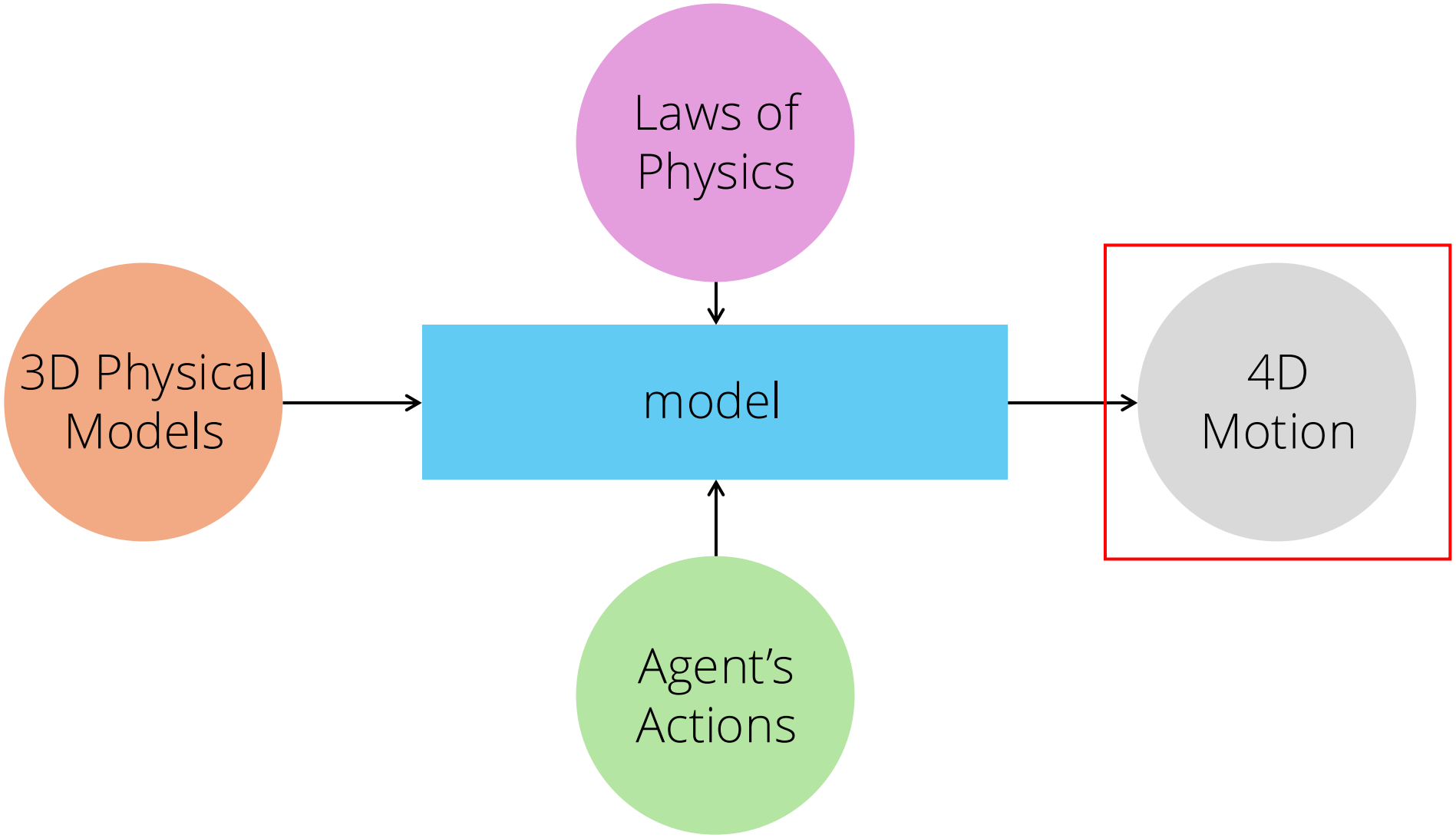


3D Gaussian Splatting

Tetrahedral Mesh

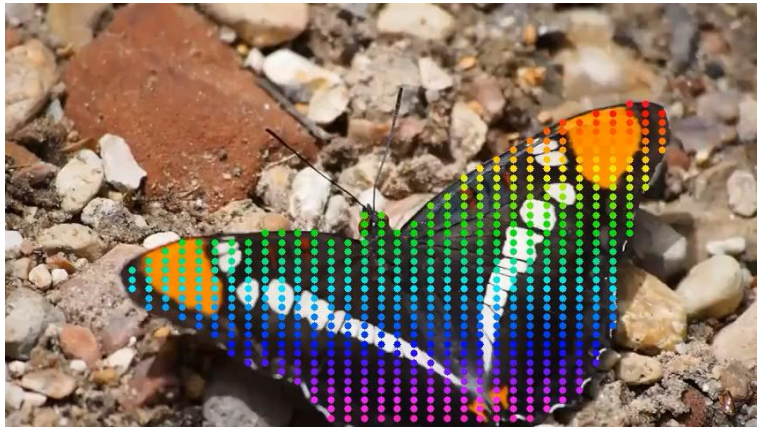


An Example of an Embodied Generative Model



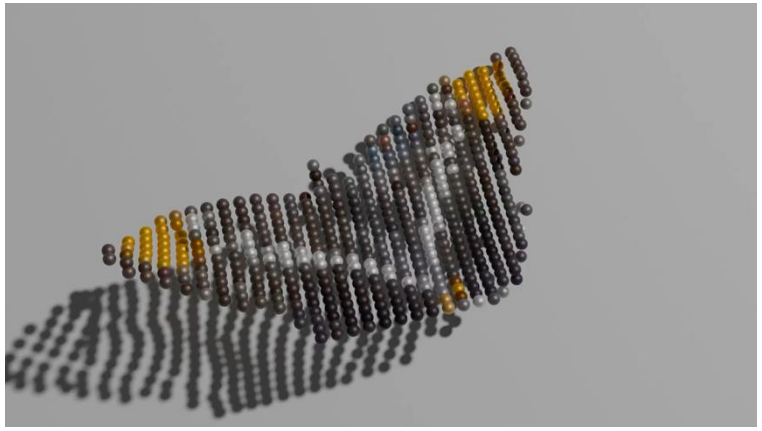
Representations of Motion

2D Pixel Motion

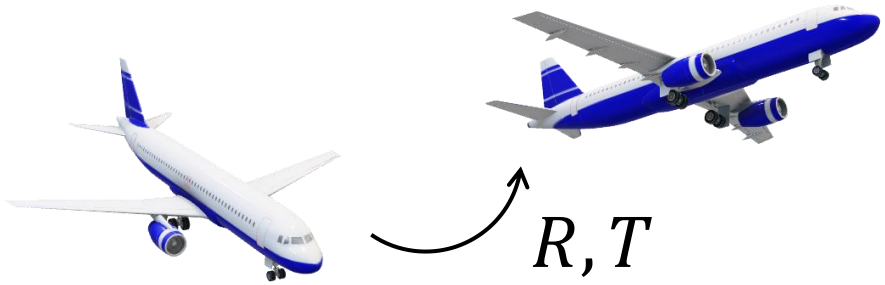


Video source SpatialTracker

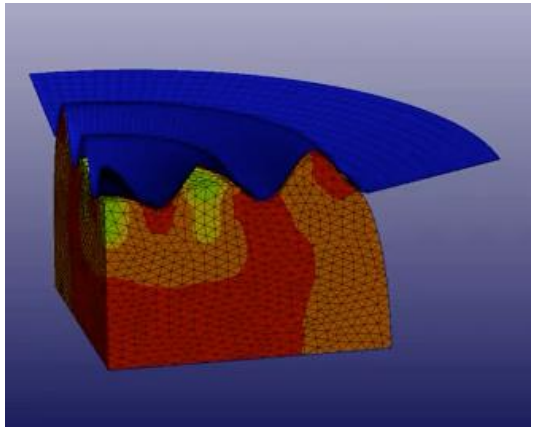
3D Particle Motion



3D Rigid Body Motion

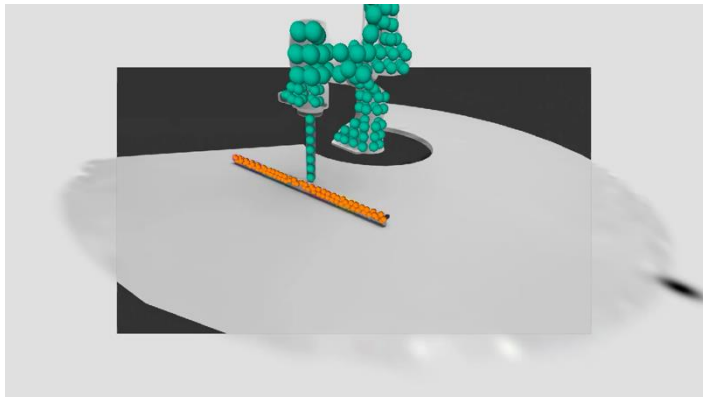


Tetrahedral Mesh Deformation



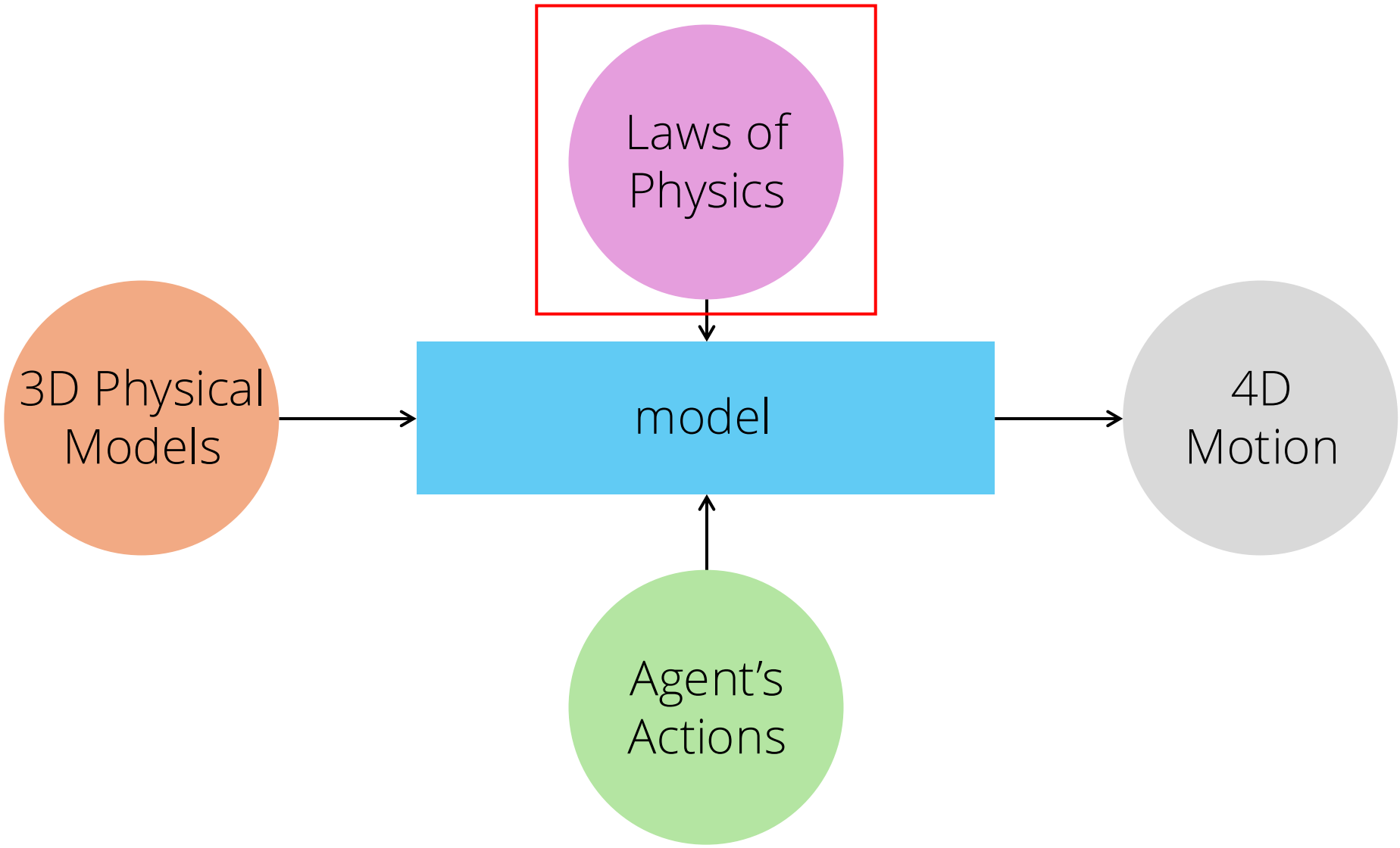
https://youtu.be/hVf2RTzv2o0?si=LEqN33vqcb12nGvZ_

3D Gaussian Splatting Motion



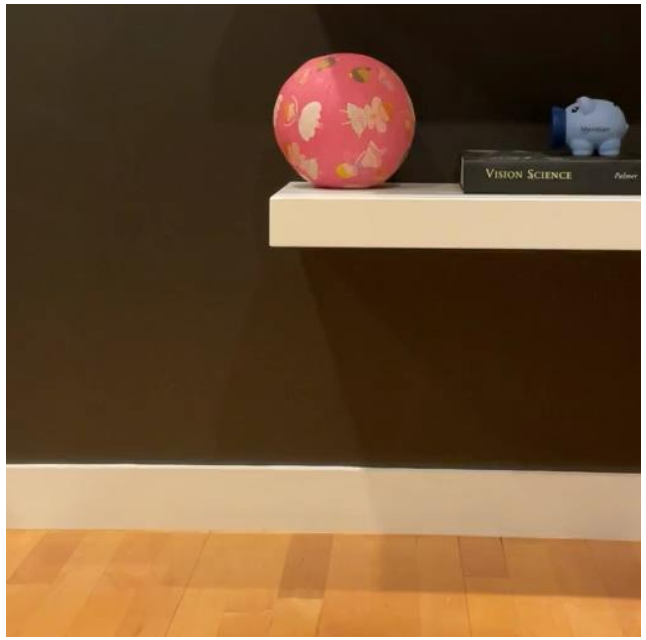
Video source EmbodiedGS

An Example of an Embodied Generative Model

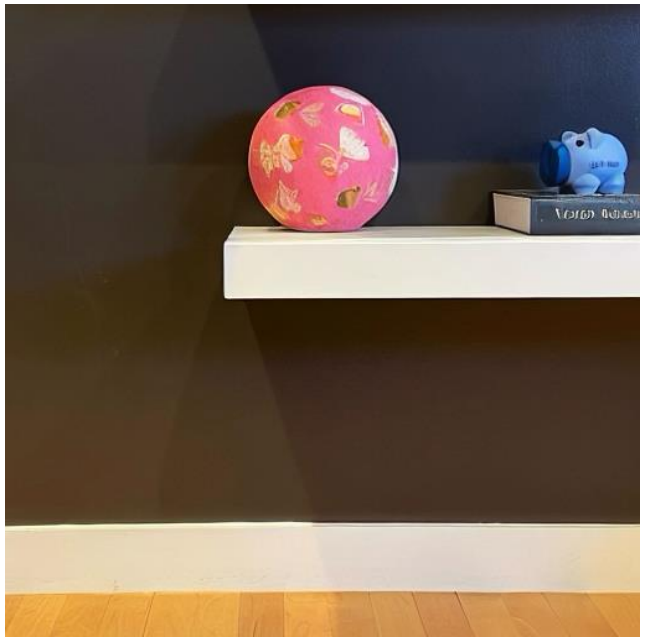


We are Embodied Agents and Physics Matter

Real physics



Fake physics



PhysGen: Rigid-Body Physics Grounded Image-to-Video Generation. Liu etal.

We are Embodied Agents and Physics Matter

We are Embodied Agents and Physics Matter

https://www.youtube.com/watch?v=RqajKat0v-4

https://www.youtube.com/watch?v=eFsT3QgIvoI

https://youtu.be/L49NBoW5Xw0?si=W8w8qi6guVdo7j_W

https://youtube.com/shorts/xSqvz-ystw4?si=M_UoOSmRS_Uahtr9

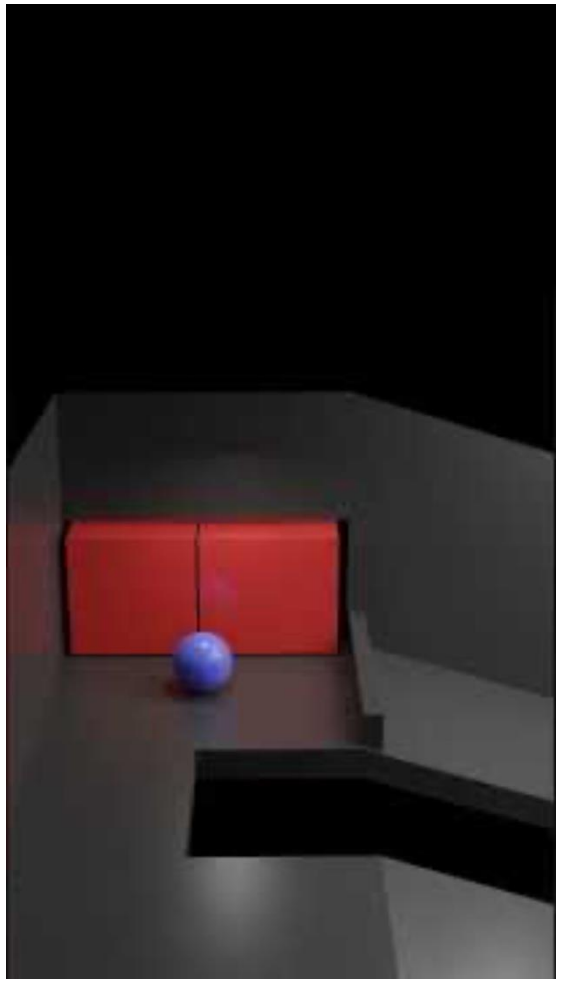
How to Get Physics Involved?
An Easy Case: Rigid Body Dynamics

$$\mathbf{q}^i(t) = \left[\mathbf{t}^i(t), \mathbf{R}^i(t), \boldsymbol{\nu}^i(t), \boldsymbol{\omega}^i(t)\right]$$

$$\frac{d}{dt}\mathbf{q}(t) = \frac{d}{dt}\begin{bmatrix}\mathbf{t}(t)\\ \mathbf{R}(t)\\ \boldsymbol{\nu}(t)\\ \boldsymbol{\omega}(t)\end{bmatrix} = \begin{bmatrix}\mathbf{v}(t)\\ \boldsymbol{\omega}(t)\times\mathbf{R}(t)\\ \frac{\mathbf{F}(t)}{M}\\ \mathbf{I}(t)^{-1}(\boldsymbol{\tau}-\boldsymbol{\omega}(t)\times\mathbf{I}(t)\boldsymbol{\omega}(t))\end{bmatrix}$$

$$\mathbf{q}(t) = \mathbf{q}(0) + \int_0^t \frac{d}{dt}\mathbf{q}(t)dt = \mathbf{q}(0) + \sum_{i=1}^{T} \frac{d}{dt}\mathbf{q}(t)|_{t=t_i}\Delta_t$$

How to Get Physics Involved?
An Easy Case: Rigid Body Dynamics



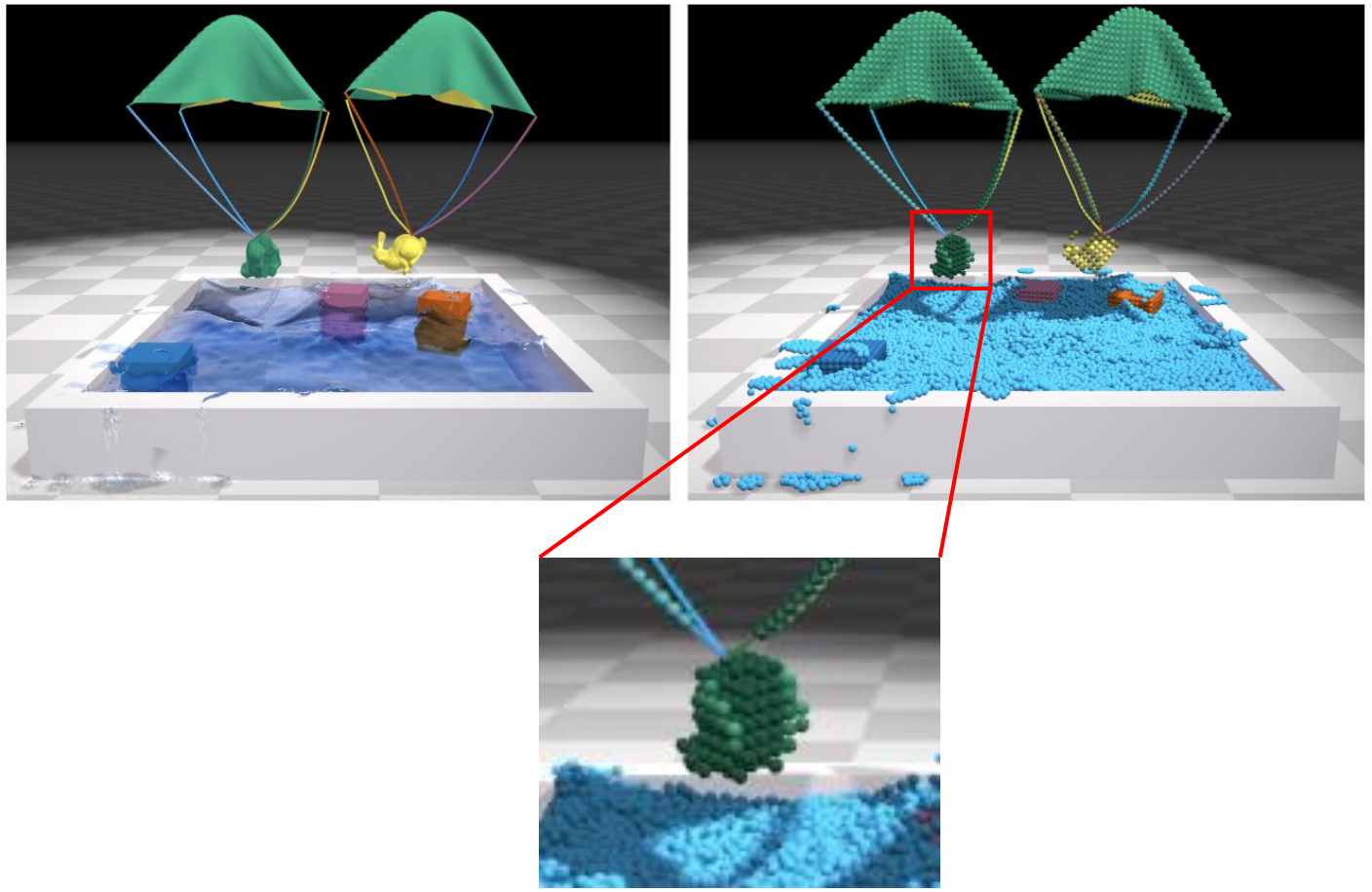
https://www.youtube.com/shorts/_jkeH4N93FY?feature=share

https://youtu.be/_XKWFSecrXs?si=R_5en86yWTFYi2Qk

How to Get Physics Involved?
A Harder Case: Deformable Object / Fluid Dynamics

Physical Modeling with Position Based Dynamics

Position Based Dynamics. Müller et al.



Continuum Mechanics

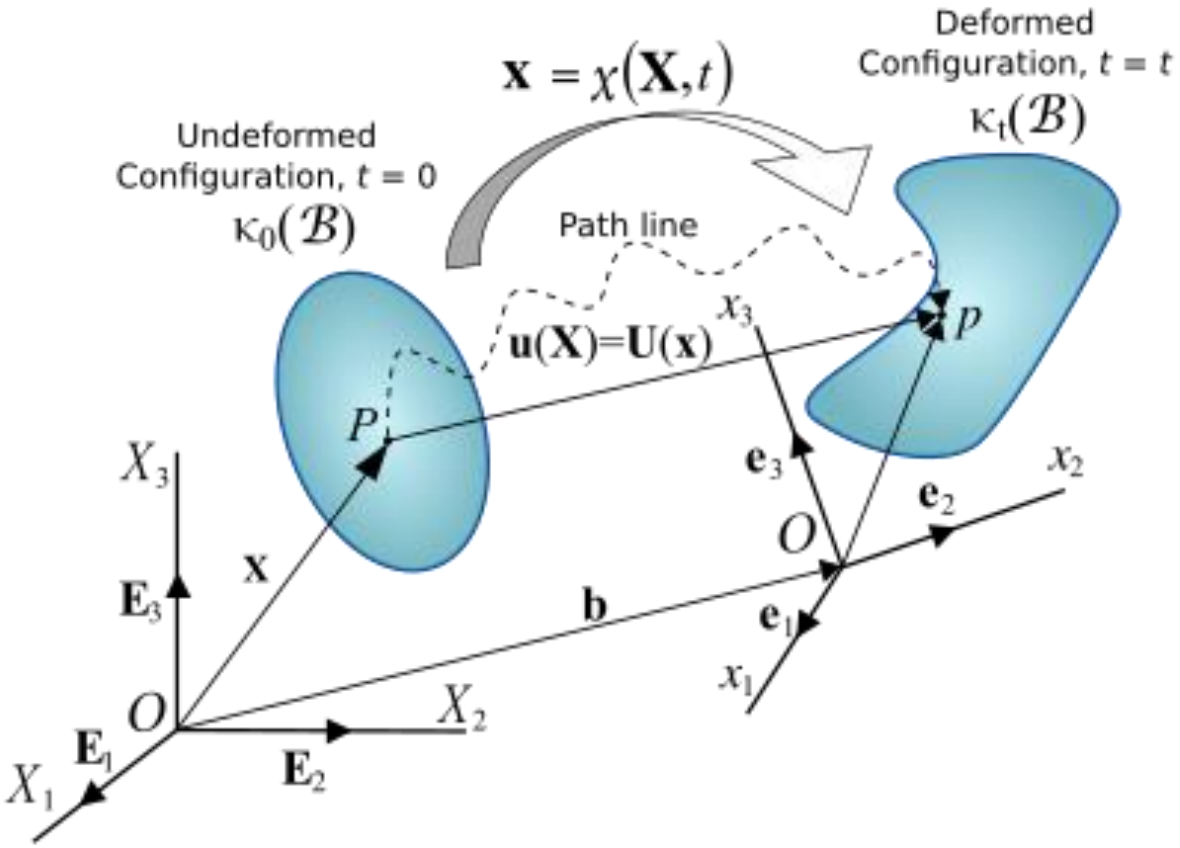


Image source: wiki

$$\frac{D}{Dt}\rho(\mathbf{x},t) + \rho(\mathbf{x},t)\nabla^{\mathbf{x}} \cdot \boldsymbol{\nu}(\mathbf{x},t) = 0 \quad \text{Conservation of mass,}$$

$$\rho(\mathbf{x},t)\frac{D\boldsymbol{\nu}}{Dt} = \nabla^{\mathbf{x}} \cdot \boldsymbol{\sigma} + \rho(\mathbf{x},t)\mathbf{g} \quad \text{Conservation of momentum,}$$

Weak form of force balance:

$$\int_{\Omega^t} q_i(\mathbf{x},t)\rho(\mathbf{x},t)a_i(\mathbf{x},t)d\mathbf{x} = \int_{\partial\Omega^t} q_i t_i ds(\mathbf{x}) - \int_{\Omega^t} q_{i,k}\sigma_{ik} d\mathbf{x}$$

Physical Modeling with Material Point Method

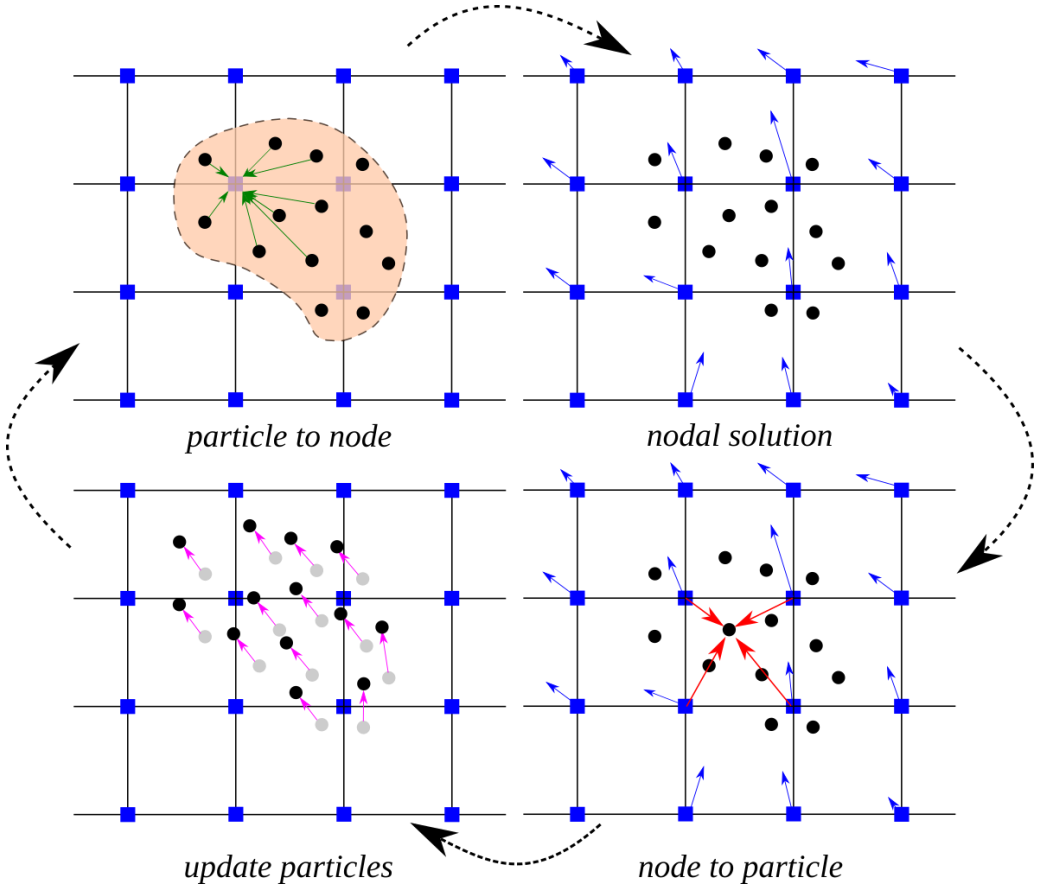
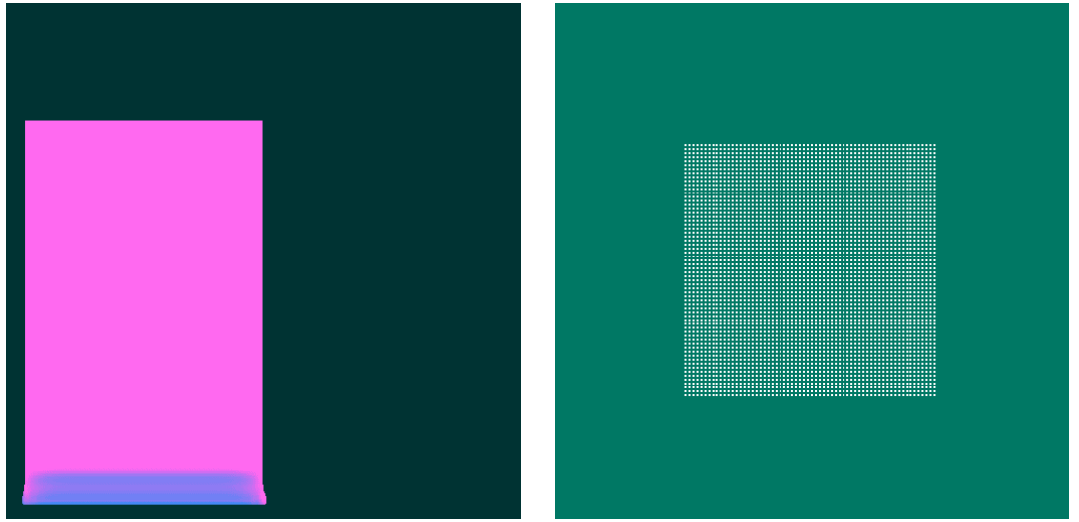


Image source
https://www.geoelements.org/LearnMPM/mpm.html



Video source:
https://nialltl.neocities.org/articles/mpm_guide

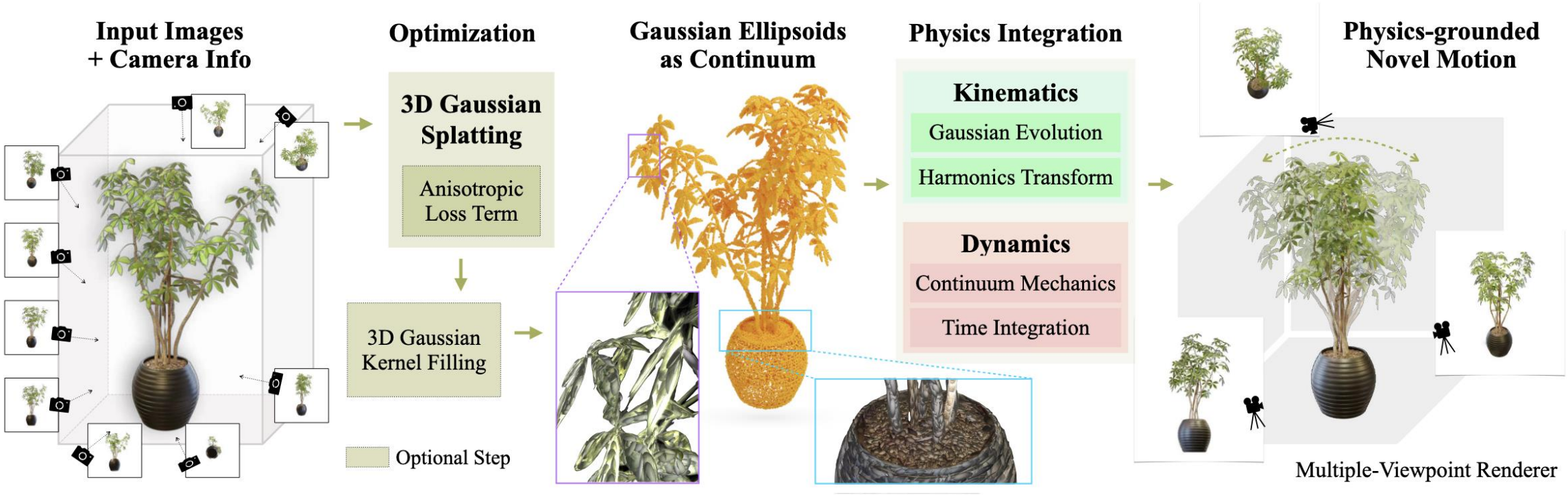
How to Get Physics Involved?
A Much Harder Case: Contact Modeling

https://youtu.be/o94Pm-Cty3M?si=Oxn1Q68Edou4XLrn

Physical Modeling with Incremental Contact Modeling

Codimensional Incremental Potential Contact. Li et al

How to Get Physics Involved?
Animate Reconstructed 3D Models with Physics



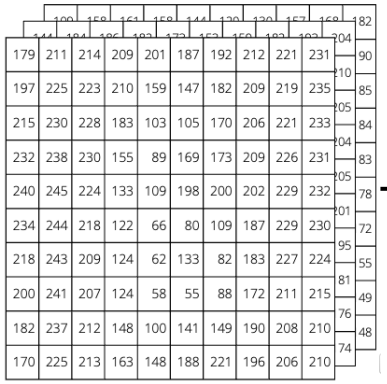
PhysGaussian: Physics-Integrated 3D Gaussians for Generative Dynamics. Xie et al. CVPR 2024

How to Get Physics Involved?
Animate Reconstructed 3D Models with Physics

How Far We are from Solving Physical Modeling?

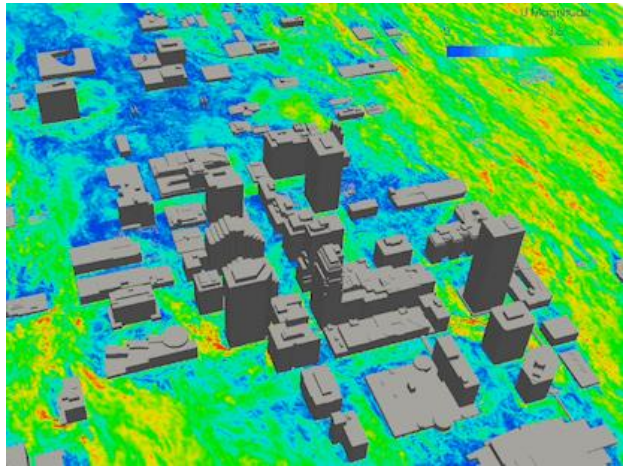
Precisely modeling the physical world is one of the core scientific problems.
SOTA simulation is coarse but computationally expensive…

Can We Reproduce the Success Driven by Data for Physical Modeling?

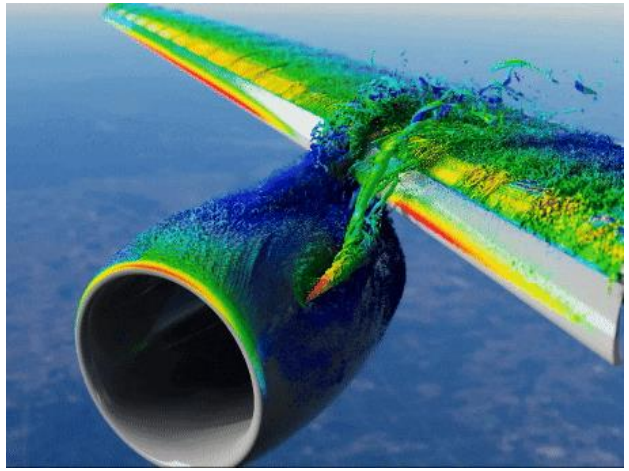


179	211	214	209	201	187	192	212	221	231
197	225	223	210	159	147	182	209	219	235
215	230	228	183	103	105	170	206	221	233
232	238	230	155	89	169	173	209	226	231
240	245	224	133	109	198	200	202	229	232
234	244	218	122	66	80	109	187	229	230
218	243	209	124	62	133	82	183	227	224
200	241	207	124	58	55	88	172	211	215
182	237	212	148	100	141	149	190	208	210
170	225	213	163	148	188	221	196	206	210

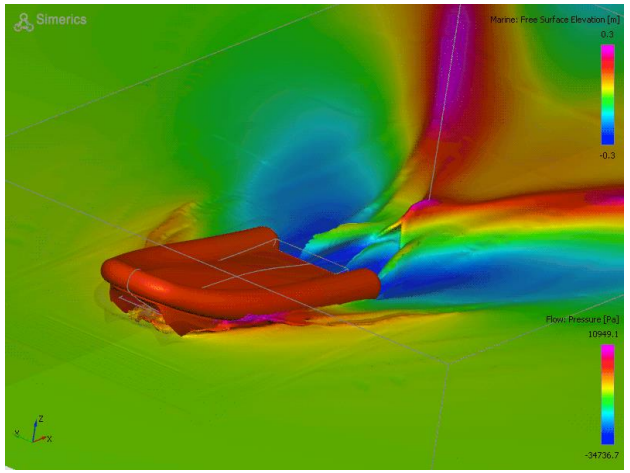
Neural Network → Dynamics



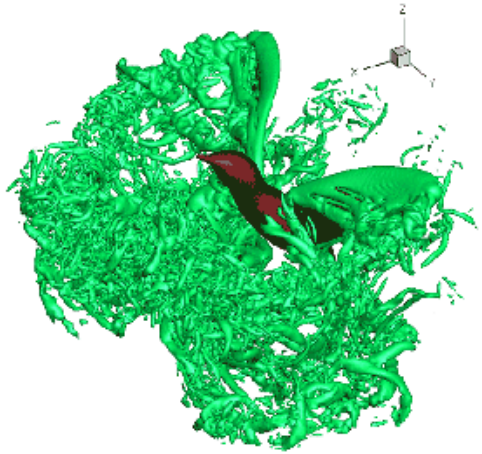
https://3d.bk.tudelft.nl/gsclara/projects/les_oklahoma/



https://blogs.nvidia.com/blog/cadence-millennium-nvidia-blackwell/

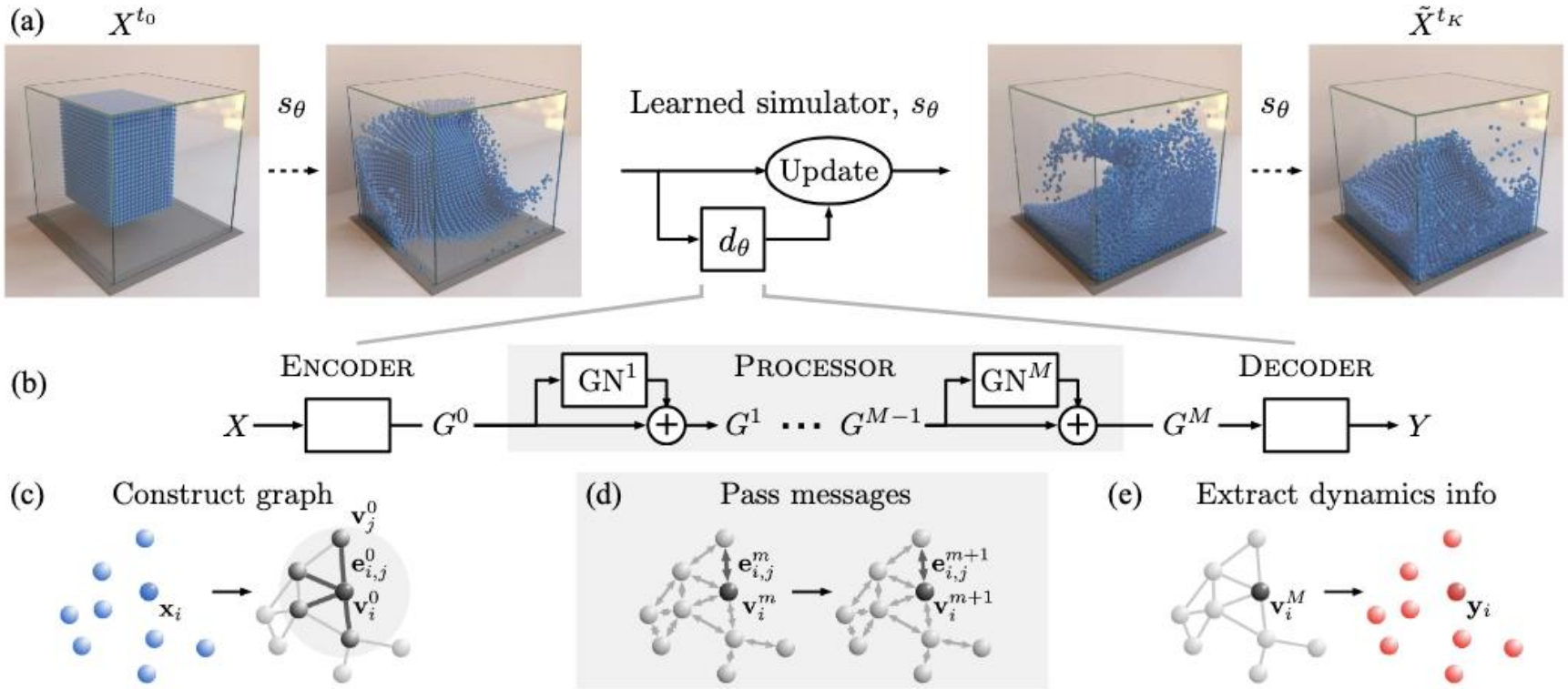


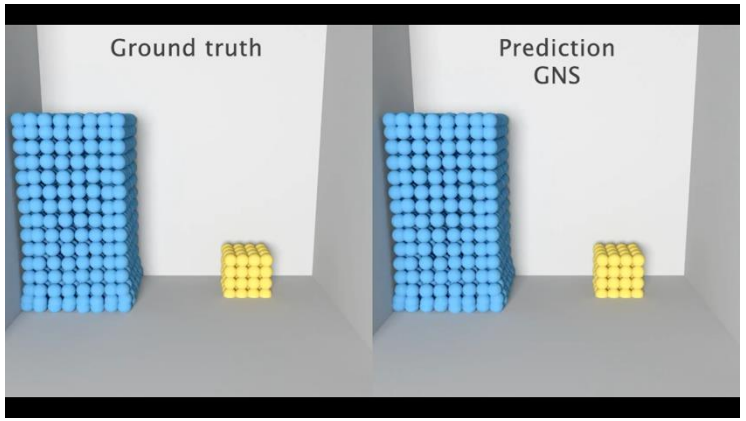
https://8020engineering.com/80-20-engineering-events/



Data-driven Framework of Dynamic Modeling

Recipe: (1) collect data, and (2) train neural networks to emulate motion labels

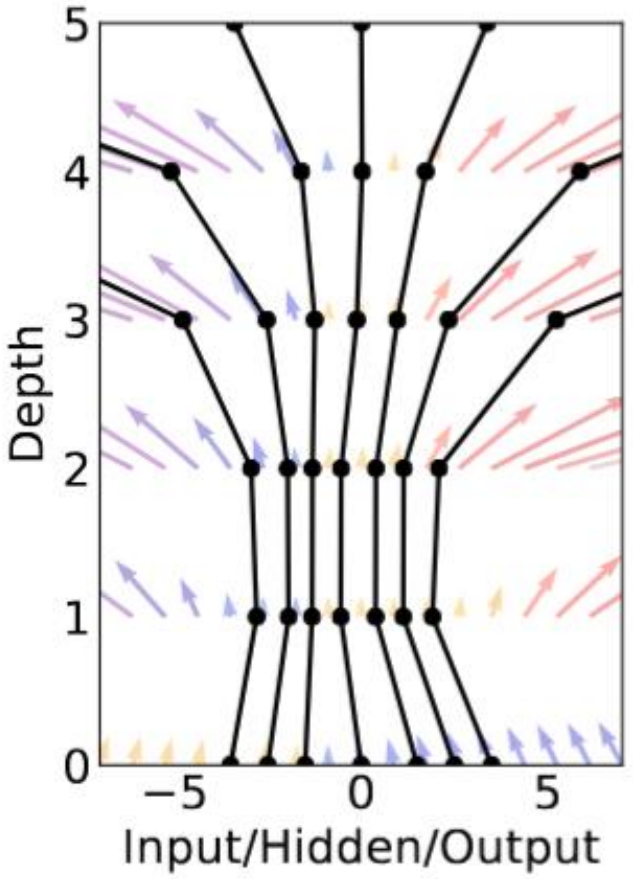




Learning to Simulate Complex Physics with Graph Networks. DeepMind

Issues of Naïve Data-driven Approach: Discretized Output

Discretized output



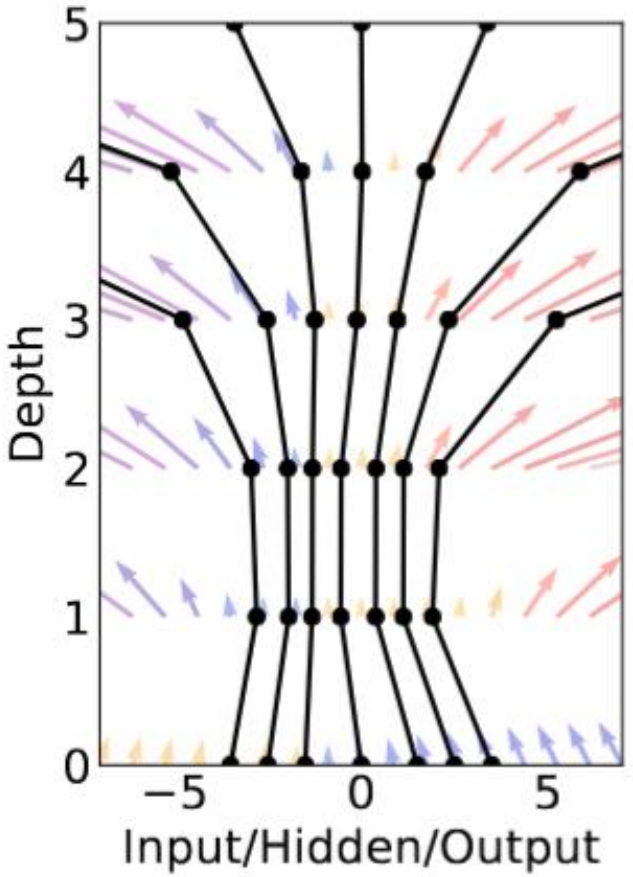
Problems: Motions are predicted sparsely at specific time

1. Intermediate motions are unknown
2. Physical accuracy degrades as the output sparsity increases

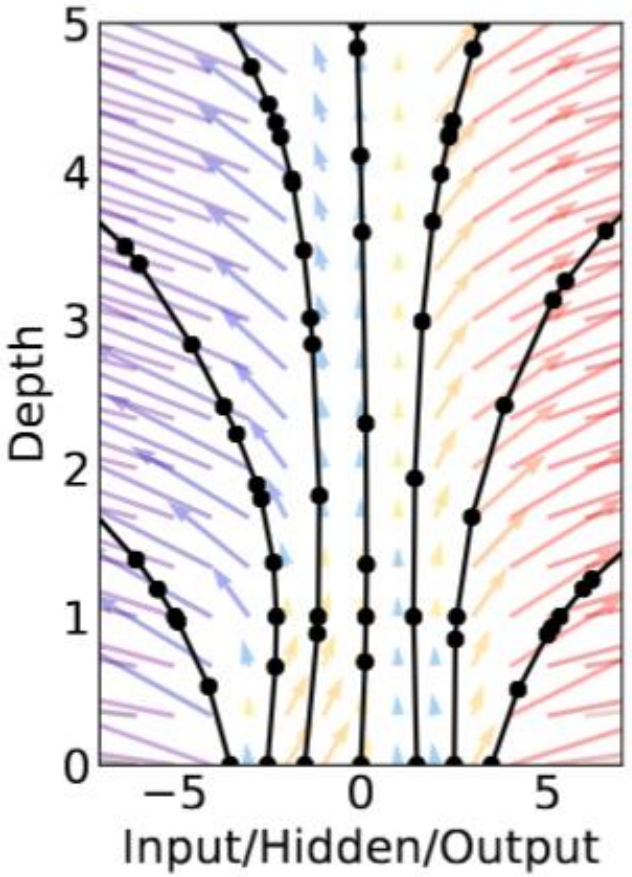
Neural Ordinary Differential Equations. Chen et al.

Continuous Output is Desired

Discretized output

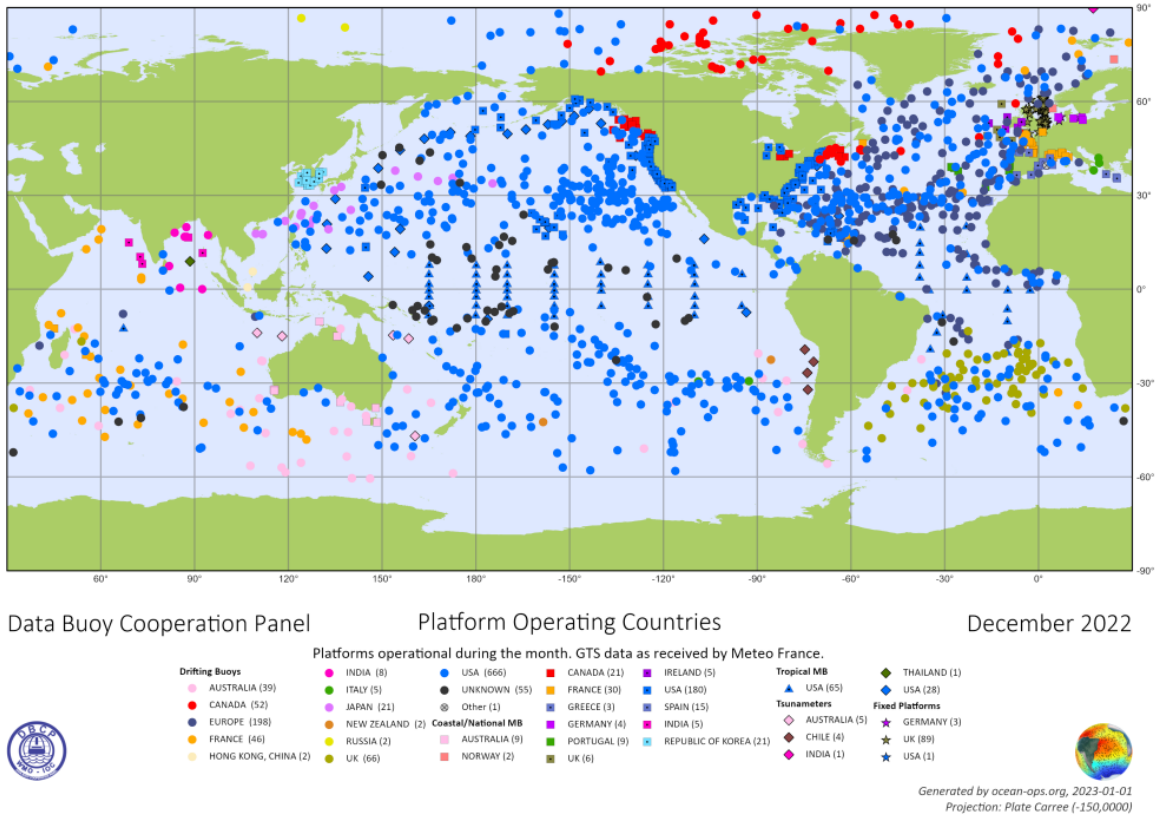


Continuous output

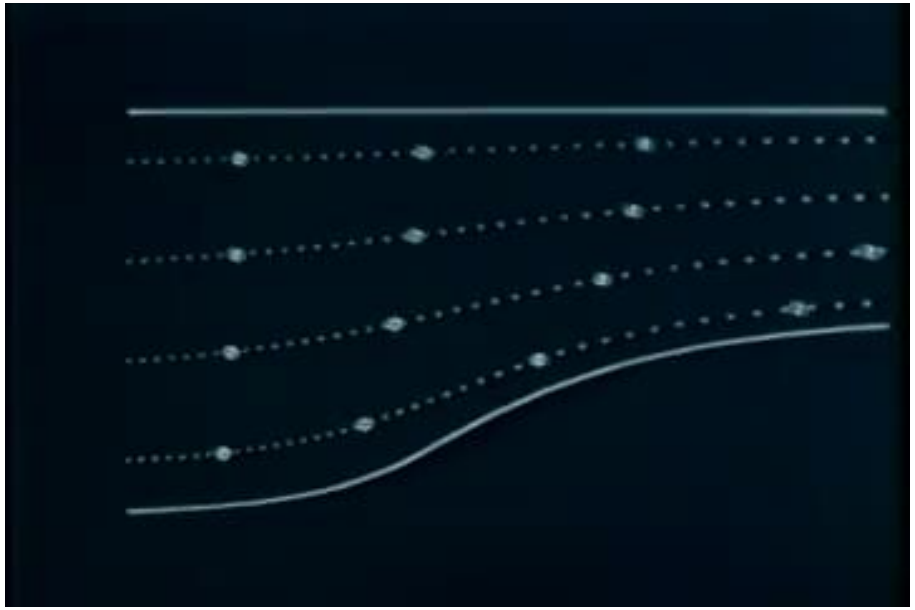


Neural Ordinary Differential Equations. Chen et al.

Issues of Naïve Data-driven Approach: Sparse Real-world Data



https://www.comc.ncku.edu.tw/最新消息/觀測系統/海氣象資料浮標



Idea: Supervise with Physics Equations when Data is Sparsely Available

$$\theta''(t) + 2\beta\theta'(t) + \omega_0^2\theta(t) = 0 \quad \text{Pendulum equation}$$
$$\theta(t_0) = \theta_0 \quad \text{Initial conditions}$$
$$\theta'(t_0) = v_0 \quad \text{Initial conditions}$$

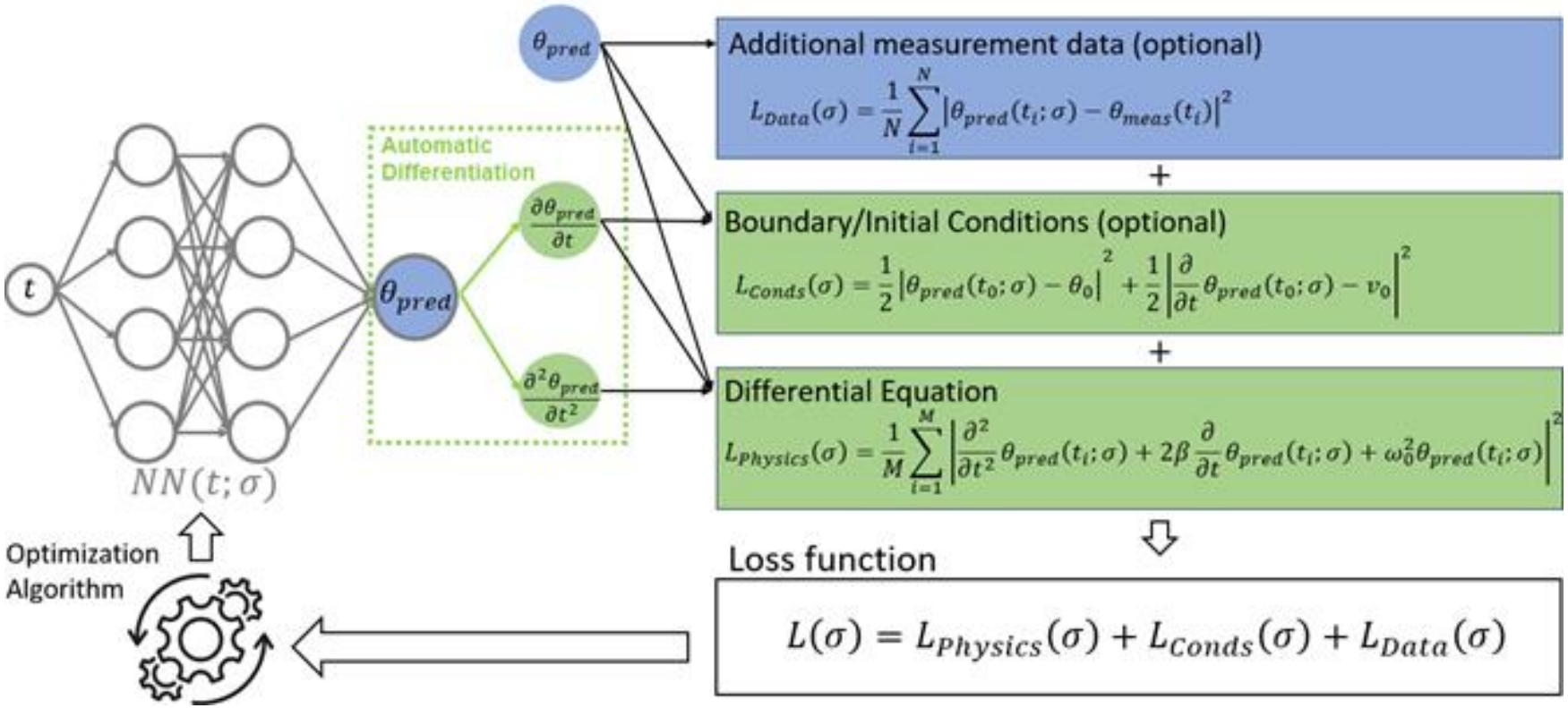
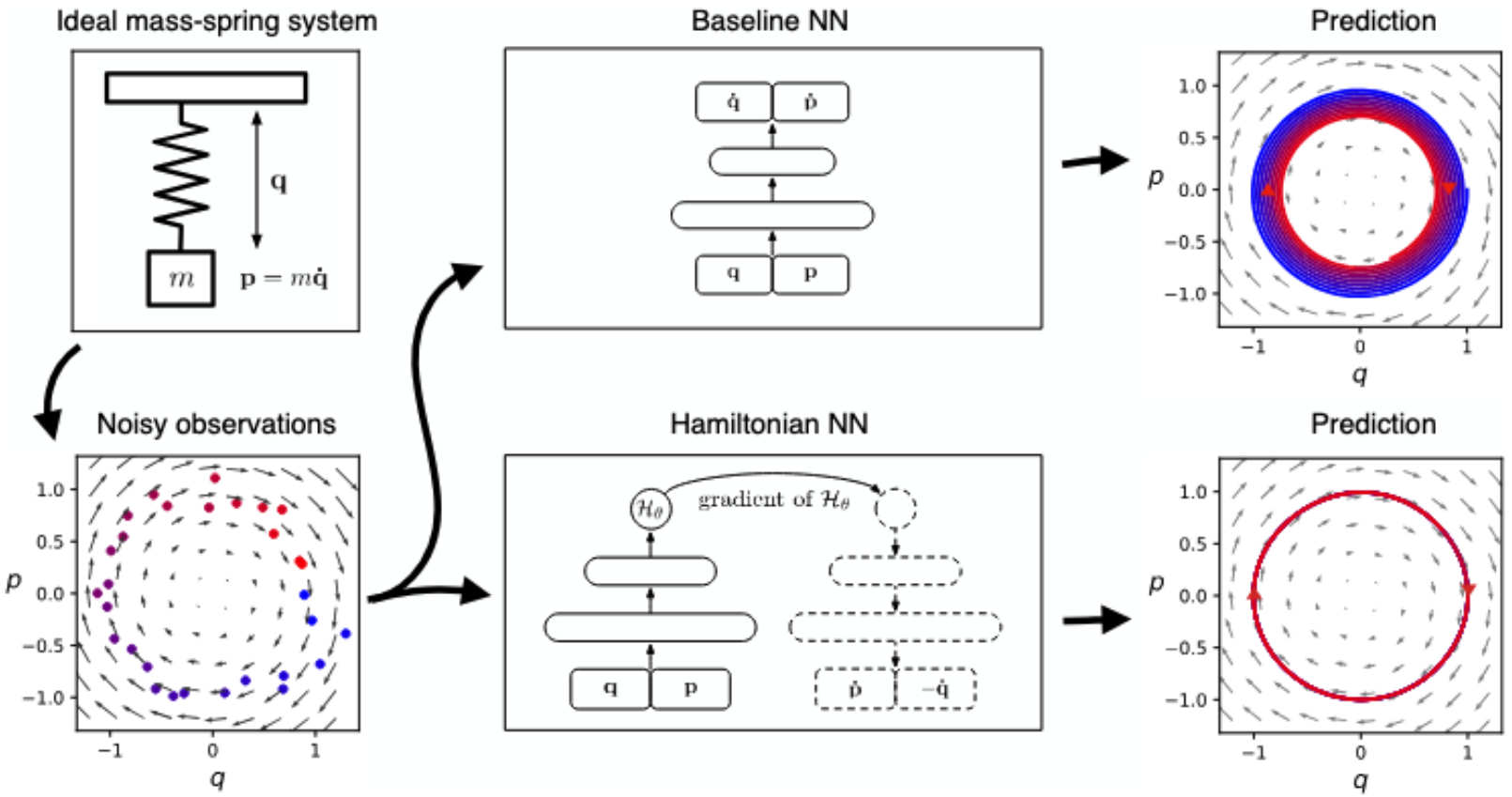


Image source: https://www.mathworks.com/discovery/physics-informed-neural-networks.html

Idea: Supervise with Physics Equations when Data is Sparsely Available



Constraint:

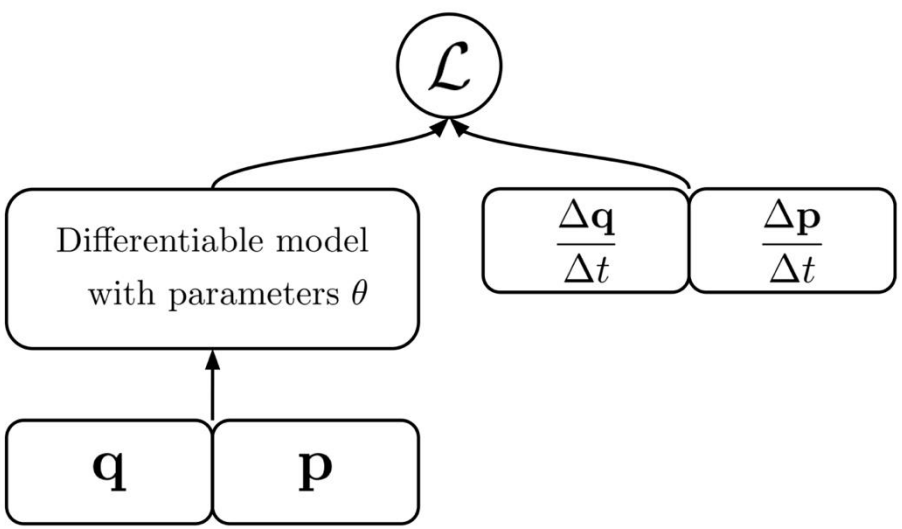
$$\frac{d\mathbf{q}}{dt} = \frac{\partial \mathcal{H}}{\partial \mathbf{p}}, \quad \frac{d\mathbf{p}}{dt} = -\frac{\partial \mathcal{H}}{\partial \mathbf{q}}$$

Learning loss:

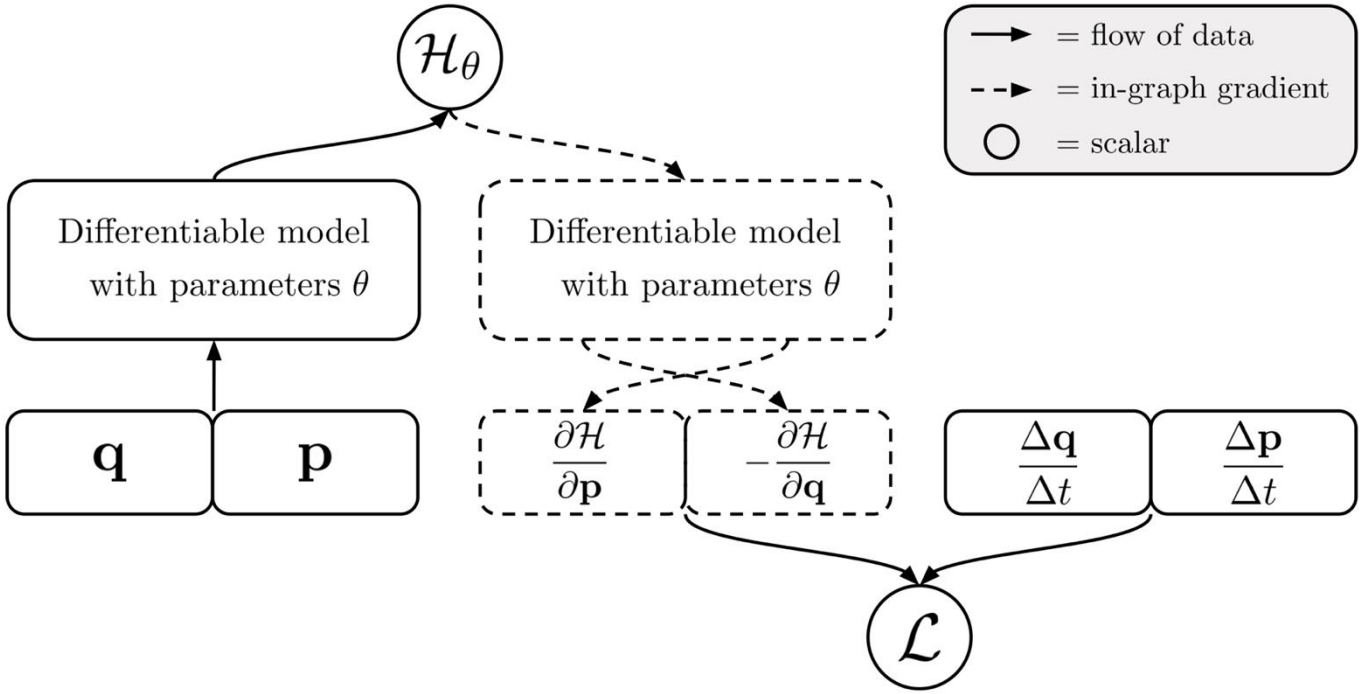
$$\underset{\theta}{\operatorname{argmin}} \left\| \frac{d\mathbf{q}}{dt} - \frac{\partial \mathcal{H}_\theta}{\partial \mathbf{p}} \right\|^2 + \left\| \frac{d\mathbf{p}}{dt} + \frac{\partial \mathcal{H}_\theta}{\partial \mathbf{q}} \right\|^2$$

Hamiltonian Neural Networks. Greydanus et al.

Idea: Supervise with Physics Equations when Data is Sparsely Available



Baseline NN



Hamiltonian NN

Idea: Supervise with Physics Equations when Data is Sparsely Available

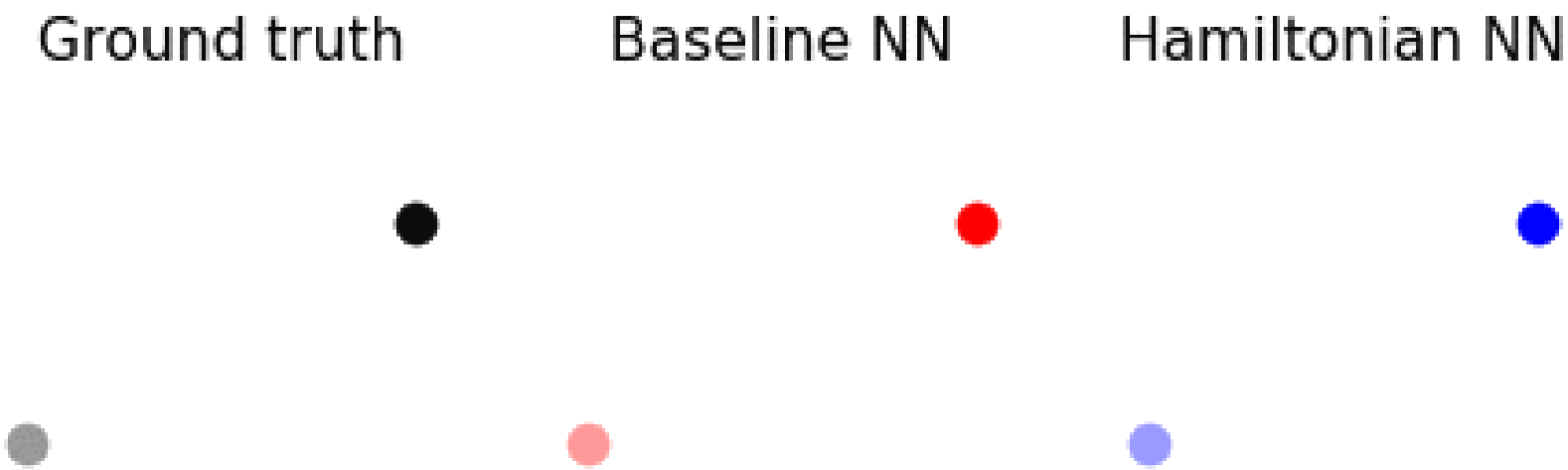


Image source: https://www.mathworks.com/discovery/physics-informed-neural-networks.html

Perception Is More Than Image / Video

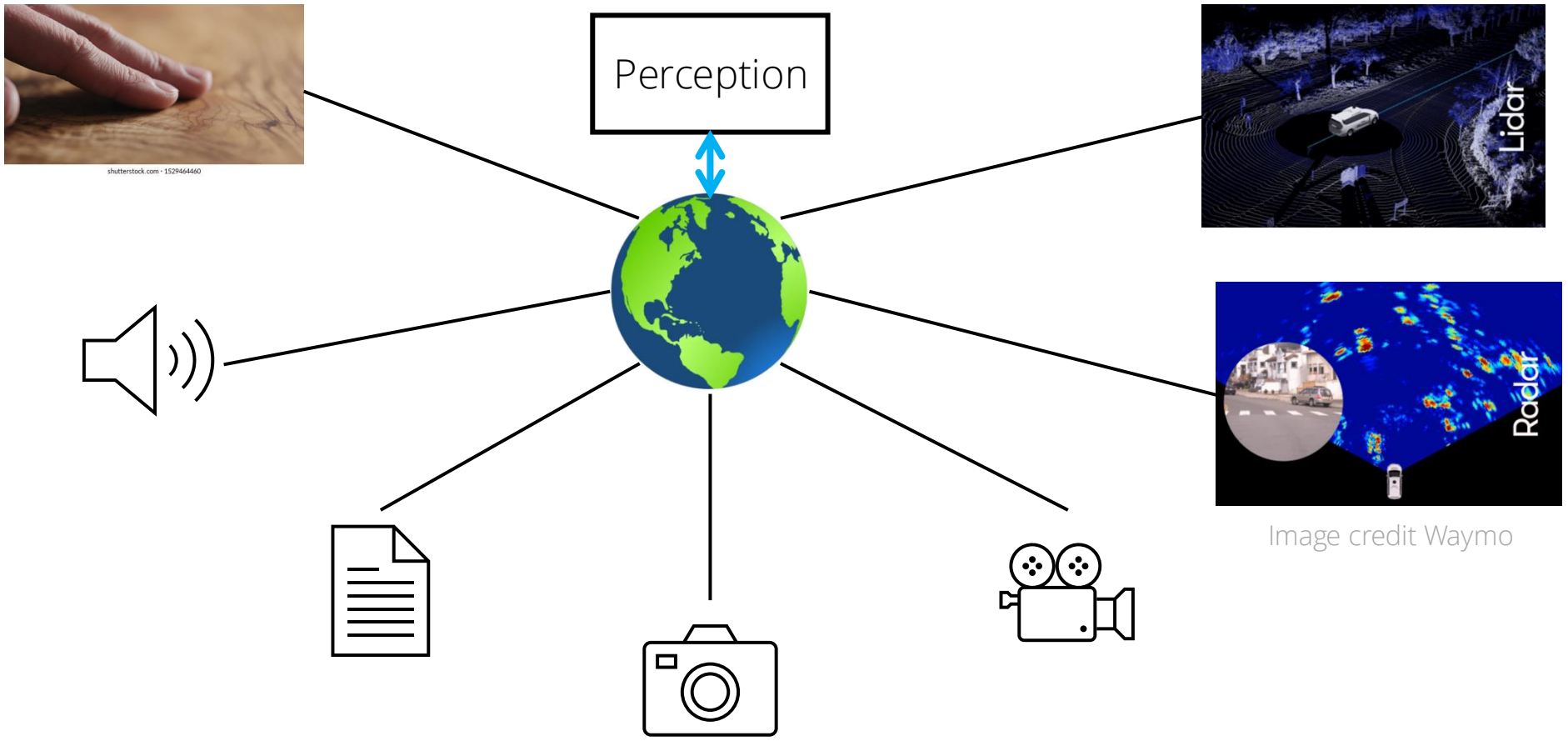
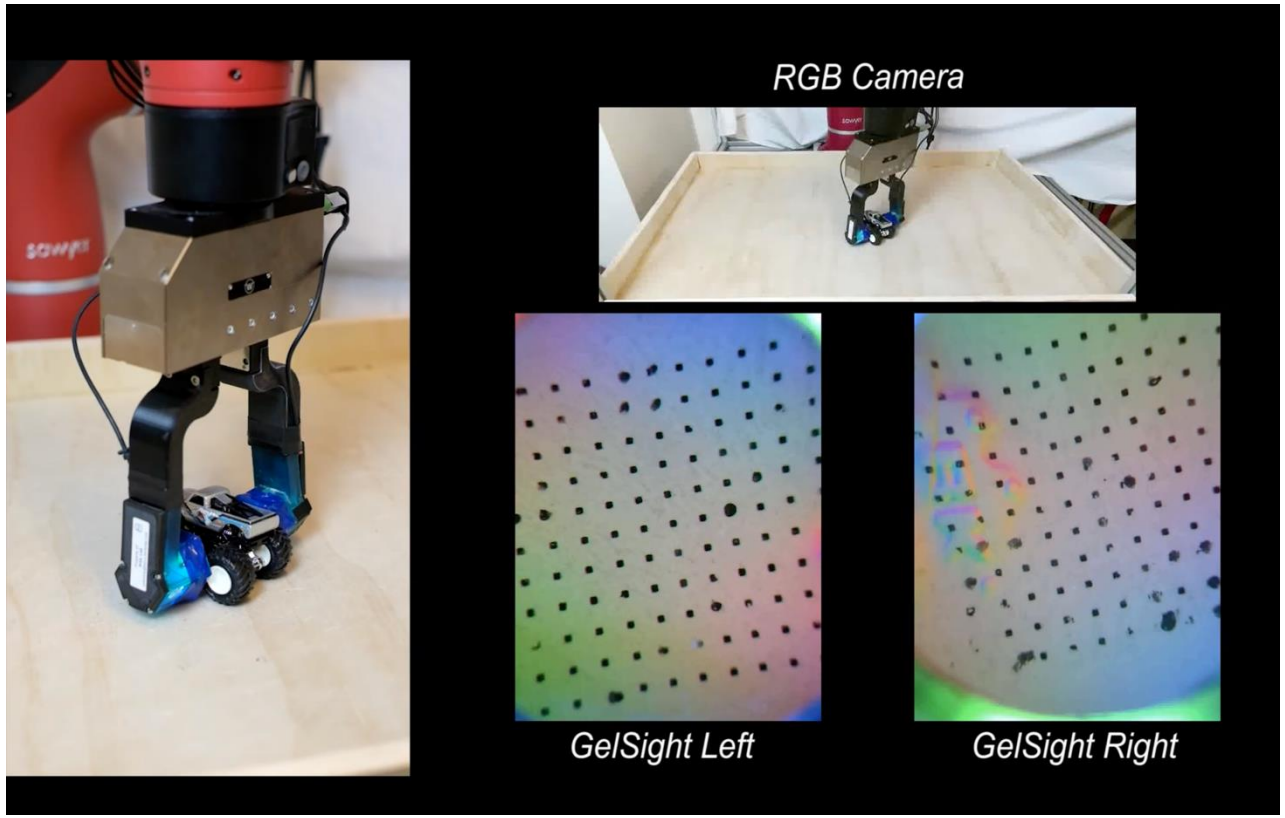
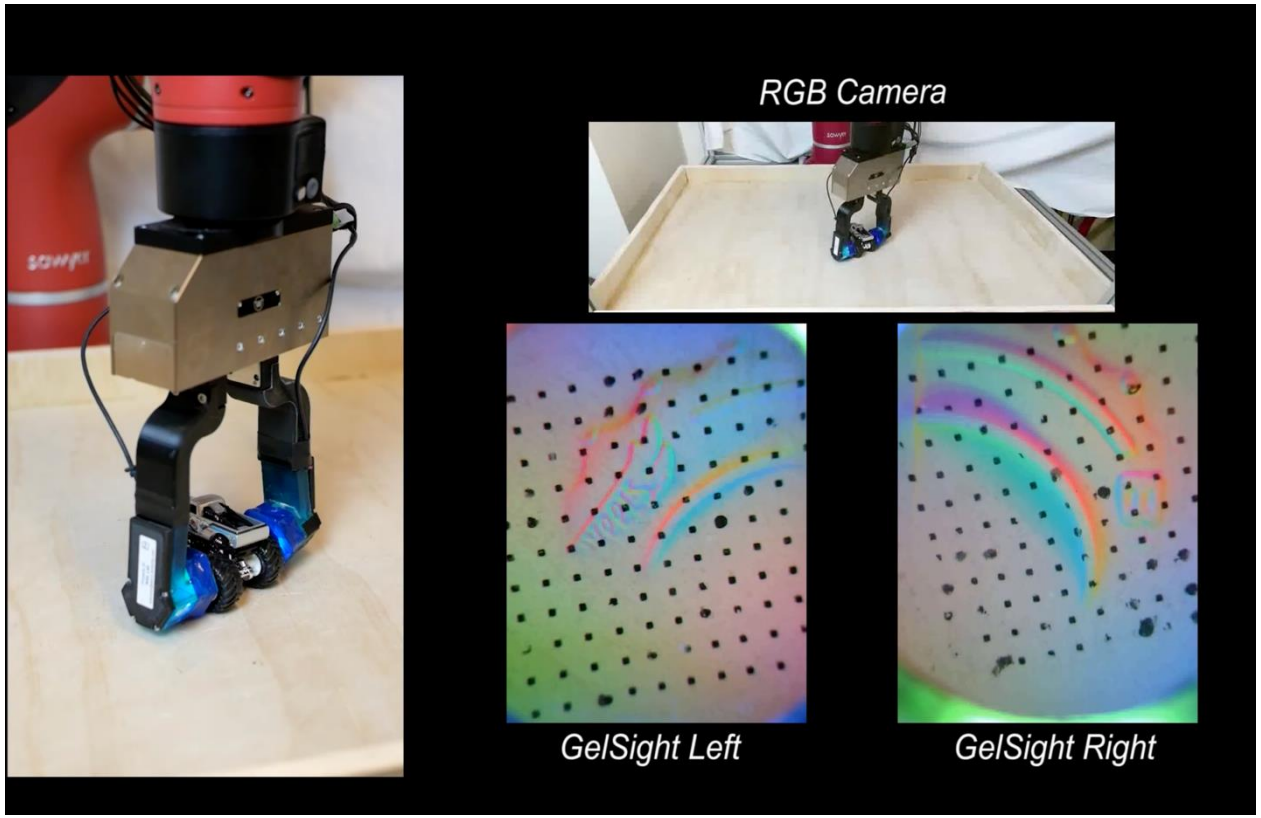


Image credit Waymo

Physical Perception: Tactile Sensing

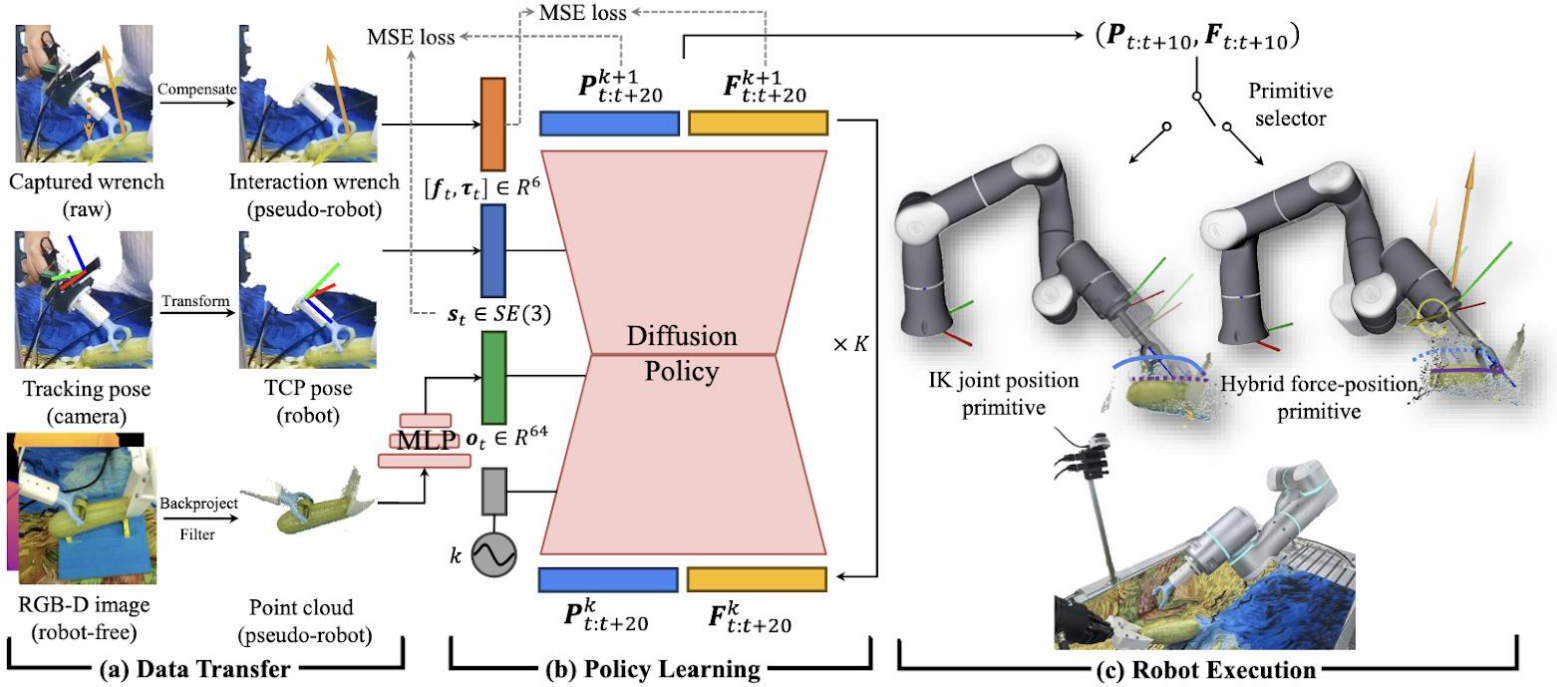


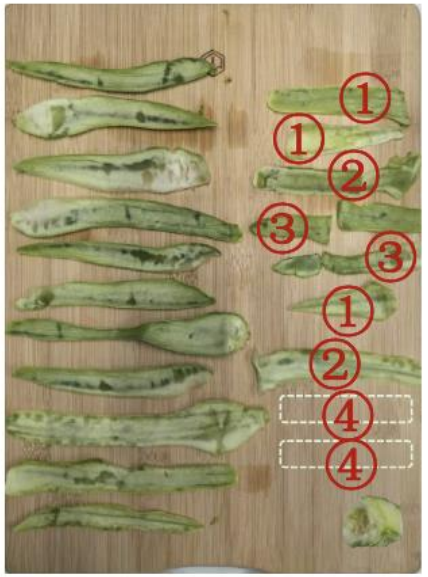


Model	Accuracy (mean ± std. err.)
Chance	62.80% ± 0.85%
Vision (+ action)	73.03% ± 0.24%
Tactile (+ action)	79.34% ± 0.66%
Tactile + Vision (+ action)	**80.28%** ± **0.68%**
Tactile + Vision (no action)	76.43% ± 0.42%

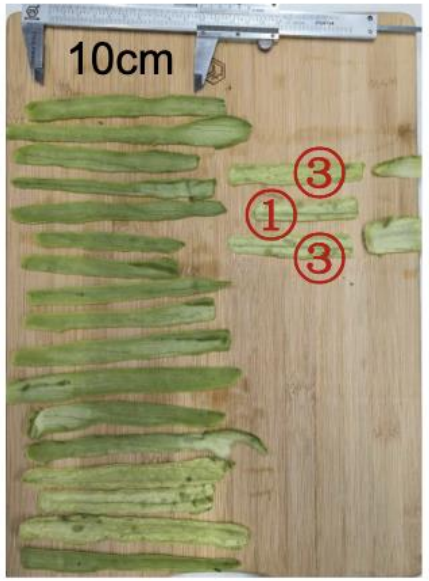
Learning to Grasp and Regrasp using Vision and Touch. Calandra et al. IRAL 2018.

Physical Feedback Enables Dexterous and Safe Manipulation





w\o force feedback

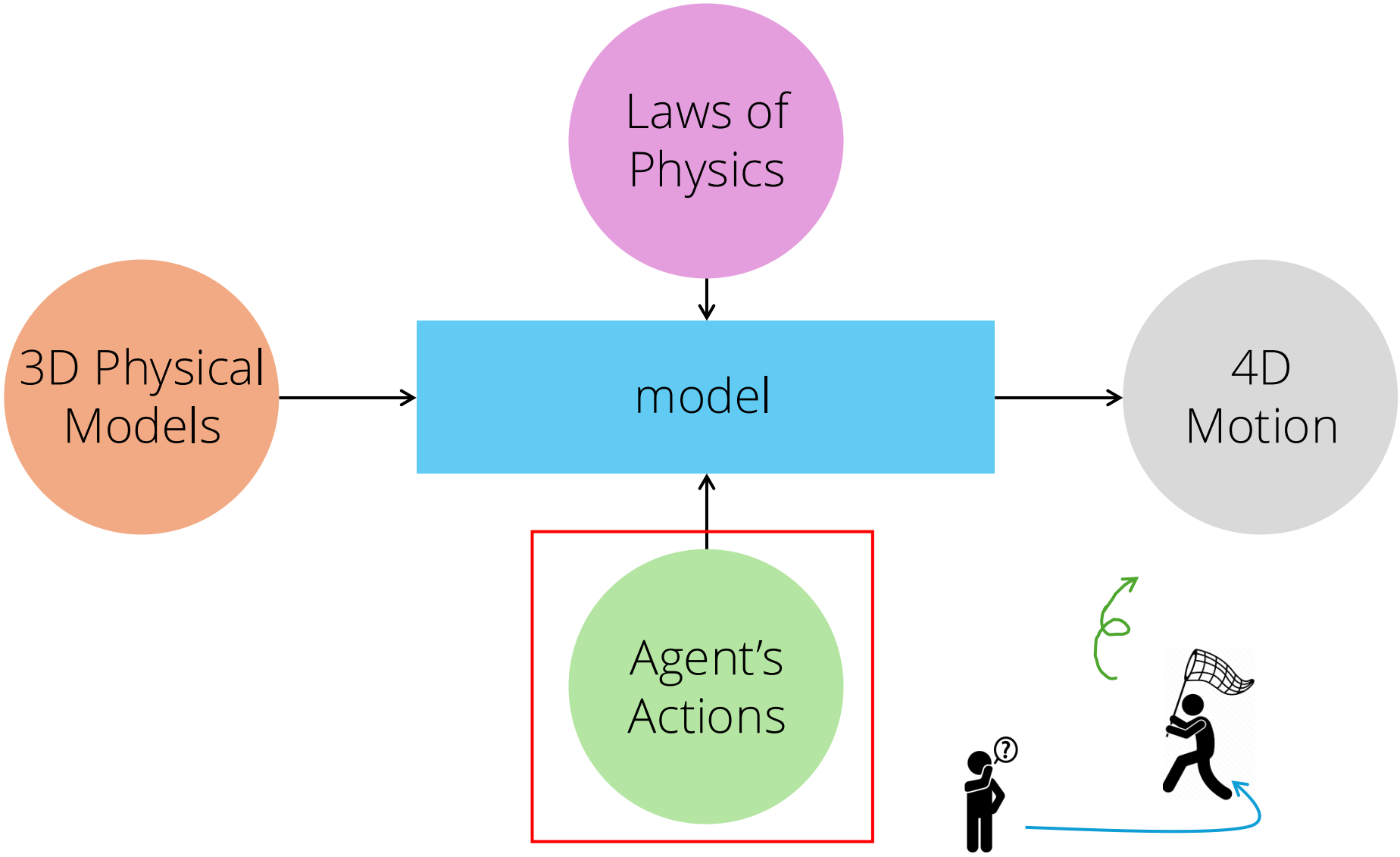


with force feedback

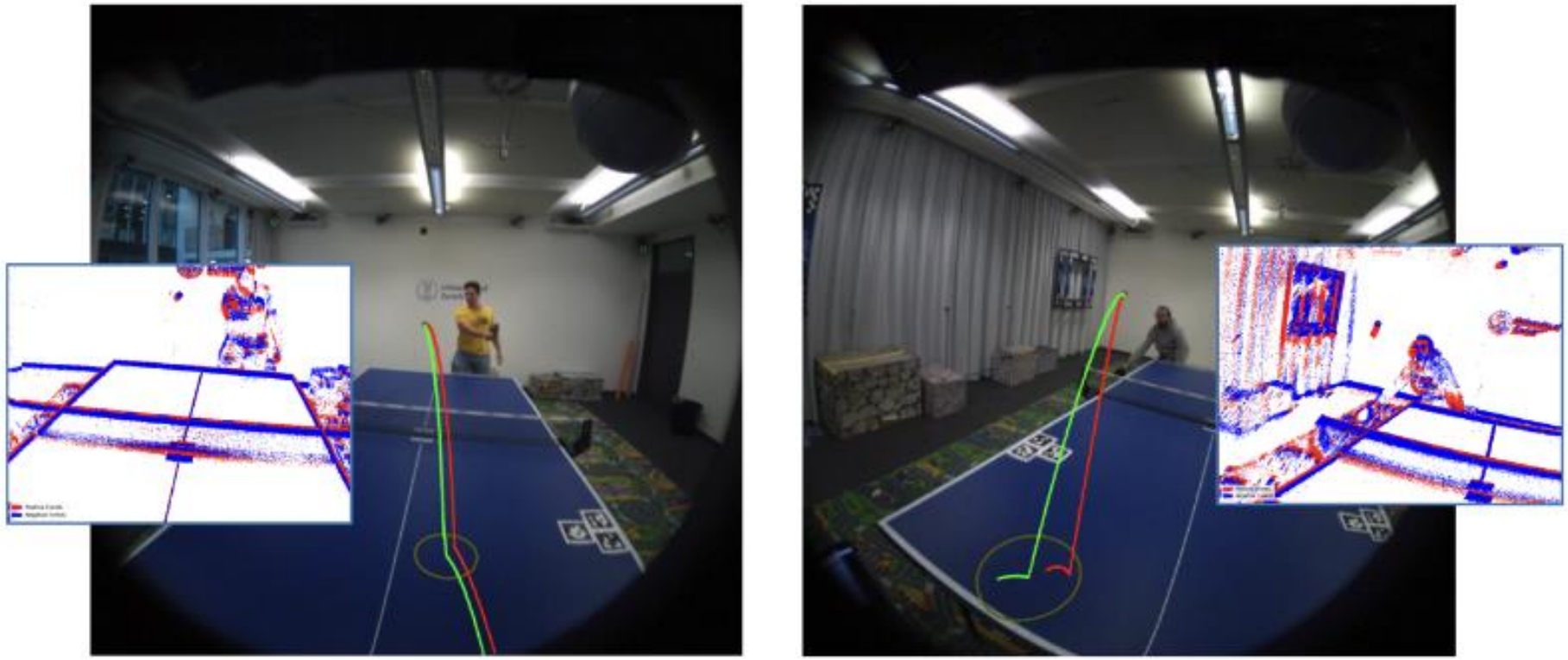
Reconstructing the “True” World with Physical Sensing

Tactile-Augmented Radiance Fields. Dou et al.

We are Embodied Agents: We Act, We Perceive and We Predict

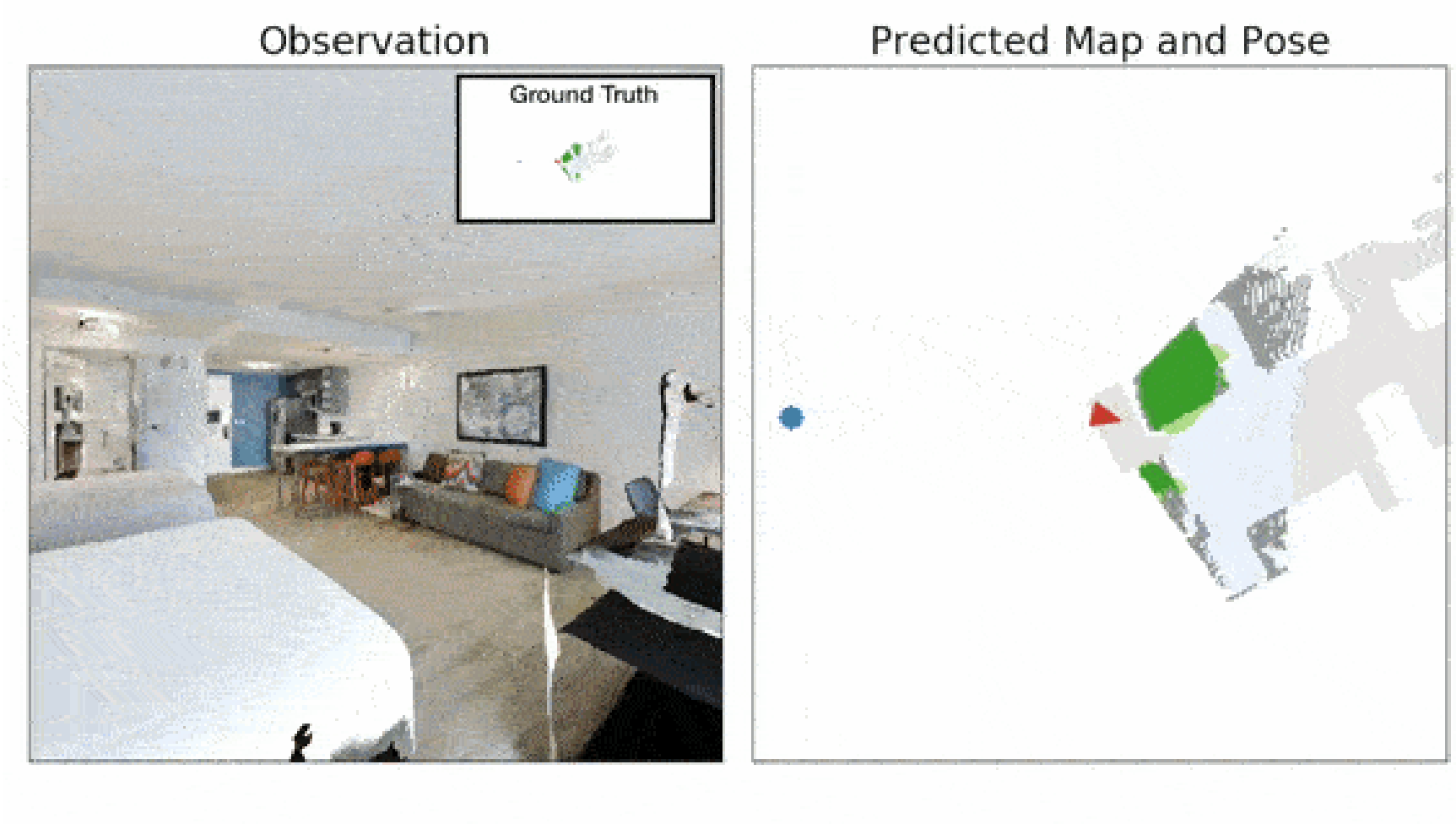


Action-centric Perception



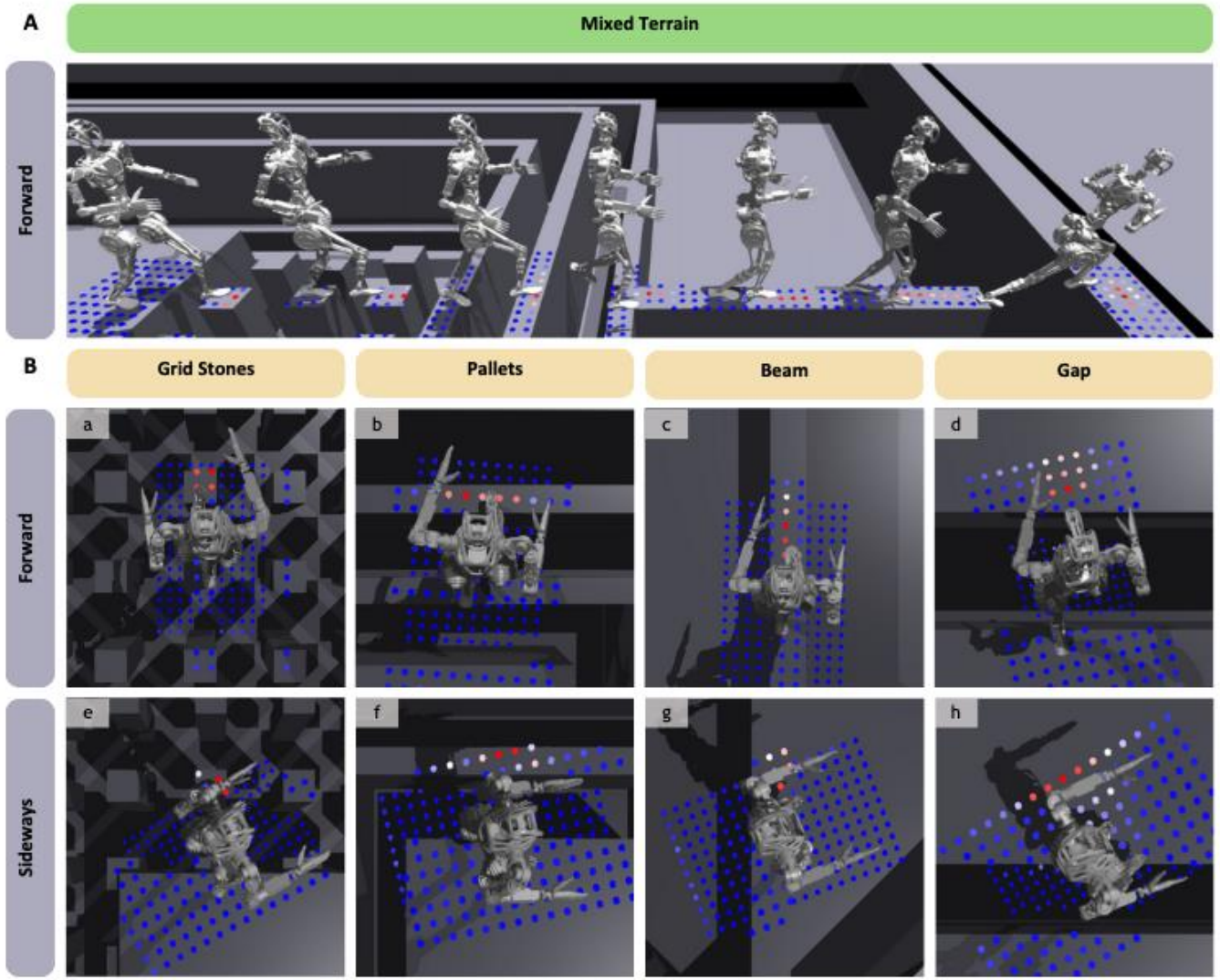
Egocentric Event-Based Vision for Ping Pong Ball Trajectory Prediction. Alberico et al.

Simultaneous Localization and Mapping



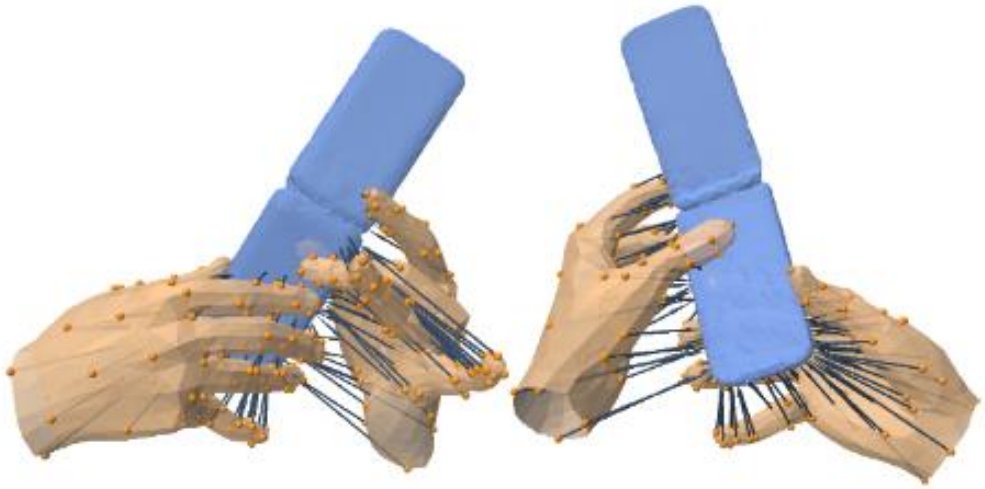
Learning to Explore using Active Neural SLAM. Chaplot et al.

Perceptive Locomotion

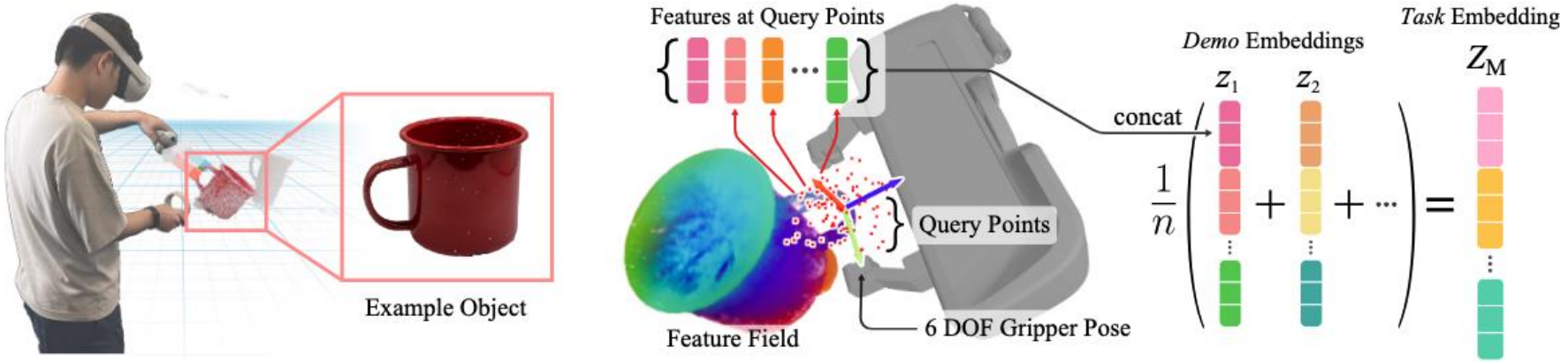


Attention-Based Map Encoding for Learning Generalized Legged Locomotion. He et al.

Perceptive Manipulation



BimArt: A Unified Approach for the Synthesis of 3D Bimanual Interaction with Articulated Objects. Zhang et al.



Distilled Feature Fields Enable Few-Shot Language-Guided Manipulation. Shen et al.

Action Prediction

- Imitation learning by fitting many action-labeled demonstrations
- Control trajectory optimization if the dynamic rules is known
- Reinforcement learning by trial and error

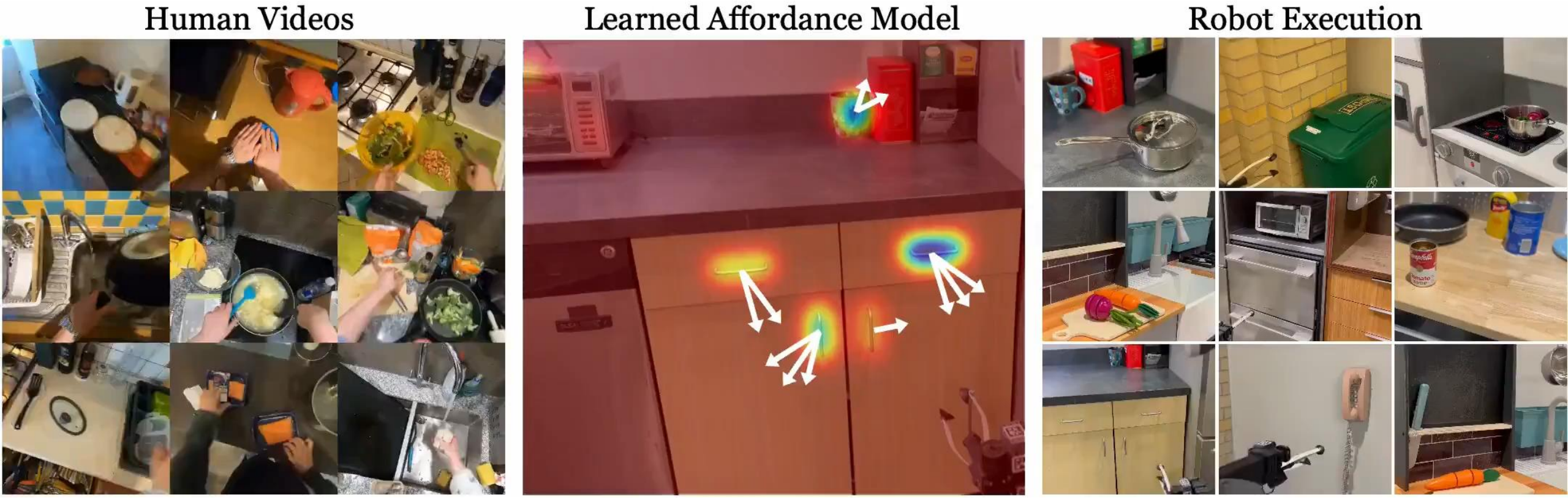
→ We won't cover these approaches, since policy learning is not the focus of this course. If you are interested, take the course of "Robot Perception and Learning" or "(Deep) Reinforcement Learning"

Learning Approaches for Action Prediction

- Distill prior knowledge from unlabeled human visual demonstrations

Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. Damen et al.

Learning Action Prediction from Human Visual Demonstrations



Affordances from Human Videos as a Versatile Representation for Robotics. Bahl et al.

Approaches for Action Prediction

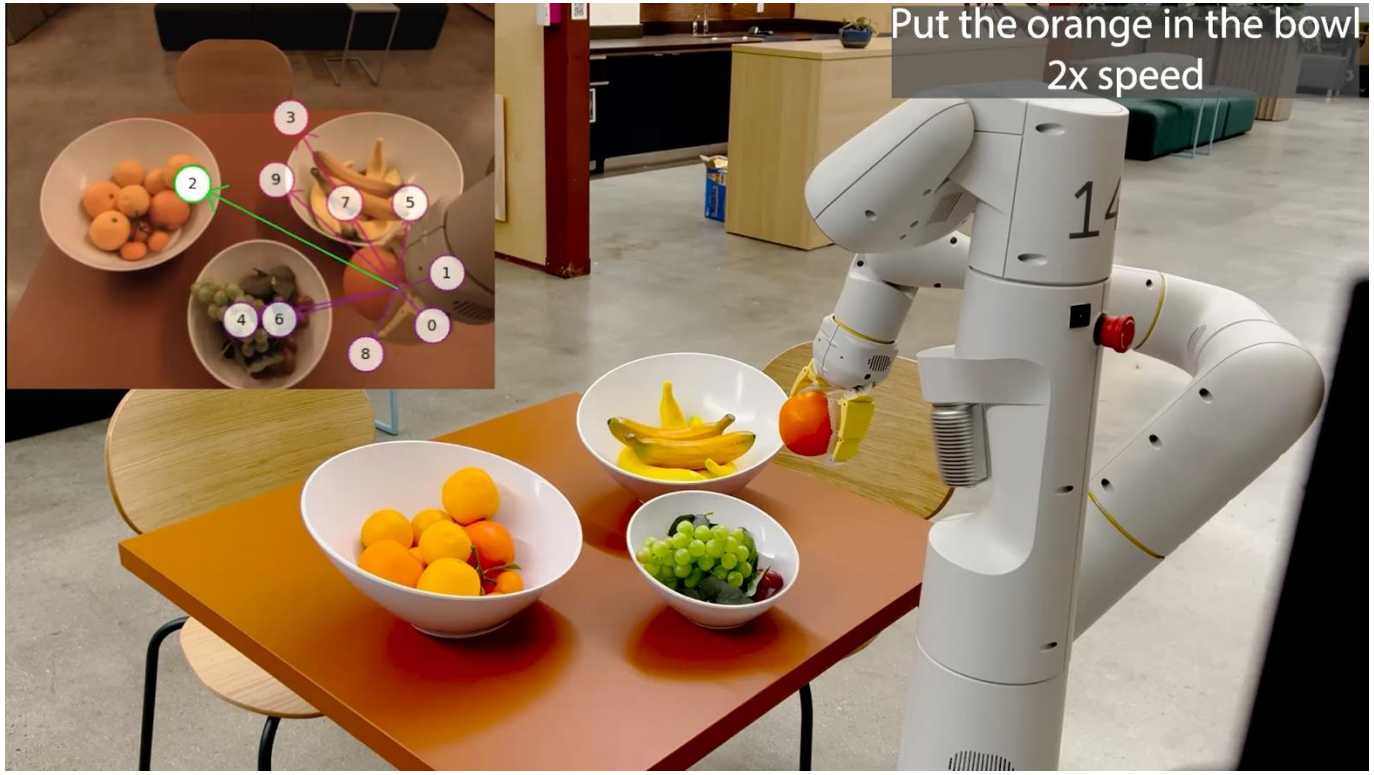
- Distill prior knowledge from visual-language models

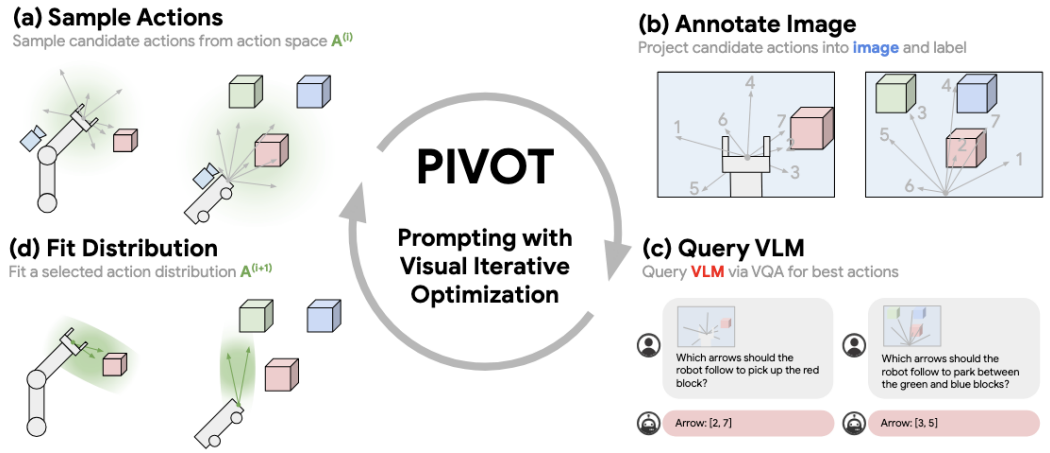
Task: "Open the top drawer"

AHA: A Vision-Language-Model for Detecting and Reasoning Over Failures in Robotic Manipulation. Duan et al.

Predicting Actions with Visual-Language Models

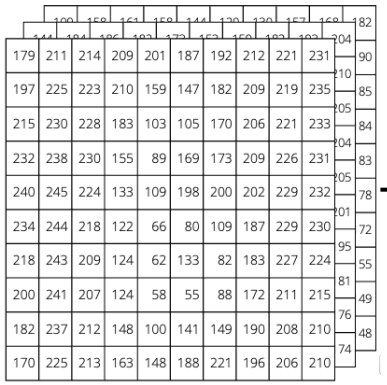
- Distill prior knowledge from visual-language models



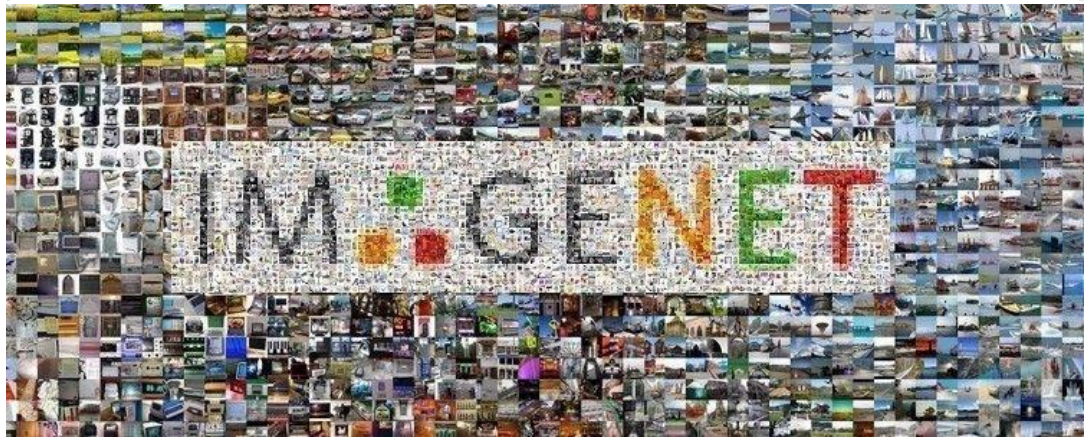


PIVOT: Iterative Visual Prompting Elicits Actionable Knowledge for VLMs. Google DeepMind.

Remember that Pre-training Visual Encoders Facilitates Downstream Visual Tasks

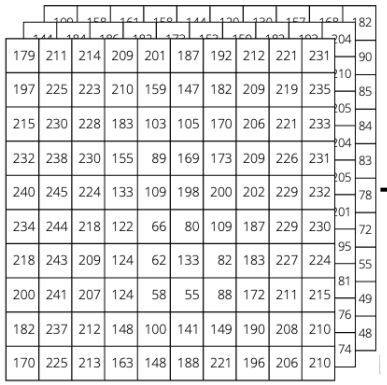


CNN + classifier → Labels



1,000,000 images with 1,000,000 labels

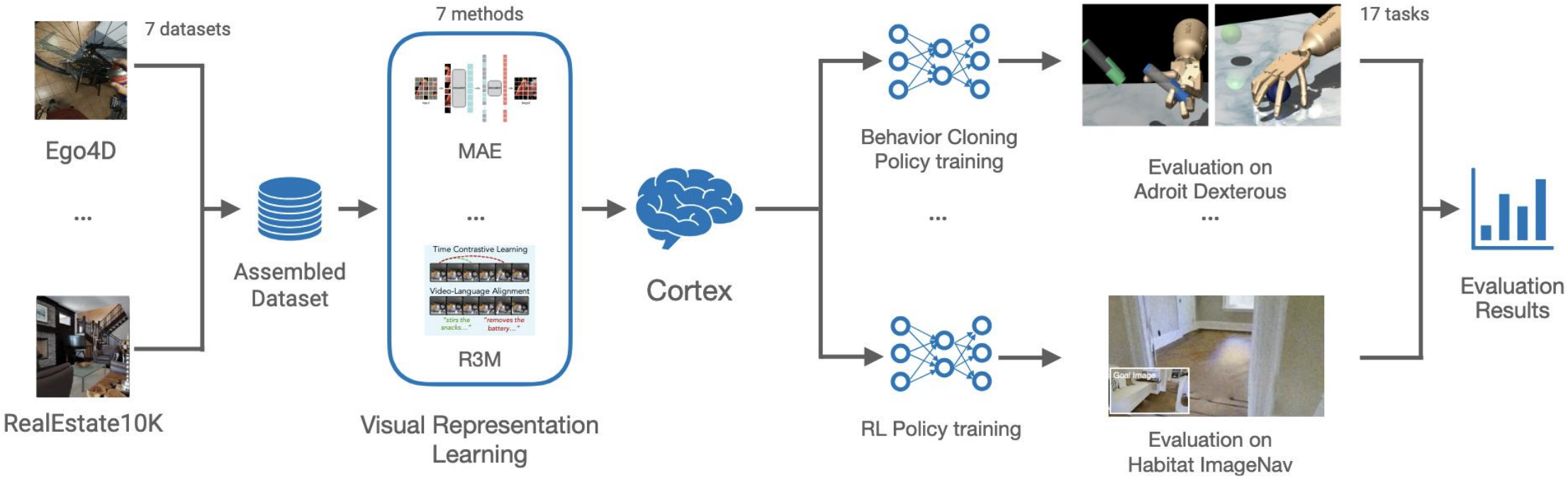
A Generic Formulation of Action Prediction Models



CNN + regressor → Actions

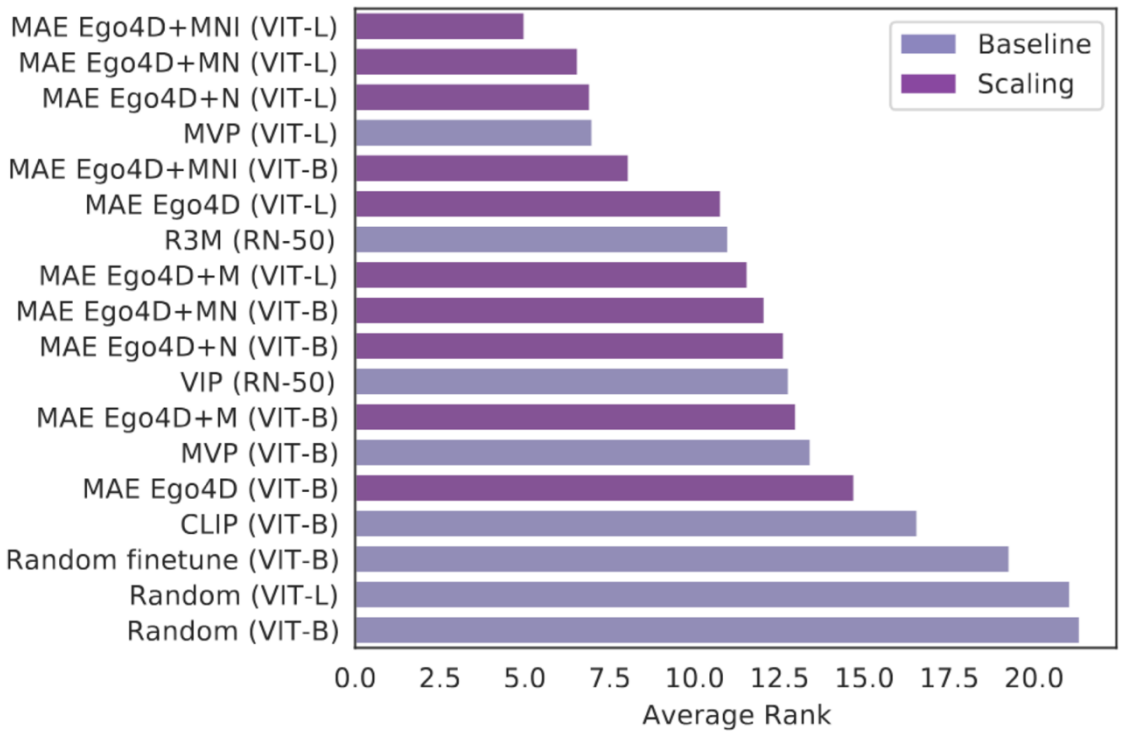
- Does pre-training the visual encoder facilitate action prediction tasks?

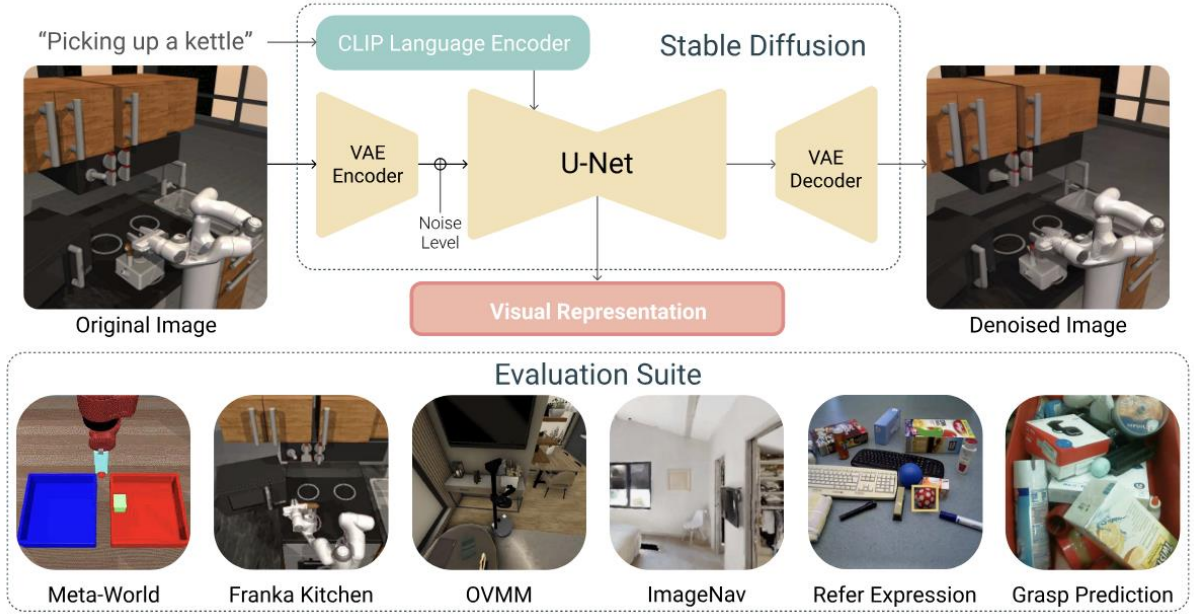
Representation Learning Facilitates Action Prediction

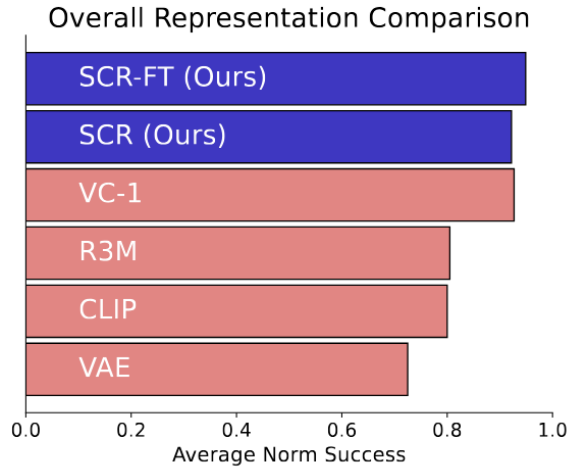


Where are we in the search for an Artificial Visual Cortex for Embodied Intelligence? Majumdar et al. NeuRIPS 2023.

Representation Learning Facilitates Action Prediction







Where are we in the search for an Artificial Visual Cortex for Embodied Intelligence? Majumdar et al. NeuRIPS 2023.

Pre-trained Text-to-Image Diffusion Models Are Versatile Representation Learners for Control. Gupta et al

The Three Components of Embodied Vision

4D Motion

- Goals:
  - 3D/4D reconstruction
  - 3D/4D generation
- Represented by:
  - <x, y, z, pitch, roll, yaw, t> of rigid bodies
  - <x, y, z, t> of particles
  - …

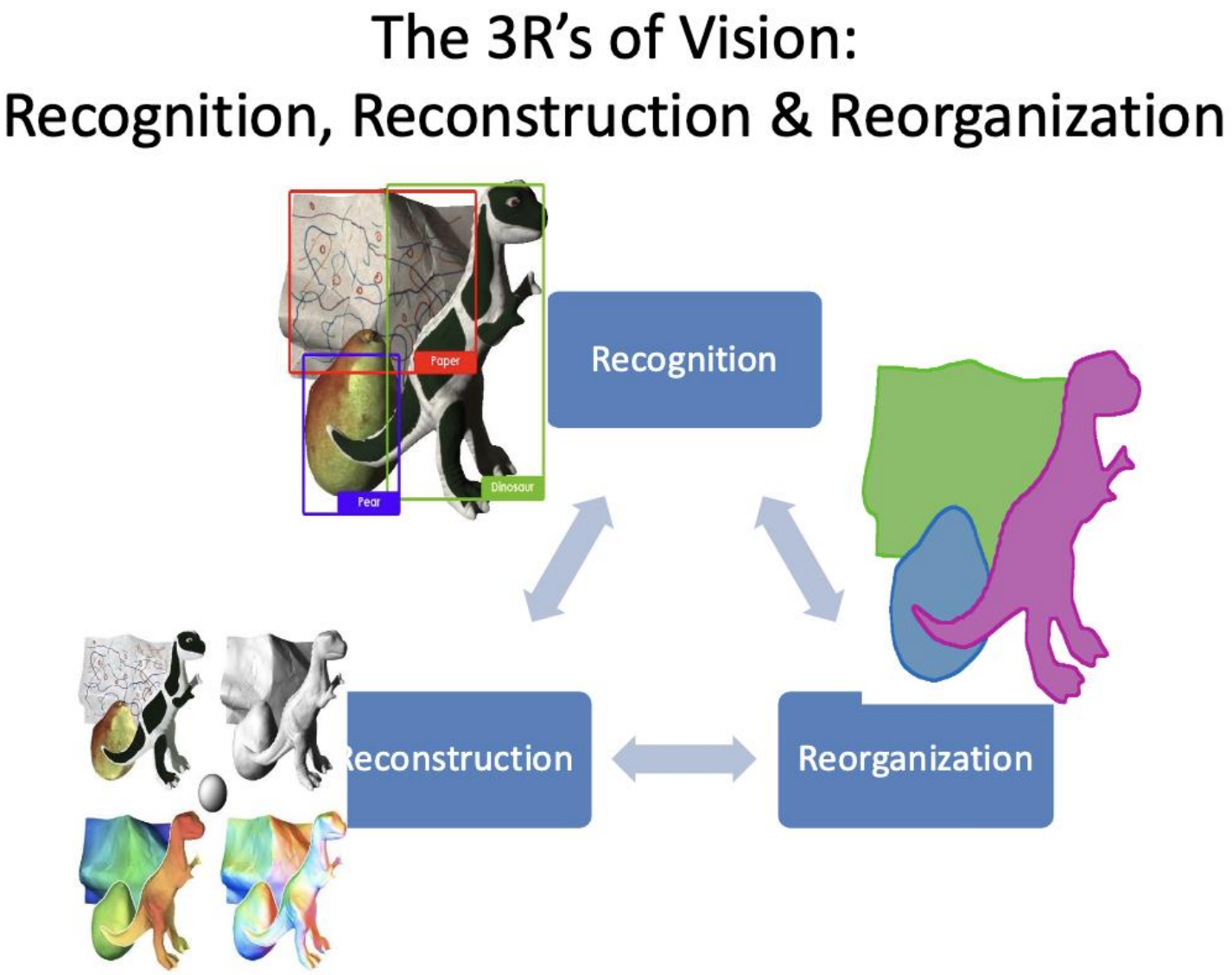
Dynamics

- Goals:
  - Physical simulation
- Represented by:
  - Laws of physics
  - Learned dynamics: $f(agent_{action}, obj_{motion}, env_{param})$
  - …

Action

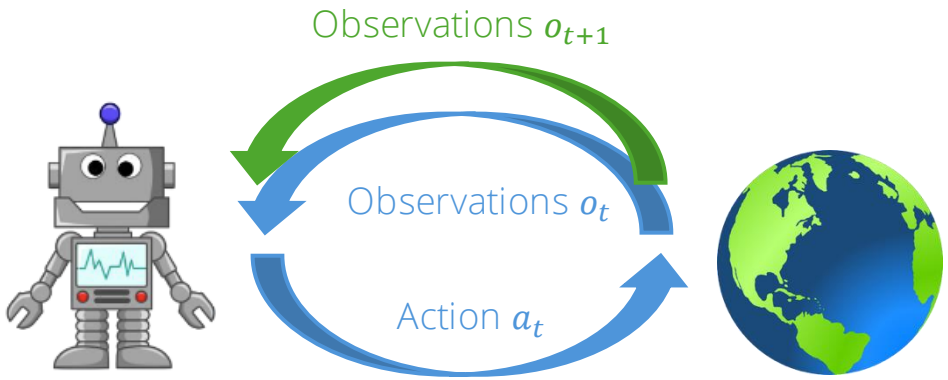
- Goals:
  - Learning from human visual demonstrations
  - Knowledge distillation from visual language models
  - Representation learning for action prediction
  - Multi-sensory input
- Represented by:
  - Agent's end-effector poses
  - Agent's joint angles
  - …

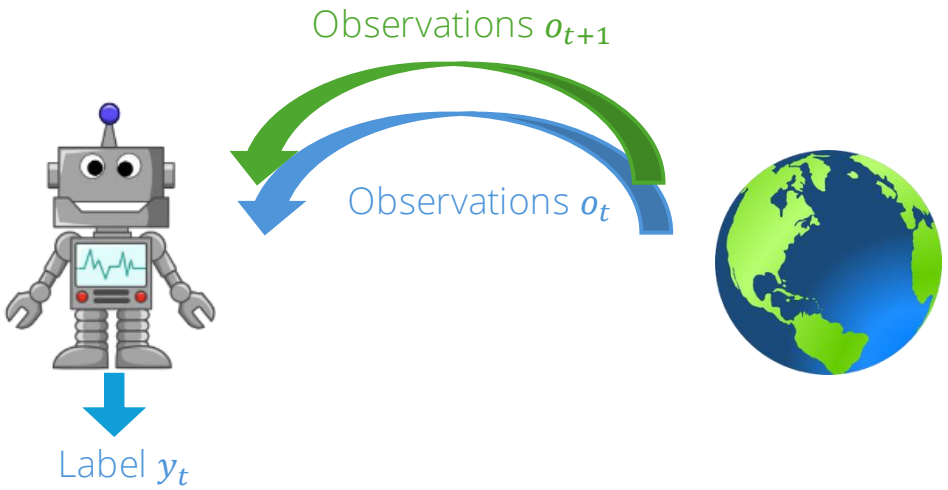
The Three Components of Computer Vision

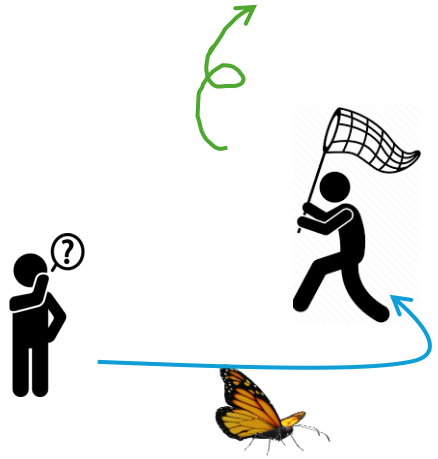


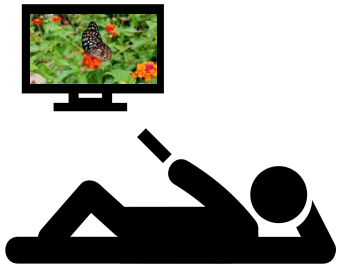
Slide credit J. Malik

Embodied Vision vs. Computer Vision

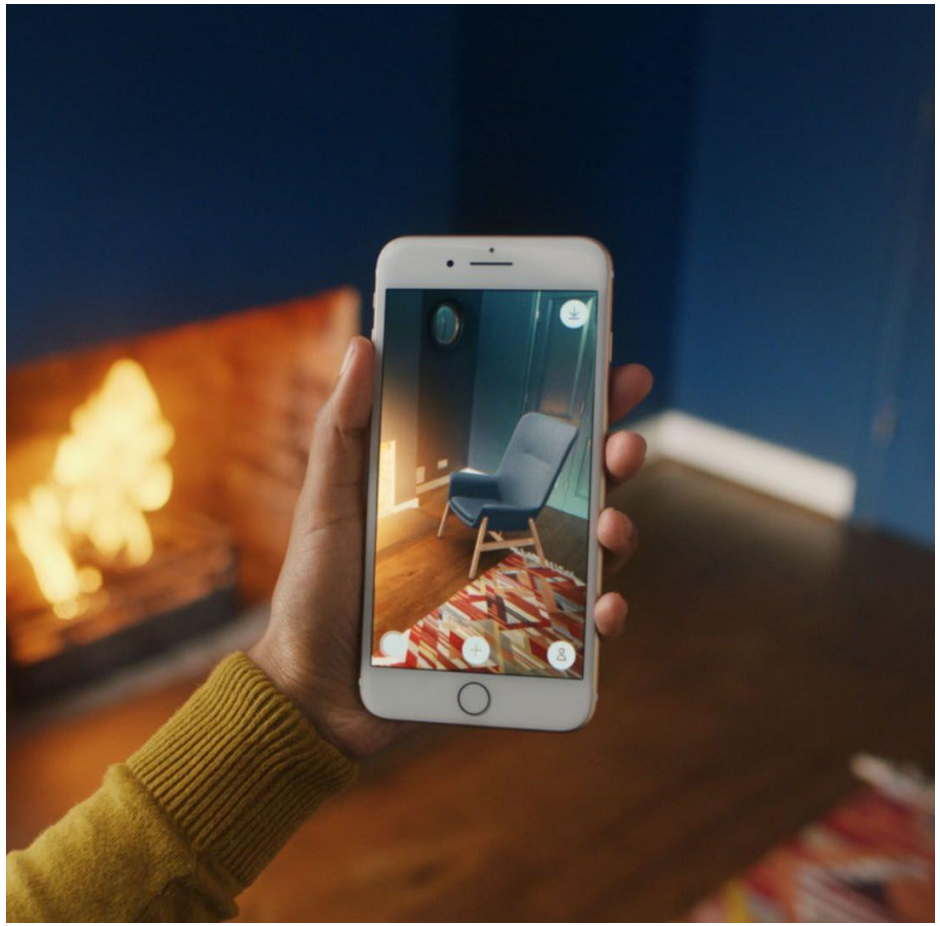








Embodied Vision and AR/VR



https://www.ikea.com/global/en/newsroom/innovation/ikea-launches-ikea-place-a-new-app-that-allows-people-to-virtually-place-furniture-in-their-home-170912/

Image credit Davide Scaramuzza

Embodied Vision and AR/VR

https://youtu.be/38W3qaVSy4Q?si=O9yvN546JpvnuAqO

Embodied Vision and Robotics

https://waymo.com/blog/2021/07/simulation-city/

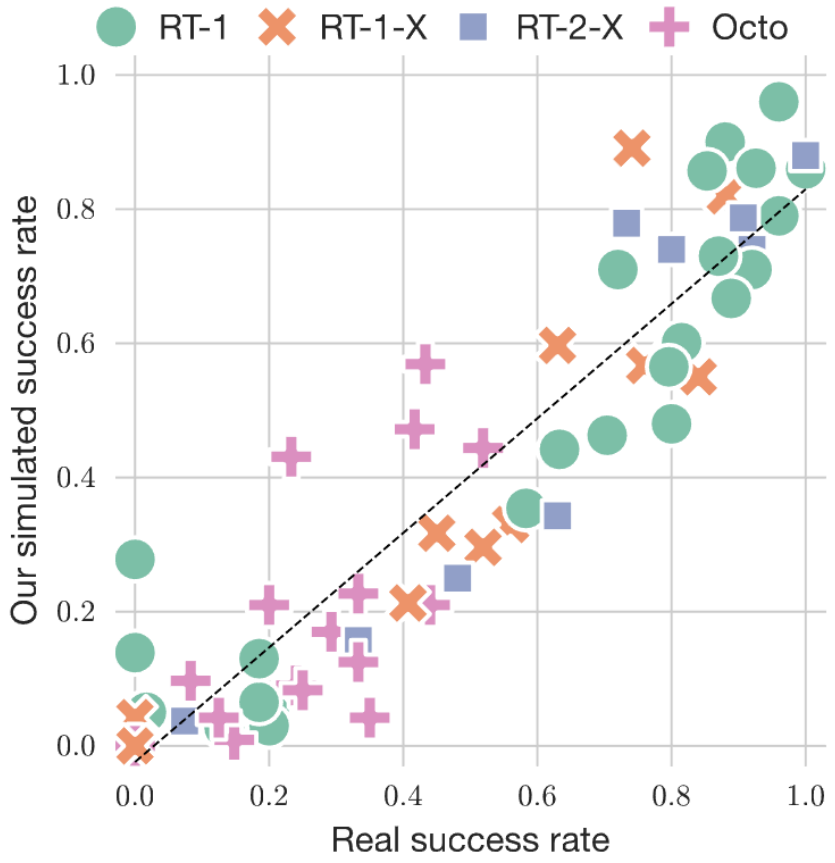
Embodied Vision and Robotics

Video source Boston Dynamics

Embodied Vision and Robotics

Real world

Simulation



Evaluating Real-World Robot Manipulation Policies in Simulation. Li et al.

Related Courses: Computer Vision

- Learning for 3D Vision by Shubham Tulsiani at CMU
- Machine Learning for Inverse Graphics by Vincent Sitzmann at MIT
- Computer Vision by Andreas Geiger at University Tübingen
- Learning for 3D Vision by Angjoo Kanazawa at Berkeley
- 3D Vision by Derek Hoiem at UIUC
- Advances in Computer Vision by Sara Beery, Kaiming He, Mina Konaković Luković and Vincent Sitzmann at MIT

Related Courses: Animation and Simulation

- Physics-based Animation by David I.W. Levin at U Toronto
- Simulation Methods for Animation and Digital Fabrication by Stelian Coros at CMU
- Computer Graphics by Nancy Pollard at CMU
- Physics-based Animation of Solids and Fluids by Minchen Li at CMU
- Computer Graphics: Animation and Simulation by Doug James at Stanford
- Topics in Computing : Physics-Based Modeling and Simulation by Eftychios Sifakis at University of Wisconsin–Madison

Related Courses: Physics Informed Machine Learning

- Physics Informed Machine Learning by Steven Brunton at UW

Reminder

- I am not enrolled in but I would like to enroll in this class. What should I do?
  - If you're interested in enrolling, please fill out and submit this form.

https://forms.gle/ZS64KrqJmoeLzxK67

- Any application of course withdrawal after 4/20 **will not** be accepted!