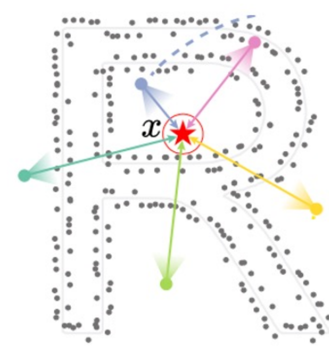
Functional and Perceptual Inductive Biases in 3D Representation Learning



Hao (Richard) Zhang, Simon Fraser University (SFU)

CVPR Workshop on Enforcing Inductive Biases in 3D Generation (Ind3D), June 12, 2025

What is inductive bias?

Wikipedia
The Free Encyclopedia

Search Wikipedia | Search

Inductive bias

7 languages

Contents hide

Article Talk

Read Edit View history Tools

From Wikipedia, the free encyclopedia

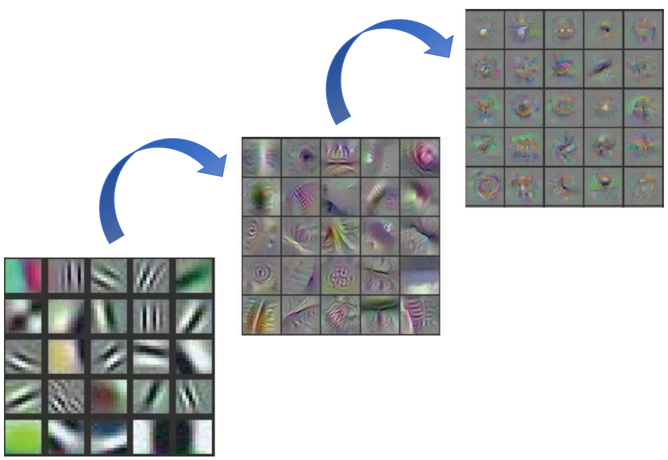
The **inductive bias** (also known as **learning bias**) of a learning algorithm is the set of assumptions that the learner uses to predict outputs of given inputs that it has not encountered.[1] Inductive bias is anything which makes the algorithm learn one pattern instead of another pattern (e.g., step-functions in decision trees instead of continuous functions in linear regression models). Learning involves searching a space of solutions for a solution that provides a good explanation of the data. However, in many cases, there may be multiple equally appropriate solutions.[2] An inductive bias allows a learning algorithm to prioritize one solution (or interpretation) over another, independently of the observed data.[3]

Inductive bias (also known as learning bias) of a learning algorithm is the set of assumptions that the learner uses to predict outputs of given inputs that it has not encountered.

It helps the learner predict and generalize better from limited training data to unseen data.

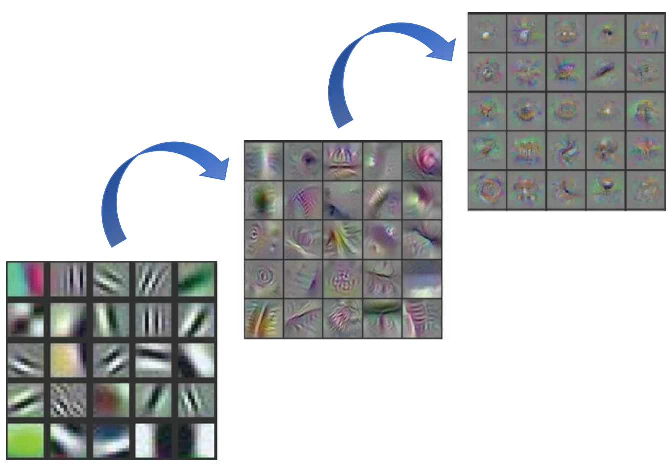
Where to embed inductive biases?

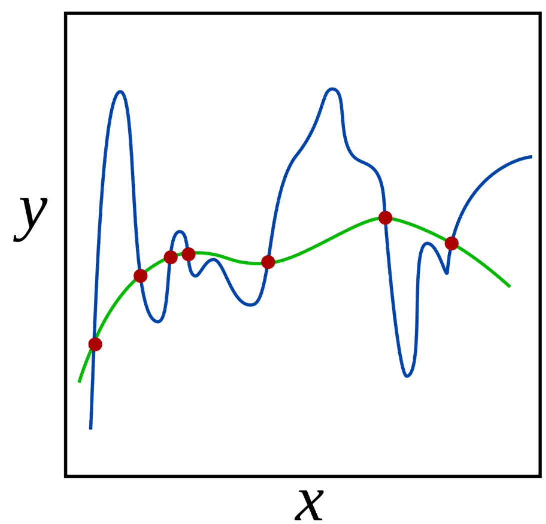
❖ Network architecture, e.g., hierarchical feature learning in CNNs, sequential data dependencies in RNNs, etc.



Where to embed inductive biases?

- Network architecture, e.g., hierarchical feature learning in CNNs , sequential data dependencies in RNNs, etc.
- Network losses, e.g., regularization terms (vs. data terms)

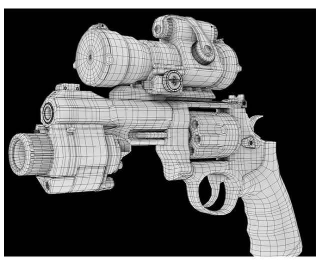


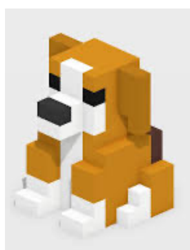


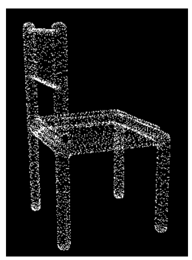
To avoid overfitting

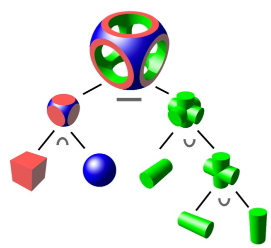
Where to embed inductive biases?

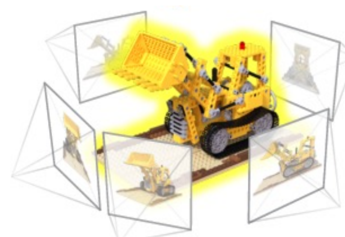
- Network architecture, e.g., hierarchical feature learning in CNNs , sequential data dependencies in RNNs, etc.
- Network losses, e.g., regularization terms (vs. data terms)
- More fundamentally, shape the data representation learned
  - For 3D models, there is no unique choice; there are many choices


Mesh


Voxels


Point cloud


CSG

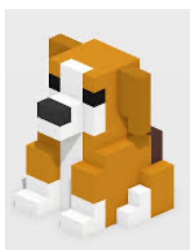

NeRF

3DGS

Examples of 3D representation bias

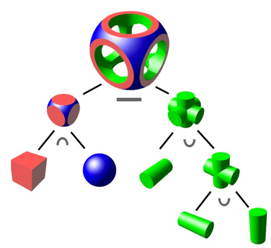
❖ Compactness, e.g., voxels/SDFs vs. CAD primitives (e.g., CSG)

Mesh



Voxels

Point cloud



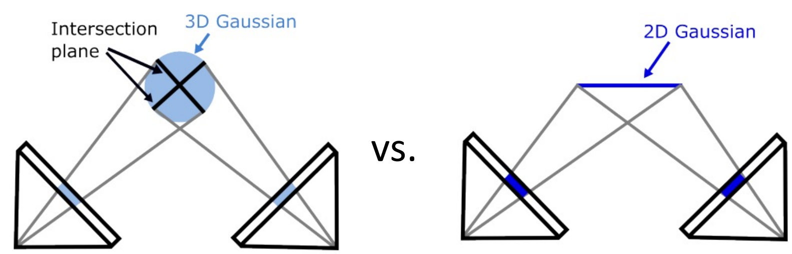
CSG

NeRF

3DGS

Examples of 3D representation bias

- ❖ Compactness, e.g., voxels/SDFs vs. CAD primitives (e.g., CSG)
- ❖ Surface bias, e.g., 3D [Kerbl et al. 2023] vs. 2D GS [Huang et al. 2024], etc.



Mesh

Voxels

Point cloud

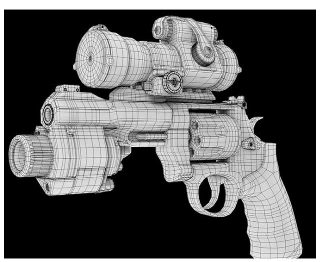
CSG

NeRF

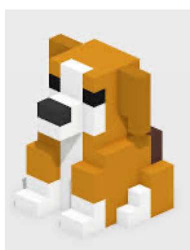
3DGS

Examples of 3D representation bias

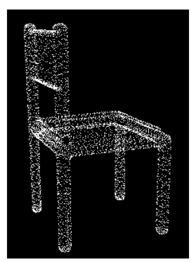
- Compactness, e.g., voxels/SDFs vs. CAD primitives (e.g., CSG)
- Surface bias, e.g., 3D [Kerbl et al. 2023] vs. 2D GS [Huang et al. 2024], etc.
- More fundamentally, bias/shape the learned 3D rep to capture the most predicable property, e.g., from a class description
  - Why? Since predictability ⇨ transferability & generalizability



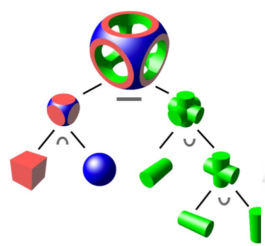
Mesh



Voxels



Point cloud



CSG

NeRF

3DGS

When you say an object class, e.g., “carts,” “lamps,” or “chairs,” etc., what properties or attributes you are most sure of about it?

What is most predicable about chairs?

- Shape?
- Topology?

Image taken from dreamstimes.com

What is most predicable about chairs?

- Shape?
- Topology?
- Color?
- Texture?
- Material?

Image taken from pinterest.ca

Chair or not?

Why is this not a chair? Or is it?

Ask Chamfer Distance and it would probably say yes ☺

"What makes a chair a chair?"

CVPR 2011

What Makes a Chair a Chair?

Helmut Grabner[1] Juergen Gall[1] Luc Van Gool[1,2]

[1]Computer Vision Laboratory
ETH Zurich
{grabner,gall,vangool}@vision.ee.ethz.ch

[2]ESAT - PSI / IBBT
K.U. Leuven
luc.vangool@esat.kuleuven.be

Abstract

Many object classes are primarily defined by their functions. However, this fact has been left largely unexploited by visual object categorization or detection systems. We propose a method to learn an affordance detector*. It identifies locations in the 3d space which "support" the particular function. Our novel approach "imagines" an actor performing an action typical for the target object class, instead of relying purely on the visual object appearance. So, function is handled as a cue complementary to appearance, rather than being a consideration after appearance-based detection. Experimental results are given for the functional category "sitting". Such affordance is tested on a 3d representation of the scene, as can be realistically obtained through SfM or depth cameras. In contrast to appearance-based object detectors, affordance detection requires only very few training examples and generalizes very well to other sittable objects like benches or sofas when trained on a few chairs.*

1. Introduction

"An object is first identified as having important functional relations, [...] perceptual analysis is derived of the functional concept [...]."
Nelson, 1974, [17]

"Affordances relate the utility of things, events, and places to the needs of animals and their actions in fulfilling them [...]. Affordances themselves are perceived and, in fact, are the essence of what we perceive."
Gibson, 1982, [8, p. 60]

"There's little we can find in common to all chairs – except for their intended use."
Minsky, 1986, [16, p. 123]

"[...] objects like coffee cups are artifacts that were created to fulfill a function. The function of an object plays a critical role in processing that object [... for] categorization and naming."
Carlson-Radvansky *et al.*, 1999, [4]

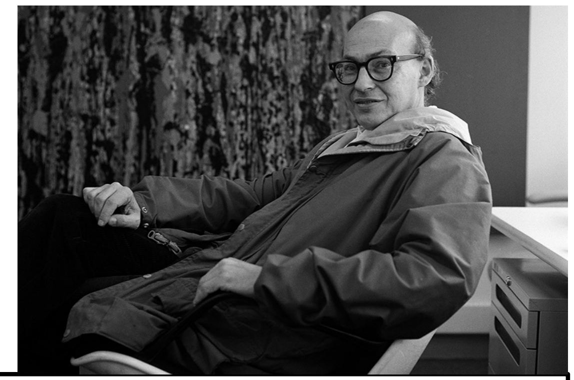
Figure 1. The "chair-challenge" by I. and H. Bülthoff [3] (reprint with the author's permission).

These quotes emphasize that functional properties or affordances[1] are essential for forming concepts and learning object categories. Experiments (*e.g.* [18, 4]) have demonstrated that both appearance and function are strong cues for learning by infants. Initially they attend only to the form of an object. Later they use form and function and finally (by the age of 18 months) they attend to the relationships between form and function. Furthermore, Booth and Waxman [2] have identified two salient cues that facilitate categorization in infancy, namely (i) object functions and (ii) object names. Moreover, names of objects most often evolve on the basis of function[2].

Whereas all this is well known for a long time, it has been left mostly unused for object detection in computer vision. Taking a look at the results of the recent Pascal VOC Challenge [5], the performance still strongly depends

[1]"*Affordance:* A situation where an object's sensory characteristics intuitively imply its functionality and use. [...] A chair, by its size, its curvature, its balance, and its position, suggests sitting on it.", http://www.usabilityfirst.com/glossary/affordance, 2010/07/28. Introduced in 1979 by Gibson [9, p. 127] based on the verb *afford*.

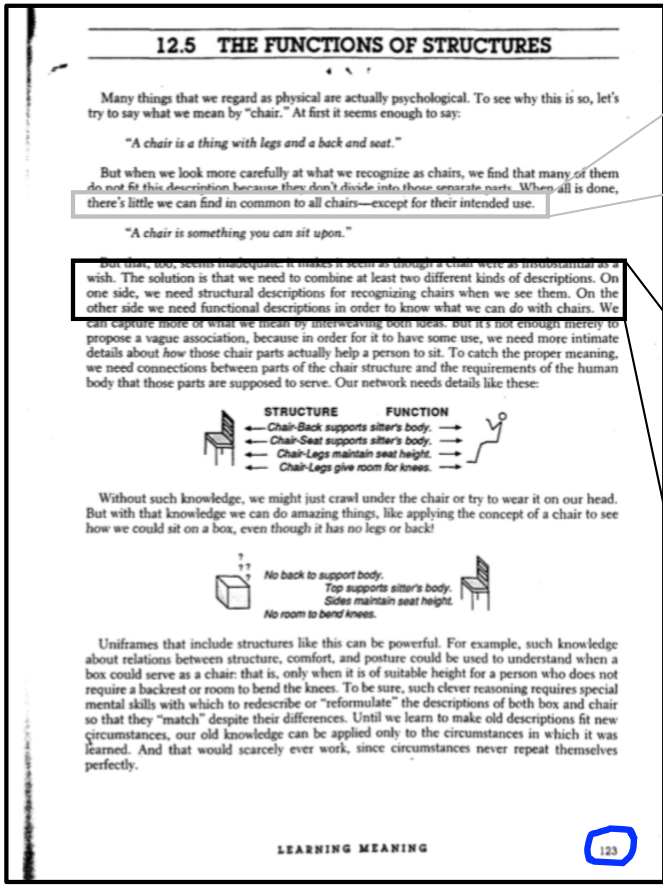
[2]When considering the evolution of a word for an object, most of the time it is based on its function. For example the word "chair": PIE base **sed-* (to sit) → Latin *sedentarius* (sitting, remaining in one place) → *sedentary* (meaning "not in the habit of exercise") → *cathedral* → *chair*. http://www.etymonline.com, 2010/10/02.



"There's little we can find in common to all chairs – except for their intended use."

Marvin Minsky: "The Society of Mind" [1986]

From Minsky's "The Society of Mind"



12.5 THE FUNCTIONS OF STRUCTURES

Many things that we regard as physical are actually psychological. To see why this is so, let's try to say what we mean by "chair." At first it seems enough to say:

"A chair is a thing with legs and a back and seat."

But when we look more carefully at what we recognize as chairs, we find that many of them do not fit this description because they don't divide into those separate parts. When all is done, there's little we can find in common to all chairs—except for their intended use.

"A chair is something you can sit upon."

But that, too, seems inadequate: it makes it seem as though a chair were as insubstantial as a wish. The solution is that we need to combine at least two different kinds of descriptions. On one side, we need structural descriptions for recognizing chairs when we see them. On the other side we need functional descriptions in order to know what we can *do* with chairs. We can capture more of what we mean by interweaving both ideas. But it's not enough merely to propose a vague association, because in order for it to have some use, we need more intimate details about *how* those chair parts actually help a person to sit. To catch the proper meaning, we need connections between parts of the chair structure and the requirements of the human body that those parts are supposed to serve. Our network needs details like these:

Without such knowledge, we might just crawl under the chair or try to wear it on our head. But with that knowledge we can do amazing things, like applying the concept of a chair to see *how we could sit on a box, even though it has no legs or back!*

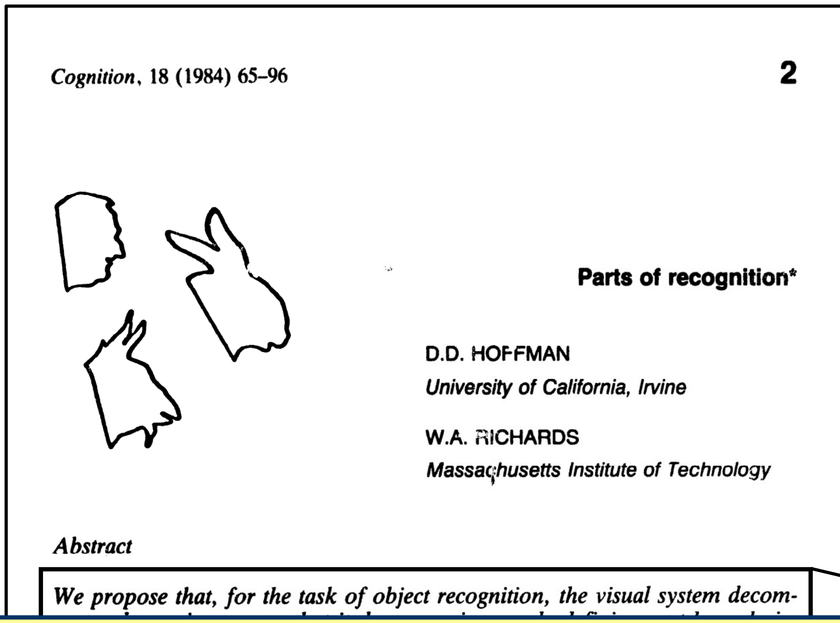
Uniframes that include structures like this can be powerful. For example, such knowledge about relations between structure, comfort, and posture could be used to understand when a box could serve as a chair: that is, only when it is of suitable height for a person who does not require a backrest or room to bend the knees. To be sure, such clever reasoning requires special mental skills with which to redescribe or "reformulate" the descriptions of both box and chair so that they "match" despite their differences. Until we learn to make old descriptions fit new circumstances, our old knowledge can be applied only to the circumstances in which it was learned. And that would scarcely ever work, since circumstances never repeat themselves perfectly.

LEARNING MEANING 123

"There's little we can find in common to all chairs – except for their intended use."

"… we need to combine at least two different kinds of descriptions (of objects). On one side, we need structural descriptions for recognizing chairs when we see them. On the other side, we need functional descriptions in order to know what we can **DO** with chairs."

Structured models mimic human perception



Cognition, 18 (1984) 65–96 **2**

Parts of recognition*

D.D. HOFFMAN
University of California, Irvine

W.A. RICHARDS
Massachusetts Institute of Technology

Abstract

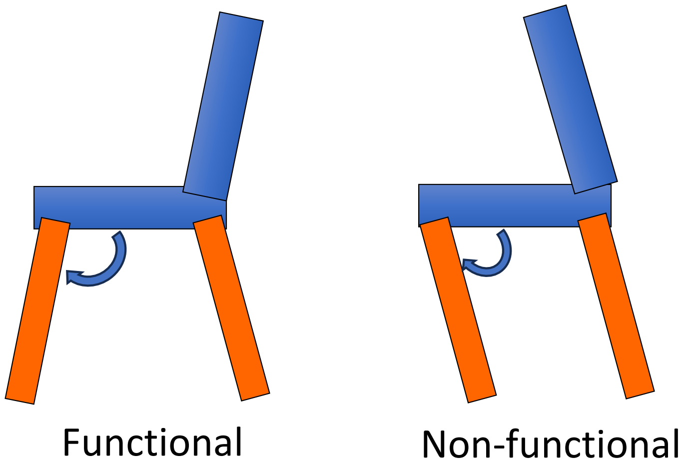
We propose that, for the task of object recognition, the visual system decom-

"For the task of object recognition, the visual system decomposes shapes into parts, . . ., parts with their descriptions and spatial relations provide a first index into a memory of shapes …

Structured models reflect our perception of the world, leading to higher degrees of transferability and controllability (e.g., editability).

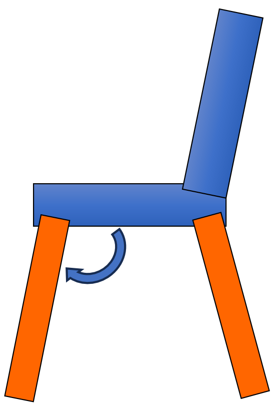
Object functions and structures

❖ Object functions are mainly characterized by object structures, i.e., parts + relations,

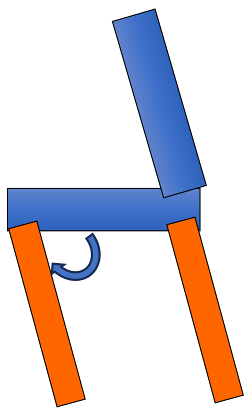


Object functions and structures

❖ Object functions are mainly characterized by object structures, i.e., parts + relations, manifested in motion



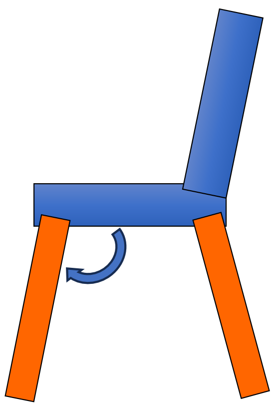
Functional



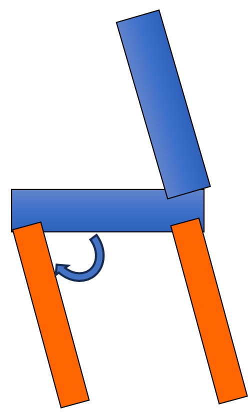
Non-functional

Object functions and structures

❖ Object functions are mainly characterized by object structures, i.e., parts + relations, manifested in motion, thru interactions



Functional



Non-functional

Human and bicycle

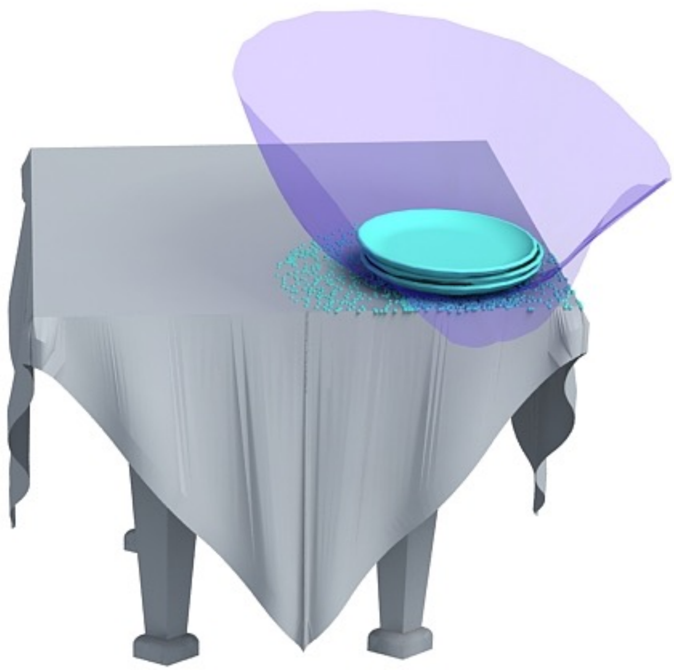
Hanger and jackets

Human and backpack

Hand and racket

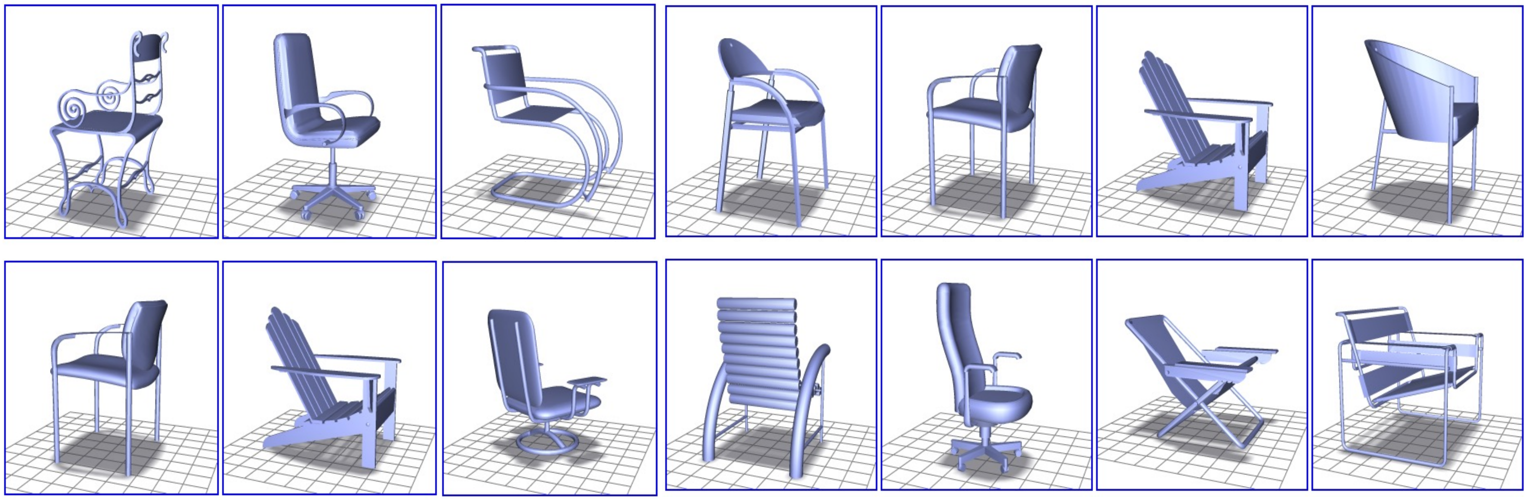
Representing object-object interactions

- **IBS: I**ntersection **B**isector **S**urface (to describe the interaction)
  = an encoding of trimmed Voronoi boundary



Key take-away

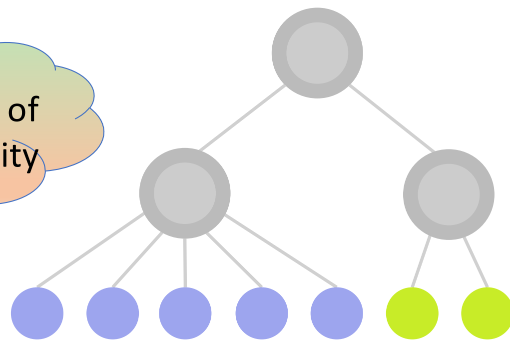
❖ A representation (e.g., IBS for interactions) that emphasizes functional understanding is more robust/invariant than any representation of a 3D object's *intrinsic itself*, whether it is shape, topology, color, or texture



IBS: chair and table

ICON (Interaction CONtext) series

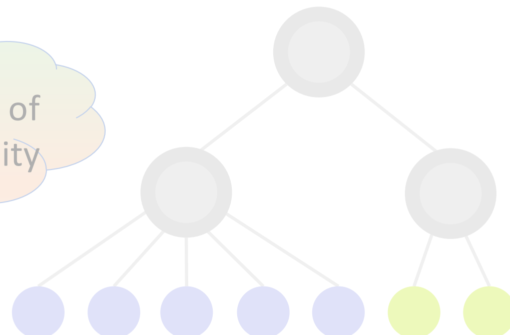
A representation of object functionality



ICON: [Hu et al. SIGGRAPH 2015]

ICON (Interaction CONtext) series

A representation of object functionality



ICON: [Hu et al. SIGGRAPH 2015]

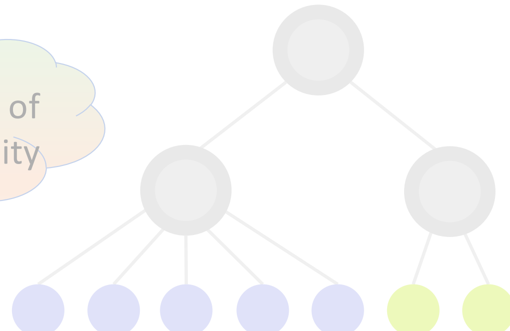
Learning functionality of an object category

"What makes a handcart a handcart, functionally?"

ICON2: [Hu et al. SIGGRAPH 2016]

ICON (Interaction CONtext) series

A representation of object functionality



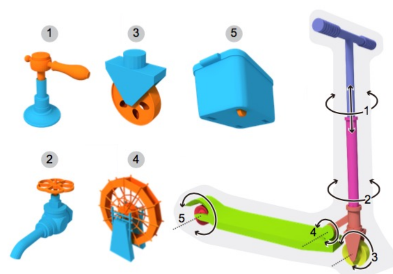
ICON: [Hu et al. SIGGRAPH 2015]

Learning functionality of an object category

"What makes a handcart a handcart, functionally?"

ICON2: [Hu et al. SIGGRAPH 2016]

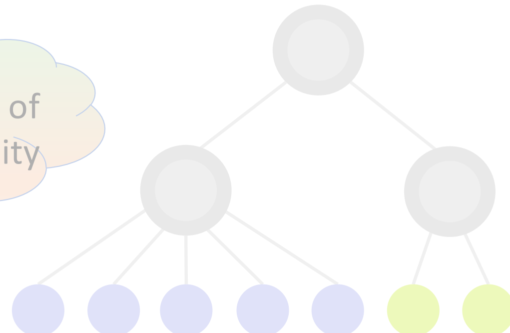
Learning part mobility from few static images



ICON3: [Hu et al. SIGGRAPH Asia 2017]

ICON (Interaction CONtext) series

A representation of object functionality



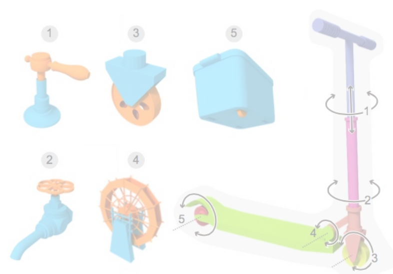
ICON: [Hu et al. SIGGRAPH 2015]

Learning functionality of an object category

"What makes a handcart a handcart, functionally?"

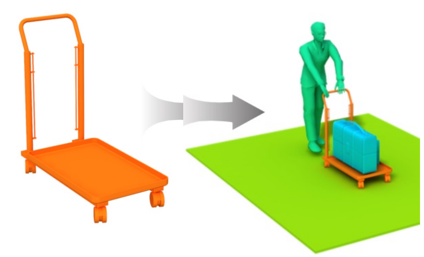
ICON2: [Hu et al. SIGGRAPH 2016]

Learning part mobility from few static images



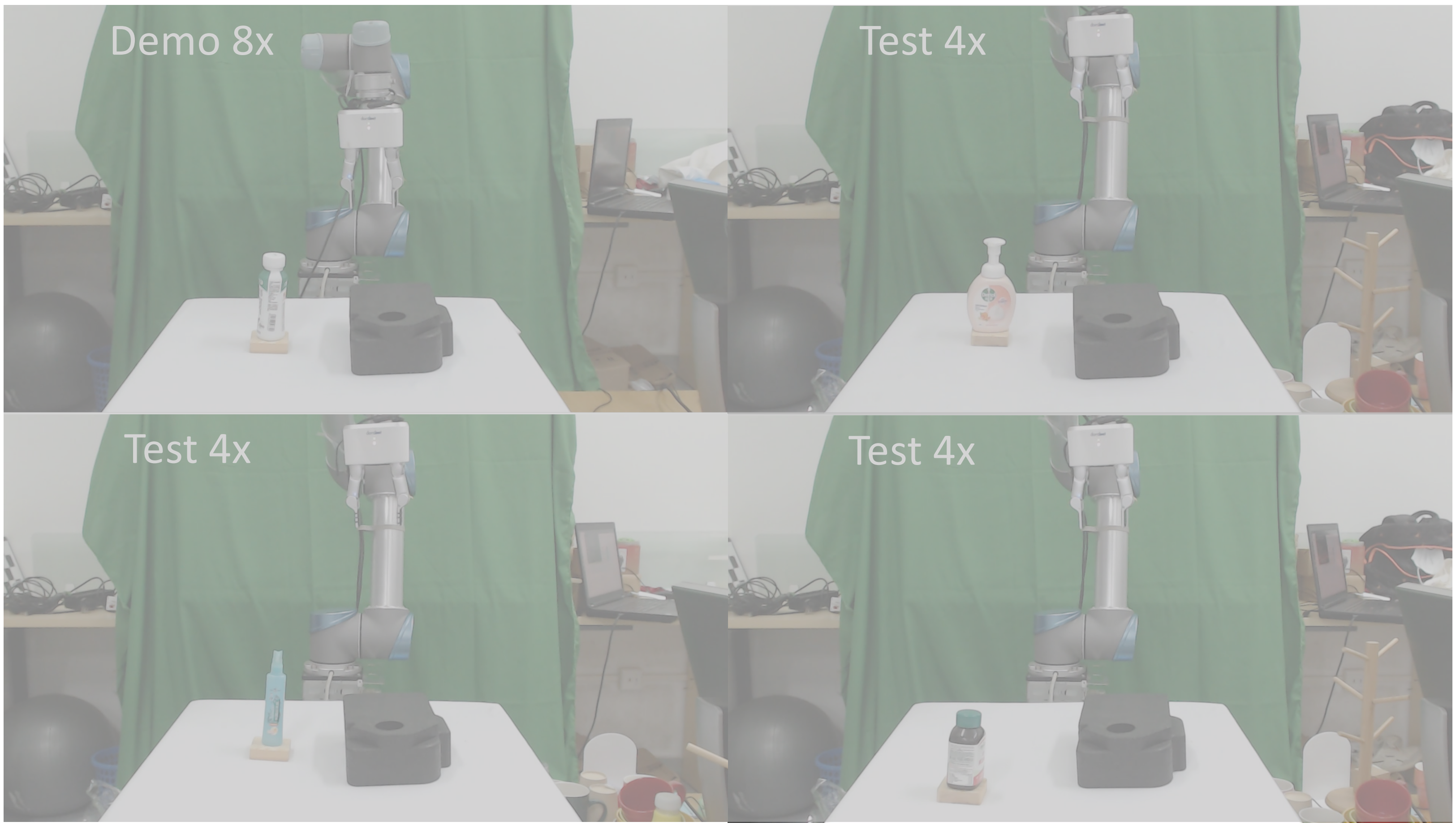
ICON3: [Hu et al. SIGGRAPH Asia 2017]

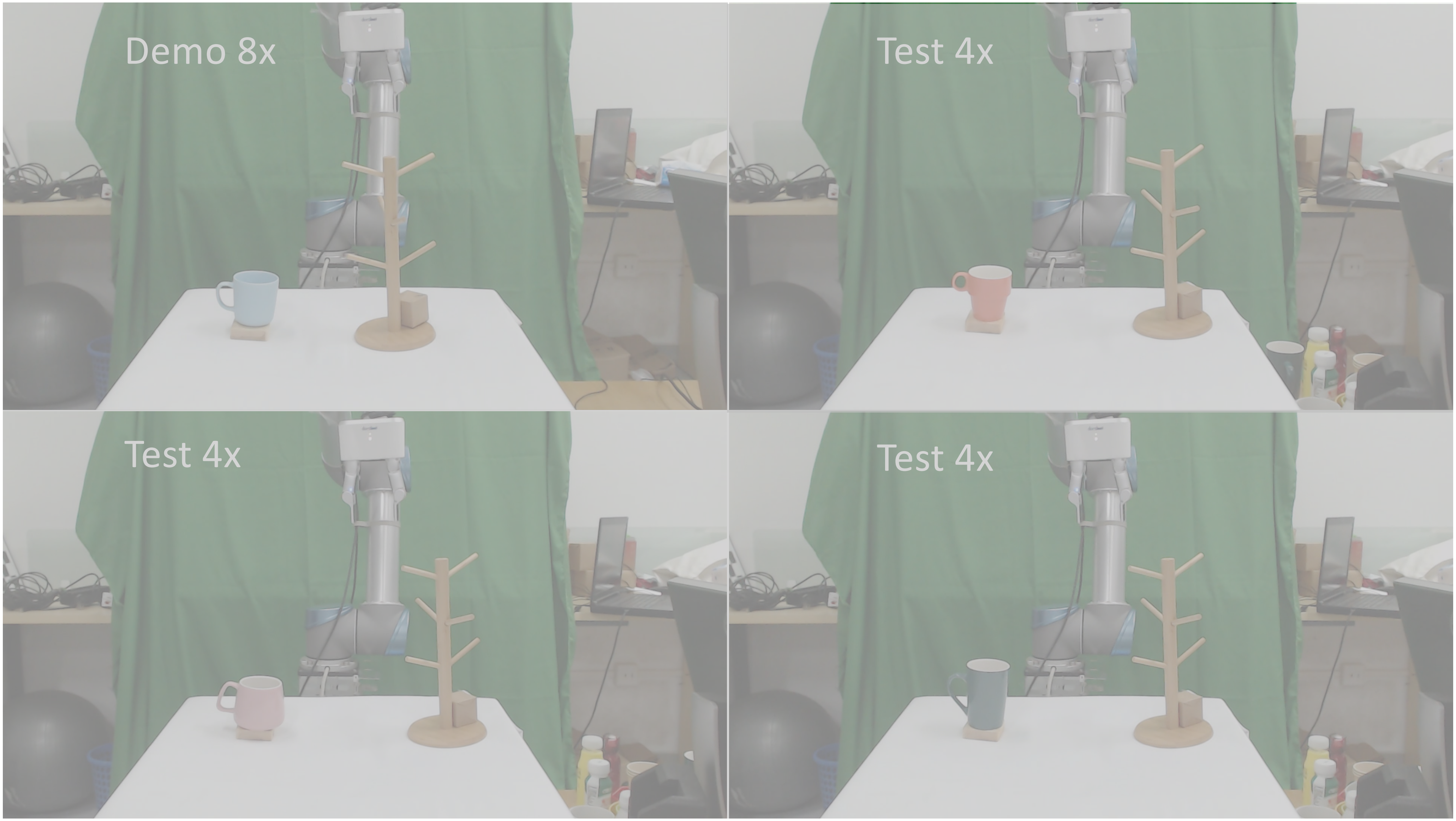
A generative network for usage scenarios



ICON4: [Hu et al. SIGGRAPH 2018]

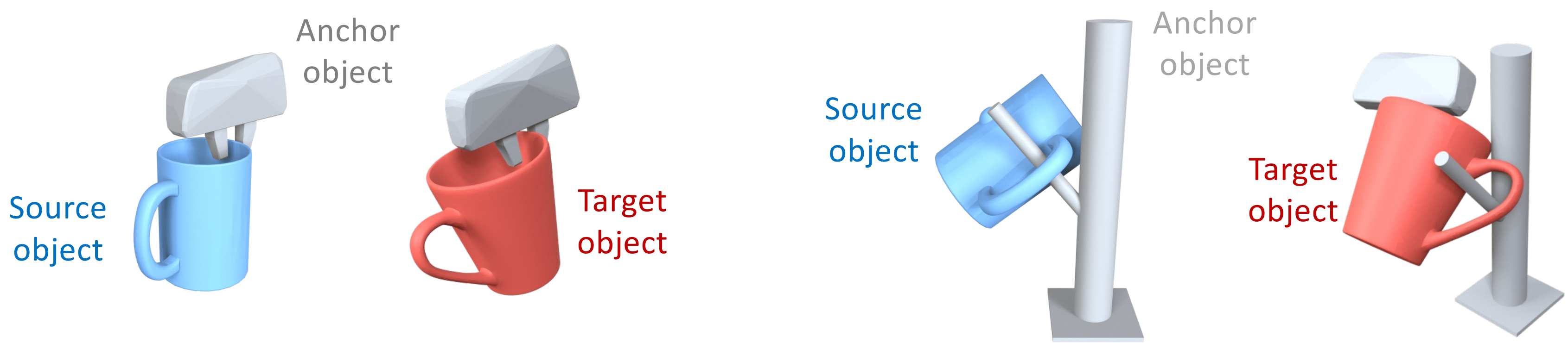
Robot pick-and-place by imitation learning




Demo 8x
Test 4x
Test 4x
Test 4x

The problem

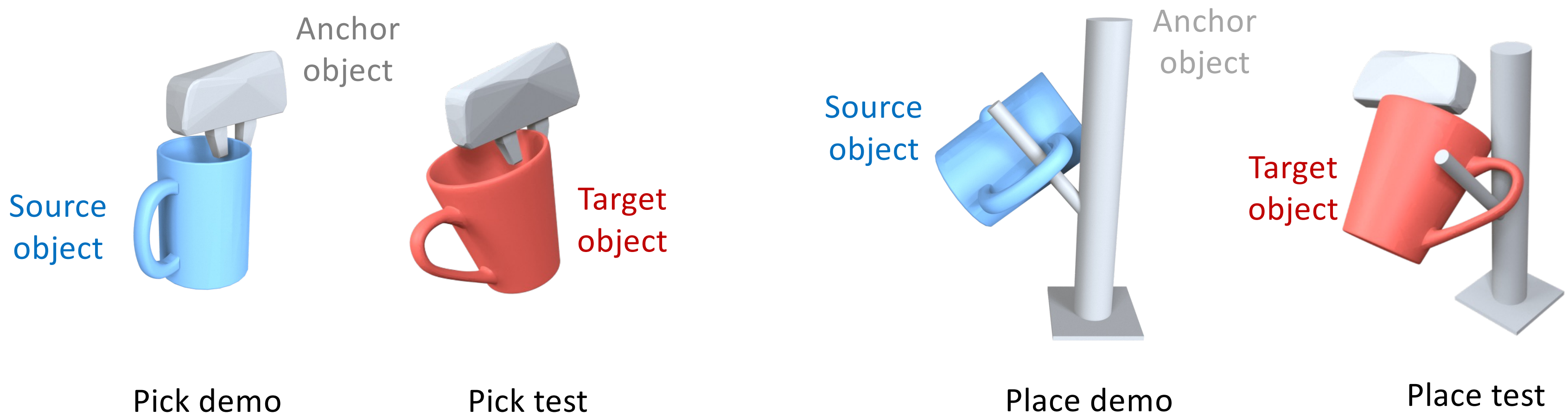
- Given one or few demo manipulations of pick-and-place, learn to perform the task on a new (target) object in arbitrary pose



Pick demo Pick test Place demo Place test

Key question

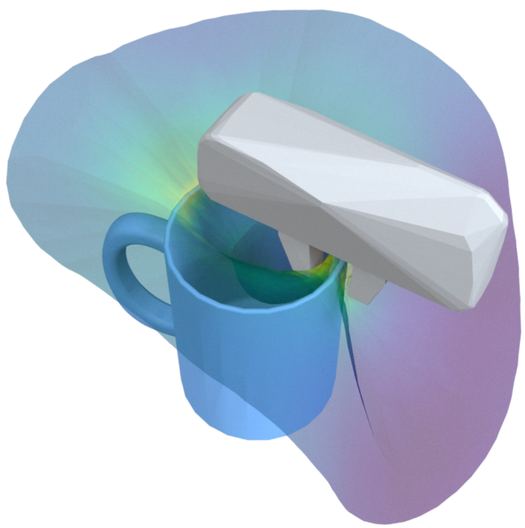
- Given one or few demo manipulations of pick-and-place, learn to perform the task on a new (target) object in arbitrary pose



- How to encode relative poses (i.e., interactions) between source/target objects and the anchor object to generalize well to new targets?

Key observation

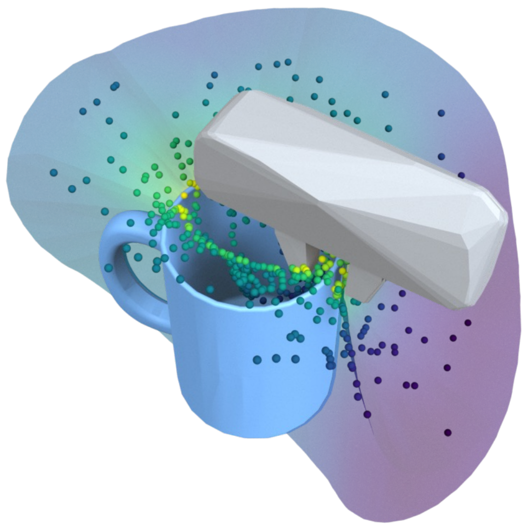
❖ The Intersection Bisector Surface (IBS) is robust against shape variations



Use interaction template

❖ The Intersection Bisector Surface (IBS) is robust against shape variations

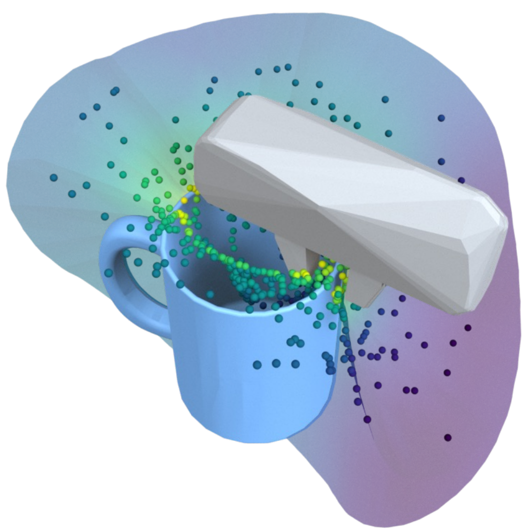
❖ Sample points from the IBS (instead of around the anchor object), and encode neural features to form an interaction template

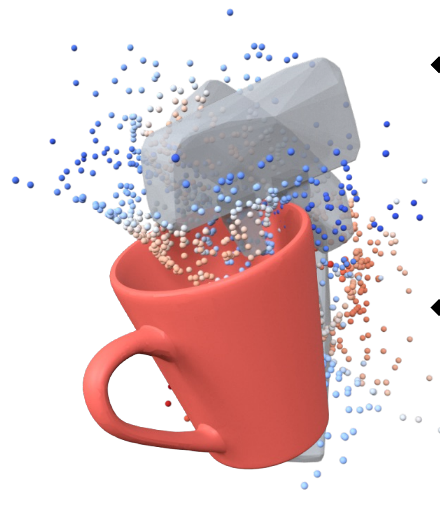


Optimize pose to match IBS template

❖ The Intersection Bisector Surface (IBS) is robust against shape variations

❖ Sample points from the IBS (instead of around the anchor object), and encode neural features to form an interaction template





Target object

❖ For pick test, re-pose gripper to match the interaction templates

❖ For place test, transform target object to match interaction templates

[Huang et al. ICRA 2023]

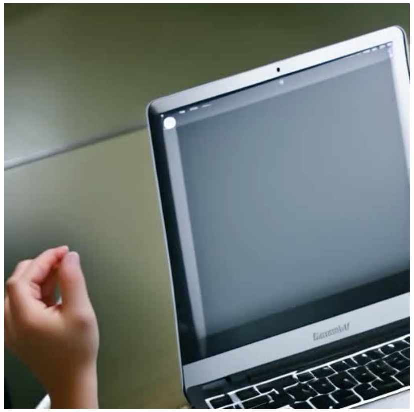
Motion generation for 3D objects

- Functions of daily objects often performed through part articulation
- Goal: generate part articulations for an input mesh without 3D annotation by leveraging open-vocabulary capabilities of video diffusion models
- The foundation model provides the inductive bias to avoid 3D annotations

Motion generation for 3D objects

- ❖ Functions of daily objects often performed through part articulation
- ❖ Goal: predict part articulations on an input mesh without 3D annotation by leveraging open-vocabulary capabilities of video diffusion models
- ❖ But existing text2video models (e.g., SVD) do not handle articulations well

"A person opening the door of dishwasher"



"A person opening the lid of the laptop"

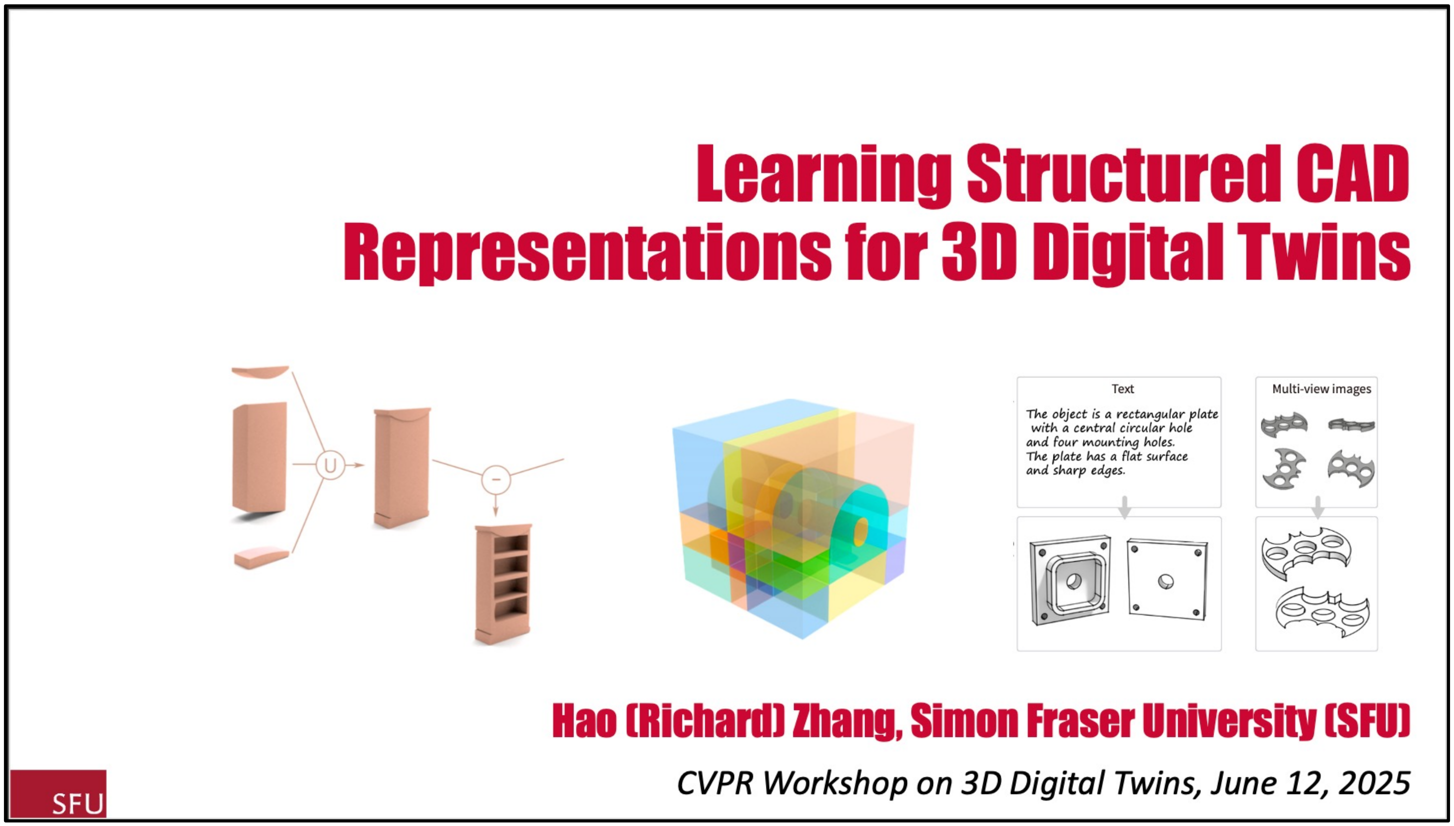
Results from Stable Video Diffusion (SVD) [Blattmann et al. 2023]

Motion generation for 3D objects

- Few-shot finetuning of SVD with category-specific motion videos
- Video motion personalization to input 3D mesh, then motion transfer
- Training of foundational models can be limited by own inductive bias …

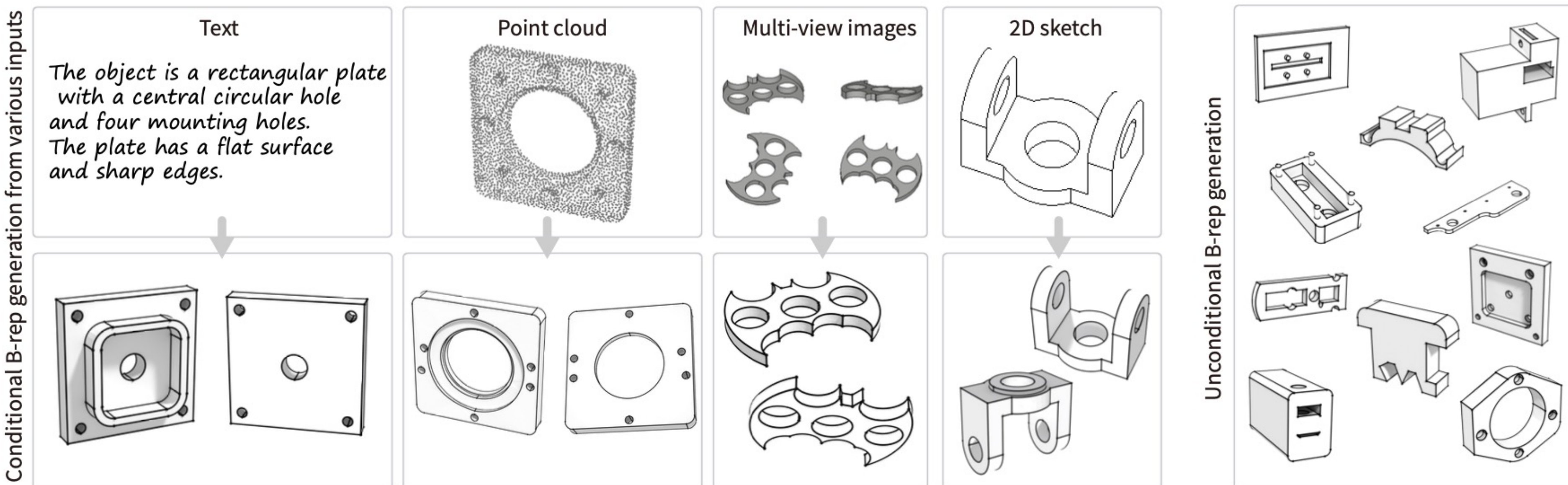
ATOP: Articulate That Object Part [Vora et al. 2025]

Learning structured 3D representations



Example 1: 1st multi-modal B-Rep generation

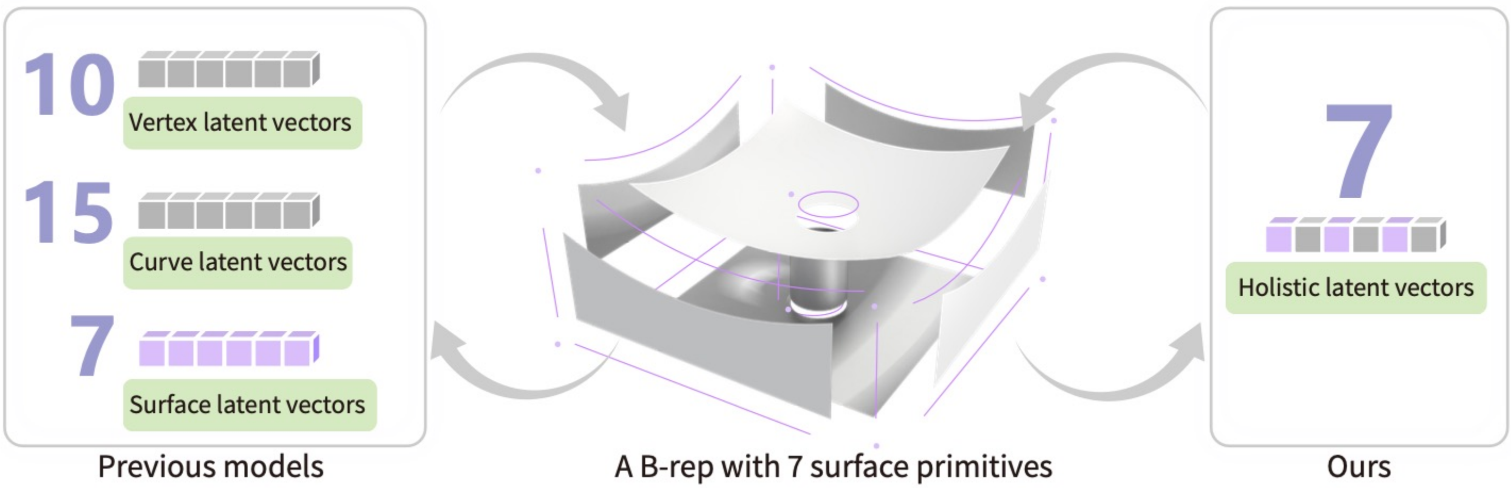
❖ B-Rep (boundary representation): *de facto* standard in CAD



Holistic Latent (HoLa) space + diffusion-based generator [Liu et al. SIG 2025]

Example 1: key idea = holistic latent

❖ Instead of having separate latents (and generators) for each primitive, learn a single surface-centric holistic latent



Holistic Latent (HoLa) space + diffusion-based generator [Liu et al. SIG 2025]

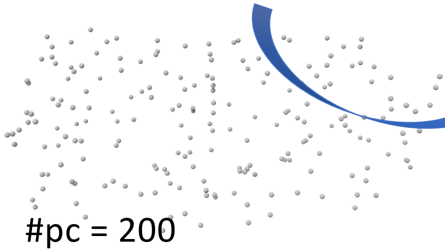
Example #2: use of CAD programs

- Programs, like languages, are inherently structured
- Easy to inject inductive biases suitable for CAD or architecture
- Program-based learning builds on token prediction, which can leverage the power of modern-day transformers

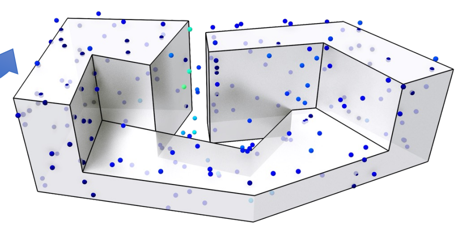
Example #2: use of CAD programs

- Programs, like languages, are inherently structured
- Easy to inject inductive biases suitable for CAD or architecture
- Program-based learning builds on token prediction, which can leverage the power of modern-day transformers

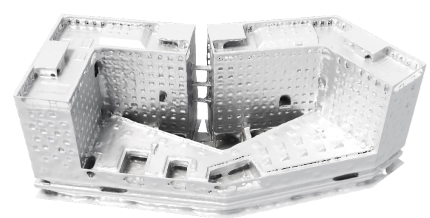
Architectural Programs for structured 3D abstraction [Huang et al. CVPR 2025]



Highly sparse, incomplete, and noisy point cloud



Output
(3D Abstraction)



Reference
(Dense Mesh)

```
φ = SetGround(-.28)
L1 = CreateLayer(
        parent=φ,
        height=.09,
        contour=[(-.43, .22), (.25, .35),
                 (.33, -.10), (.11, -.14),
                 (.064, .082), (-.16, .04),
                 (-.12, -.19), (-.35, -.23)])
L2 = CreateLayer(
        parent=L1,
        height=.46,
        contour=[(-.43, .22), (-.21, .27),
                (-.16, .039), (-.39, -.0041)])
L3 = CreateLayer(
        parent=L1,
        height=.14,
        contour=[(.021, .31), (.25, .35),
                (.33, -.10), (.11, -.14)])
L4 = CreateLayer(
        parent=L3,
        height=.23,
        contour=[(.021, .31), (.25, .35),
                (.29, .13), (.064, .082)])
```

CVPR’25 Highlight: Poster Session 2, Exhibition Hall D, Poster #114, 4-6PM, June 13

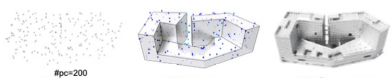
ArcPro: Architectural Programs for Structured 3D Abstraction of Sparse Points

Qirui Huang[1,2], Runze Zhang[1], Kangjun Liu[2], Minglun Gong[3], Hao Zhang[4], Hui Huang1*
[1]Shenzhen University [2]Pengcheng Laboratory [3]University of Guelph [4]Simon Fraser University

Motivation

Problem: Recover structured 3D abstractions from highly sparse and low-quality point clouds of architectures.

Sparse / Incomplete / Noisy / Outlier / SfM / Non-uniform density



Challenges: Optimization-based methods fail at plane detection on diverse, low-quality data, while learning-based approaches require suitable 3D representations and lack sufficient training data.
Our Solution: Represent the architecture as a shape program using a domain-specific language (DSL).
Motivations of DSL: 1) Inject architectural priors; 2) Provide a more compact representation space; 3) Leverage mature procedural modeling research to synthesize training data.
More Advantages: 1) Editability via adjustable program parameters; 2) Scalability through DSL extensions; 3) Compatibility with natural-language modality

Contributions

- The first program-based method for structured representation learning from sparse architectural point clouds.
- We connect feedforward and inverse procedural modeling by applying a feedforward process to synthesize training data, enabling the network to make reverse predictions.
- Comprehensive experiments demonstrate that ArcPro outperforms existing architecture proxy reconstruction and learning-based 3D abstraction methods.

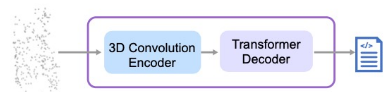
Method

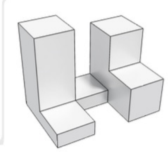
DSL Definition: Represent Architecture as a Shape Program

```
⟨Program⟩     ⊨ ⟨Statement⟩ ⟨Program⟩ | λ
⟨Statement⟩   ⊨ ⟨SetGround⟩ | ⟨CreateLayer⟩
⟨SetGround⟩   ⊨ SetGround(z=⟨Float⟩)
⟨CreateLayer⟩ ⊨ CreateLayer(parent=⟨LayerID⟩, h=⟨Float⟩, c=⟨Contour⟩)
⟨Contour⟩     ⊨ ⟨Polygon⟩
⟨Polygon⟩     ⊨ ⟨PointList⟩
⟨PointList⟩   ⊨ ⟨Point⟩ | ⟨Point⟩ ; ⟨PointList⟩
⟨Point⟩       ⊨ (⟨Float⟩, ⟨Float⟩)
⟨Float⟩       ⊨ a real number R
⟨LayerID⟩     ⊨ a symbol in {Φ, L1, ..., Lt}
```

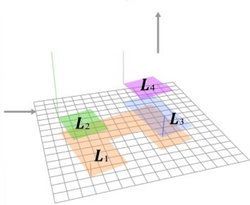
Data Synthesis: Prepare Training Data

Training: Next Token Prediction Loss





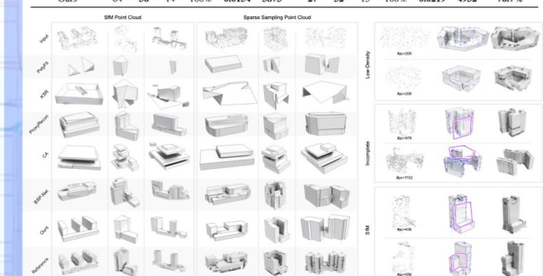
```
ϕ = SetGround(-.28)
L1 = CreateLayer(
        parent=ϕ,
        height=.09,
        contour=[(-.43, .22), (.25, .35),
                 (.33, -.10), (.11, -.14),
                 (.064, .082), (-.16, .04),
                 (-.12, -.19), (-.35, -.23)])
L2 = CreateLayer(
        parent=L1,
        height=.46,
        contour=[(-.43, .22), (-.21, .27),
                 (-.16, .039), (-.39, -.0041)])
L3 = CreateLayer(
        parent=L1,
        height=.14,
        contour=[(.021, .31), (.25, .35),
                 (.33, -.10), (.11, -.14)])
L4 = CreateLayer(
        parent=L3,
        height=.23,
        contour=[(.021, .31), (.25, .35),
                 (.29, .13), (.064, .082)])
```



Inference: Construct a 3D Model based on the Predicted Program

Experiments

Method	SfM Point Cloud #V	#F	#P	R_v	HD	LFD	Sparse Sampling Point Cloud #V	#F	#P	R_v	HD	LFD	User Study
PolyFit [21]	84	72	**11**	40%	0.0473	4365	91	78	**12**	15%	0.0458	7779	0.3%
KSR [3]	280	97	97	78%	0.0397	5905	32	42	40	54%	0.1131	8713	1.1%
ProxyRecon [8]	107	114	58	100%	0.0243	4364	60	90	34	100%	0.0256	5340	21.0%
CA [35]	**60**	180	180	100%	0.0363	6246	56	168	168	100%	0.0396	6987	0.1%
BSP-Net [6]	132	96	84	100%	0.0431	6671	102	170	67	100%	0.0487	7162	0.9%
Ours	64	**36**	14	100%	**0.0154**	**3873**	**27**	**32**	15	100%	**0.0219**	**4932**	**76.7%**



Applications

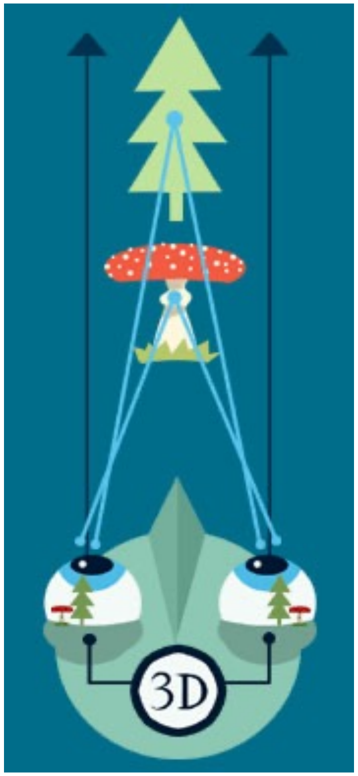
More applications of ArcPro: (a) architecture geometry structure analysis and natural language retrieval; (b) processing of raw SfM point clouds with ground points from drone aerial multi-view images; (c) processing of LiDAR-derived point clouds.

Another inductive bias via human perception

"Seeing" 3D/depth is an ill-posed task, performed by perception or extrapolation by our brain based on

- ❖ Binocular disparities: diff images from two eyes
- ❖ Monocular cues: shading, occlusion, perspectives

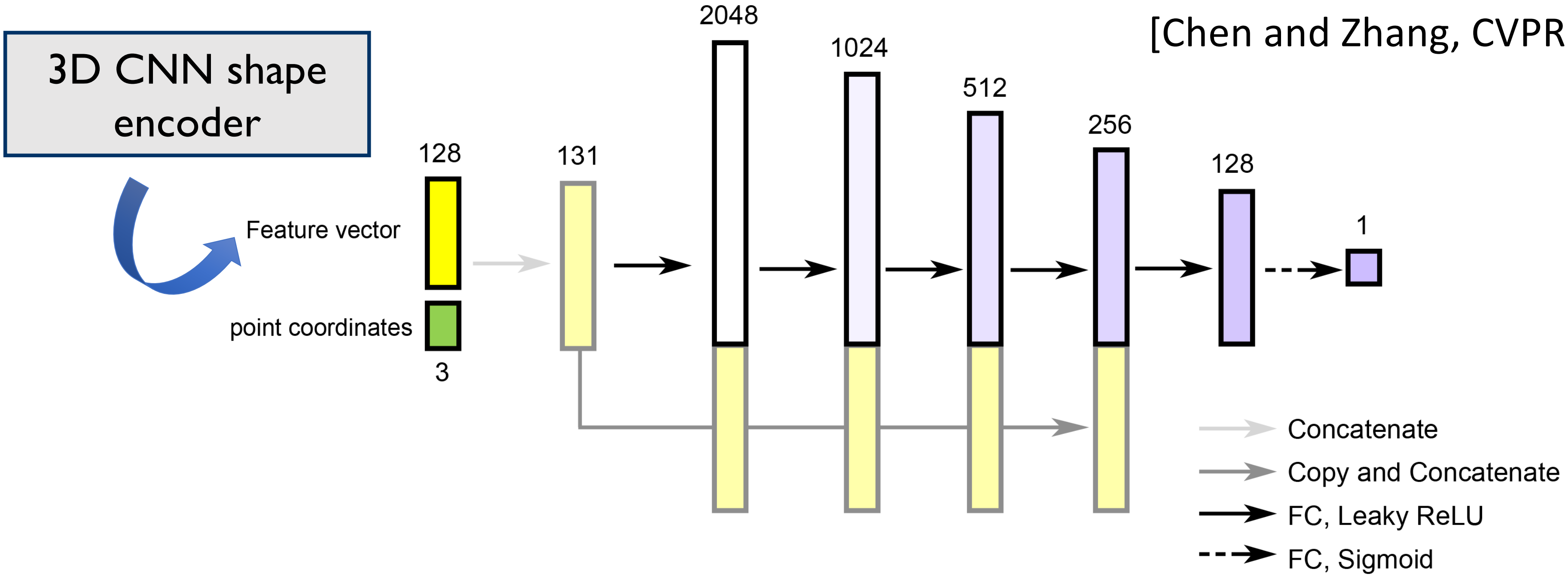
What can we learn in terms of 3D representation:
a perceptually motivated representation should be a pretty good "extrapolator" over the unseen …



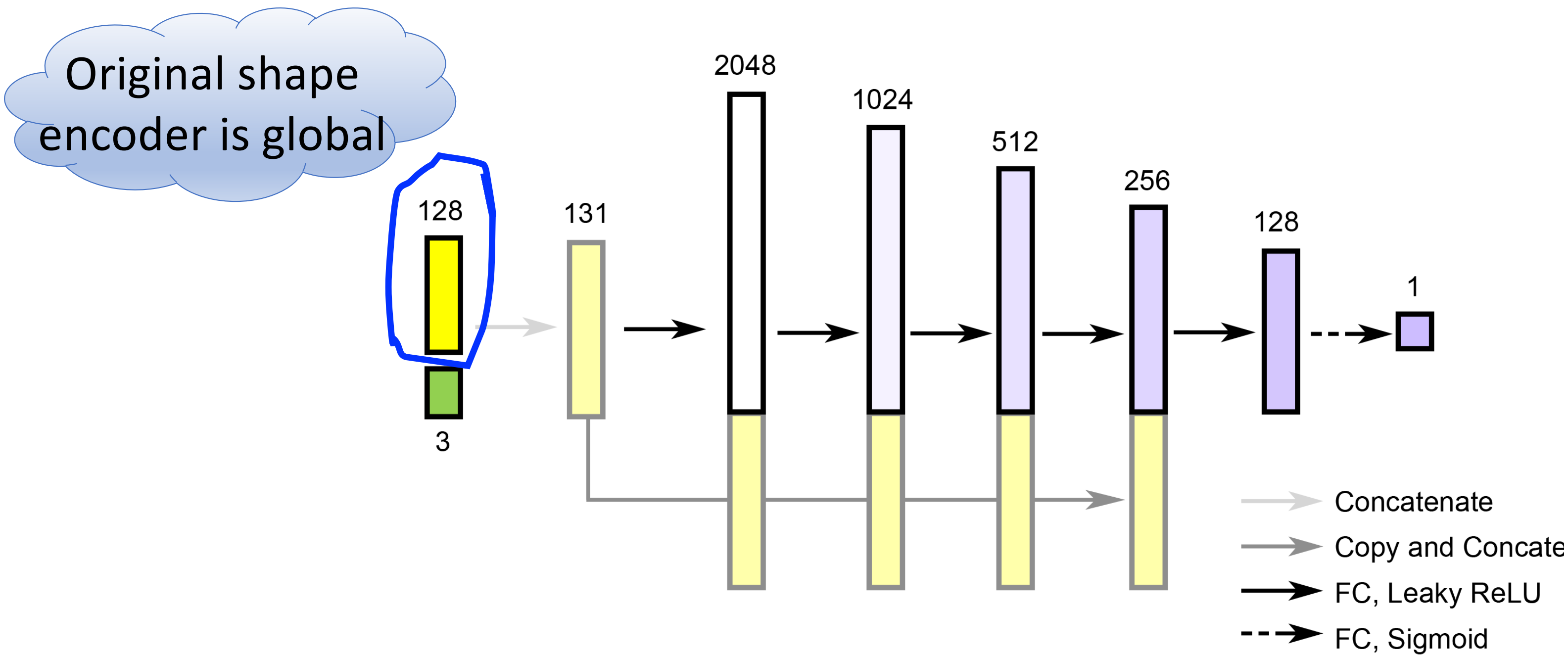
Start w/ IM-Net: an implicit field generator

- Learn a mapping from a 3D query point (x, y, z) to inside/outside status (= occupancy) with respect to shape boundary

[Chen and Zhang, CVPR 2019]

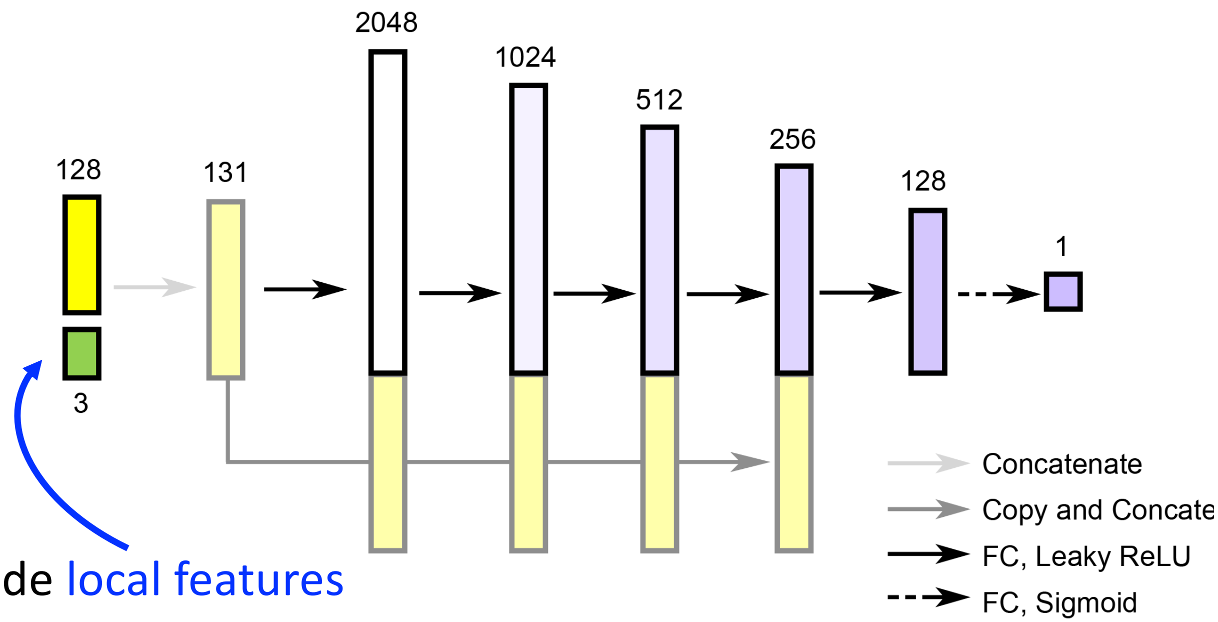


Feature encoding in IM-Net



IM-Net [Chen and Zhang, CVPR 2019]

Improved feature encoding

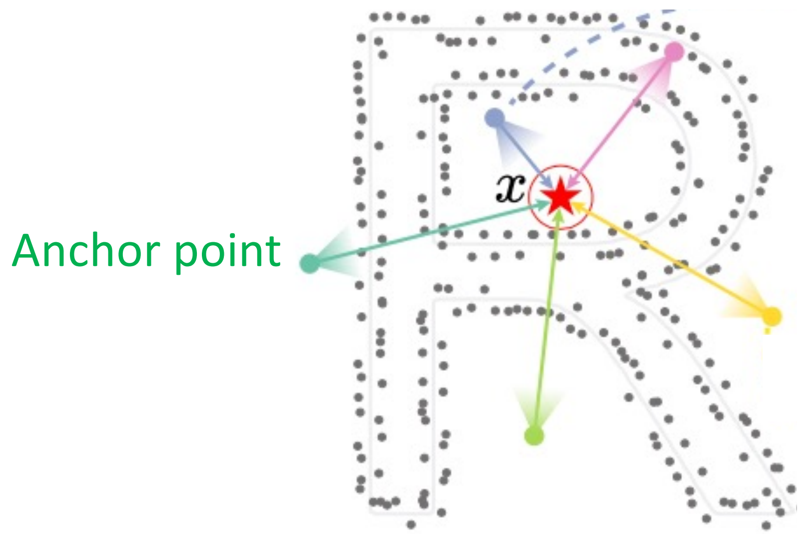


- ❖ Better to encode local features
- ❖ Even better with specificity to query point

IM-Net [Chen and Zhang, CVPR 2019]

Key: encode features perceptually

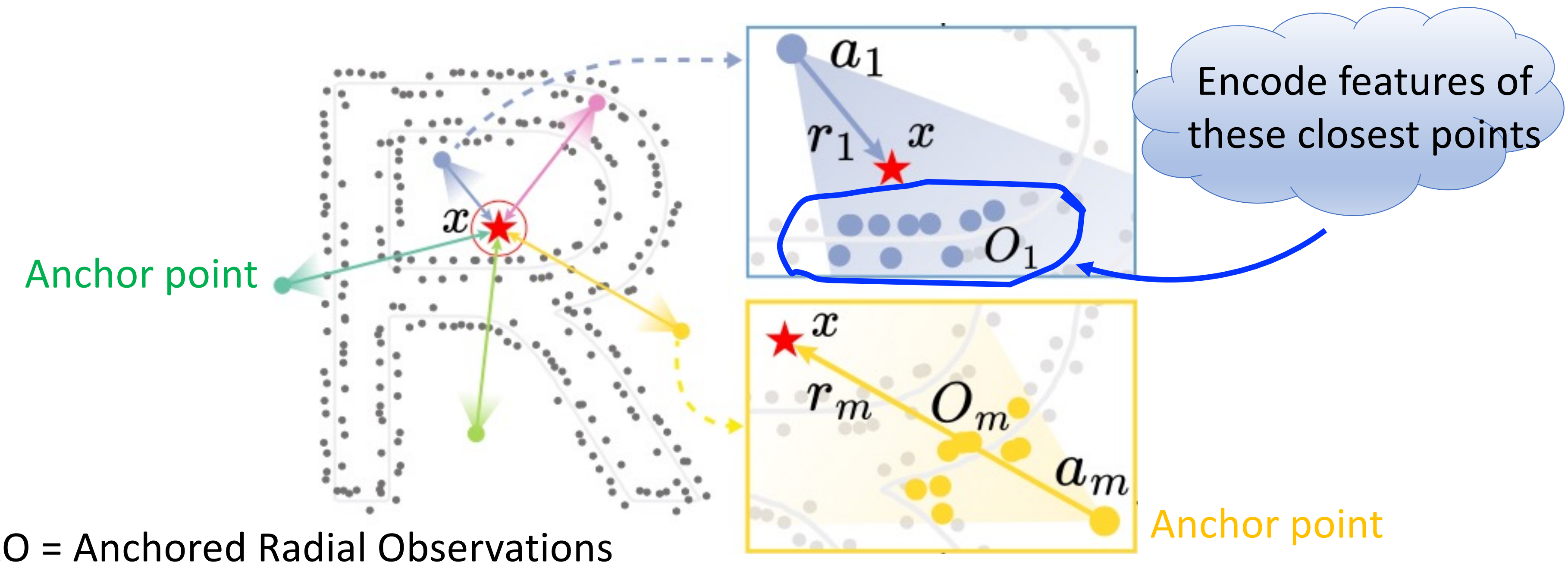
- Encode point features via multi-view perception: “What does the shape look at from various view/anchor points towards the query point (*x*)?”



ARO = Anchored Radial Observations

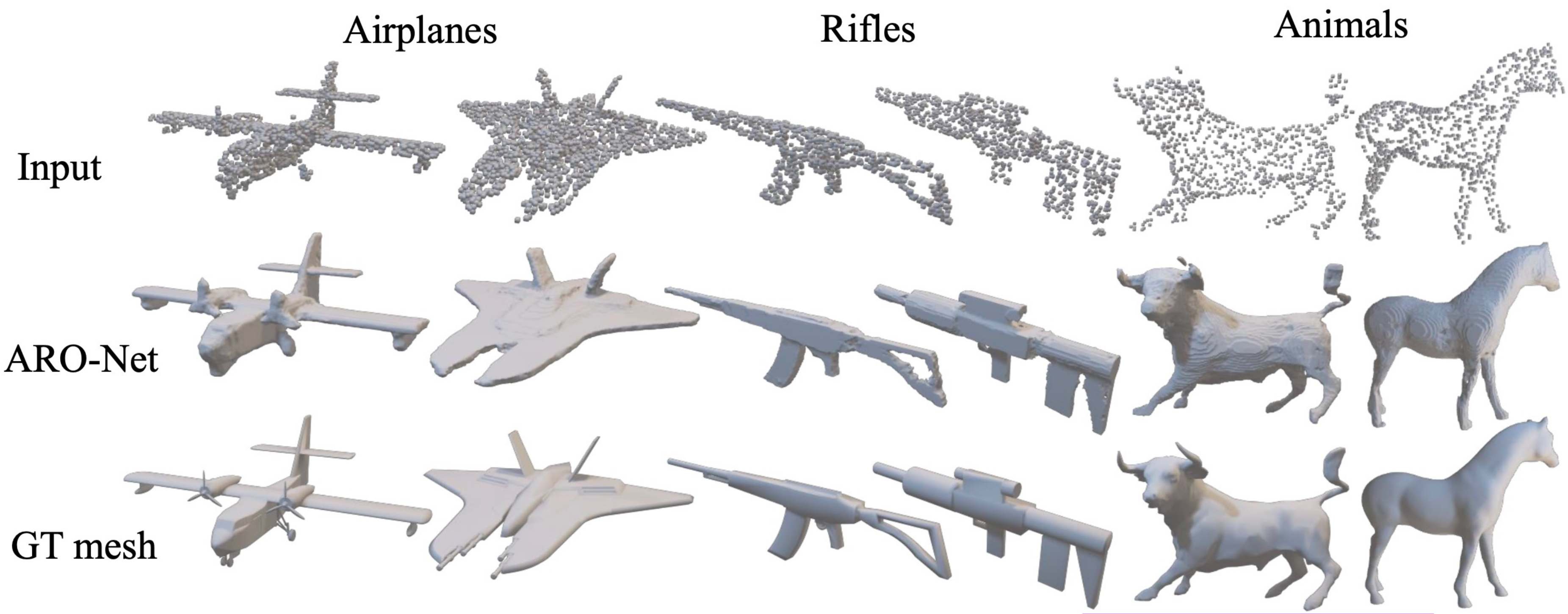
Key: encode features perceptually

- Encode point features via multi-view perception: “What does the shape look at from various view/anchor points towards the query point (*x*)?”



ARO = Anchored Radial Observations

Generality and quality of reconstruction

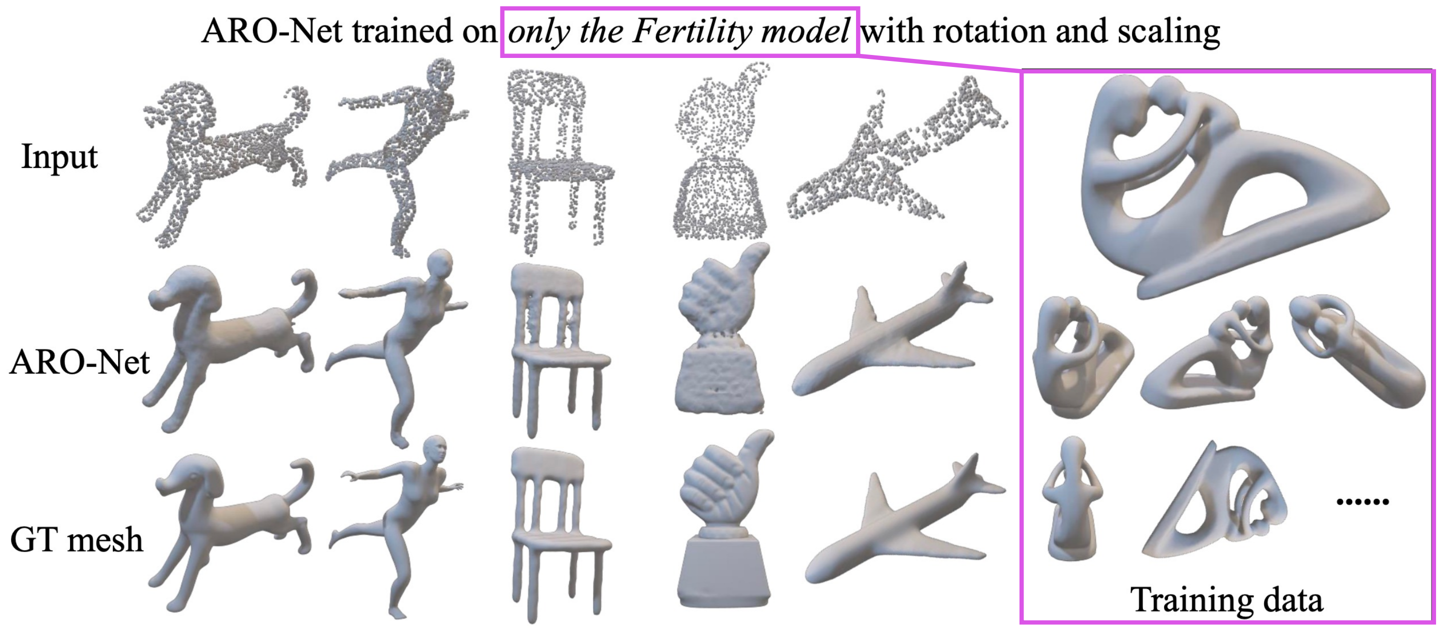


3D Reconstruction from sparse point clouds by ARO-Net *trained on 4K chairs*

Ill-posed task

An extreme example for generalizability

❖ Train with a single 3D model while attaining generalizability



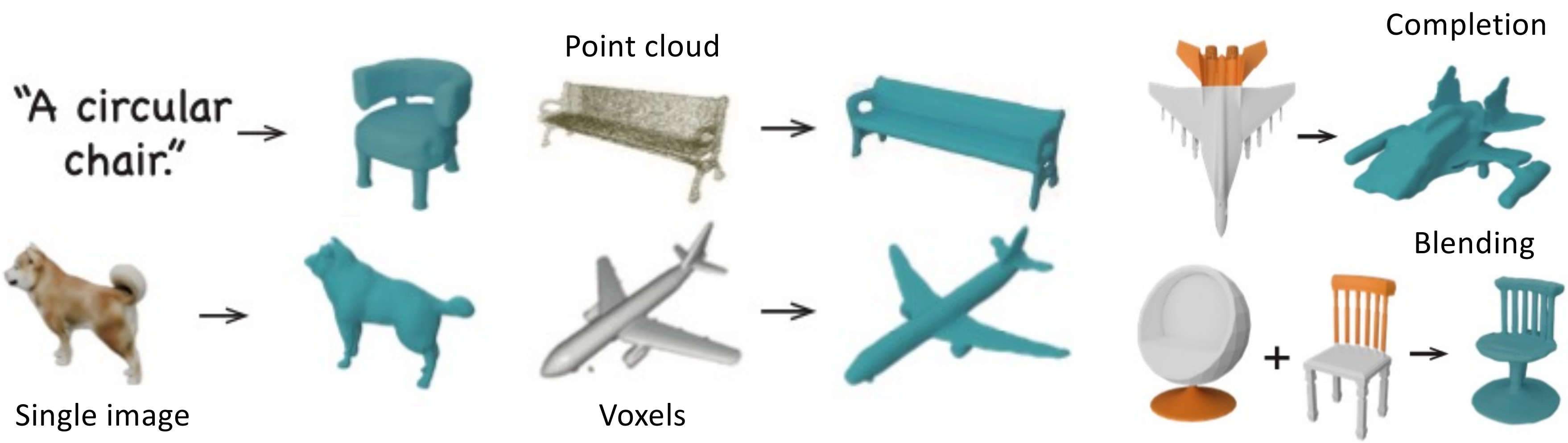
ARO-Net: neural 3D reconstruction from sparse point clouds [Wang et al. CVPR 2023]

Seek 3D reps ``biased” towards the task

- Solving ill-posed tasks (e.g., sparse reconstruction) require understanding
- A “perceptual” feature representation can understand/extrapolate better

Latest: multi-modal generation

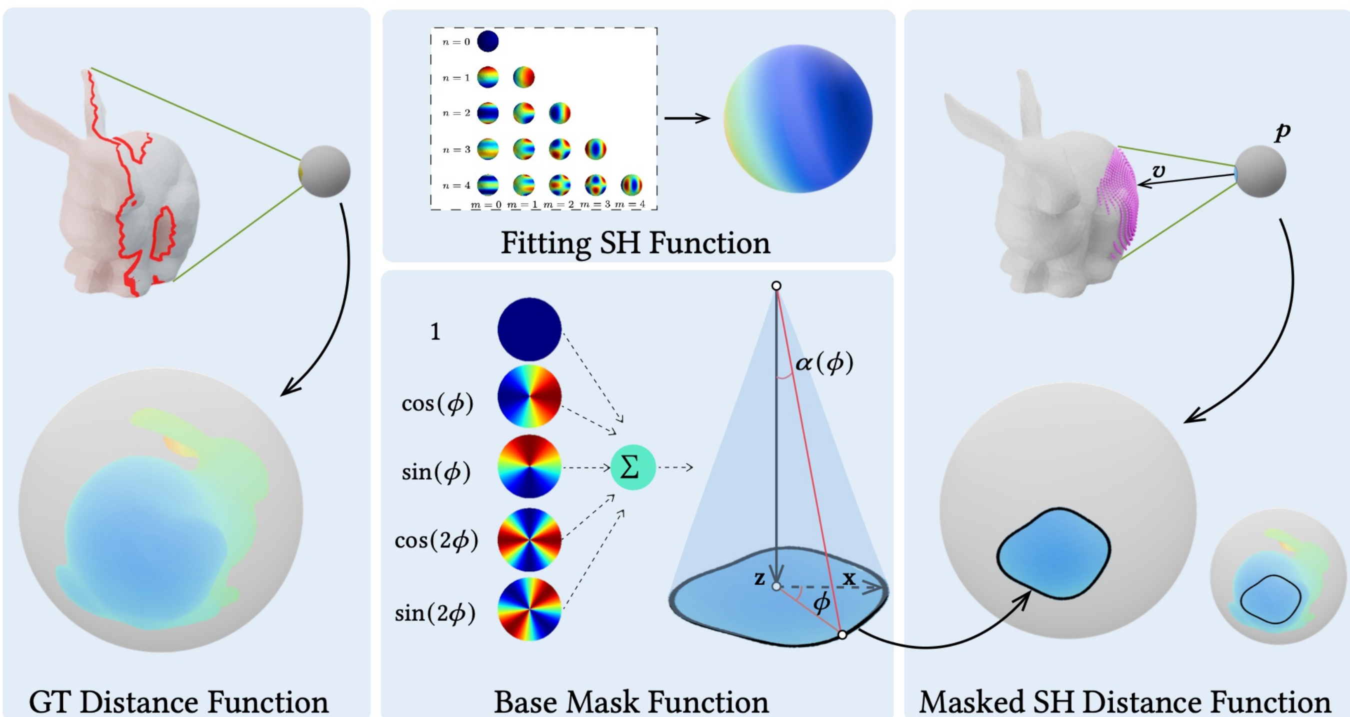
- Solving ill-posed tasks (e.g., sparse reconstruction) requires understanding
- A "perceptual" feature representation can understand/extrapolate better



An extension of ARO-Net: Masked Anchored SpHerical Distances (MASH) [Li et al. 2025]

Masked anchored spherical distances

❖ Parametric representation of MASH from a single anchor

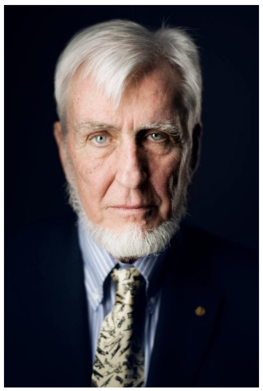


What more can we learn from the brain?

- Especially in the context of robotics and Embodied AI (EAI)
  - Many decisions are not made "on-the-fly", but on knowledge/memory
  - Our brain possesses an innate spatial awareness, e.g., for navigation
  - Our brain also possesses cognitive awareness, e.g., for action planning
  - Should AI agents form similar spatial and cognitive maps too?

How do humans do these in our brains?

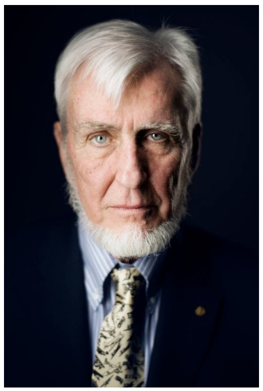
Internal GPS in human/animal brains



John O'Keefe

- ❖ Nobel Prize in 2014 for understanding neural processes in the mental representation of spatial environments to enable us to navigate
- ❖ Discovery of location-aware place cells in rats in 1971

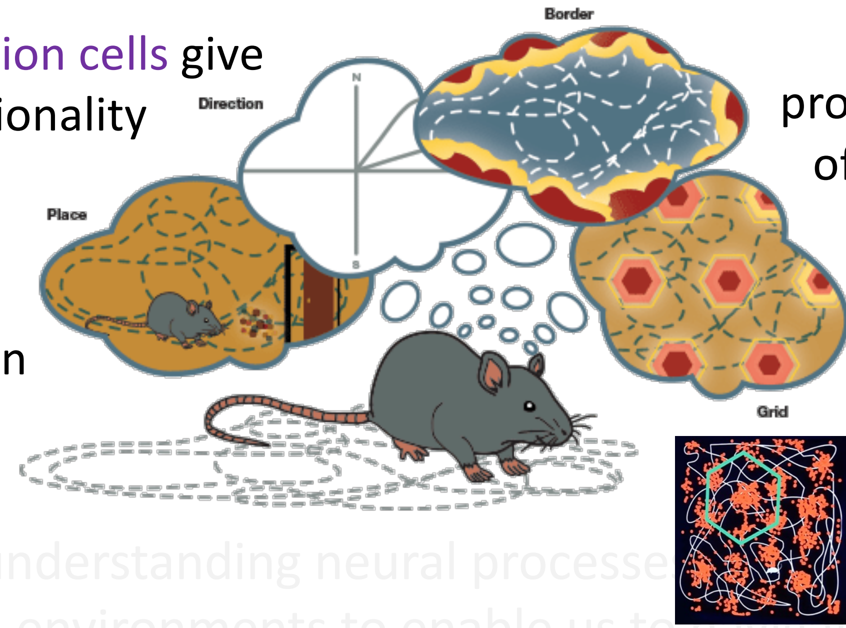
Internal GPS in human/animal brains



John O'Keefe

Head-direction cells give directionality

Place cells provide location



Border cells provide perimeters of a given space

Grid cells give sense of latitude and longitude

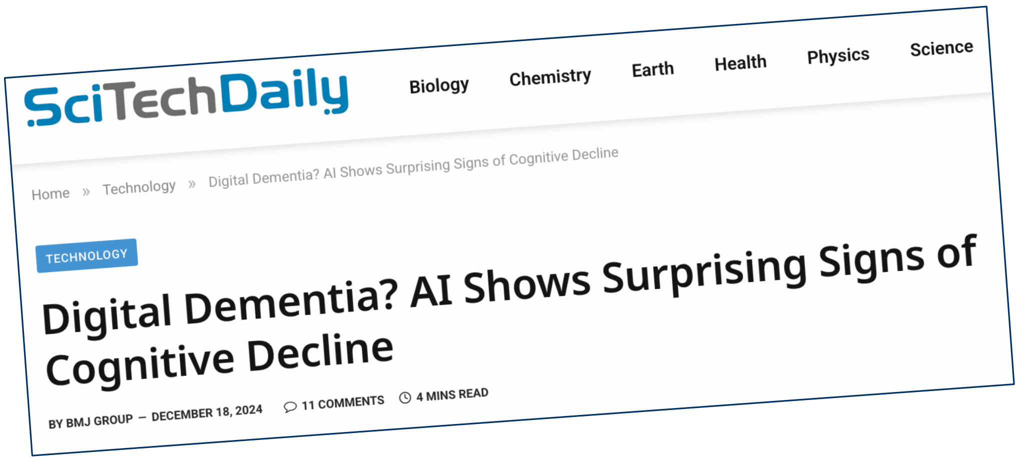
- ❖ Nobel Prize in 2014 for understanding neural processes … mental representation of spatial environments to enable us to navigate
- ❖ Discovery of location-aware place cells in rats in 1971
- ❖ Later works discovered other cells (e.g., speed cells) for the "internal GPS"

https://www.youtube.com/watch?v=216r36KCE1M

Unknowns and interesting facts

- How all the cells work together? Still on-going research …
- Dementia patients (Alzheimer's) lose these cell functions first

Missing ingredient?


SciTechDaily Biology Chemistry Earth Health Physics Science

Home » Technology » Digital Dementia? AI Shows Surprising Signs of Cognitive Decline

TECHNOLOGY

Digital Dementia? AI Shows Surprising Signs of Cognitive Decline

BY BMJ GROUP – DECEMBER 18, 2024 11 COMMENTS 4 MINS READ

- LLMs exhibit signs of similar decline as dementia patients
- Are LLMs missing these cell functions?
- Questions on representation will emerge

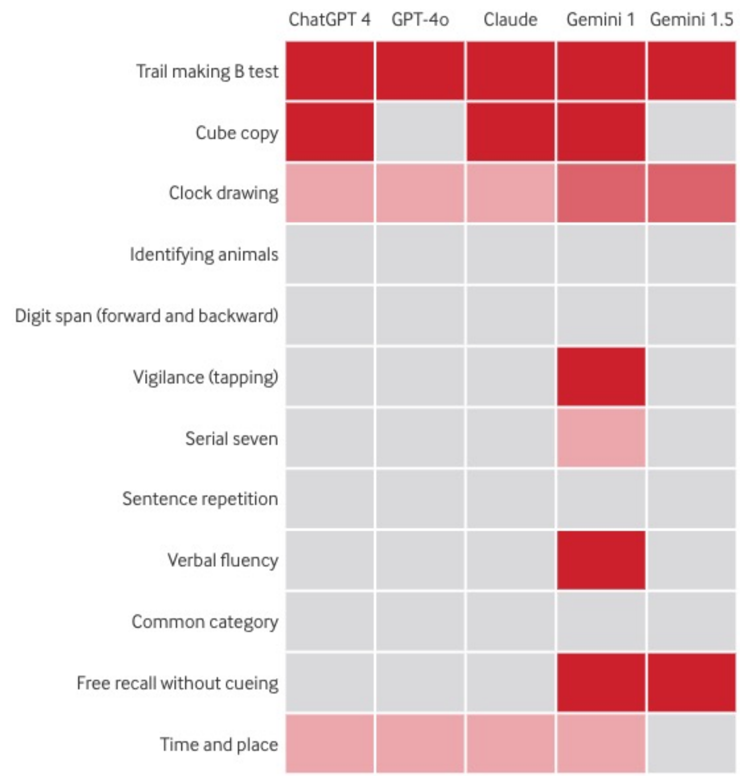
Age against the machine—susceptibility of large language models to cognitive impairment: cross sectional analysis

BMJ 2024 ; 387 doi: https://doi.org/10.1136/bmj-2024-081948 (Published 20 December 2024)
Cite this as: *BMJ* 2024;387:e081948

Cognitive assessment of AI models

How leading large language generative AI models respond to The Montreal Cognitive Assessment test

Darker red boxes show greater errors as a percentage of maximum scores. Hover boxes to show scores and click them to show details of responses

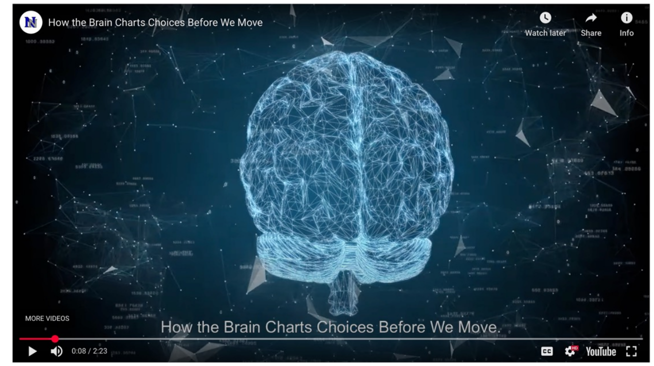


Article DOI: 10.1136/bmj-2024-081948

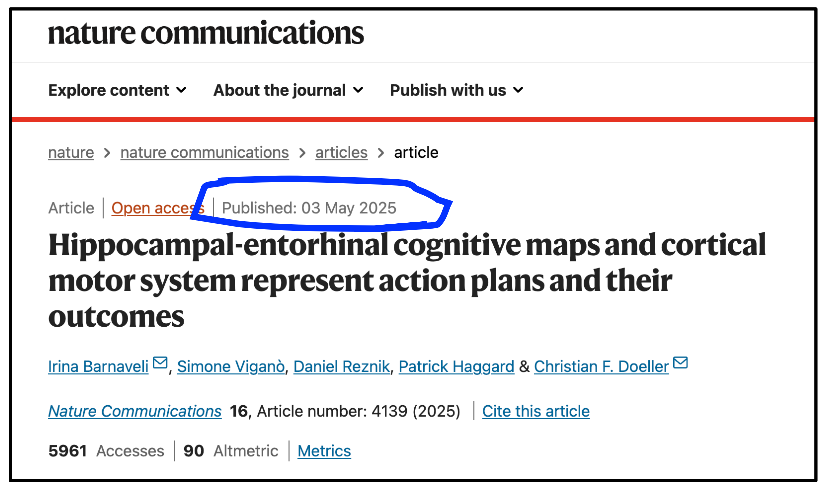
We navigate actions too!

- Our brain organizes potential actions and outcomes in a cognitive map, similar to the way we navigate spaces

- The closer two actions were on this cognitive map, the more participants perceive them as similar



https://www.youtube.com/watch?v=oNc2VU6gYcw&t=8s



nature communications

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature communications > articles > article

Article | Open access | Published: 03 May 2025

Hippocampal-entorhinal cognitive maps and cortical motor system represent action plans and their outcomes

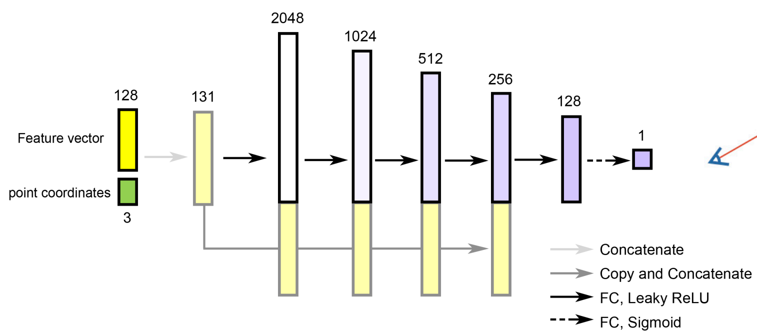
Irina Barnaveli, Simone Viganò, Daniel Reznik, Patrick Haggard & Christian F. Doeller

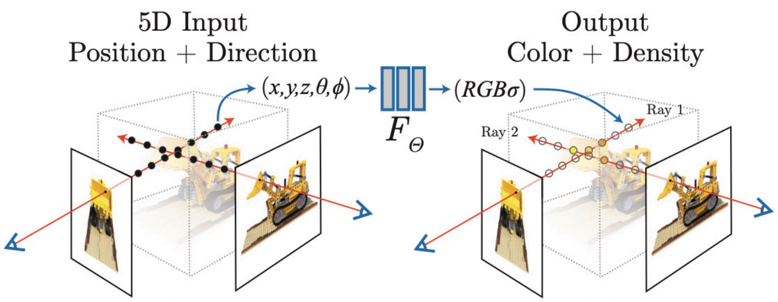
Nature Communications **16**, Article number: 4139 (2025) | Cite this article

5961 Accesses | **90** Altmetric | Metrics

Summary

- Many representation choices were computationally motivated:
  - Low-level (rather than structural) reps make differentiability easier
  - NeRF/IM-Net/OccNet/DeepSDF motivated by continuity of volume rep
  - 3DGS (also Instant NGP earlier) popularized due to rendering efficiency





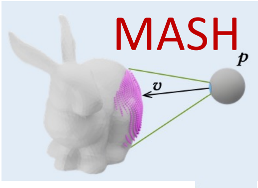
Summary

❖ Should consider inductive biases for 3D gen & rep learning

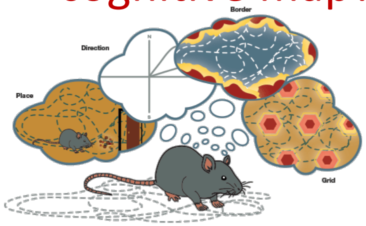
Human perception,
what can we learn?

Representations: perceptual

ARO



Robot's spatial &
cognitive map?



Summary

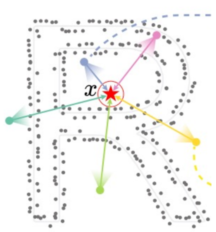
- Should consider inductive biases for 3D gen & rep learning

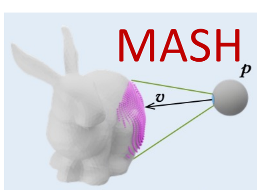
Human perception, what can we learn?

Function/motion /interaction

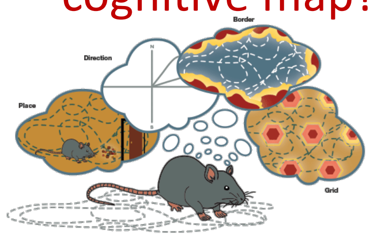
Representations: perceptual, interaction

ARO





Robot's spatial & cognitive map?



IBS/ICON

ATOP

Summary

❖ Should consider inductive biases for 3D gen & rep learning

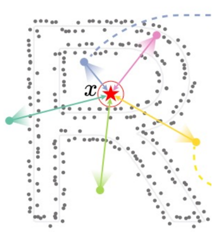
Human perception, what can we learn?

Function/motion /interaction

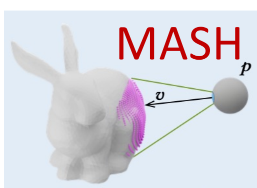
Geometry/space /physics

Representations: perceptual, interaction, structural

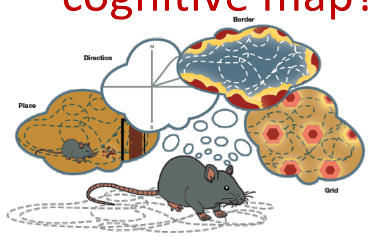
ARO



MASH

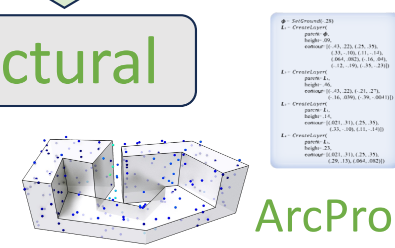


Robot's spatial & cognitive map?



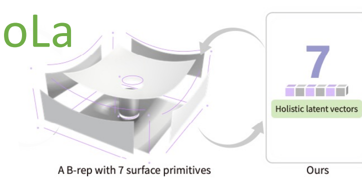
IBS/ICON

ATOP



ArcPro

HoLa



Summary

- 😀 Inductive biases improve generalizability and alleviate data scarcity
- ☹️ But the assumptions/priors can also be limiting at the same time

Summary

- 😀 Inductive biases improve generalizability and alleviate data scarcity
- ☹️ But the assumptions/priors can also be limiting at the same time
- 😀 Functional inductive biases covered are still implicit, e.g.,
  - ❖ Neural representations of object-object interactions
  - ❖ Learning structured representations: a necessity but not exactly the same
  - ❖ Motion priors from video foundational models: realization of functions
- ☹️ Differentiable functionality loss for 3D generation still elusive

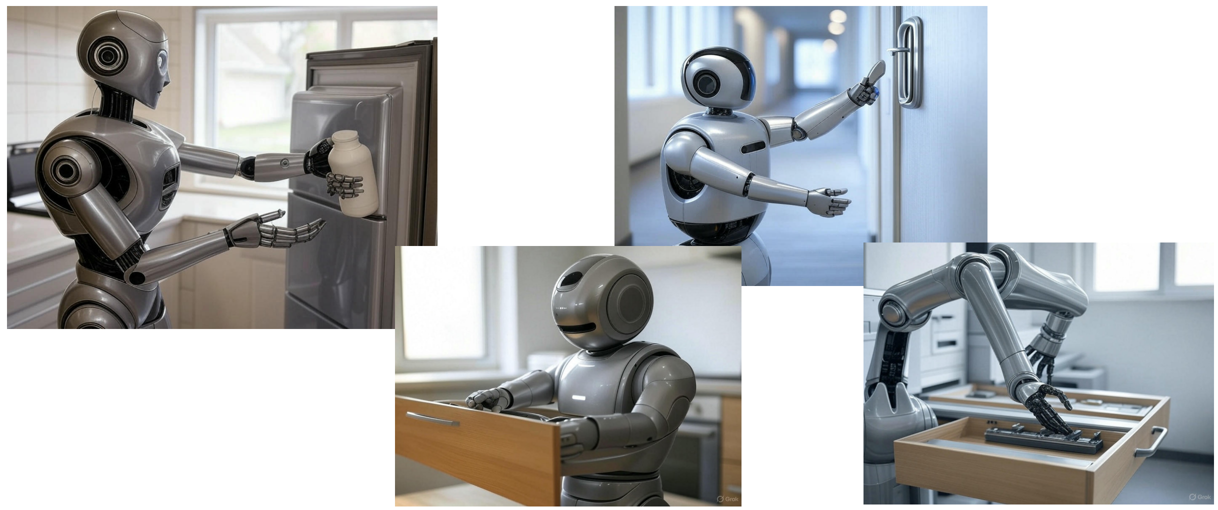
Summary

Robots opening everything and acquiring both exteriors and interiors

😀 Goal: learn structured, text-grounded, motion-enabled 3D representations

Summary

Robots opening everything and
acquiring both exteriors and interiors



😀 Goal: learn structured, text-grounded, motion-enabled 3D representations

😀 Build foundation models with spatial intelligence, encompassing 3D, text, & image, beyond Q&A and NTP, to do things in physical worlds

Papers covered

[Chen and Zhang 2019] (IM-Net) Zhiqin Chen and **HZ**, "Learning Implicit Field for Generative Shape Modeling," *CVPR 2019*.

[Hu et al. 2015] (ICON) Ruizhen Hu, Chenyang Zhu, Oliver van Kaick, Ligang Liu, Ariel Shamir, and **HZ**, "Interaction Context (ICON): Towards a Geometric Functionality Descriptor," *SIGGRAPH* 2015.

[Hu et al. 2016] (ICON2) Ruizhen Hu, Oliver van Kaick, Bojian Wu, Hui Huang, Ariel Shamir, and **HZ** "Learning How Objects Function via Co-Analysis of Interactions," *SIGGRAPH* 2016.

[Hu et al. 2017] (ICON3) Ruizhen Hu, Wenchao Li, Oliver van Kaick, Ariel Shamir, **HZ**, and Hui Huang, "Learning to Predict Part Mobility from a Single Static Snapshot," *SIGGRAPH Asia* 2017.

[Hu et al. 2018] (ICON4) Ruizhen Hu, Zihao Yan, Jingwen Zhang, Oliver van Kaick, Ligang Liu, Ariel Shamir, and **HZ**, "Predictive and Generative Neural Networks for Object Functionality," *SIGGRAPH* 2018.

[Huang et al. 2025] (ArcPro) Qirui Huang, Runze Zhang, Kangjun Liu, Minglun Gong, **HZ**, and Hui Huang, "ArcPro: Architectural Programs for Structured 3D Abstraction of Sparse Points," *CVPR* 2025 (Highlight).

[Huang et al. 2023] (NIFT) Zeyu Huang, Juzhan Xu, Sisi Dai, Kai Xu, **HZ**, Hui Huang, and Ruizhen Hu, "NIFT: Neural Interaction Field and Template for Object Manipulation," *ICRA 2023*.

[Li et al. 2025] (MASH) Changhao Li, Xin Yu, Xiaowei Zhou, Ariel Shamir, **HZ**, Ligang Liu, and Ruizhen Hu, "MASH: Masked Anchored SpHerical Distances for 3D Shape Representation and Generation," conditionally accepted to *SIGGRAPH* 2025.

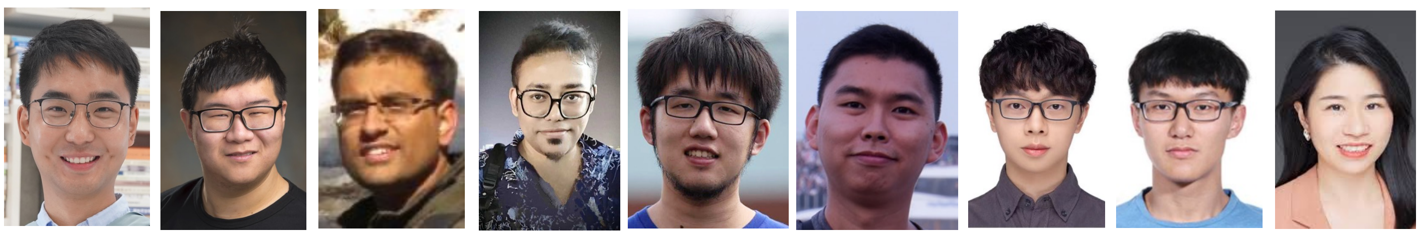
[Liu et al. 2025] (HoLa) Yilin Liu, Duoteng Xu, Xingyao Yu, Xiang Xu, Daniel Cohen-Or, **HZ**, and Hui Huang, "HoLa: B-Rep Generation using a Holistic Latent Representation," *SIGGRAPH* 2025.

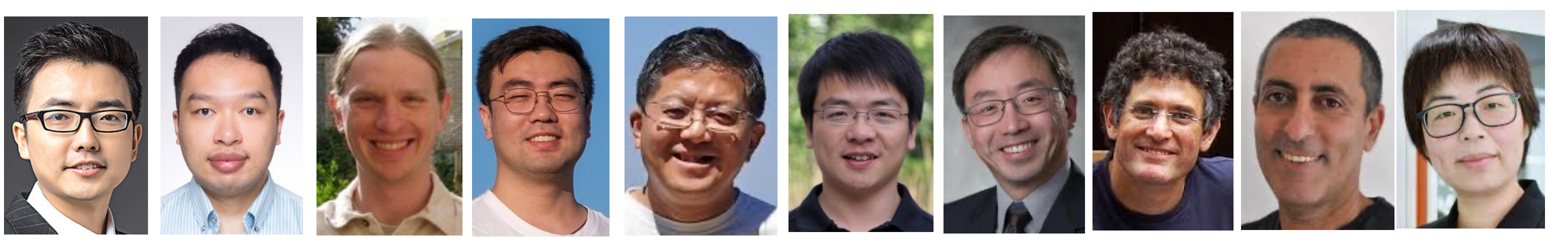
[Vora et al. 2025] (ATOP) Aditya Vora, Sauradip Nag, Kai Wang, and **HZ** "ATOP (Articular That Object Par): 3D Part Articulation from Text and Motion Personalization," *arXiv* 2025.

[Wang et al. 2023] (ARO-Net) Yizhi Wang, Zeyu Huang, Ariel Shamir, Hui Huang, **HZ**, and Ruizhen Hu, "ARO-Net: Learning Implicit Fields from Anchored Radial Observations," *CVPR* 2023.

Thank you and acknowledgment

❖ Students, postdocs, and collaborators on covered works





Amazon, Autodesk, Carleton University, University of Guelph, Reichman University, NUDT, SFU, Shenzhen University, Tel-Aviv University, Zhejiang University, and University of Science and Technology