

Geometric Deep Learning for Virtual Humans

Intro & Geometric DL

Riccardo Marin



20th November 2025

About me

3D Matching/Shape Analysis



Verona

PhD Computer Science



Paris

École Polytechnique
(Visiting)



Rome

Sapienza



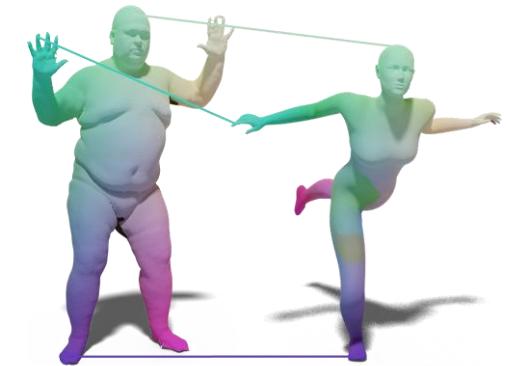
Tuebingen

AI Center

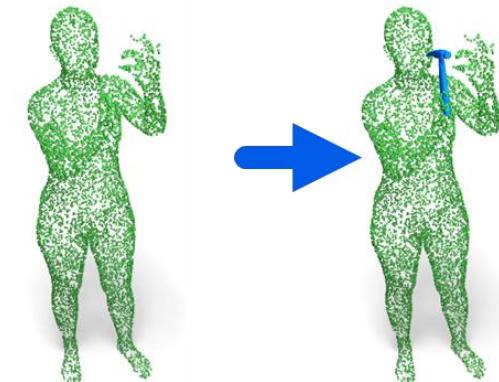


Munich

TUM/MCML



Virtual Humans





One of six national AI competence centers in Germany
that is permanently funded since July 2022



 ≈ 80
Principal Investigators

 460+
Junior Members

 2210+
Publications

Opportunities ▾

News

For Junior Members

For Research X-Change

For PhD Applicants

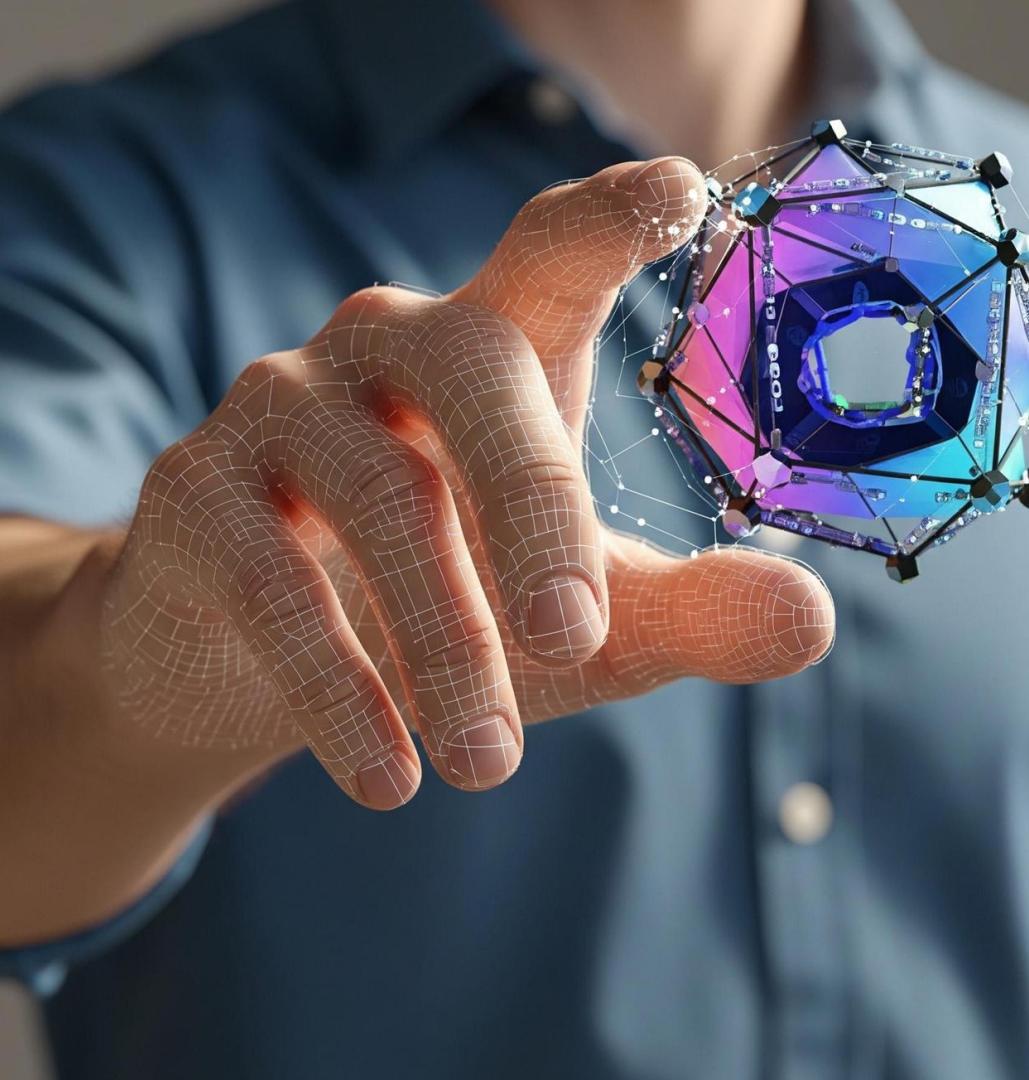
For Schools

For Potential Partners

For ML Consulting Services

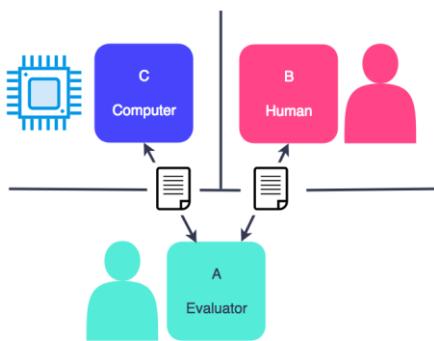
For Job Candidates

Feel free to reach
out!

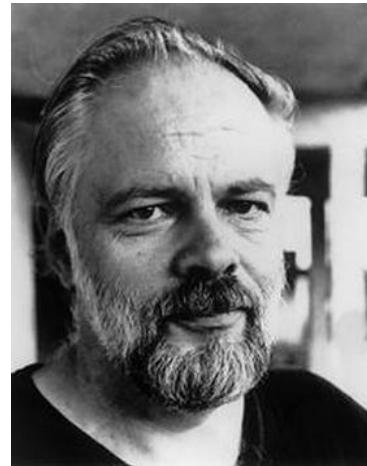


Motivation

Alan Turing



Philip K. Dick



Steve Wozniak



2 Your little boy shows you his butterfly collection, plus the killing jar. What do you say?

"Oh, lovely!"

"That's nice, but why don't you keep the killing jar for yourself?"

Nothing. I take my boy to the doctor

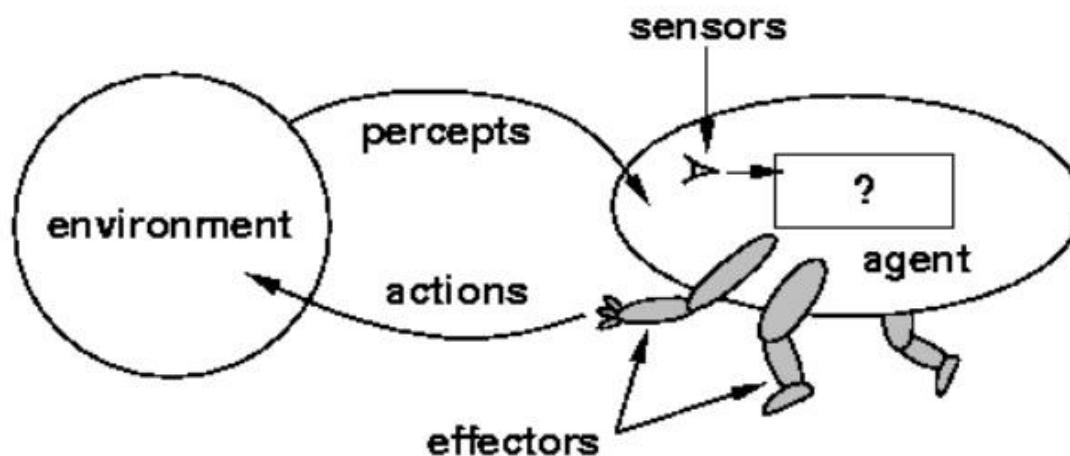


Artificial intelligence

From Wikipedia, the free encyclopedia

"AI" redirects here. For other uses, see [AI \(disambiguation\)](#) and [Artificial intelligence \(disambiguation\)](#).

Artificial intelligence (AI) is [intelligence](#) demonstrated by [machines](#), as opposed to [natural intelligence](#) displayed by [animals](#) including [humans](#). Leading AI textbooks define the field as the study of "intelligent agents": any system that perceives its environment and takes actions that maximize its chance of achieving its goals.^[a] Some popular accounts use the term "artificial intelligence" to describe machines that mimic "cognitive" functions that humans associate with the [human mind](#), such as "learning" and "problem solving", however, this definition is rejected by major AI researchers.^[b]



Sight

Touch

Hearing Smell

1250 MB/sec

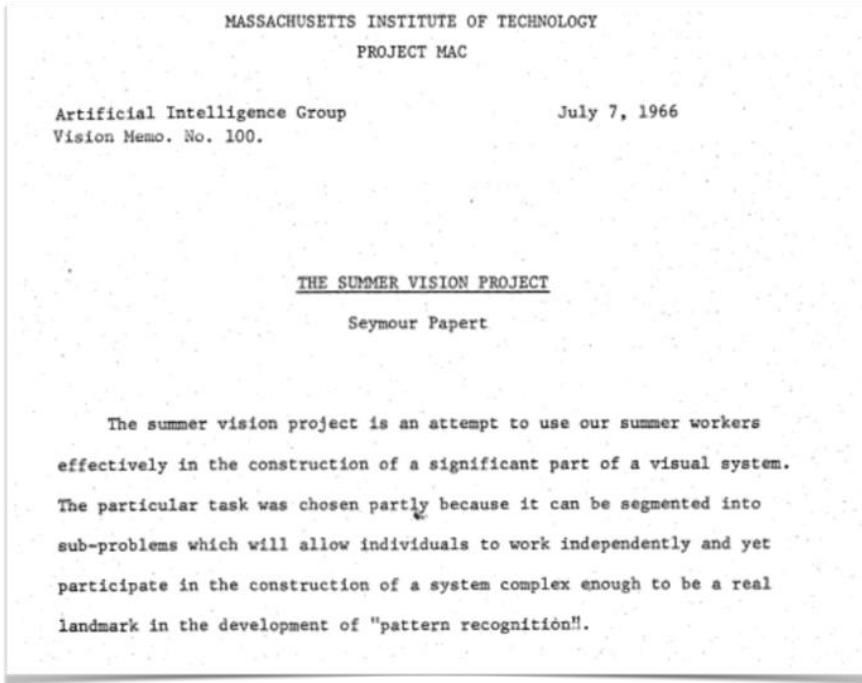
125 MB/sec

12.5 MB/sec

Taste

The start of research in CV

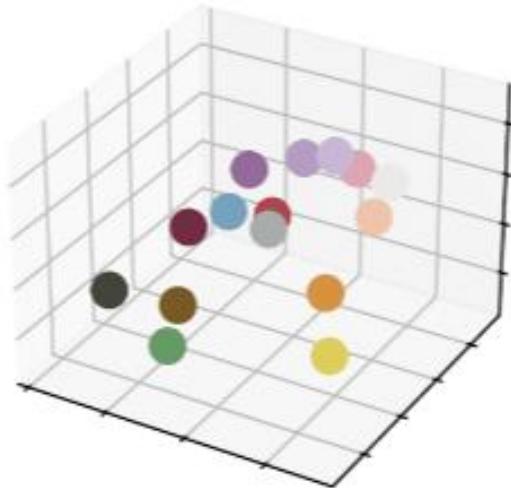
○ Project MAC (est. 1963) – CSAIL today



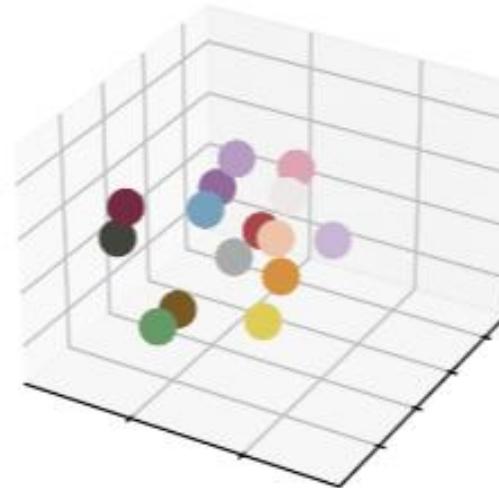
About LLM perception – Yes...

Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color,
Abdou et al., 2021

CIELAB



BERT, controlled context



About LLM perception – ... But

Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces

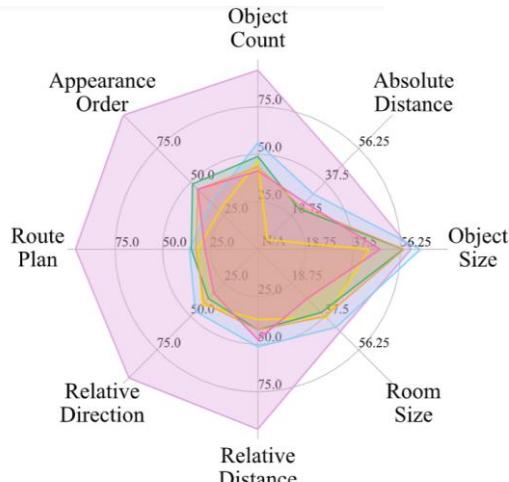
Jihan Yang^{1*} Shusheng Yang^{1*} Anjali W. Gupta^{1*} Rilyn Han^{2*} Li Fei-Fei³ Saining Xie¹

¹New York University ²Yale University ³Stanford University

 Project Page

 Evaluation Code

 VSI-Bench



8. Discussion and Future Work

We study how models see, remember, and recall spaces by building VSI-Bench and investigating the performance and behavior of MLLMs on it. Our analysis of how MLLMs think in space linguistically and visually identifies existing strengths (*e.g.*, prominent perceptual, temporal, and linguistic abilities) and bottlenecks for visual-spatial intelligence (*e.g.*, egocentric-allocentric transformation and relational reasoning). While prevailing linguistic prompting

About GenAI – Yes...



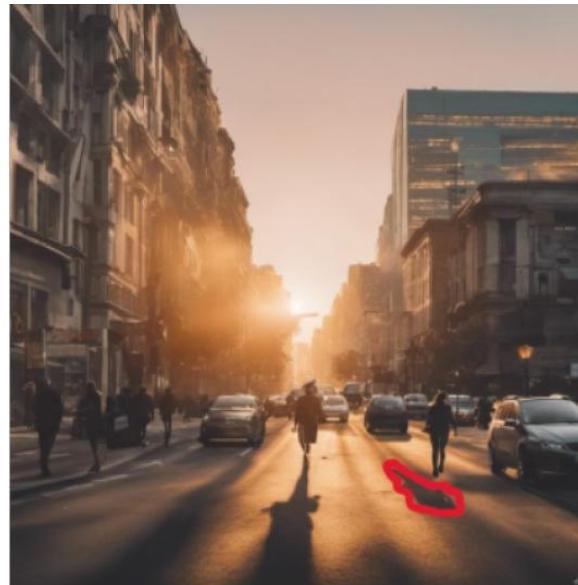
About GenAI – ...But

Shadows Don't Lie and Lines Can't Bend!

Generative Models don't know Projective Geometry...for now



Generated Image



Shadow Errors



Vanishing Point Errors

Deal with data bottleneck

Ilya Sustkever, NeurIPS 2024



Vs.

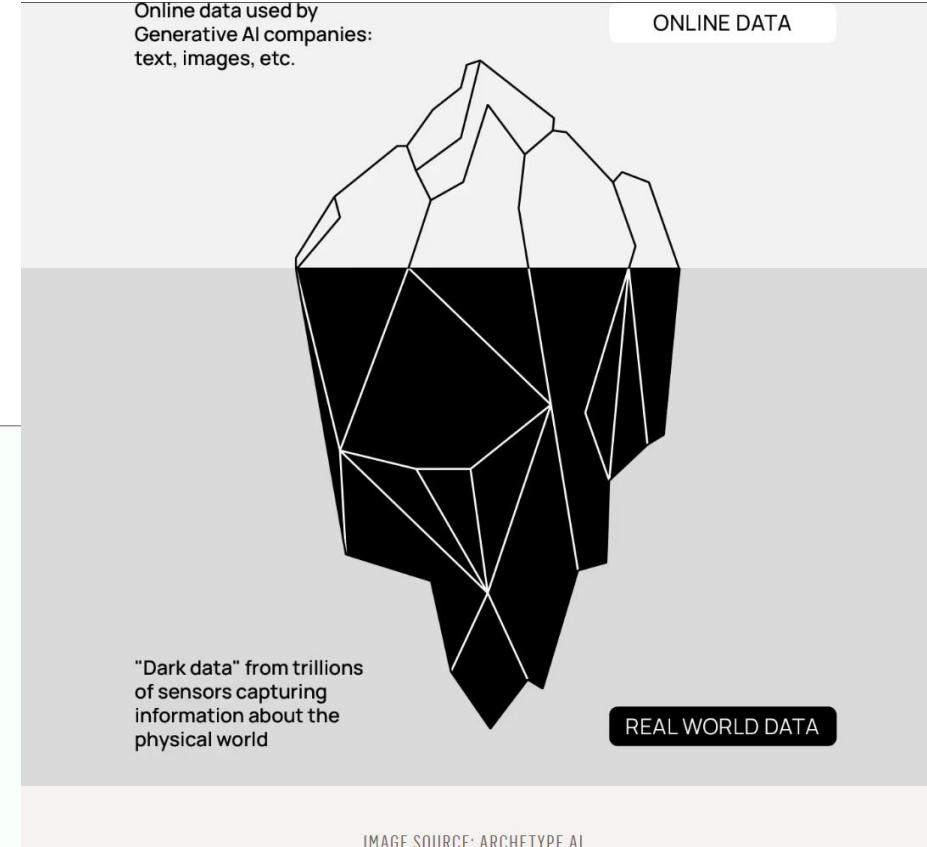
Pre-training as we know it will end

Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

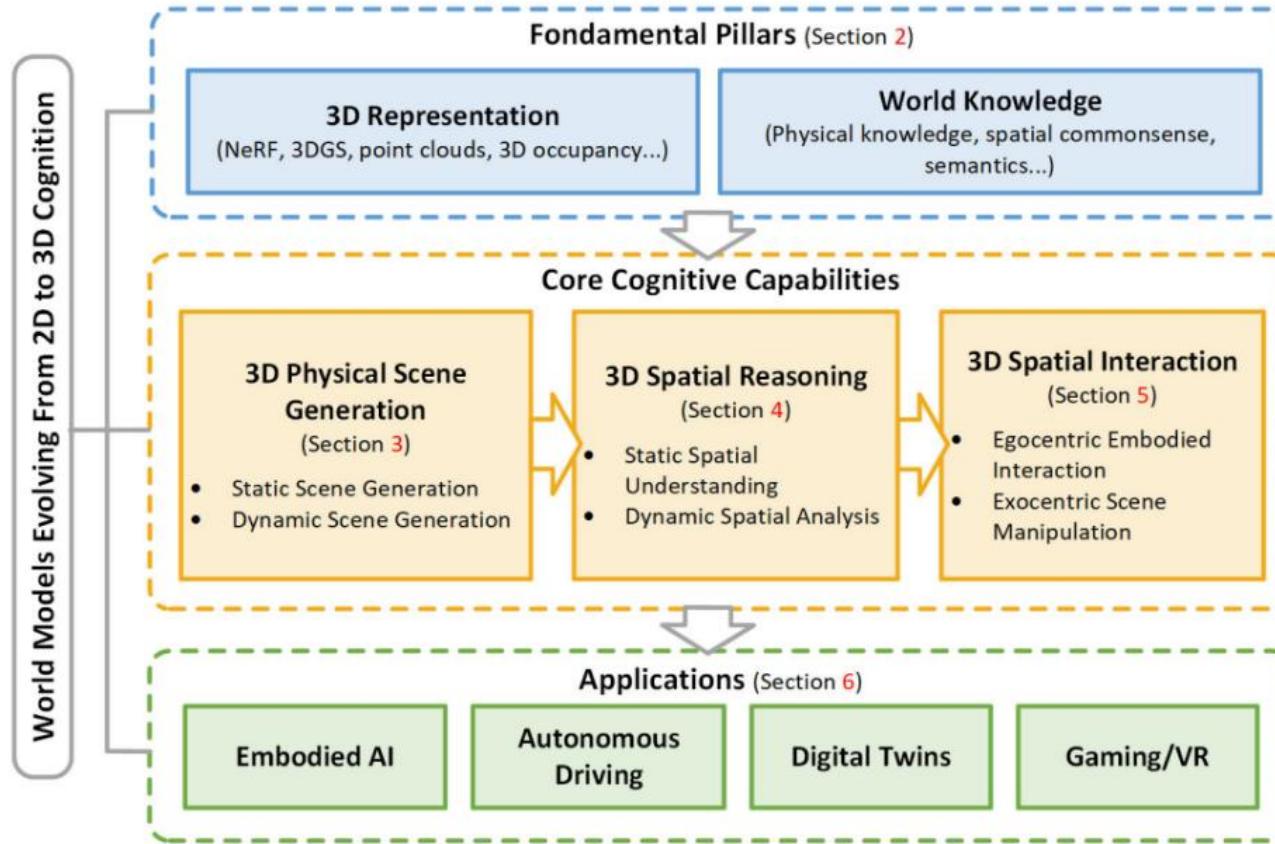
Data is not growing:

- We have but one internet
- **The fossil fuel of AI**



SpatialAI

From 2D to 3D Cognition: A Brief Survey of General World Models, Xie et al., 2025



From Words to Worlds: Spatial Intelligence is AI's Next Frontier



FEI-FEI LI

NOV 10, 2025

Tremendous progress has indeed been made in the past few years. Multimodal LLMs (MLLMs), trained with voluminous multimedia data in addition to textual data, have introduced some basics of spatial awareness, and today's AI can analyze pictures, answer questions about them, and generate hyperrealistic images and short videos. And through breakthroughs in sensors and haptics, our most advanced robots can begin to manipulate objects and tools in highly constrained environments.

Yet the candid truth is that AI's spatial capabilities remain far from human level. And the limits reveal themselves quickly. State-of-the-art MLLM models rarely perform better than chance on estimating distance, orientation, and size—or “mentally” rotating objects by regenerating them from new angles. They can't navigate mazes, recognize shortcuts, or predict basic physics. AI-generated videos—nascent and yes, very cool—often lose coherence after a few seconds.

Almost a half billion years after nature unleashed the first glimmers of spatial intelligence in the ancestral animals, we're lucky enough to find ourselves among the generation of technologists who may soon endow machines with the same capability—and privileged enough to harness those capabilities for the benefits of people everywhere. Our dreams of truly intelligent machines will not be complete without spatial intelligence.

Why care about 3D for generative media?

-  **Realism:** "The only way to make realistic images or videos of the world is to reconstruct or generate a 3D environment and render it"
-  **Robotics:** "The only way that a robot can navigate a 3D environment is to reconstruct that environment in 3D"
-  **Control:** "3D is the best/only interface for humans to control image generation"
-  **Speed:** "The way to render things cheaply is to generate them in 3D once and then render them repeatedly."
-  **Humanity:** "Humans live in 3D, so the AI should too."

But geometrical data has many (immediate) applications

Entertainment



Architecture



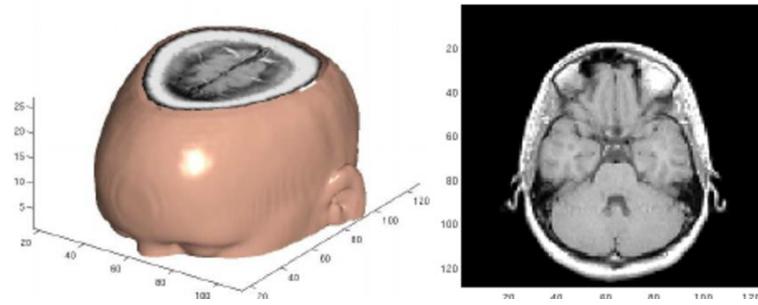
Fashion



Digital Fabrication



Medical Imaging

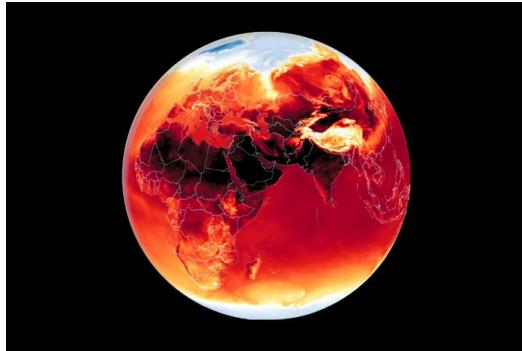


Archeology

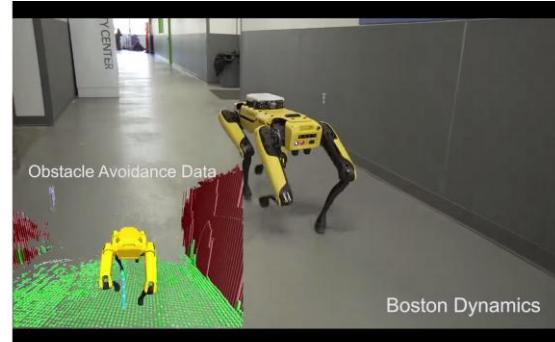


And more...

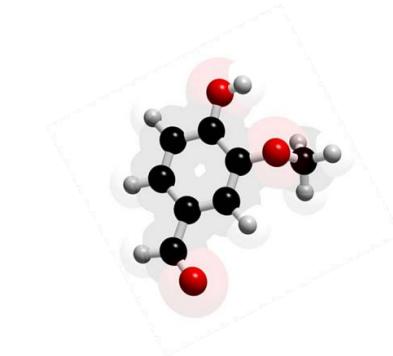
Geology/Climate



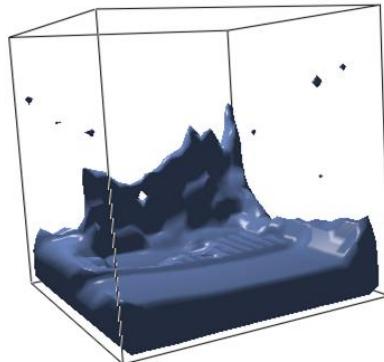
Robotics



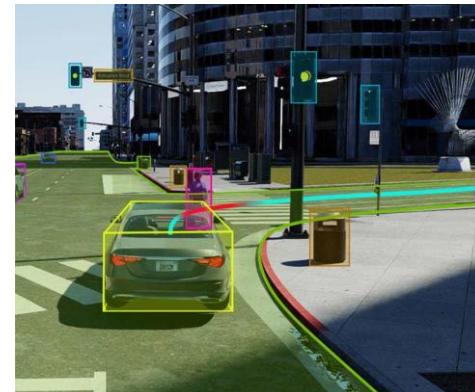
Biology



Physics/Simulations



Autonomous driving



Space Exploration



...

3D data are scaling

Objaverse-XL

A Universe of 10M+ 3D Objects

arXiv

Google Colab

Github

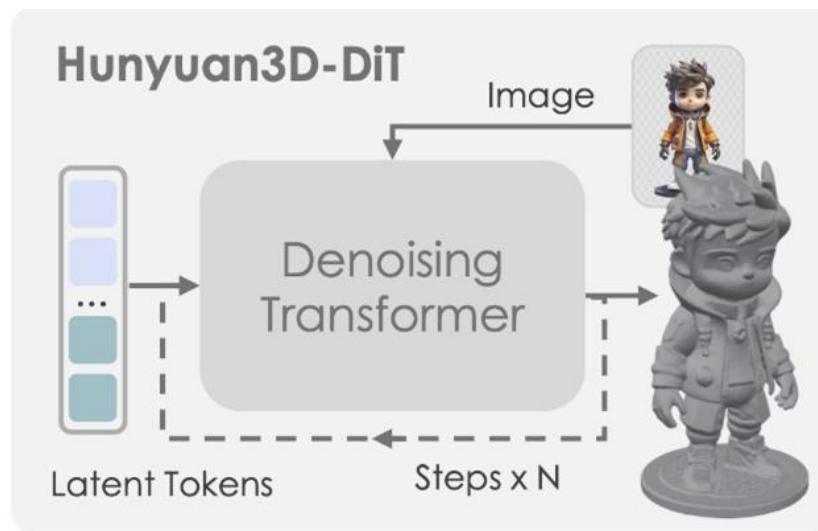
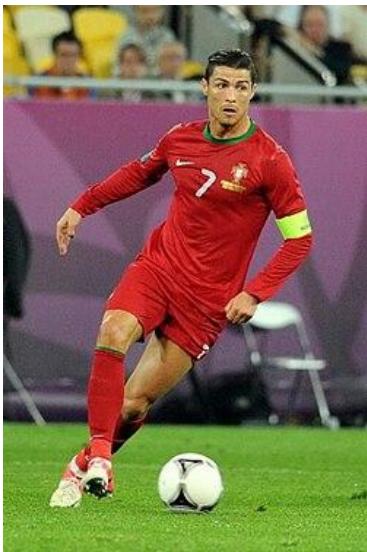
Hugging Face



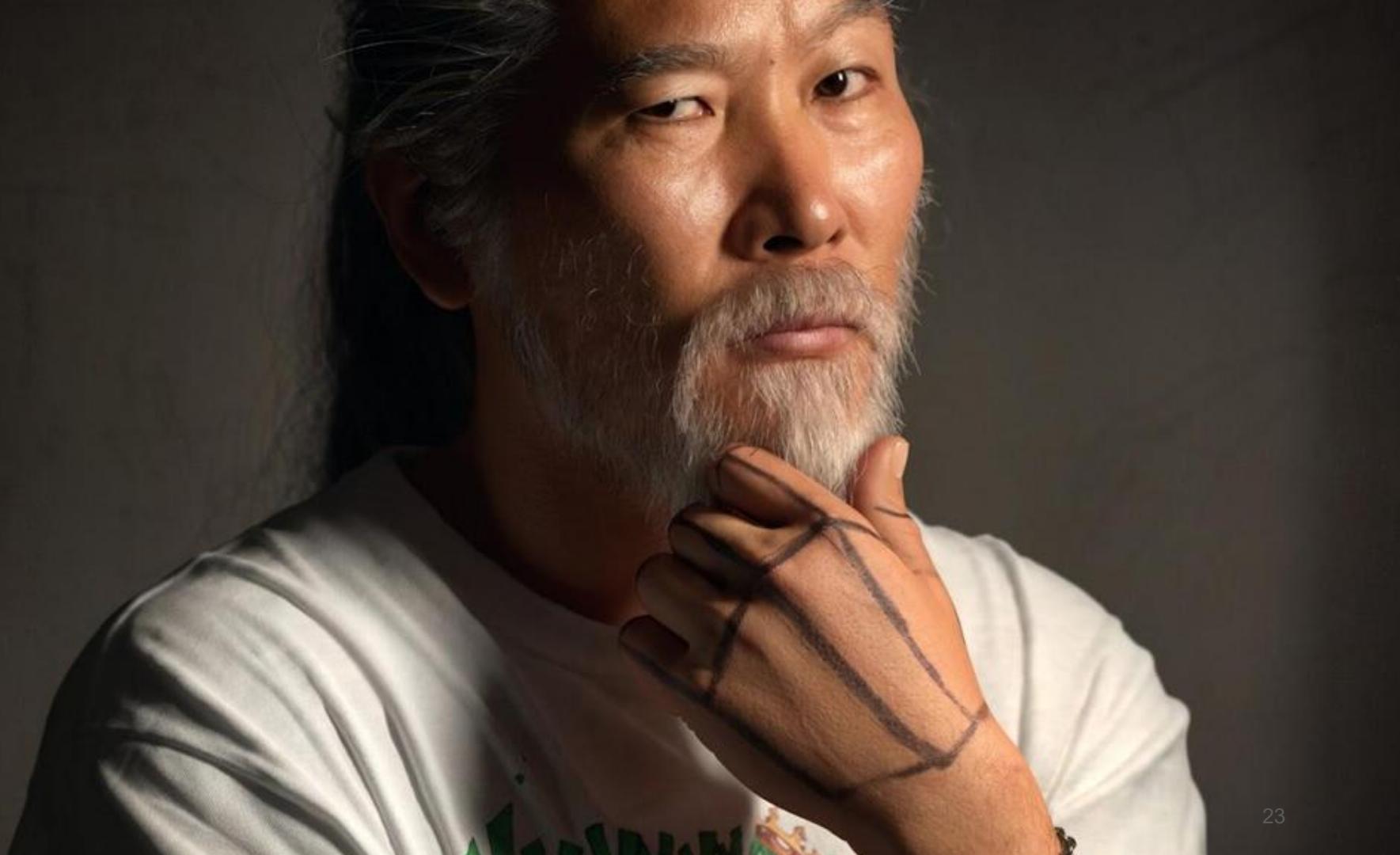
How to handle them?

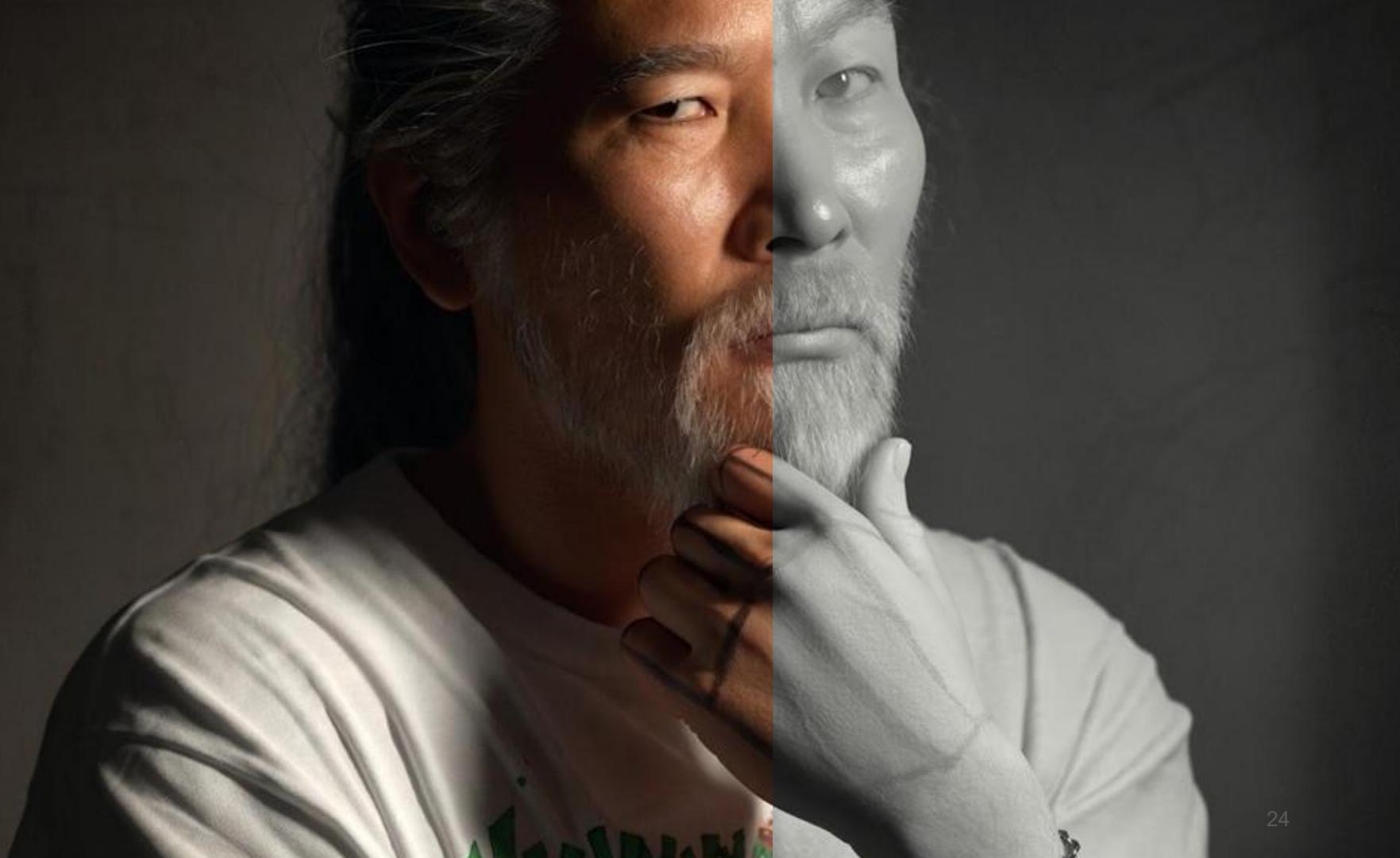
3D/4D Generative AI as a new data source

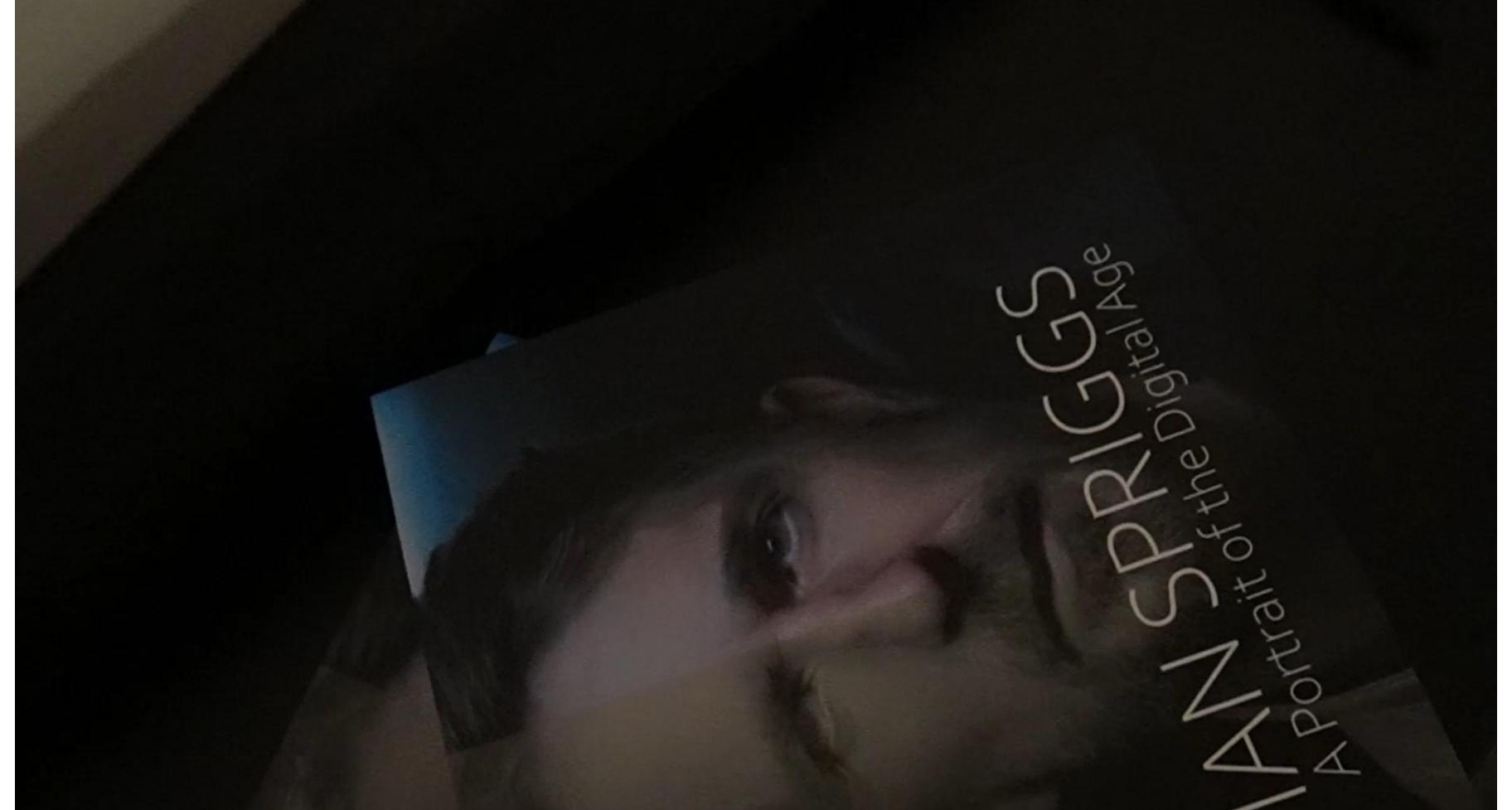
<https://huggingface.co/spaces/tencent/Hunyuan3D-2>



Disparate sources of data call for tools to relate them!



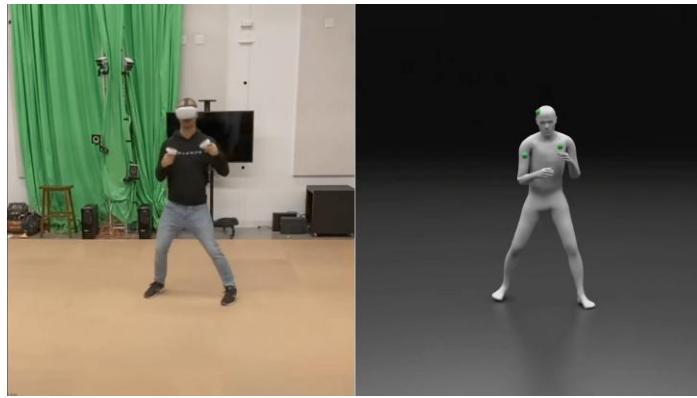




IAN SPRIGGS

A Portrait of the Digital Age

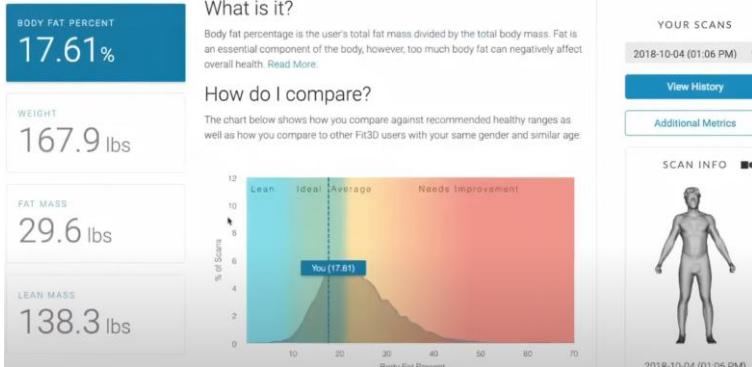
Virtual Humans as enablers of Spatial AI



VR/AR



Marketing
(Virtual Influencers)



Health, Medics, Training



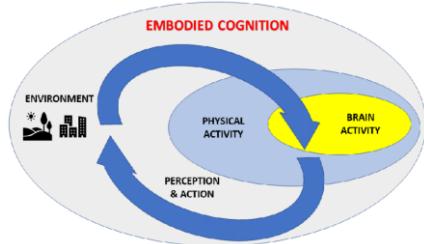
Assistance and Therapy 26

Does intelligence need a body?



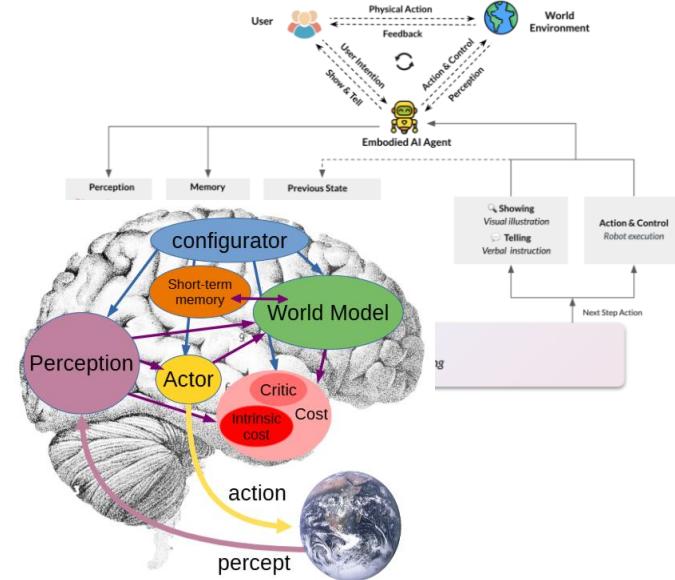
Descartes: No

<https://PMC.ncbi.nlm.nih.gov/articles/PMC3512413/>



Psychologists:
embodied cognition
(e.g., how movements relate to language and memory)

<https://pubmed.ncbi.nlm.nih.gov/20739194/>



**CV\AI Researchers
(LeCunn, Malik):**
Embodied AI

<https://openreview.net/pdf?id=BZ5a1r-kVsF>
<https://arxiv.org/pdf/2506.22355>

Industry and applications



People AI at Meta

...

We are hiring an engineering manager in Zurich, to help us shape the future of human body perception technology for AR and VR. If you are interested feel free to reach out.

#hiring #computervision #machinelearning #augmentedreality #virtualreality
#metaverse

PhD Internships



Google

Looking for Research Scientists in Visual Computing and 3D Human Modeling!

Are you keen on advancing the state-of-the-art research on Human Modeling and at the same time improving the lives of billions of users? Are you passionate about Augmented and Mixed Reality, Visual Computing or Machine Learning? Then our team might be the perfect fit for you!



meshcapade

CV/ML Engineer for 3D Digital Humans (x/f/m) Remote

Permanent employee, Full-time · Remote

Virtual Humans

Research Groups

MPI-IS Tuebingen Director

MAX-PLANCK-INSTITUT
FÜR INTELLIGENTE SYSTEME



 UNIVERSITY
OF AMSTERDAM



ETH zürich

Teaching

Seminar on Digital Humans

ETH zürich
Course Catalogue

TUM

MSc Course

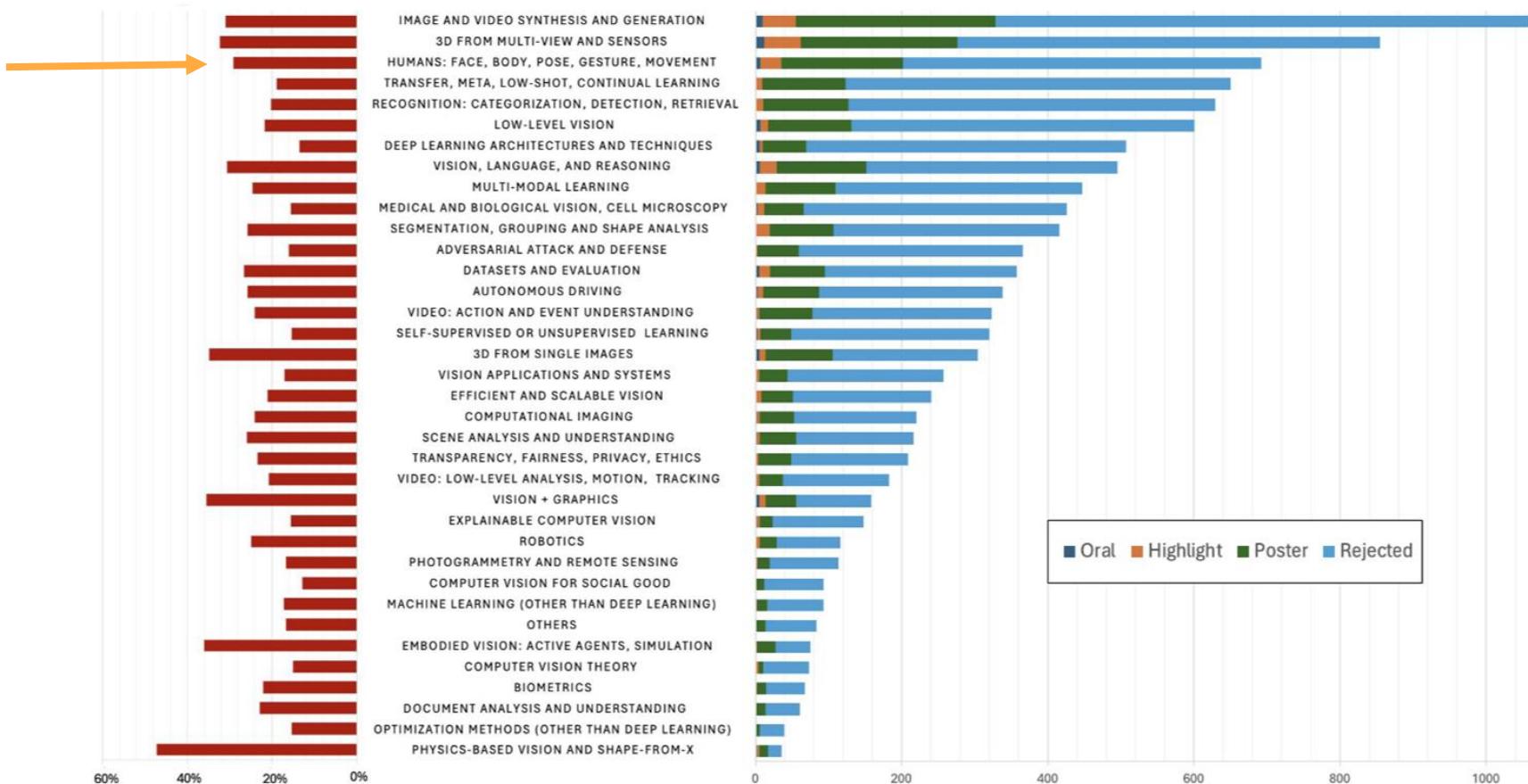


Tutorials, Workshops

CVPR
VIRTUAL JUNE 19-25

ECCV
TEL AVIV 2022

Submissions and decisions, by primary topic



Structure

Today

- Challenges of Geometrical Data
- Geometric deep learning
- Networks for point sets
 - Pointnet, Pointnet++
 - Saliency
- Networks for representing geometry
 - Neural Fields

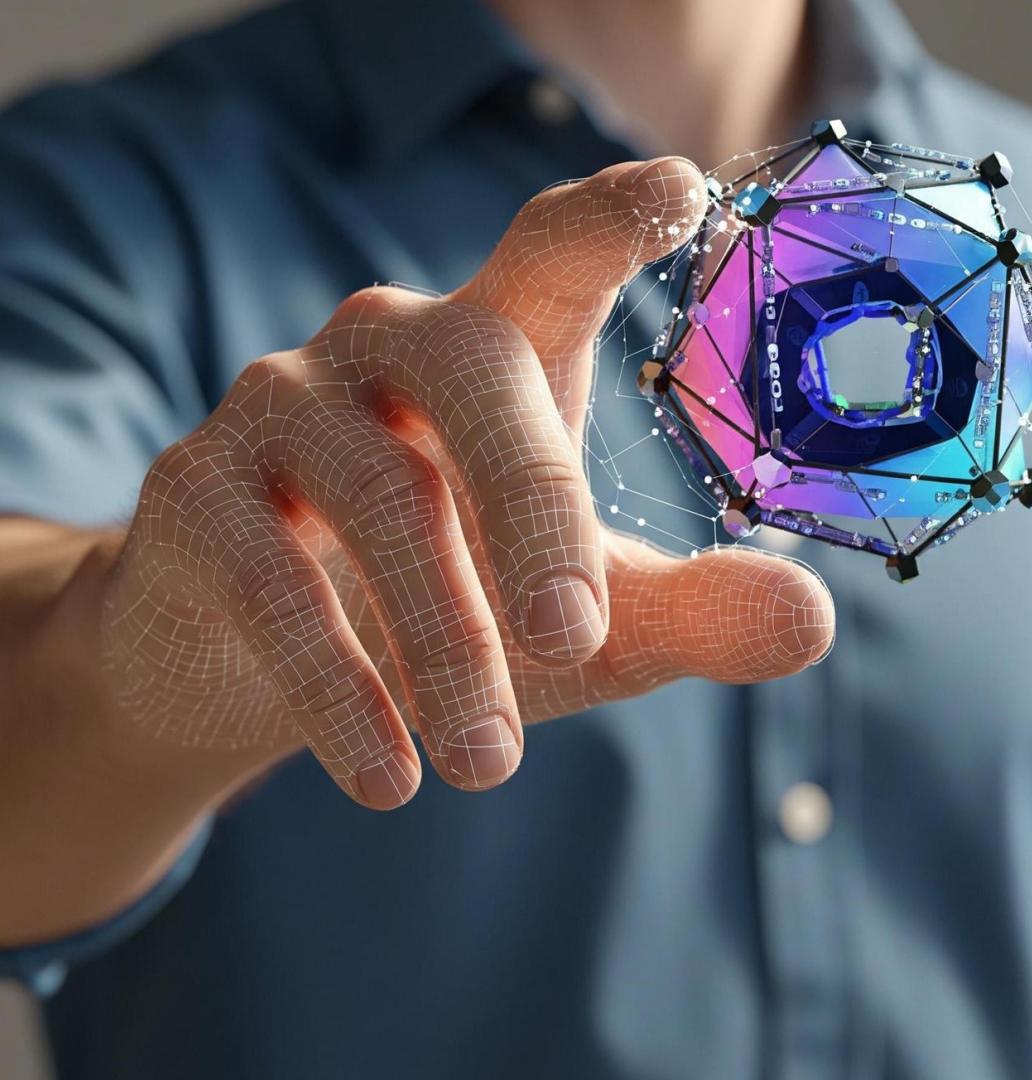
+ hands-on code Lab

Tomorrow

- Virtual Humans
- How to represent them digitally
 - SMPL
 - Extensions
- Research Topics
 - Human-Object Interactions

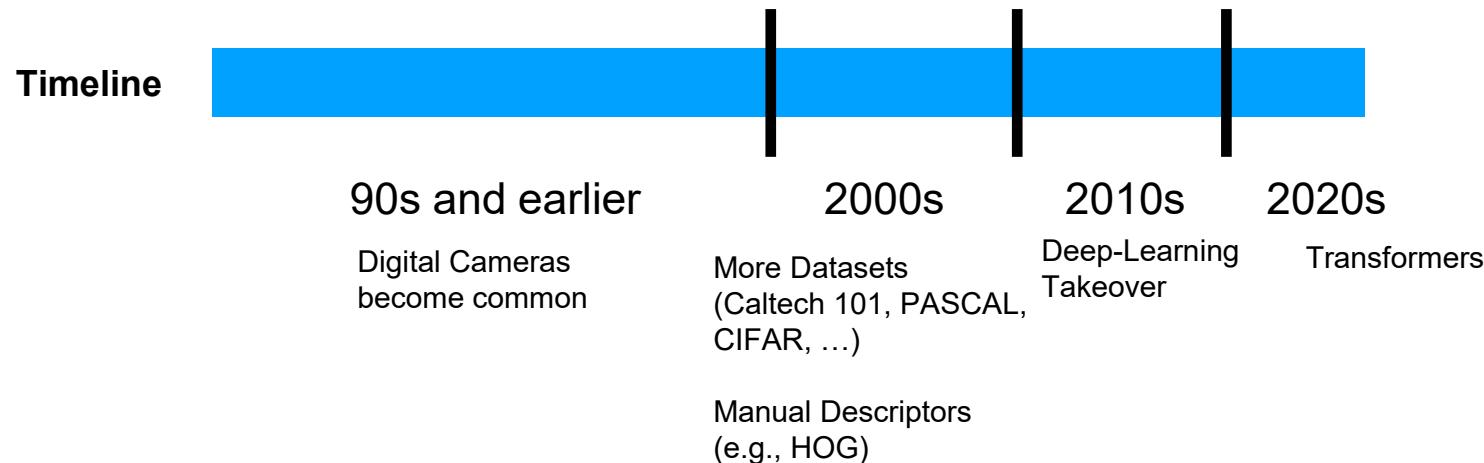
+ hands-on code Lab

<https://github.com/riccardomarin/GeoHumanUNIVR/>



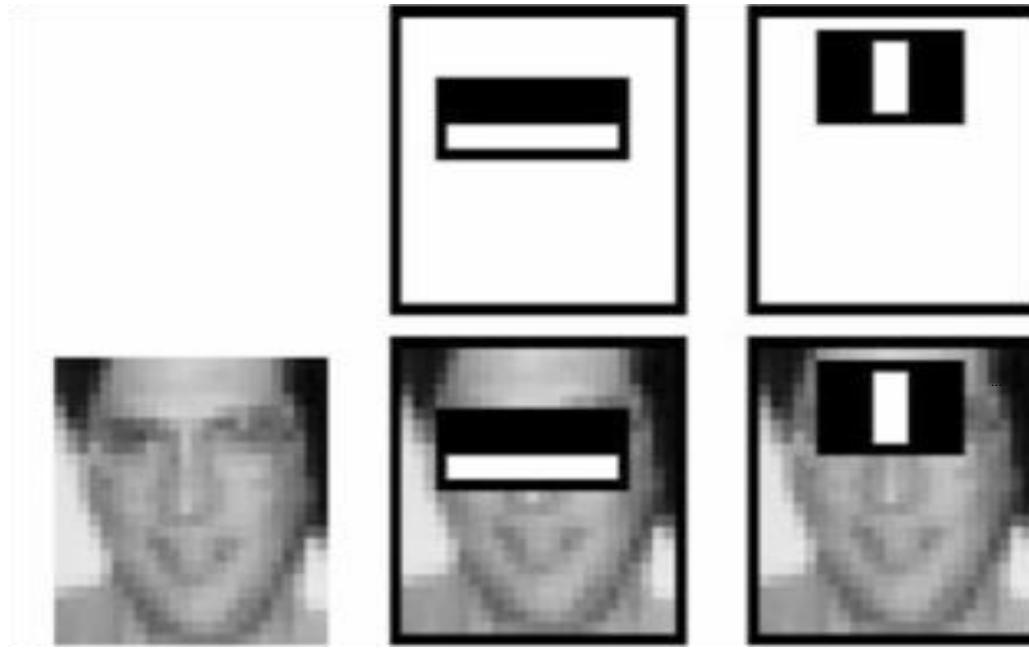
Processing Geometrical Data

Progress: Now and then



Pre-Learning era: Handcrafted features

Haar-like features



$$h_1(\cdot)$$

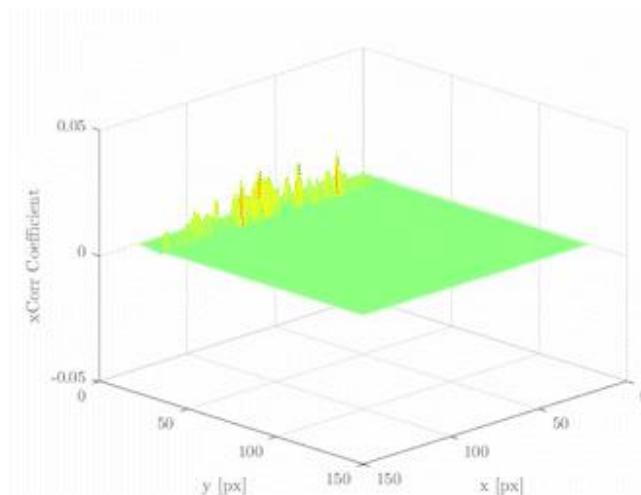
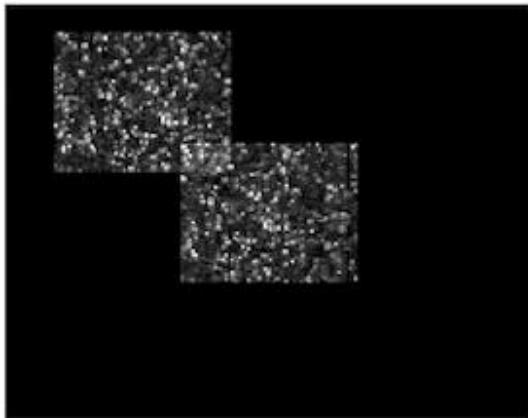
$$h_2(\cdot)$$

“weak” learners

Pre-Learning era: Handcrafted features

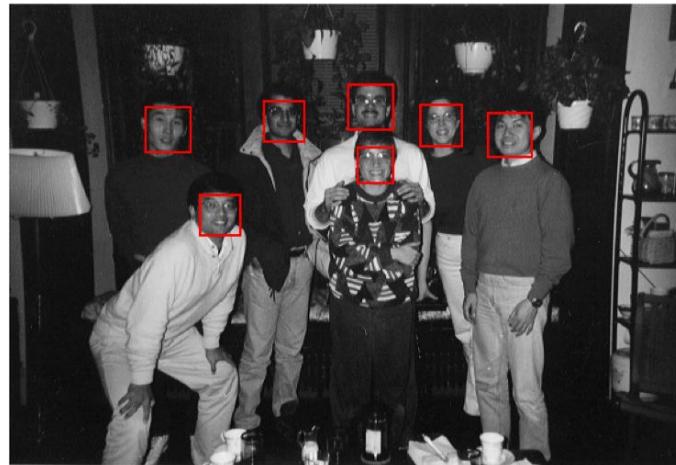
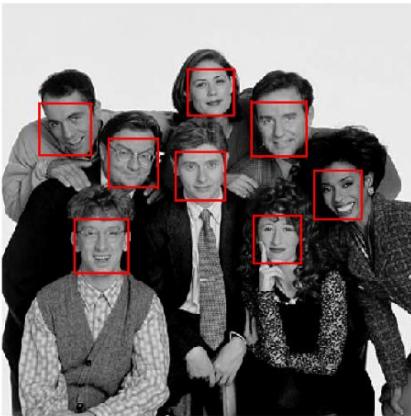
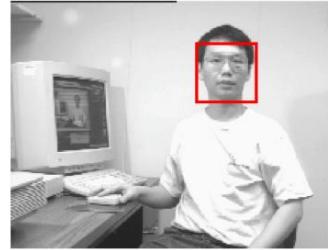
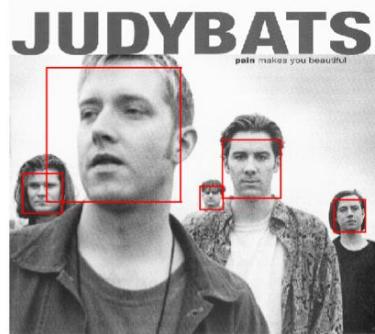
- Zero-normalised cross-correlation (ZNCC):

$$d(I_{(x_0, y_0)}, T) = \frac{1}{n} \sum_{x,y} \frac{1}{\sigma_I \sigma_T} (I_{(x_0, y_0)}(x, y) - \mu_I)(T(x, y) - \mu_T)$$



Source: <https://en.wikipedia.org/wiki/Cross-correlation>

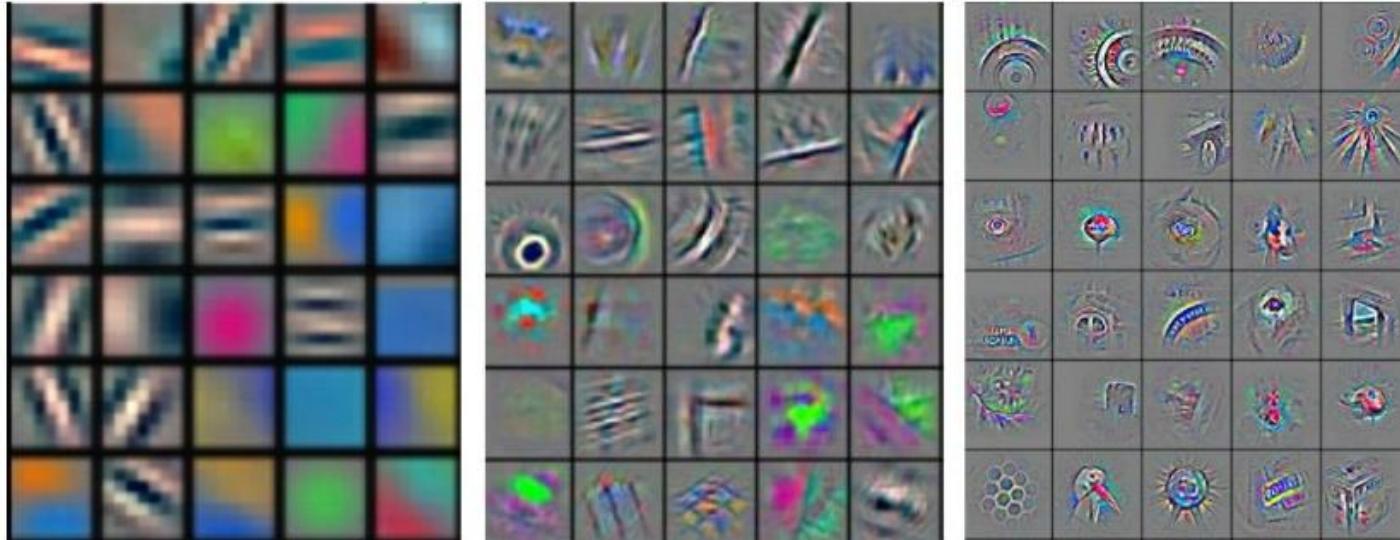
Pre-Learning era: Handcrafted features



Deep Learning:

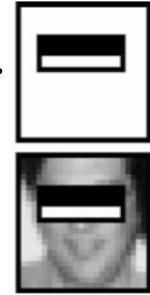


Deep learning is a paradigm to extract patterns and latent features from given observations

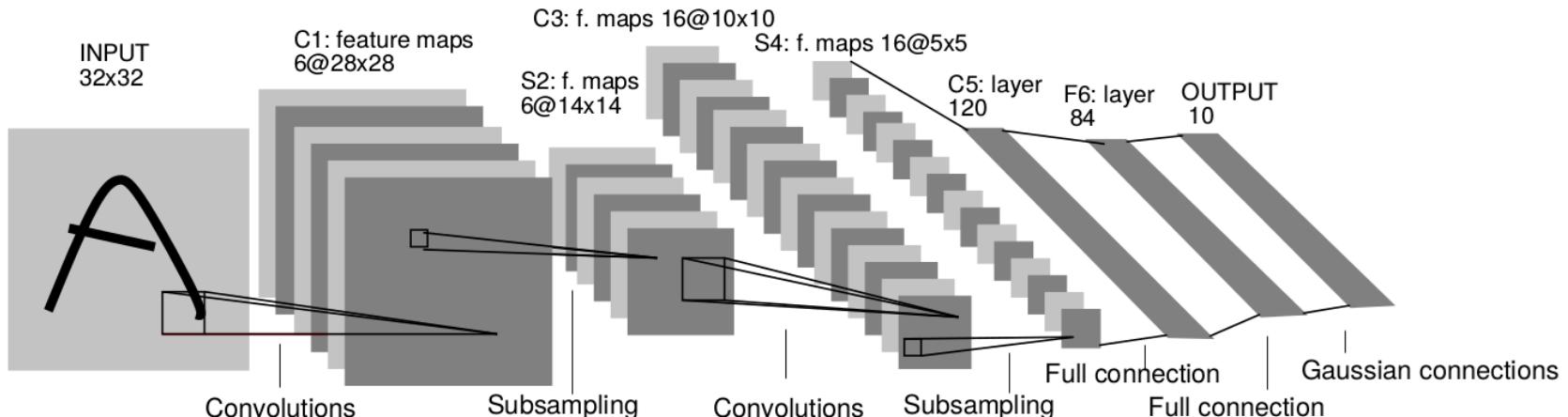


CNN:

Data are often composed of hierarchical, local, shift-invariant patterns.

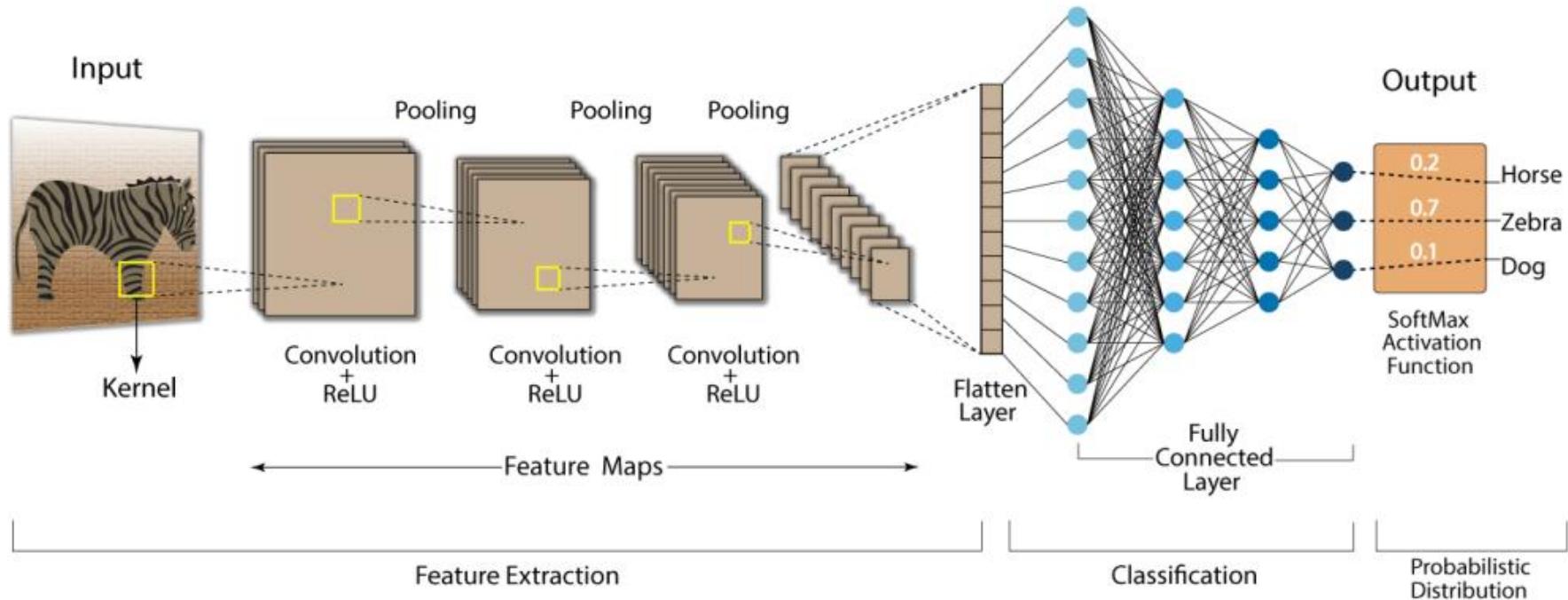


CNNs directly exploit this fact as a prior.



Convolutional network

Convolution Neural Network (CNN)

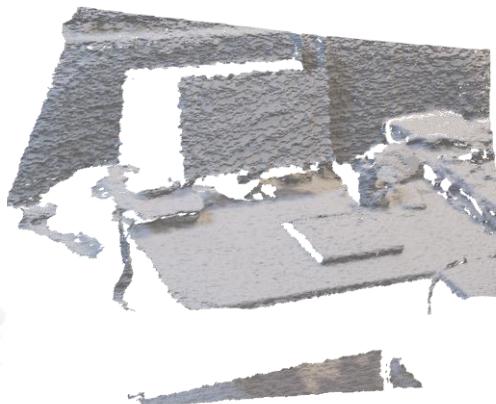


Convolution relies on the underlying structure

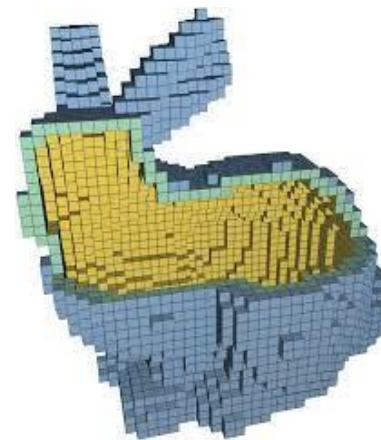
Geometrical data have different structure



Point clouds



Range map



Volumetric



Implicit

“Fast Parallel Surface and Solid Voxelization on GPUs”, M. Schwarz et al., 2010

“Implicit Geometric Regularization for Learning Shapes”, A. Gropp et al., 2020

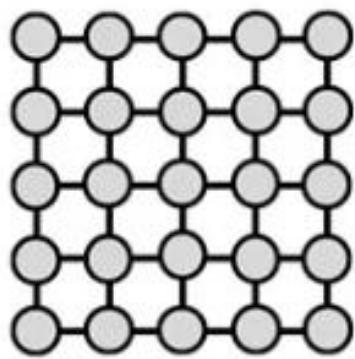
Geometric Data

1. Data are **not organized** on a fixed grid or template
2. **Different representations** are possible
3. **Rigid** transformations are possible
4. **Limited** amount of **data** available



Geometric deep learning

Grids



Euclidean samples,
e.g. image

Groups



Homogenous spaces
with global symmetries,
e.g. sphere

Graphs



Nodes and connections,
e.g. social network

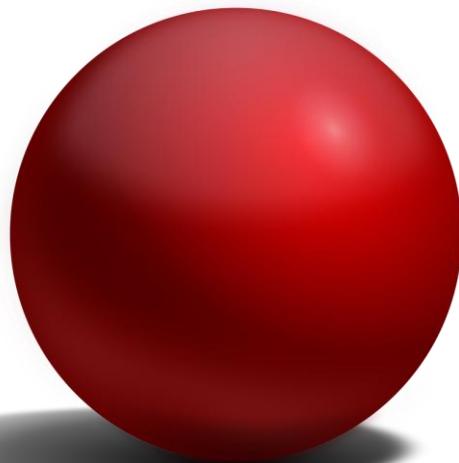
Geodesics & Gauges



Manifolds,
e.g. 3D mesh

The roles of a representation

Convey the
geometry

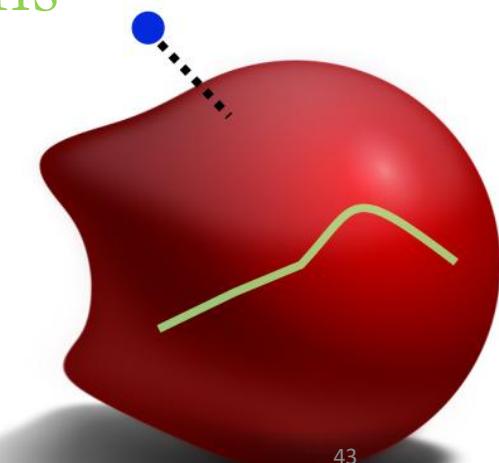


Support
computations

Geometry
Evaluations

Spatial
Queries

Modifications





Ceci n'est pas une pipe.



Example: Single RGB

- No Depth
- Occlusions
- Computations are difficult
- Cheap
- Perfect for your Computer Monitor

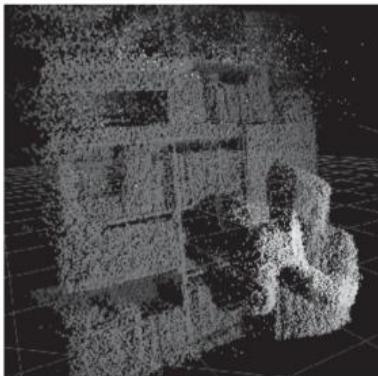
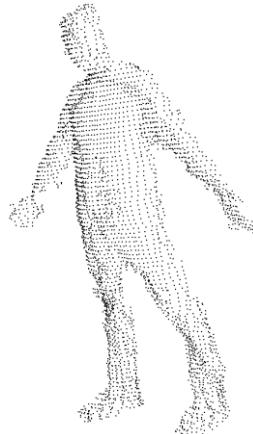


Multiple RGBs



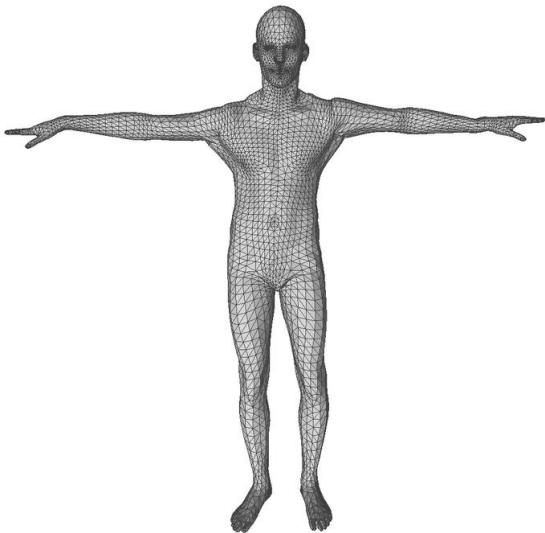
- We can infer spatial relations
- Can be processed by (CNNs)
- Often requires calibrated multi-view cameras

Point Cloud



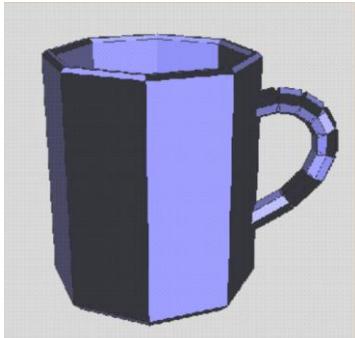
- Geometry is where points are
- Computations?
Adding or subtracting points, spatial distance, ...
- What happens in the middle?
- Output of different pipelines

(Triangular) Mesh



<https://yifita.netlify.app/project/neural-shape/>

- Geometry is a set of piecewise flat surfaces
- Points + Connectivity
- Great theoretical properties
- Not everything is easy
(e.g., spatial query, remove geometry)



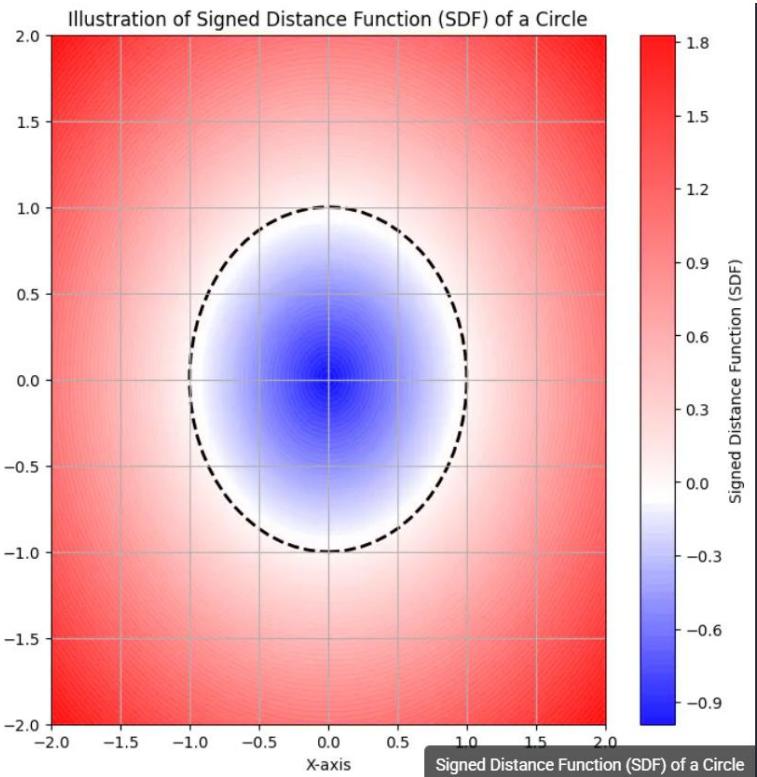
<https://www.cs.princeton.edu/courses/archive/fall00/cs426/lectures/reps/sld015.htm>

Implicit Representations (E.g., Voxelization)



- Geometry as a function in a 3D grid
- Like an image, but in 3D
- Great for learning, (e.g, CNNs), model objects with different topologies
- Expensive, not that easy to visualize

Implicit Representations (E.g., Voxelization)



- Geometry as a function (grid or not)
- Like an image, but in 3D
- Great for learning, (e.g, CNNs), model objects with different topologies
- Expensive, not that easy to visualize

To sum up, choose your representations wisely!

Do I need to edit it?

How will I visualize it?

Intrinsic or spatial queries?

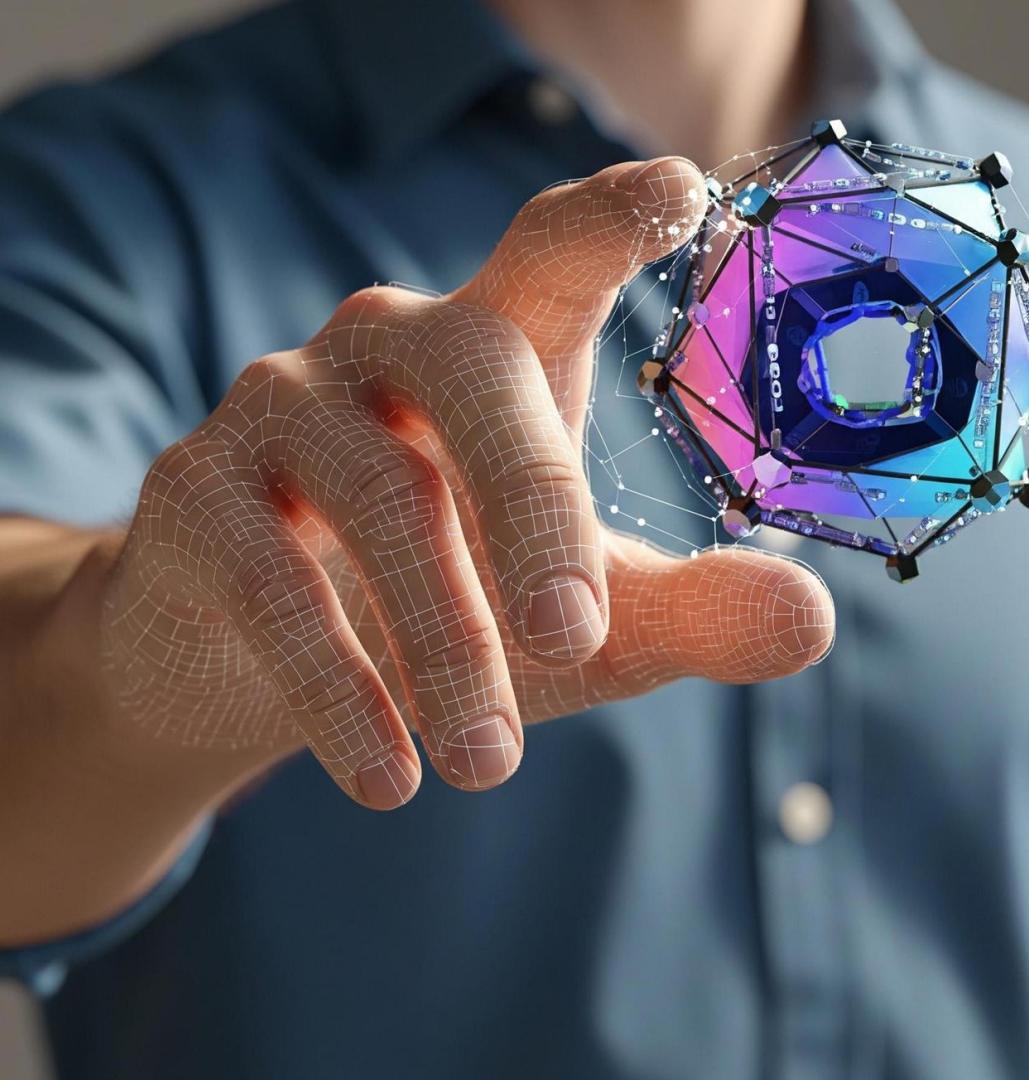
Invariance to 3D embedding?

Can I combine, extend, learn representations?

...

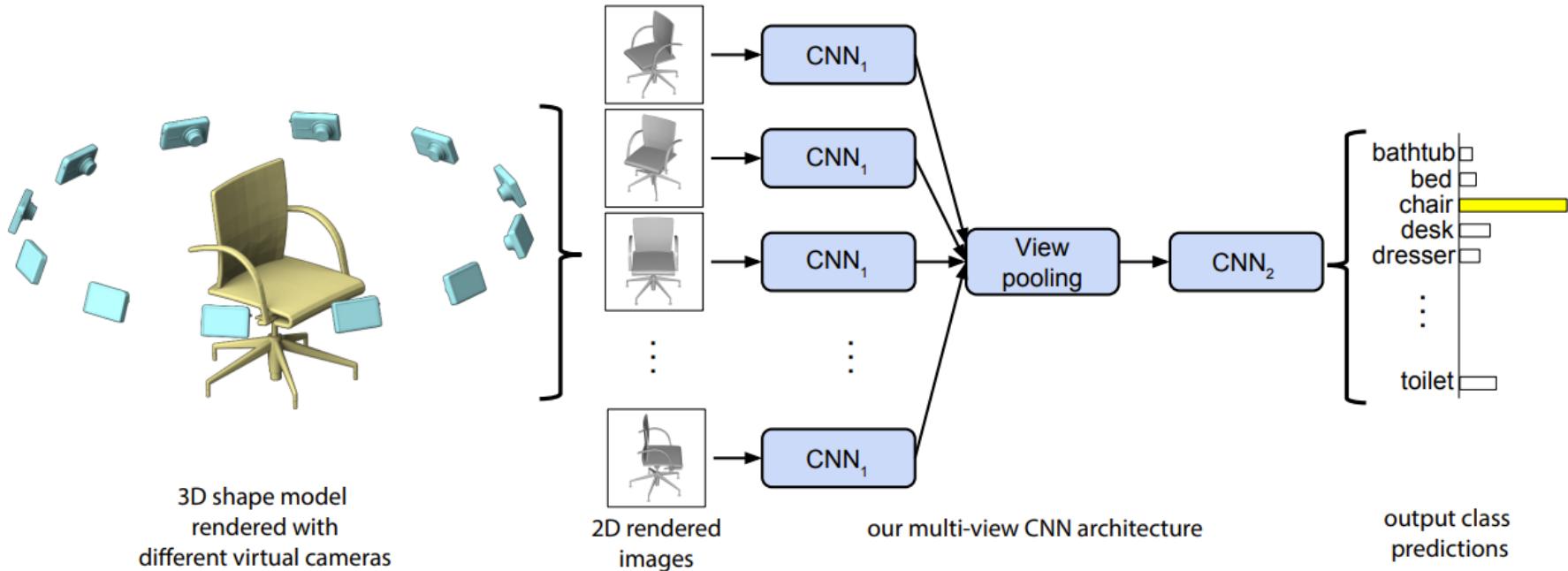
To sum up, choose your representations wisely!

What about learning?



Different
architectures
for different
representations

Multi-view CNN



Multi-view CNN - Saliency

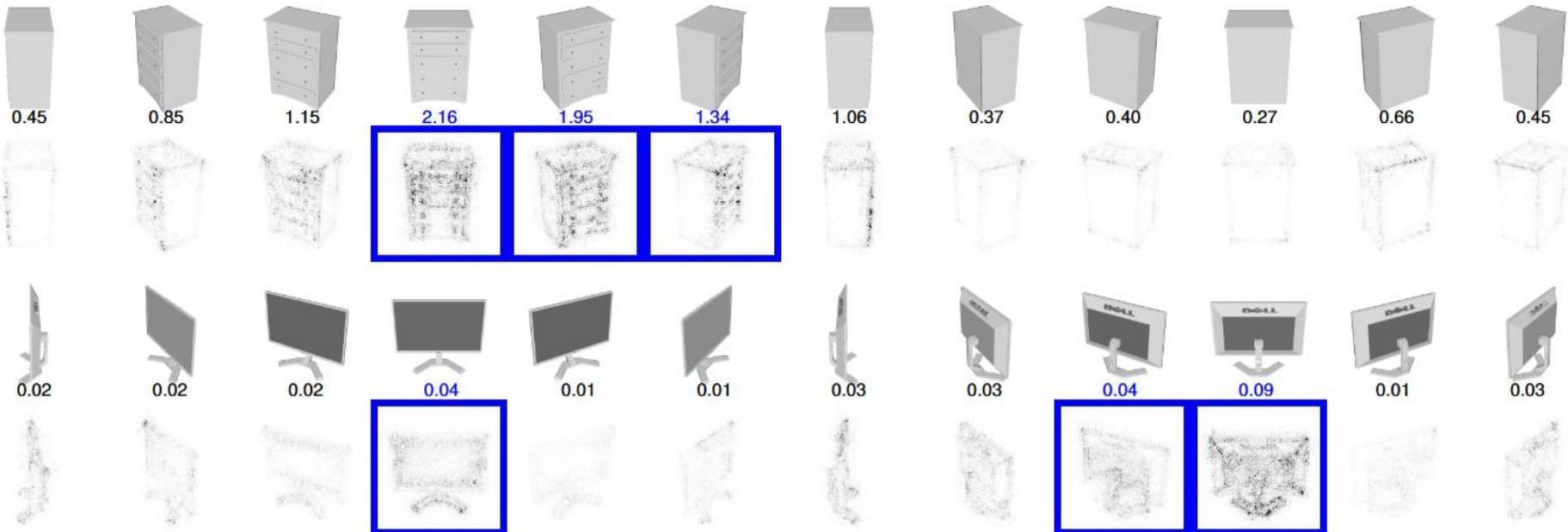
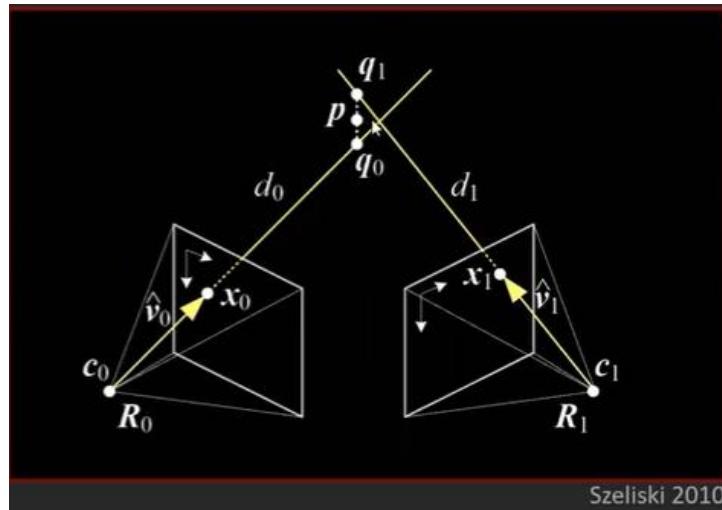
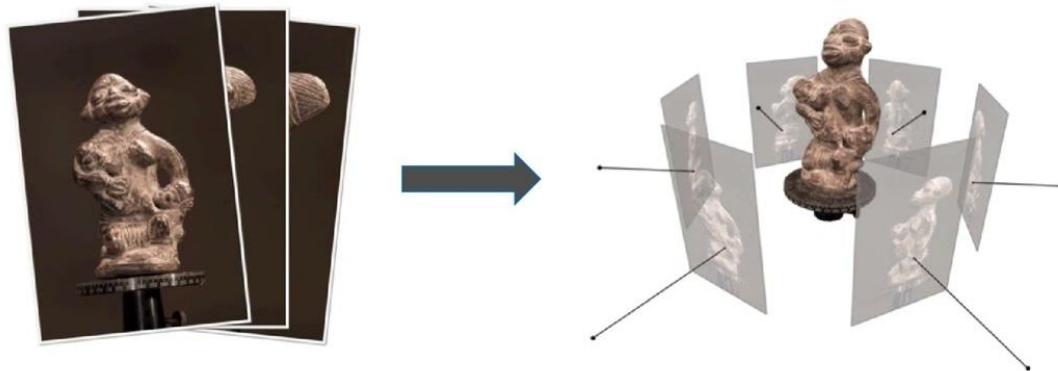
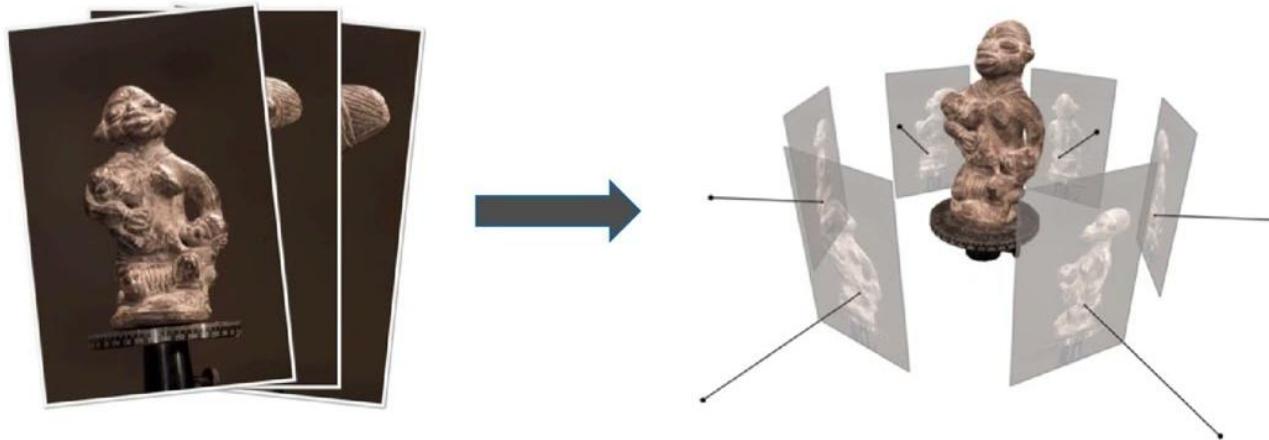


Figure 3. Top three views with the highest saliency are highlighted in blue and the relative magnitude of gradient energy for each view is shown on top. The saliency maps are computed by back-propagating the gradients of the class score onto the images via the view-pooling layer. Notice that the handles of the dresser are the most discriminative features. (Figures are enhanced for visibility.)

Multiple RGBs are the foundation for 3D reconstruction



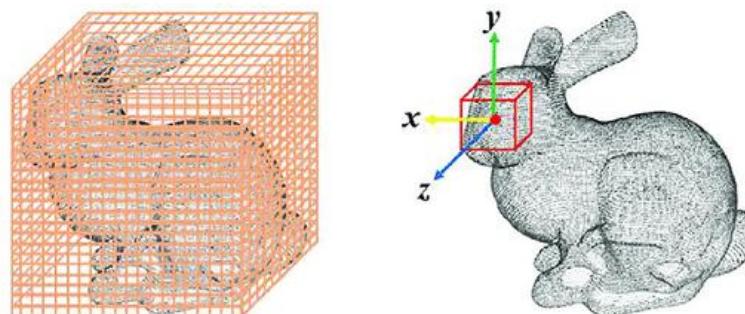
Multiple RGBs are the foundation for 3D reconstruction



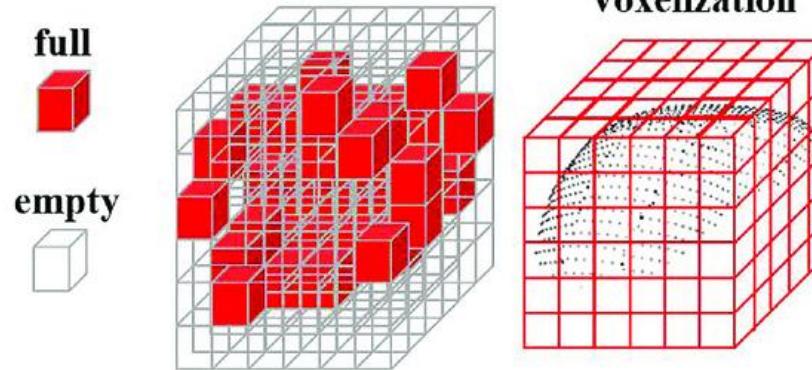
Why learning from images
when I can incorporate an explicit
structural bias (3D reconstruction)?

Converting 3D data to a fixed grid

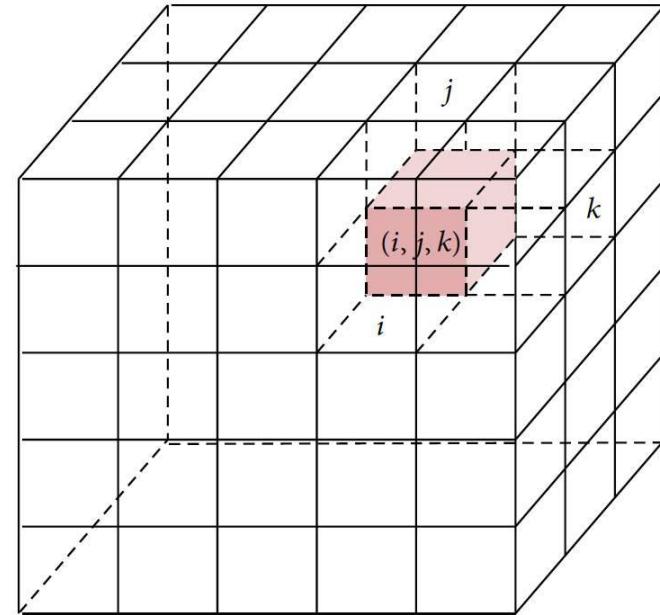
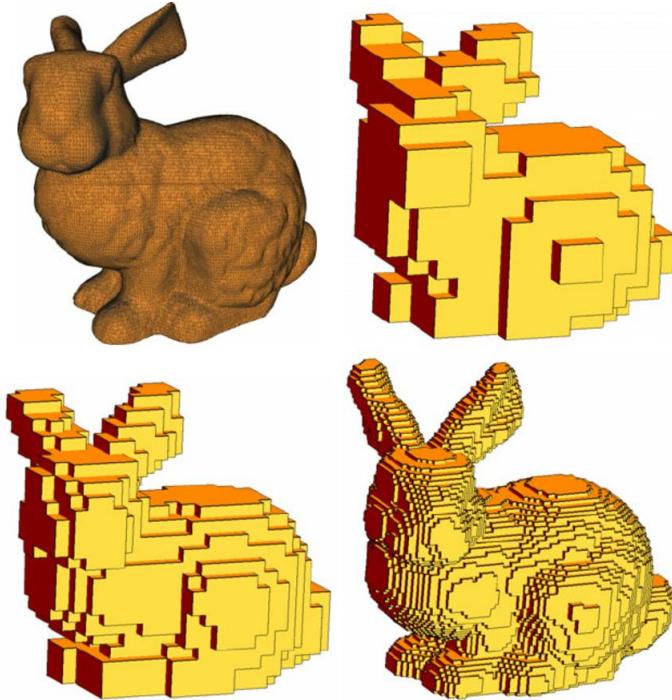
Voxel Grid



Voxelization

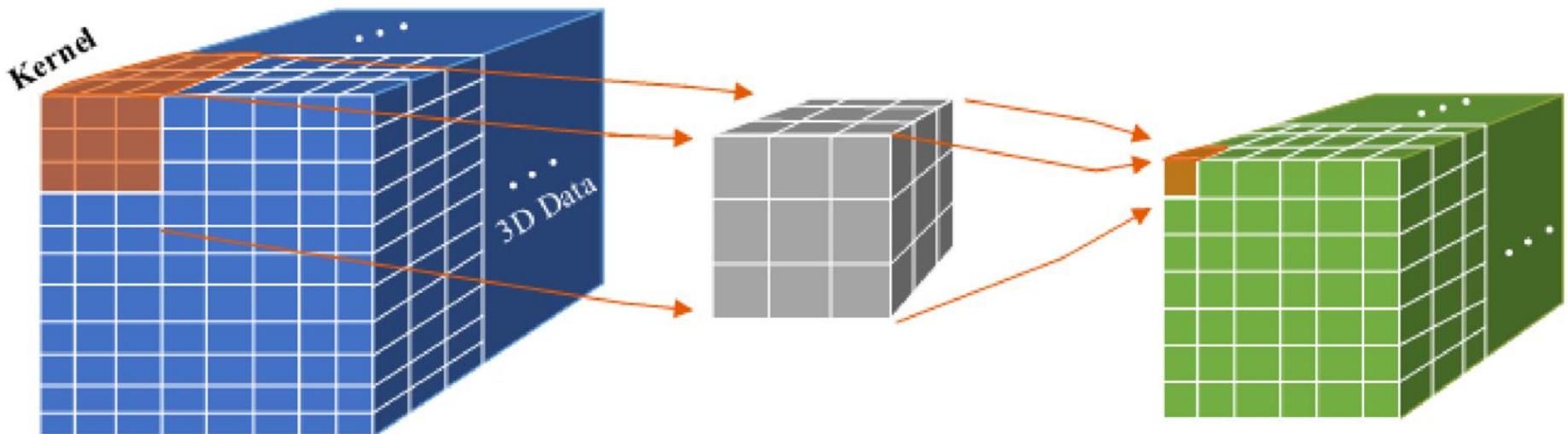


Voxel - Occupancy

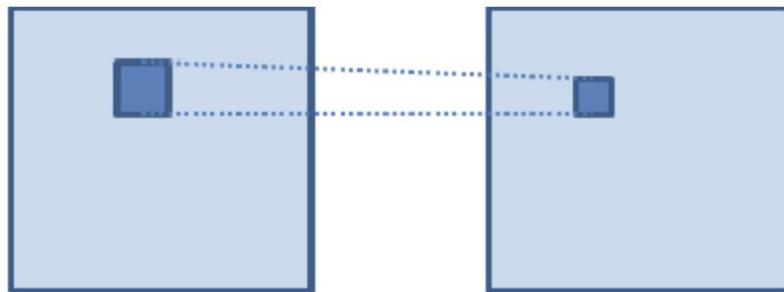


A grid ($N \times N \times N$)

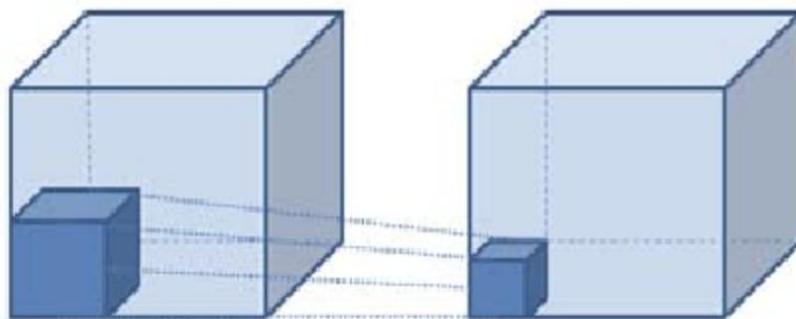
3D Convolutional network



Voxel - 3D Convolution



(a) 2D convolution



(b) 3D convolution

3. Comparison of (a) 2D convolution and (b)

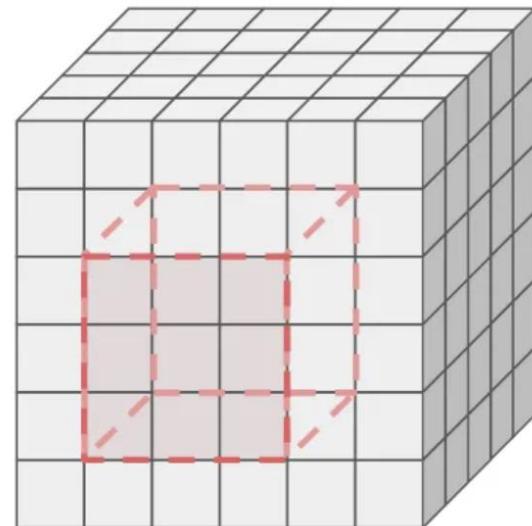


Fig. 7: Example with a $3 \times 3 \times 3$ kernel sliding across the voxel region

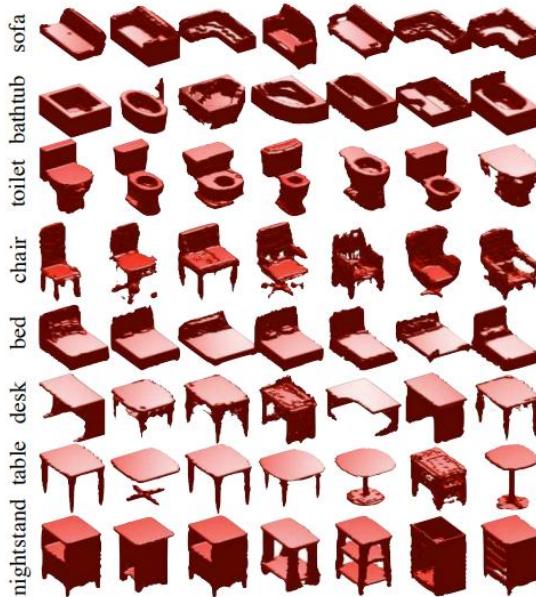
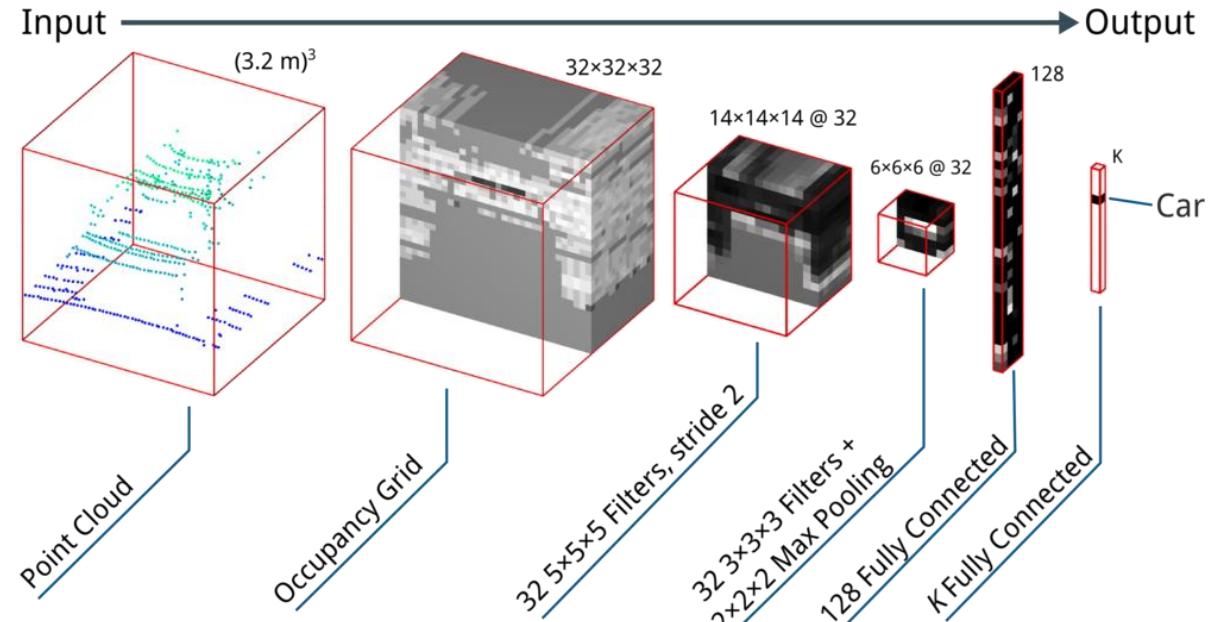


Figure 6: **Shape Sampling.** Example shapes generated by sampling our 3D ShapeNets for some categories.

ShapeNet 3D

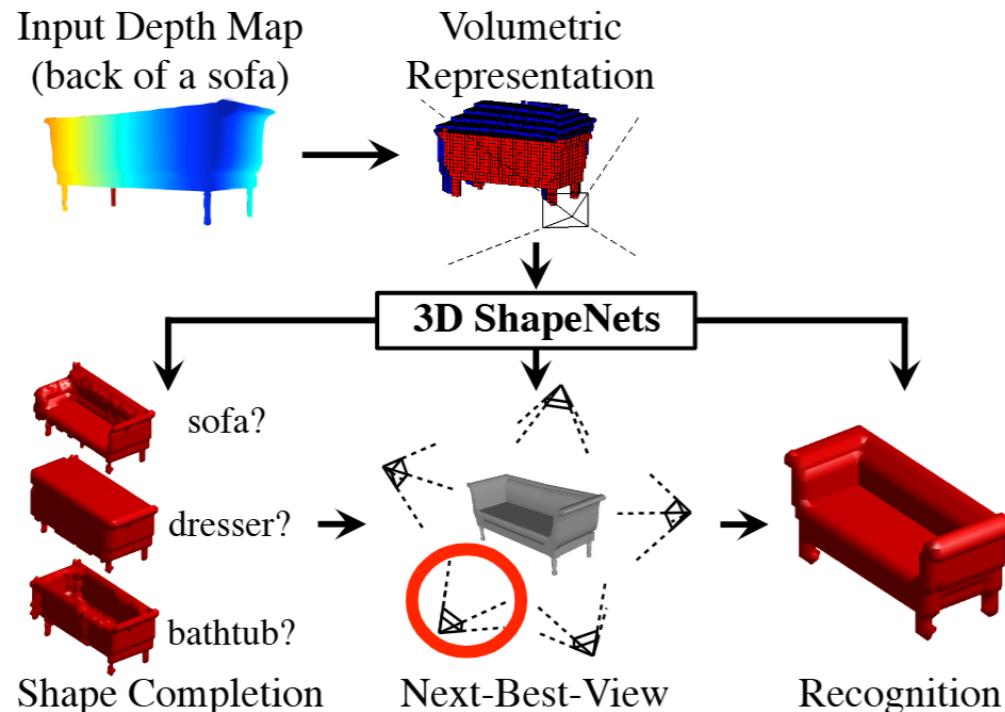
CVPR
June 2015



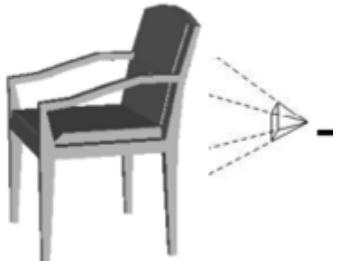
VoxNet

IROS
September 2015

ShapeNet 3D Training (classification&completion)

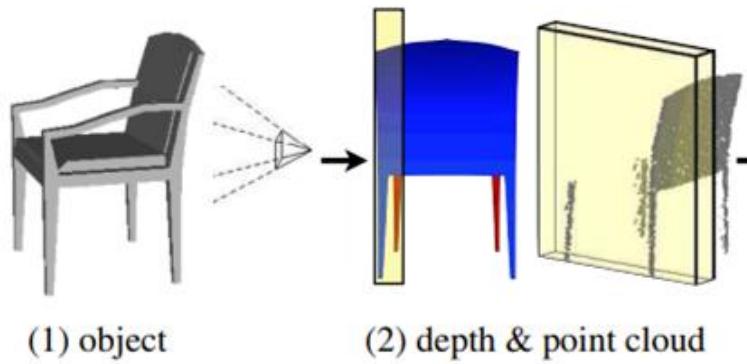


ShapeNet 3D Training (classification&completion)

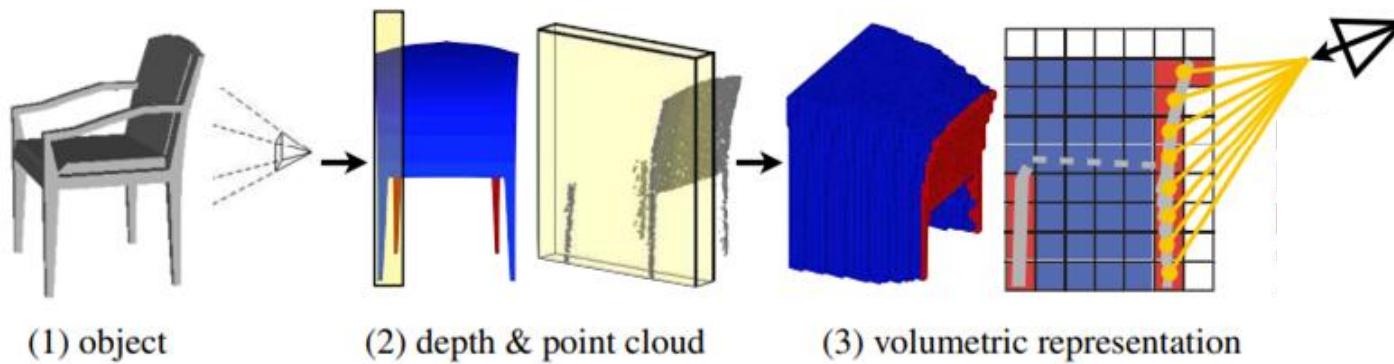


(1) object

ShapeNet 3D Training (classification&completion)



ShapeNet 3D Training (classification&completion)

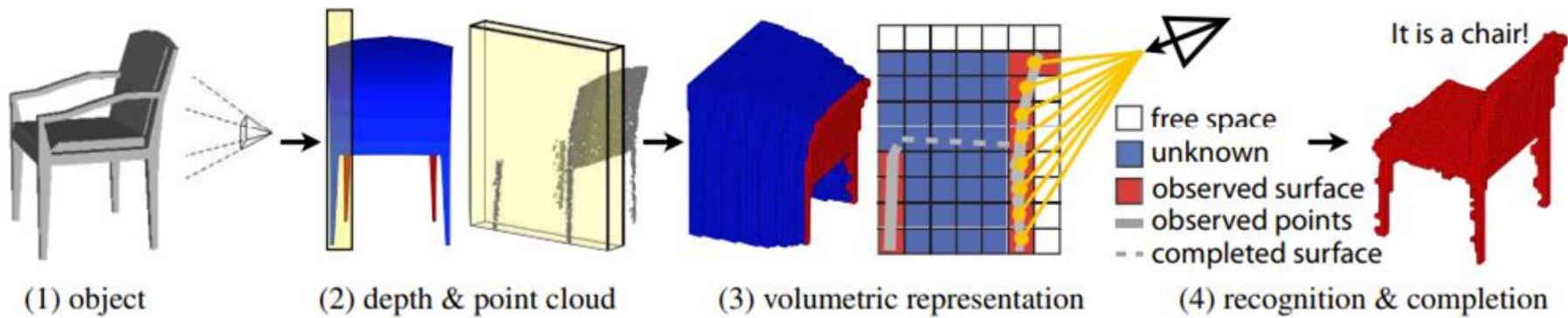


(1) object

(2) depth & point cloud

(3) volumetric representation

ShapeNet 3D Training (classification&completion)



Completion

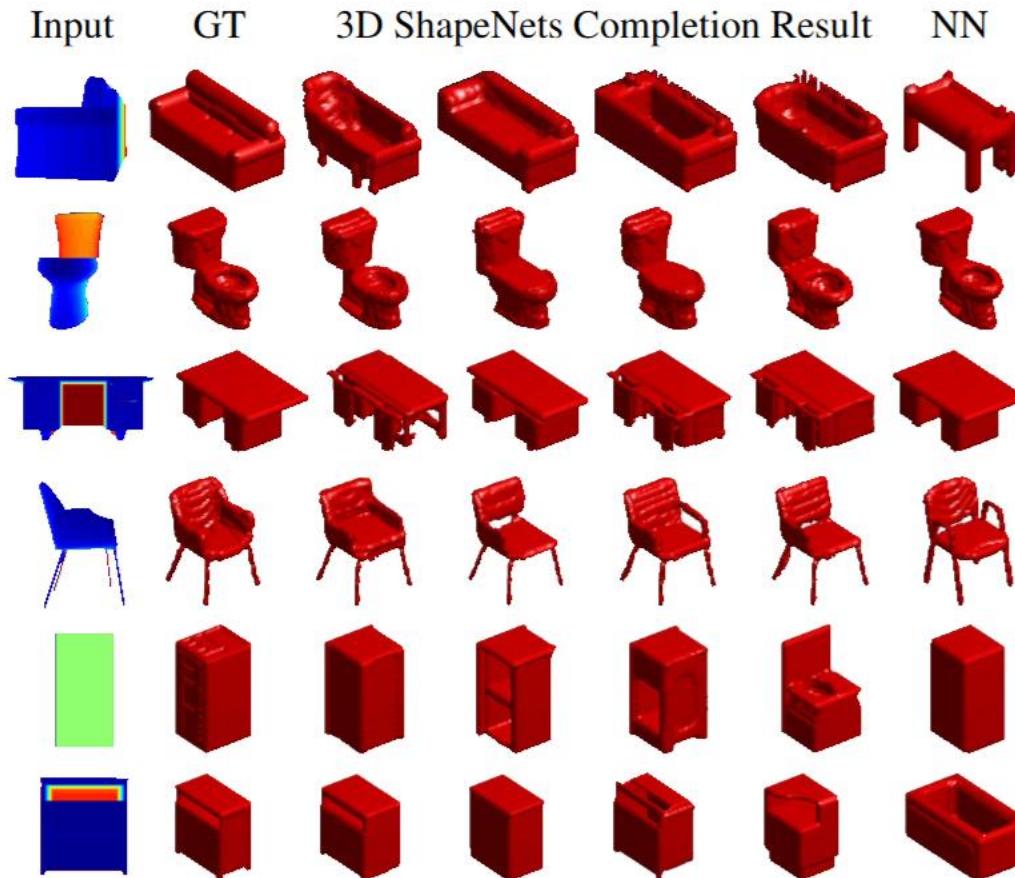
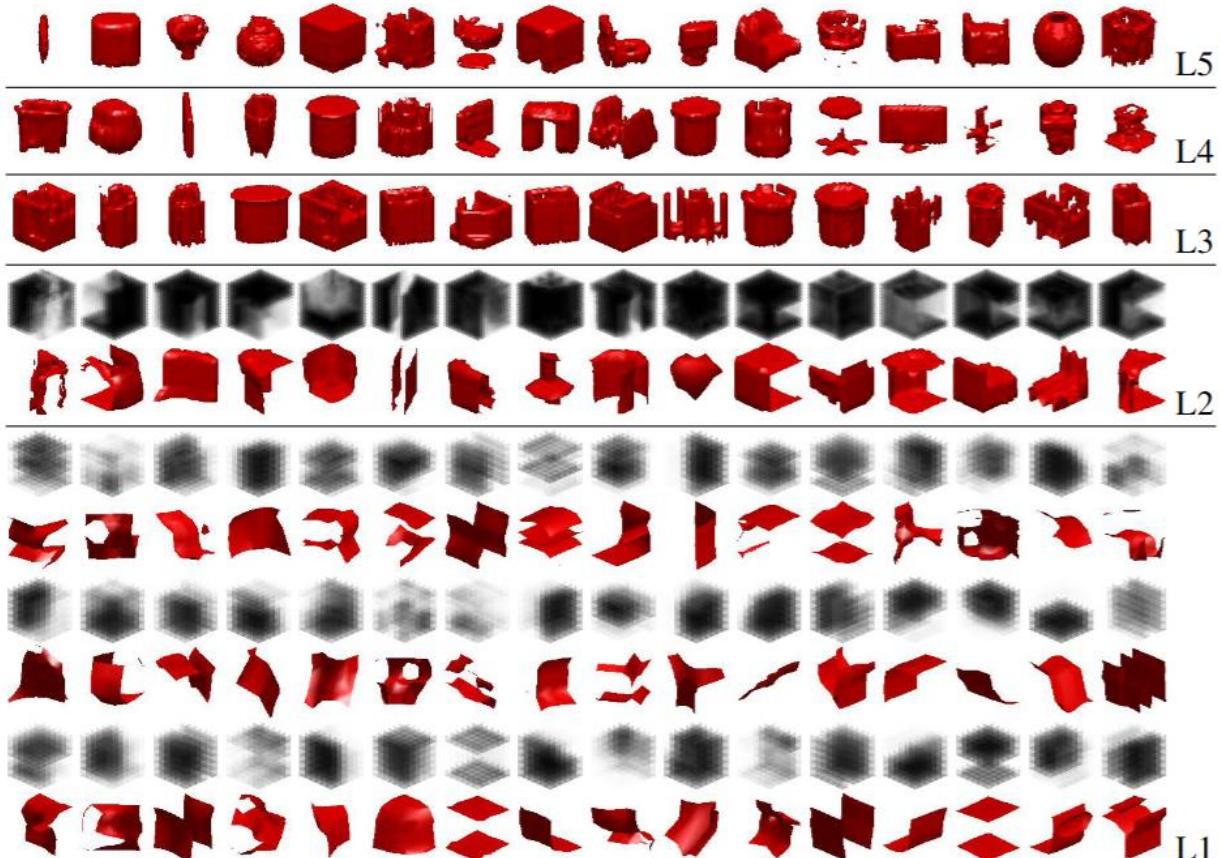


Figure 8: **Shape Completion.** From left to right: input depth map from a single view, ground truth shape, shape completion result (4 cols), nearest neighbor result (1 col).

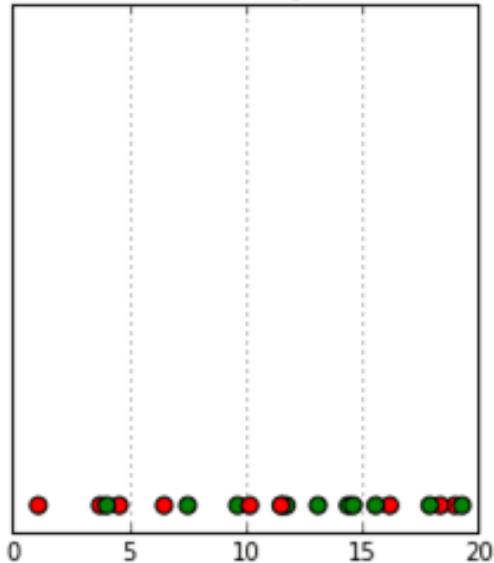
Interpretability study



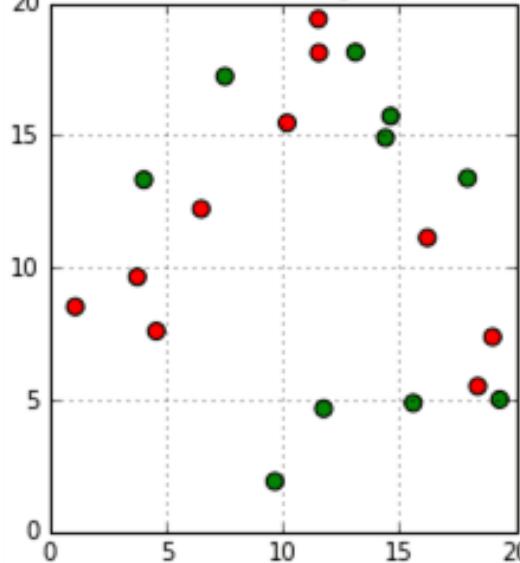
(b) Data-driven visualization: For each neuron, we average the top 100 training examples with highest responses (>0.99) and crop the volume inside the receptive field. The averaged result is visualized by transparency in 3D (Gray) and by the average surface obtained from zero-crossing (Red). 3D ShapeNets are able to capture complex structures in 3D space, from low-level surfaces and corners at L1, to objects parts at L2 and L3, and whole objects at L4 and above.

Problem of voxels: Curse of dimensionality

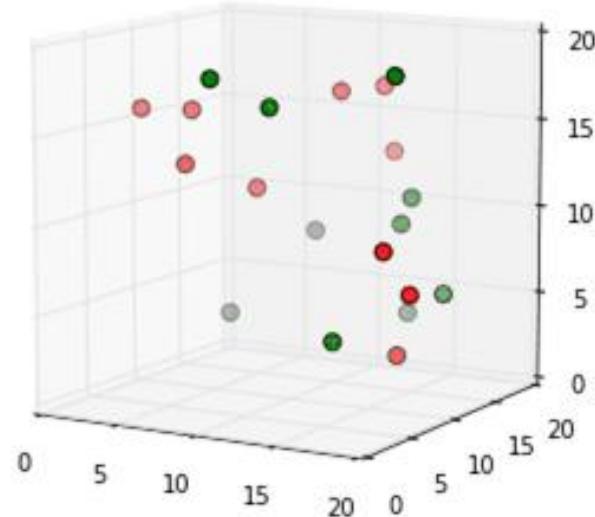
a) 1D - 4 regions



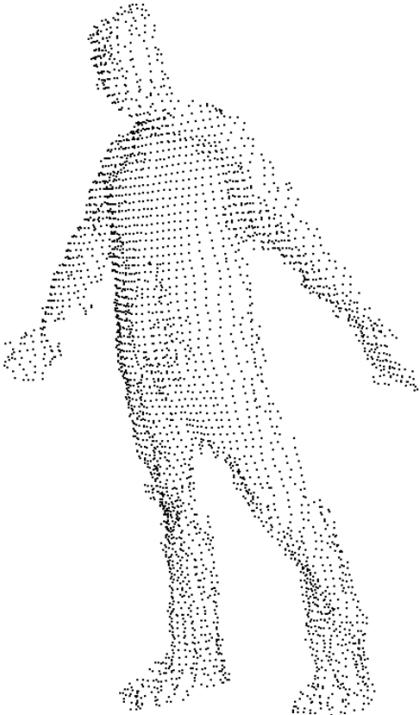
b) 2D - 16 regions



c) 3D - 64 regions



Can we directly process points?



Point Cloud

$$P \in \mathbb{R}^{N \times 3}$$

Point clouds are **sets**

All the permutations represent the same object

There is no grid\structure

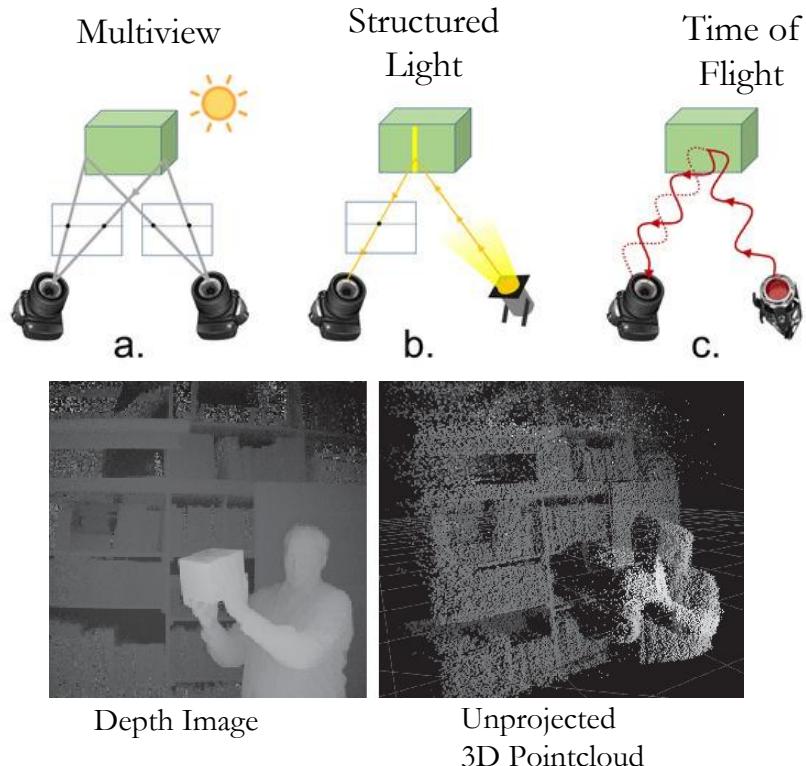
1.1	- 2	0.1
1.5	- 1	0.4
0.1	1.1	0.7
...

=

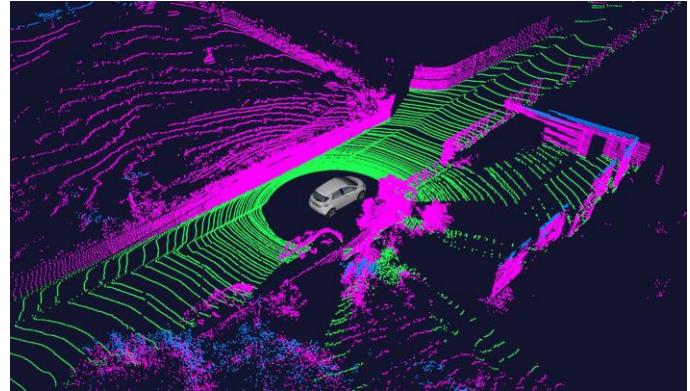
0.1	1.1	0.7
1.5	- 1	0.4
1.1	- 2	0.1
...

Point clouds are convenient

Point clouds are often the output of acquisition pipeline (raw data)
They are compact (no voxel) while in 3D (no multi-view)
They support information (e.g., colors)



Autonomous Driving



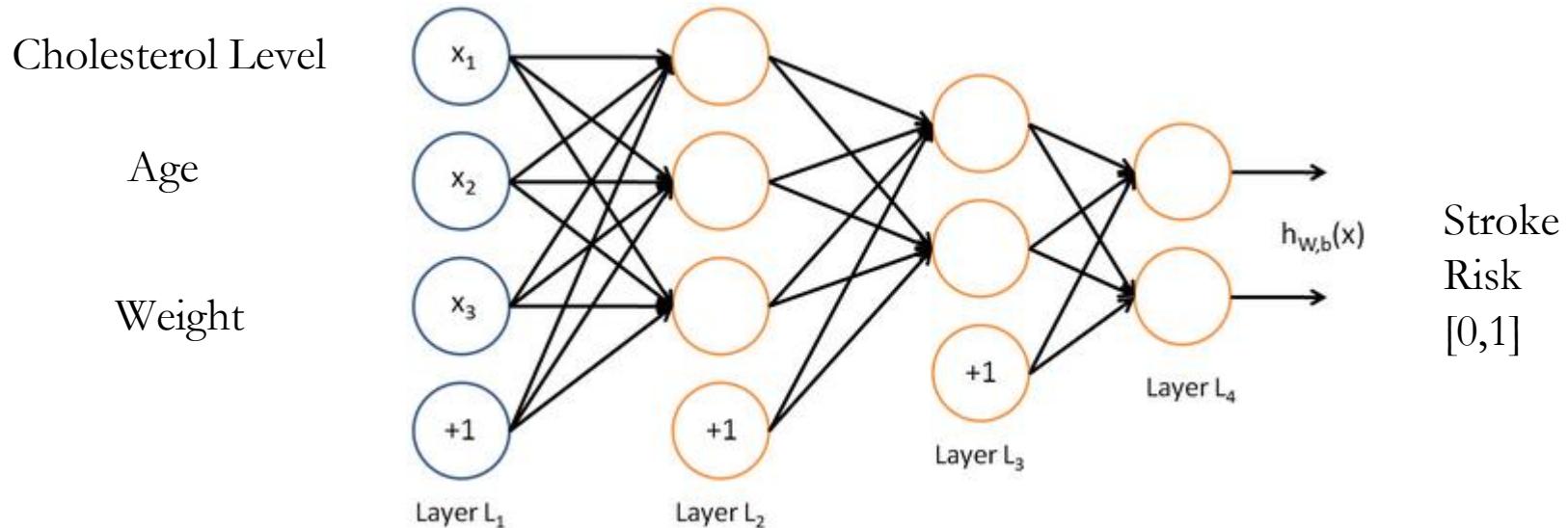
Commercial Depth Sensors



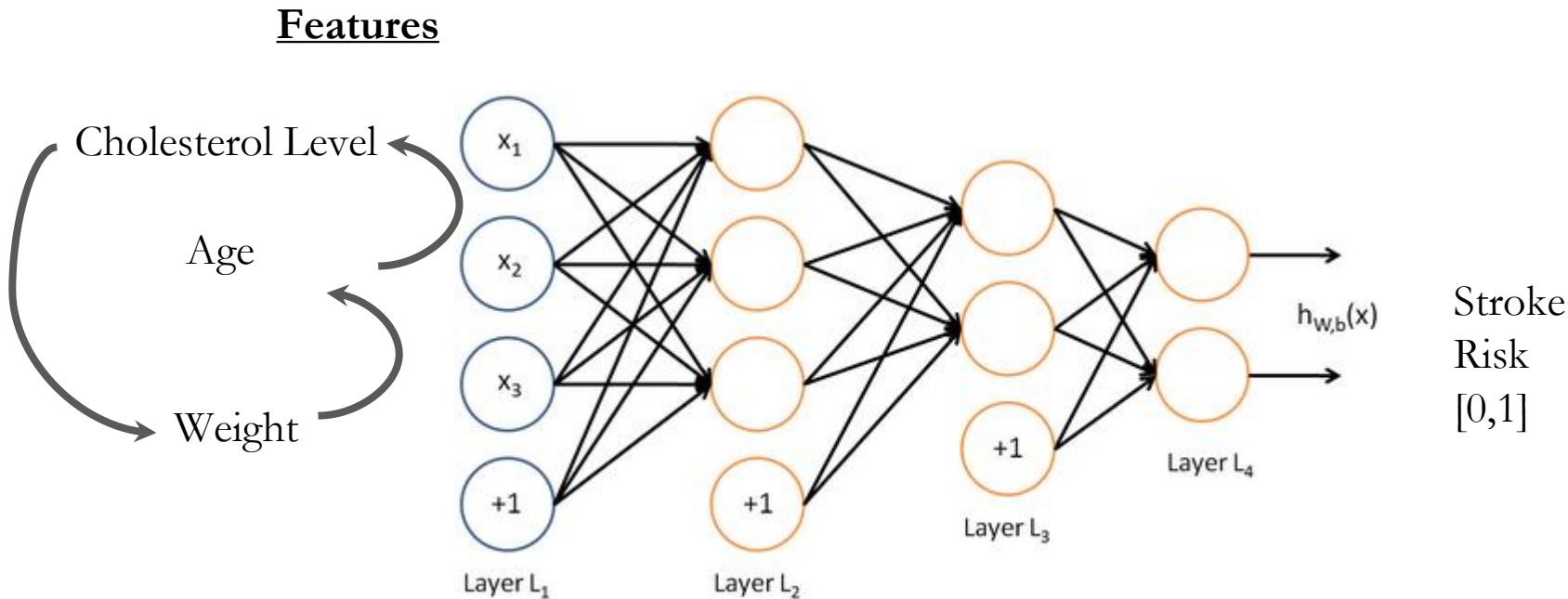
(a)
https://www.researchgate.net/figure/Overview-a-3D-point-cloud-of-a-human-in-a-cluttered-home-environment-b-Recovered_fig6_278700751
<https://keymakr.com/blog/how-3d-point-cloud-segmentation-will-make-the-future-hands-free/>
<https://www.mdpi.com/1424-8220/22/14/5448>
<https://docs.unity3d.com/Documentation/manual/author/set-up-sensors/configure-a-tof.html>

Standard MLP

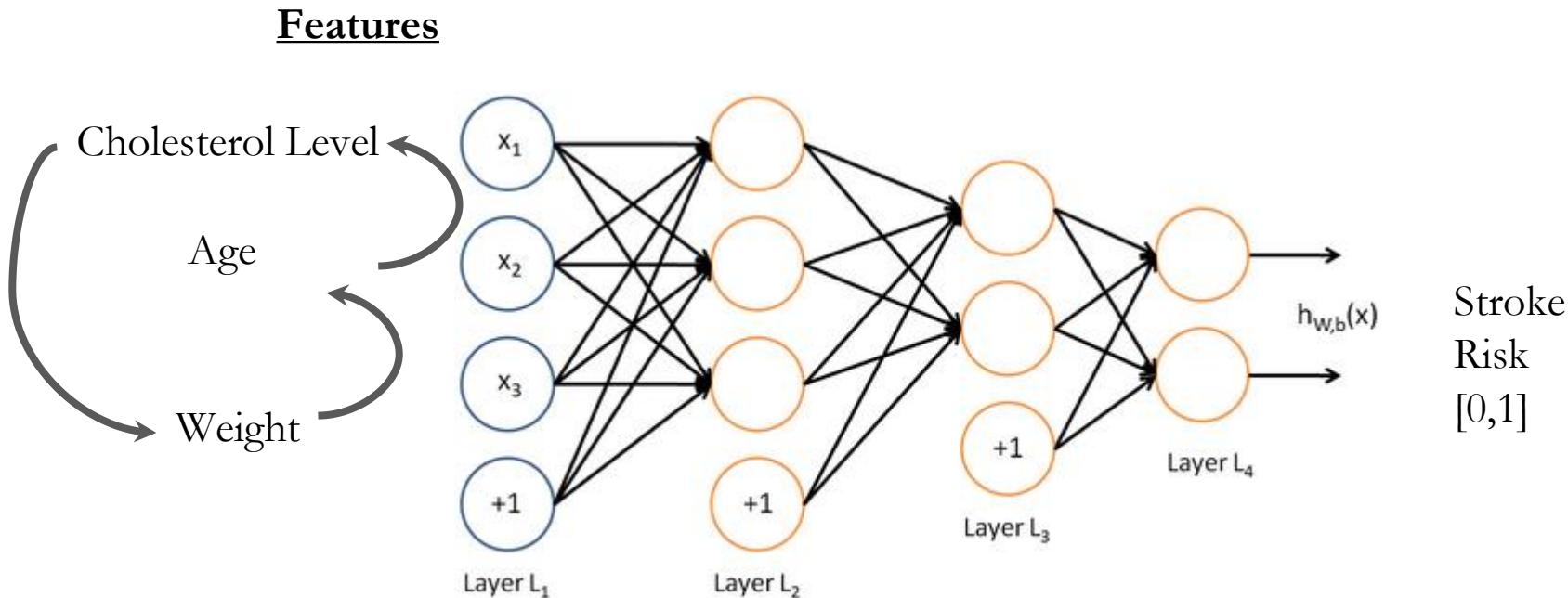
Features



What happen if I shuffle the input?



What happen if I shuffle the input?



You can't! Order matters: features are an ordered list!

PointNet

PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation

Charles R. Qi*

Hao Su*

Kaichun Mo
Stanford University

Leonidas J. Guibas

Abstract

Point cloud is an important type of geometric data structure. Due to its irregular format, most researchers transform such data to regular 3D voxel grids or collections of images. This, however, renders data unnecessarily voluminous and causes issues. In this paper, we design a novel type of neural network that directly consumes point clouds, which well respects the permutation invariance of points in the input. Our network, named PointNet, provides a unified architecture for applications ranging from object classification, part segmentation, to scene semantic

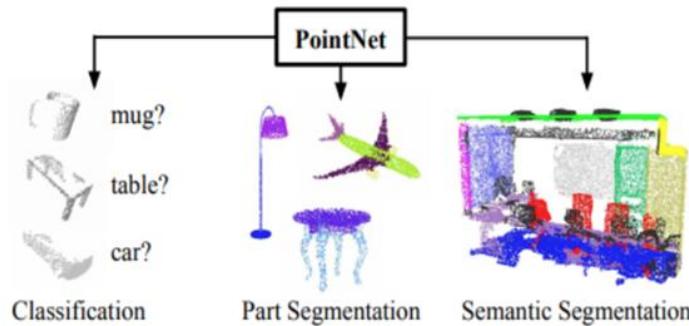
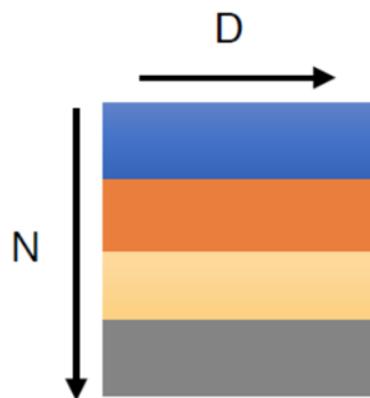


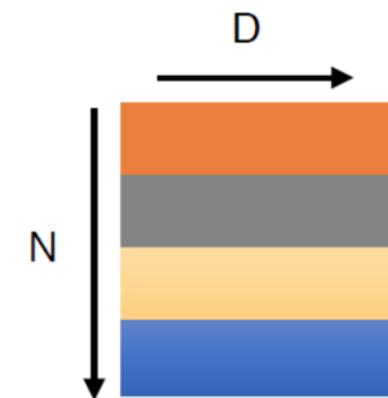
Figure 1. **Applications of PointNet.** We propose a novel deep net architecture that consumes raw point cloud (set of points) without voxelization or rendering. It is a unified architecture that learns both global and local point features, providing a simple, efficient

PointNET properties:

unorganized data = should be invariant to permutations and to different possible sampling



represents the same **set** as



Permutation invariance

Examples of symmetric functions:

$$f(x_1, x_2, \dots, x_n) = f(x_{\pi_1}, x_{\pi_2}, \dots, x_{\pi_n}), x_i \in \mathbb{R}^D, \pi \text{ permutazione}$$

$$f(x_1, x_2, \dots, x_n) = \max\{x_1, x_2, \dots, x_n\}$$

$$f(x_1, x_2, \dots, x_n) = \text{mean}\{x_1, x_2, \dots, x_n\}$$

$$f(x_1, x_2, \dots, x_n) = x_1 + x_2 + \dots + x_n$$

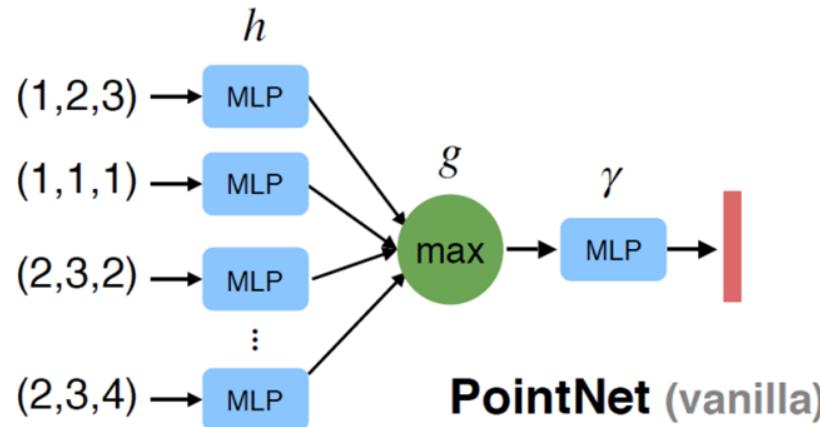
...

Can we construct
a family of
symmetric function
by
Neural Networks?

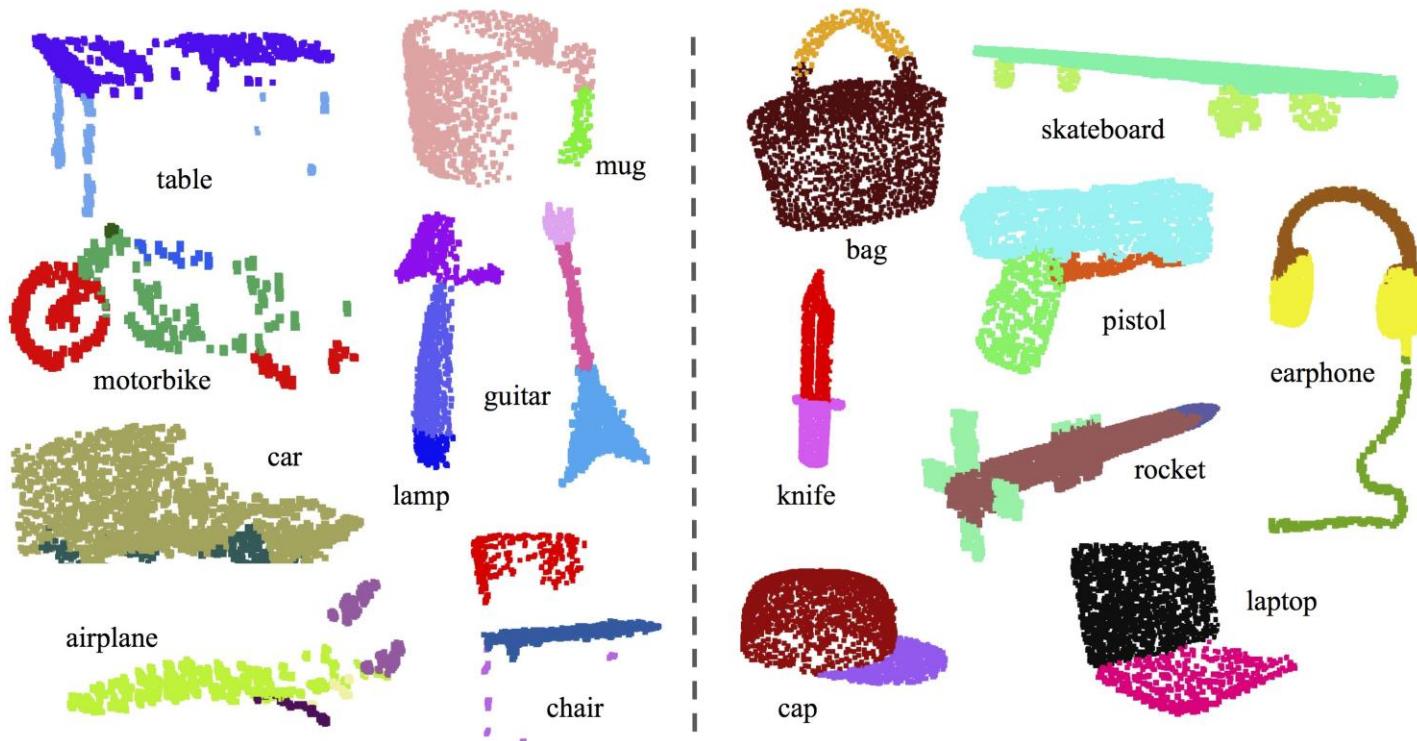
Basic PointNet

$g = \text{max pooling}$

$f, \gamma = \text{Multi-Layer perceptron}$



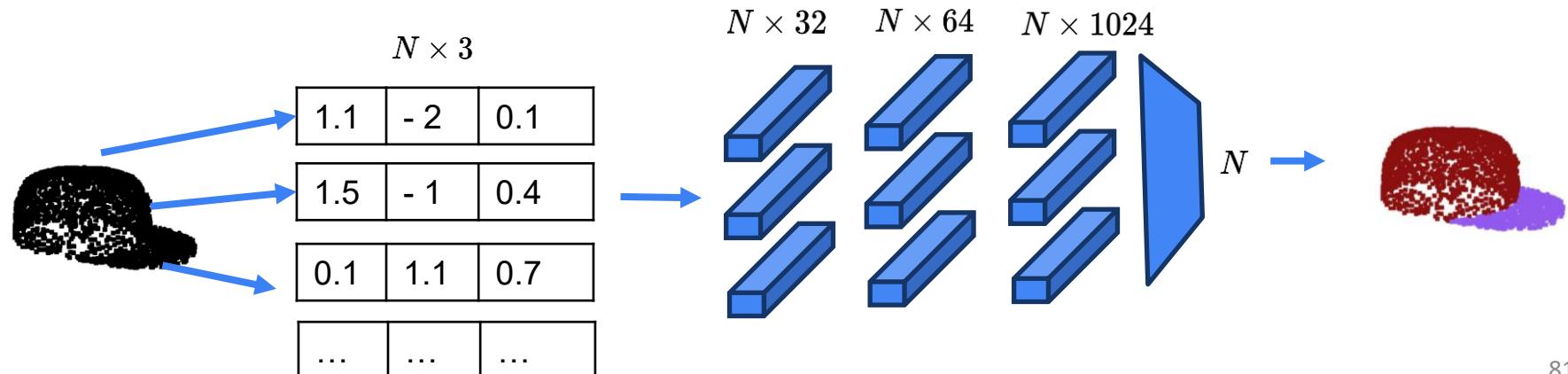
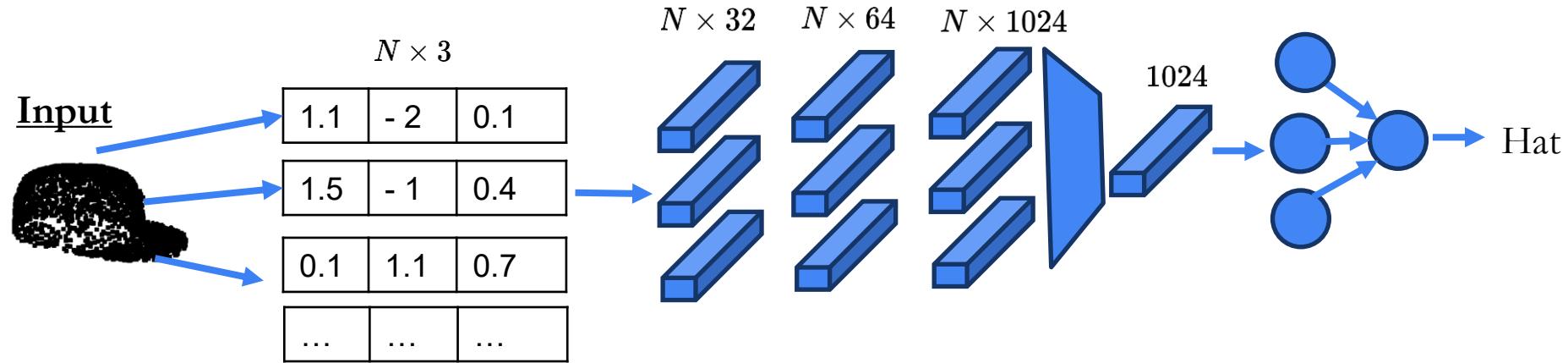
Point cloud classification and segmentation



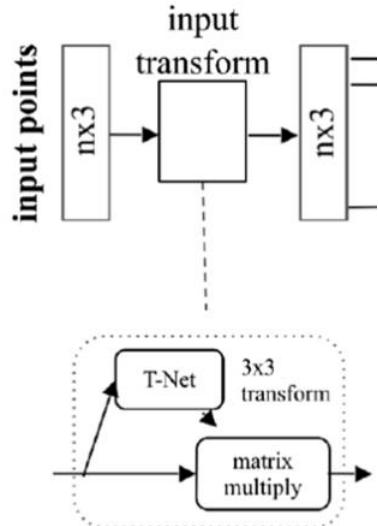
Classification = one label per point cloud

Segmentation = one label per point

Point cloud classification and segmentation

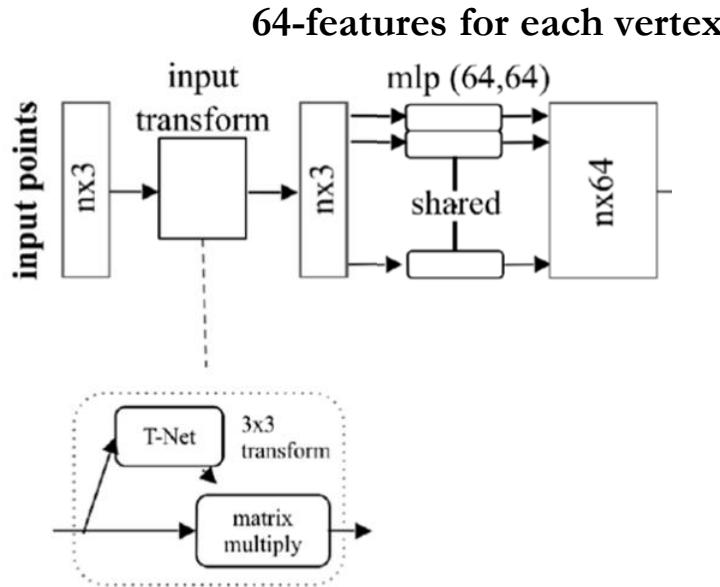


Point cloud classification

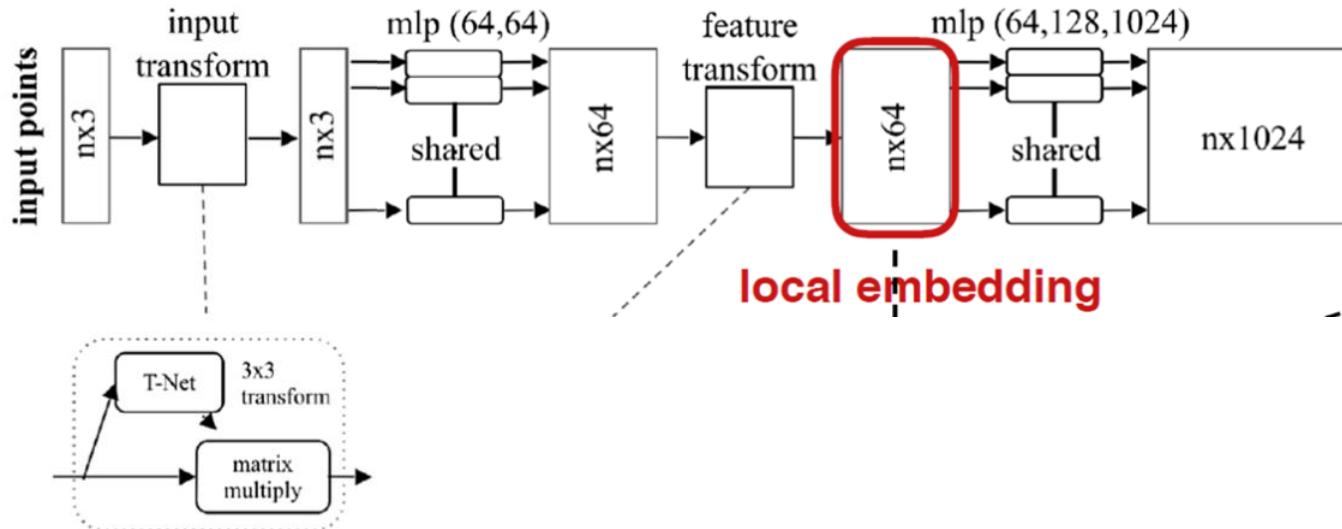


The network learns a canonicalization

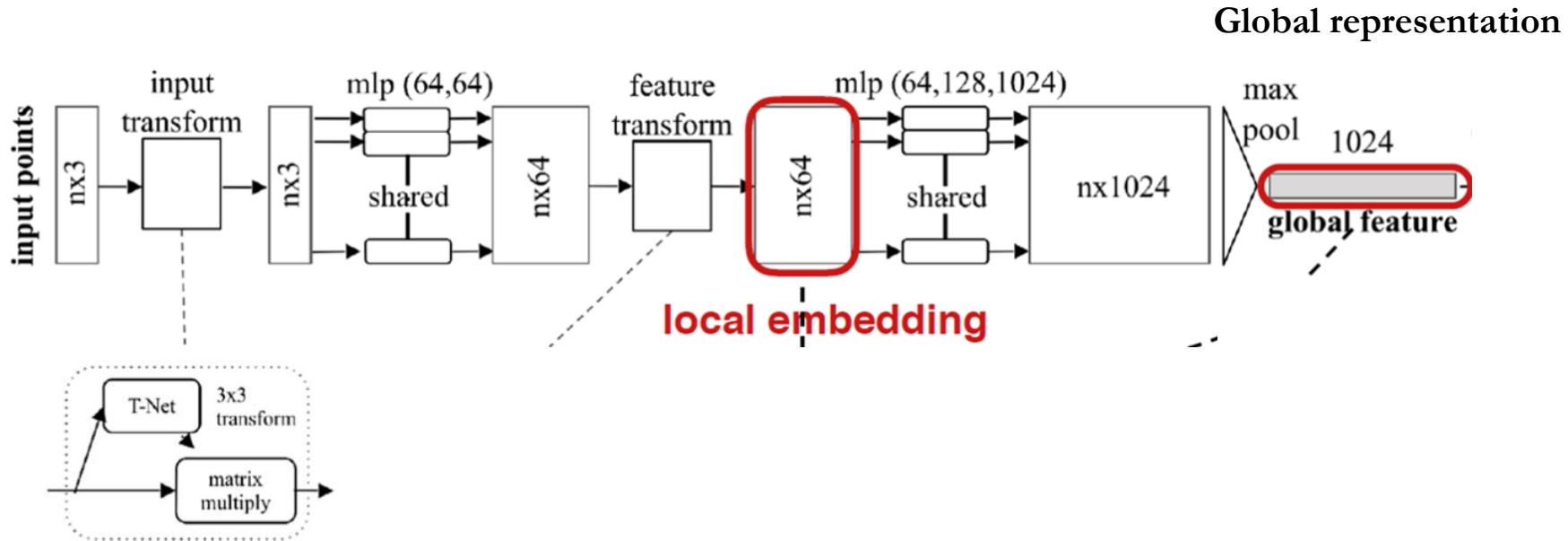
Point cloud classification



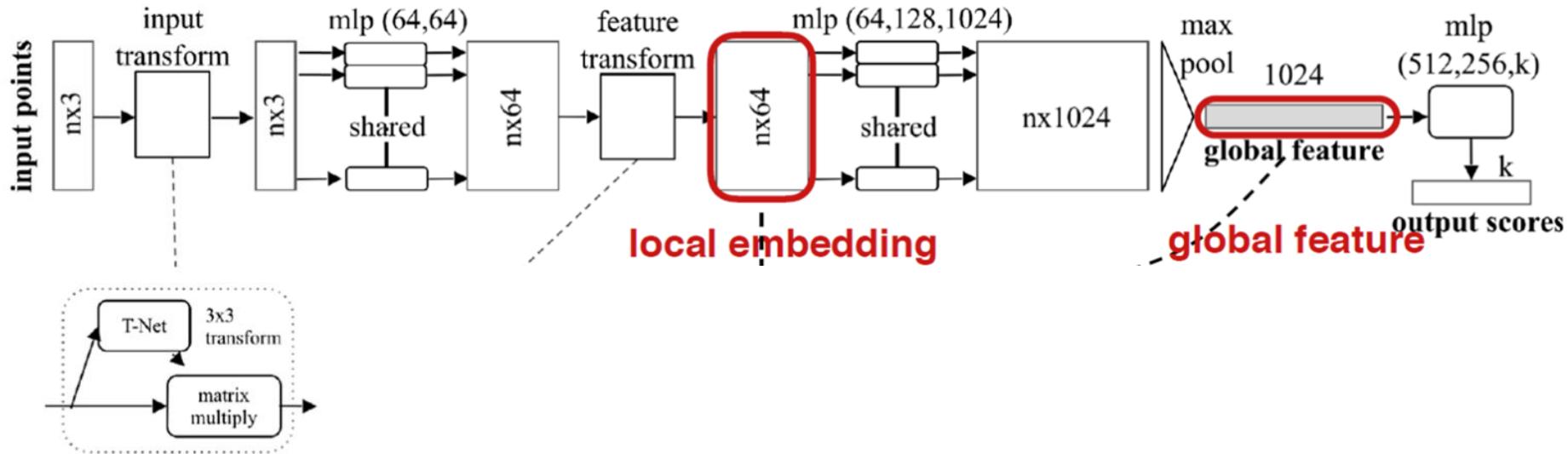
Point cloud classification



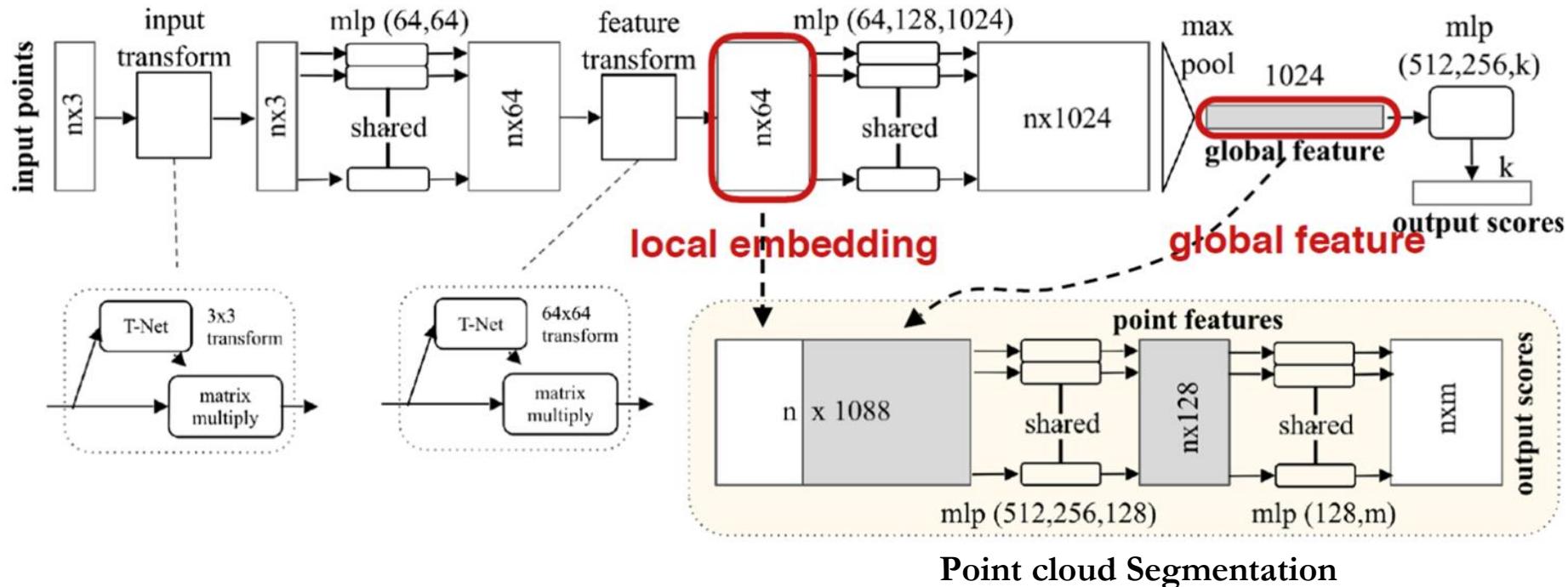
Point cloud classification



Point cloud classification



Point cloud classification and Segmentation



Important points for global features:

Original sets of points



Points represented in the global feature



Which ones are these points?

Saliency Study

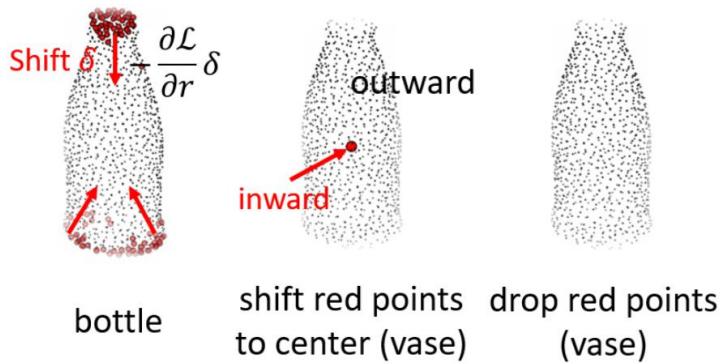


Figure 2. Approximate point dropping with point shifting toward the point-cloud center.

- Given an input point cloud P and the associated class c , first of all we compute the center of P as the median of the three individual coordinates of the points:

$$\mathbf{p}_m = (\text{median}(\mathbf{P}_x), \text{median}(\mathbf{P}_y), \text{median}(\mathbf{P}_z)). \quad (12)$$

- For each point, we compute the vector that connects the center \mathbf{p}_m to it:

$$\mathbf{r}_i = (\mathbf{p}_i - \mathbf{p}_m) \quad (13)$$

- We cast P and c through the network, obtaining the output offsets \mathbf{S}_K . We use them to compute L_{off} as reported in the main manuscript.

- For each input point \mathbf{p}_i of the point cloud, we recover the gradient by backpropagation:

$$\mathbf{g}_i = \nabla_{\mathbf{p}_i} L_{off} \quad (14)$$

- We construct the point-wise saliency map as:

$$s_i = -\|\mathbf{r}_i\|_2 (\mathbf{r}_i \cdot \mathbf{g}_i) \quad (15)$$

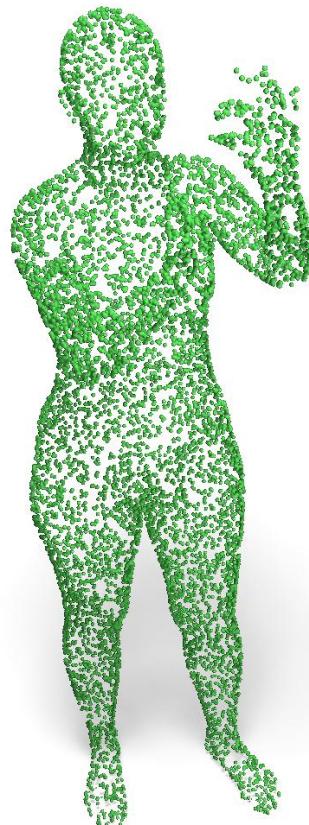
- We pick the 90 input points (1% of the point cloud) associated to the top saliency scores, and we shift the position of each of these points toward the shape median:

$$\tilde{\mathbf{p}}_j = \mathbf{p}_j - 0.05\mathbf{r}_i \quad (16)$$

- We substitute these values in the original point cloud, and we restart the procedure from the beginning for 10 times.

Input

Point Cloud



Output

Posed object



Object pop-up



Object pop-up: Can we infer 3D objects and their poses from human interaction alone?

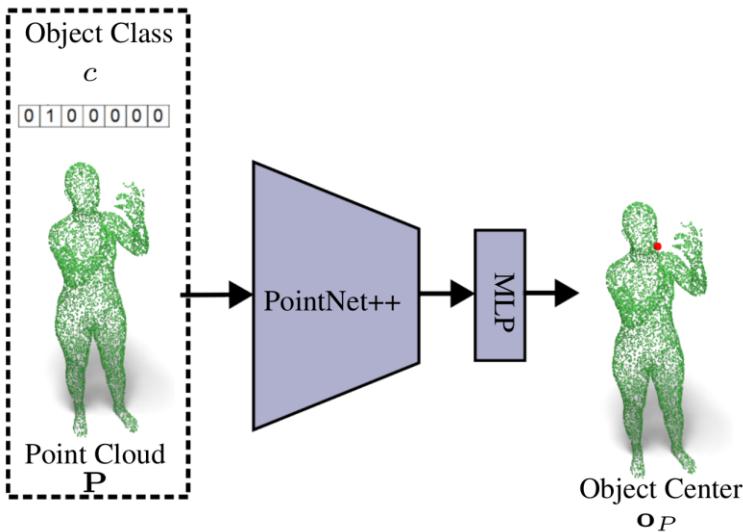
Illya A. Petrov^{1,2}Riccardo Marin^{1,2}Julian Chibane^{1,3}Gerard Pons-Moll^{1,2,3}

¹University of Tübingen; ²Tübingen AI Center;

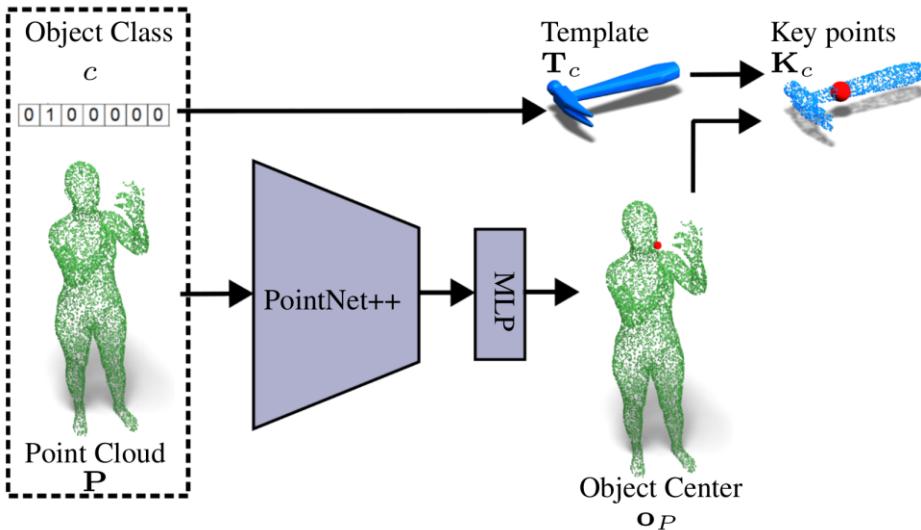
³Max Planck Institute for Informatics, Saarland Informatics Campus

https://virtualhumans.mpi-inf.mpg.de/object_popup/

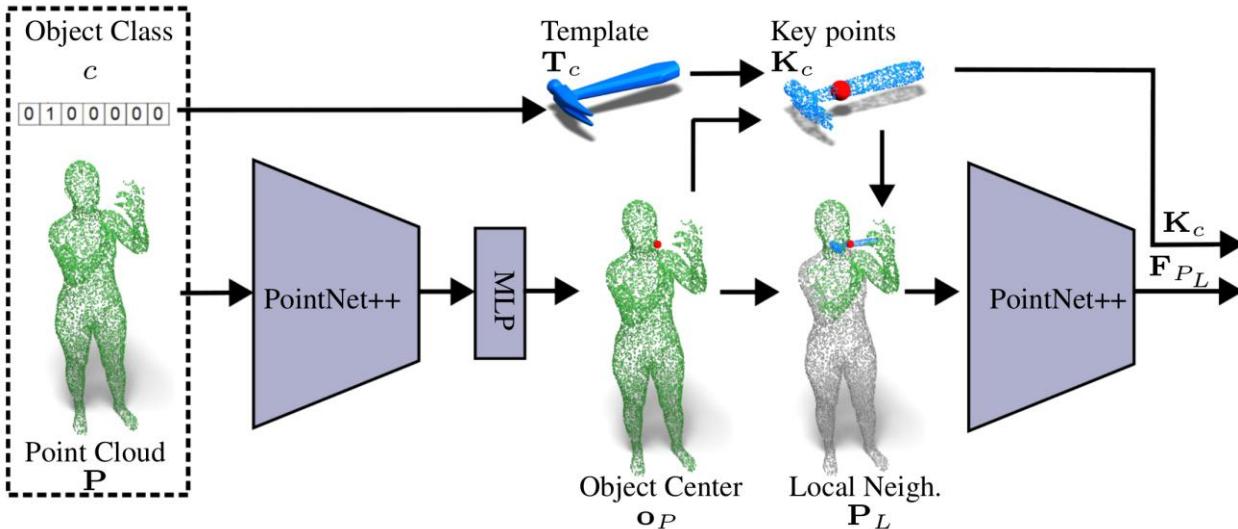
Object pop-up: whole body features and center prediction



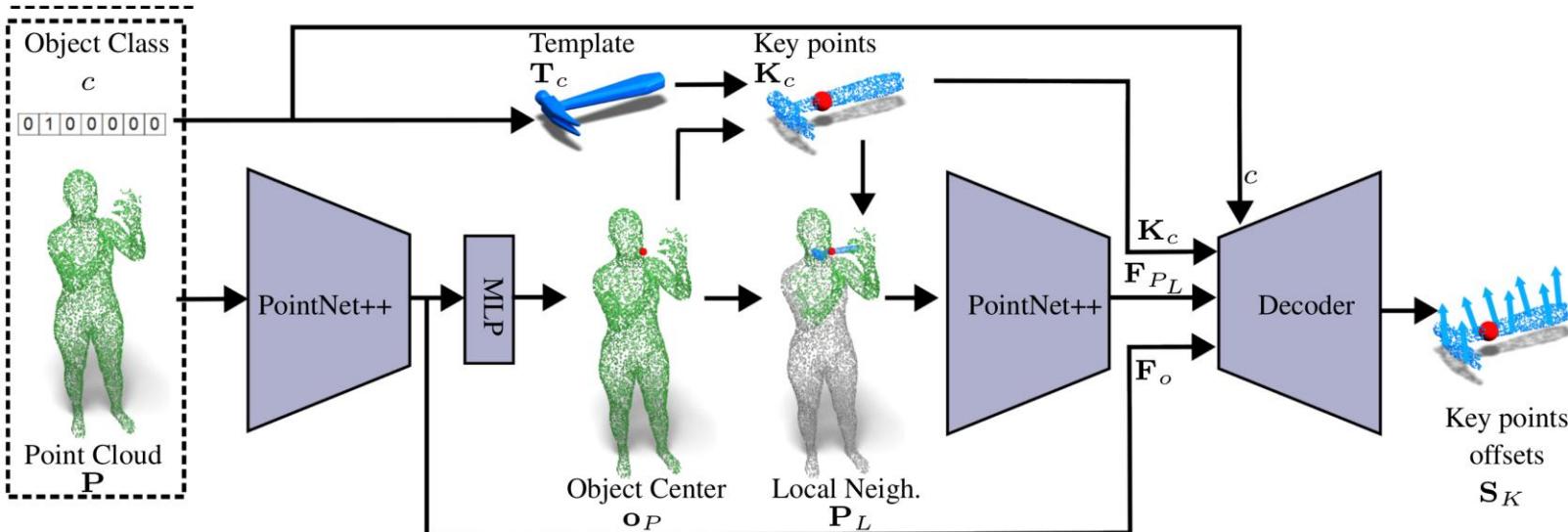
Object pop-up: object keypoints



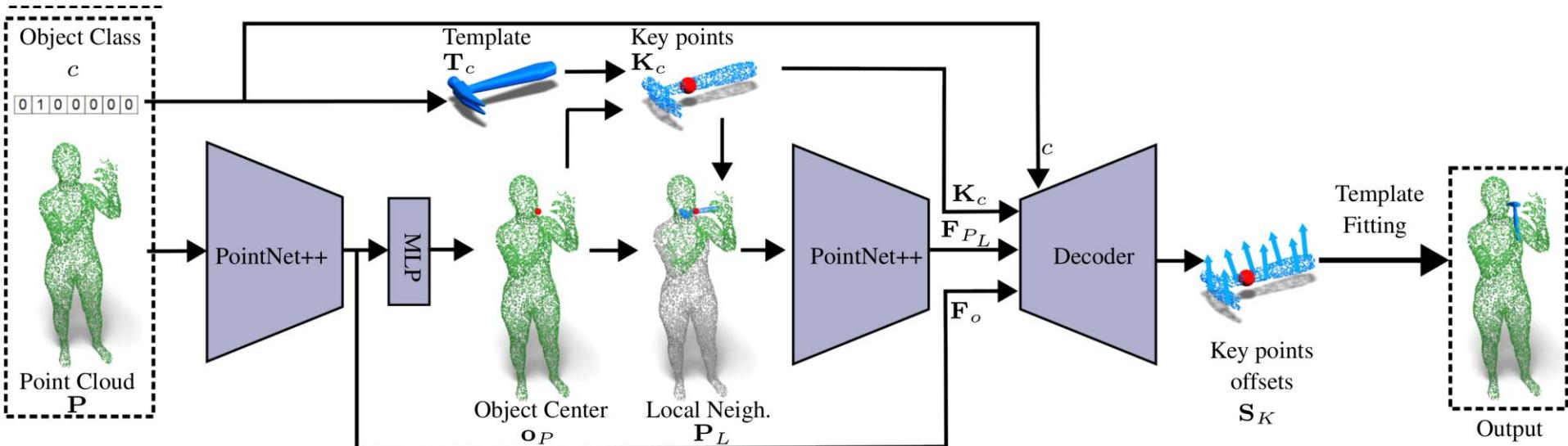
Object pop-up: local features



Object pop-up: per-point offset prediction



Object pop-up: overview



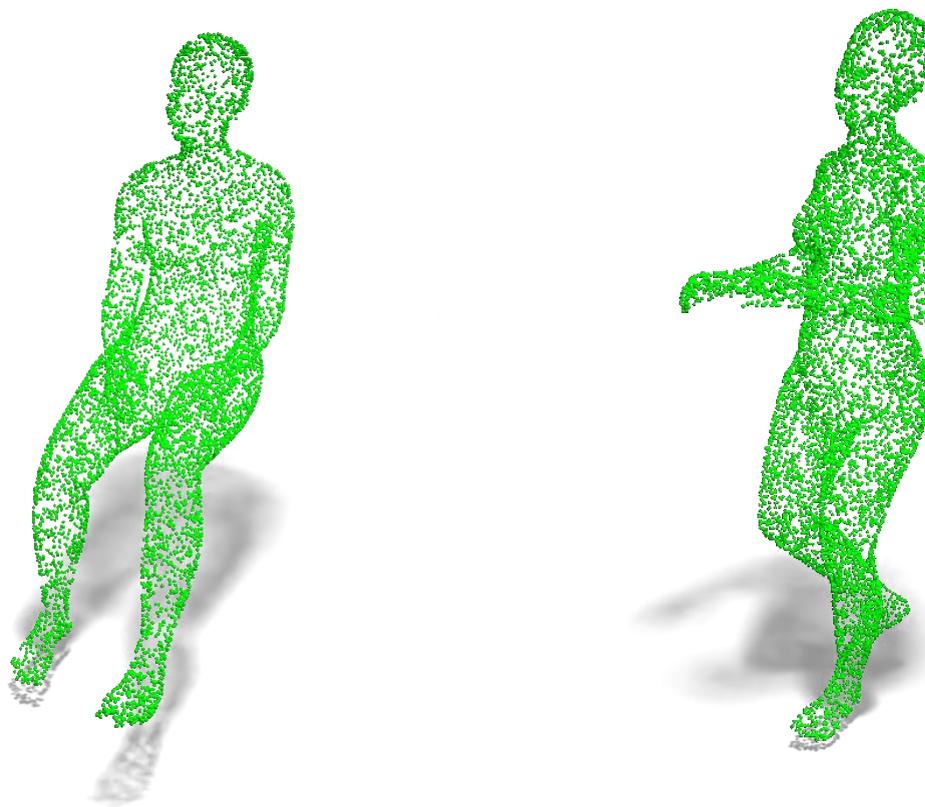


Input



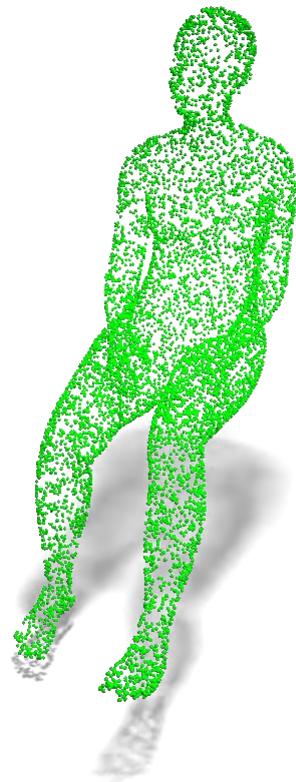
Input

Can we classify an object of interaction?

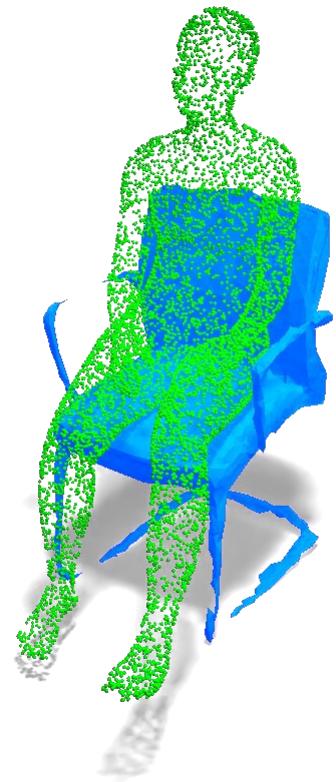


Yes

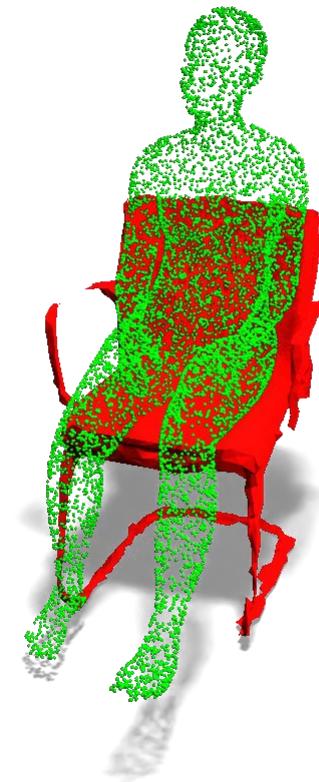
Input



Prediction

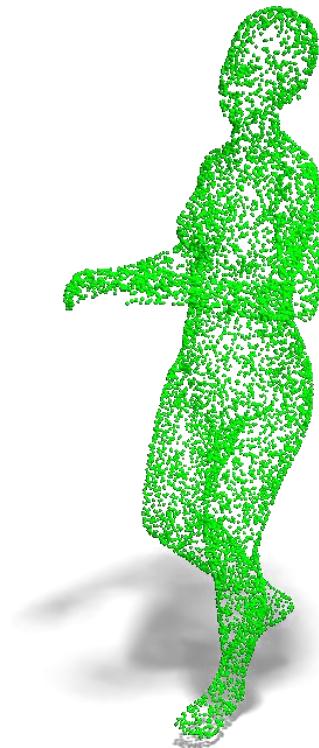


Ground-truth

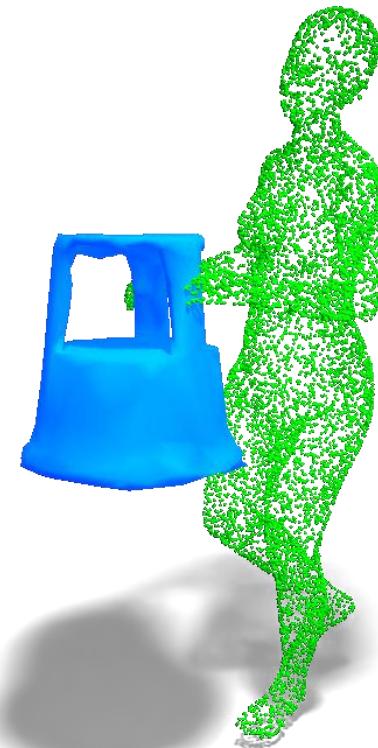


But the task is inherently ambiguous

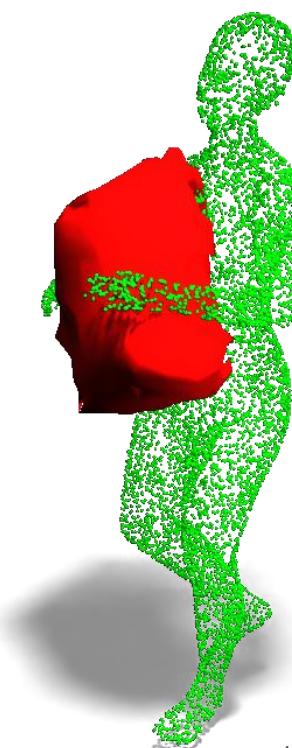
Input

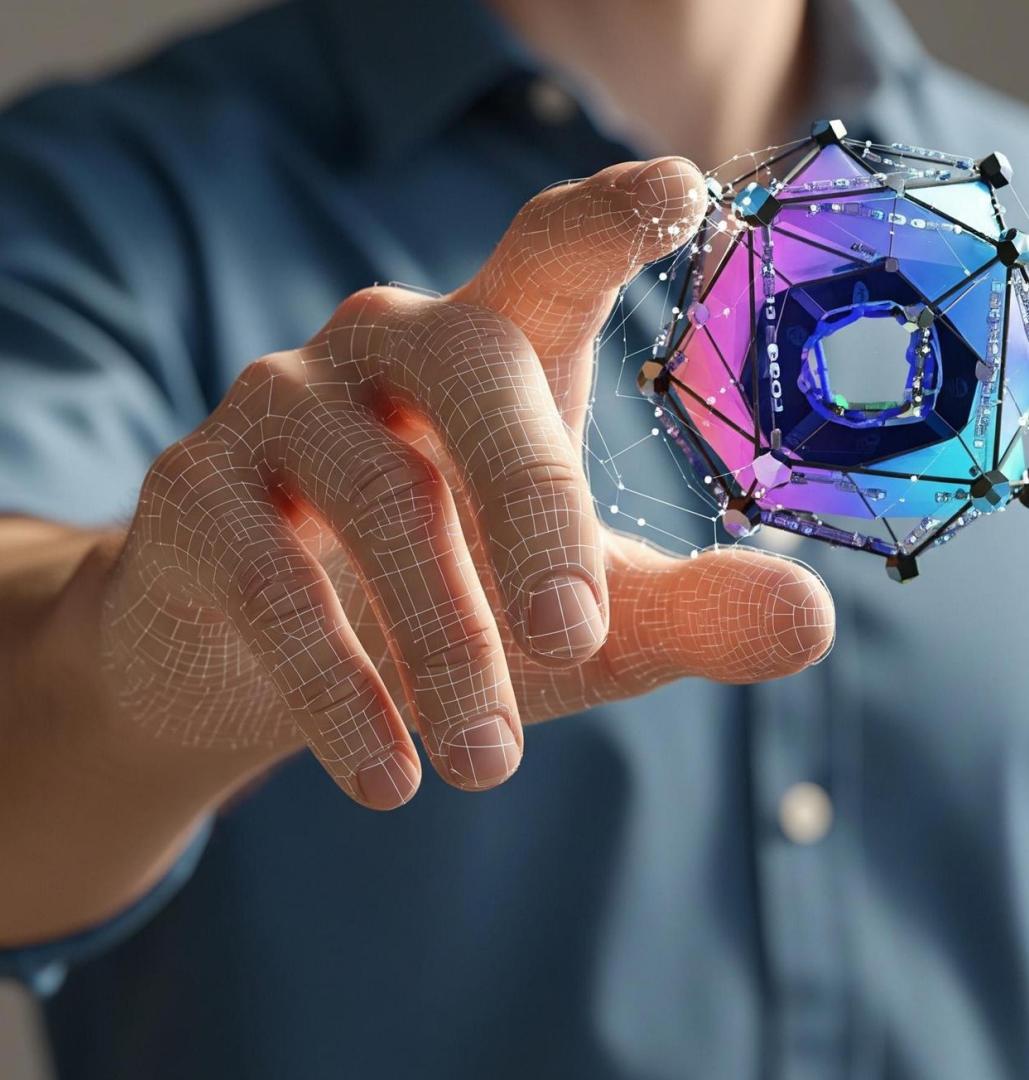


Prediction



Ground-truth

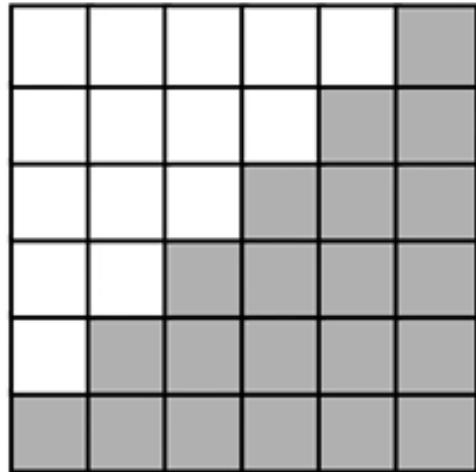




Neural Field
Geometry as networks

Voxels:

- ▶ **Discretization** of 3D space into grid
- ▶ Easy to process with neural networks
- ▶ Cubic memory $O(n^3)$ ⇒ limited resolution
- ▶ Manhattan world bias

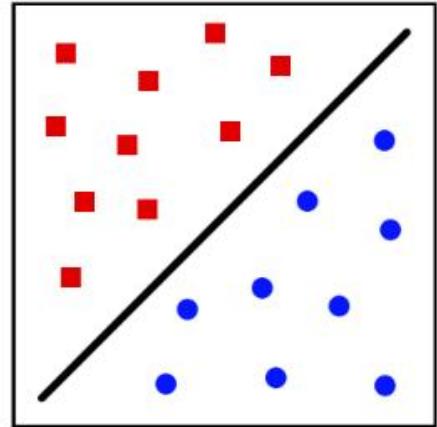


[Maturana et al., IROS 2015]

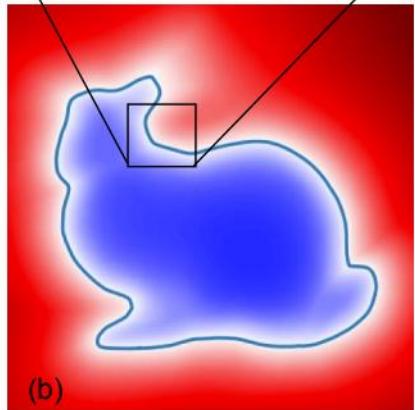
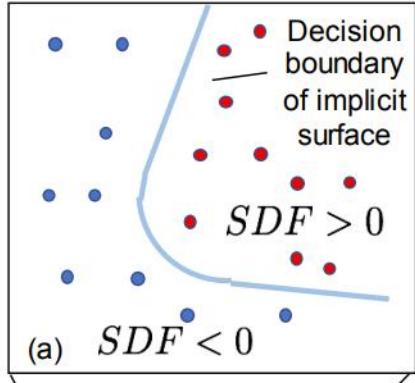


Key Idea:

- ▶ Do not represent 3D shape explicitly
- ▶ Instead, consider surface **implicitly**

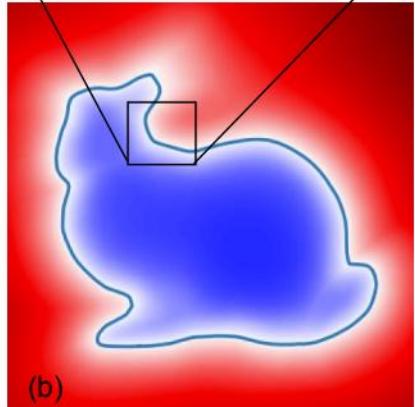
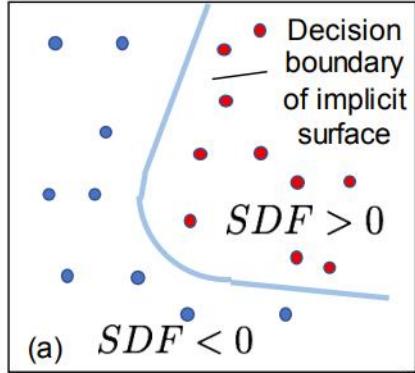


Continuous SDF



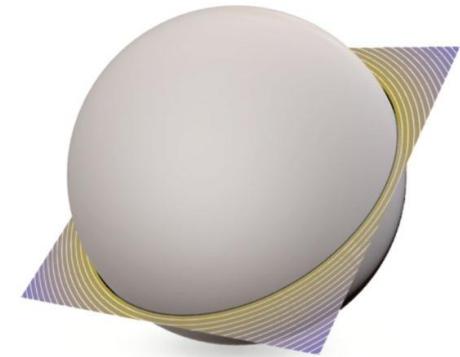
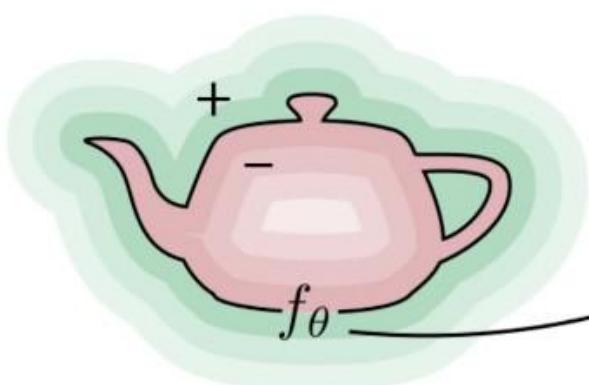
Continuous SDF

<https://pbs.twimg.com/media/EswYEU3UcAAqgsK.jpg:large>



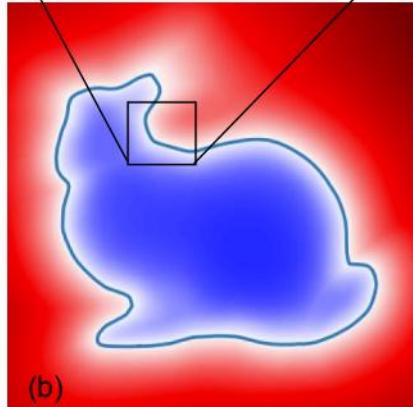
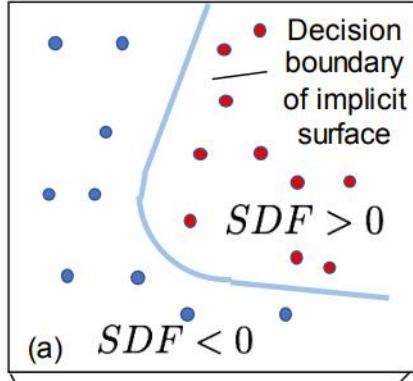
Problem
analytical form is rare

Idea:
Use a parametric function



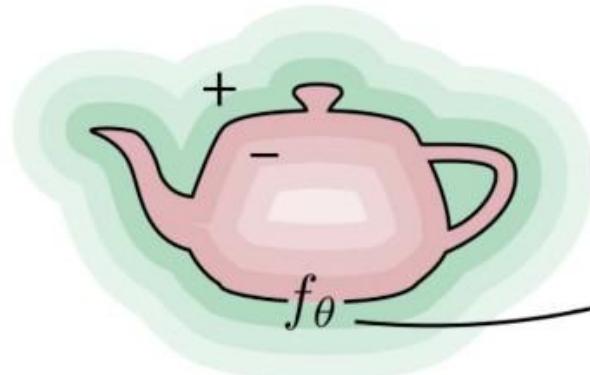
$$f(x, y, z) = \sqrt{x^2 + y^2 + z^2} - 1$$

Continuous SDF



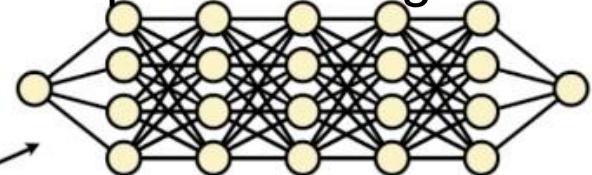
Problem
analytical form is rare

Idea:
Use a parametric function

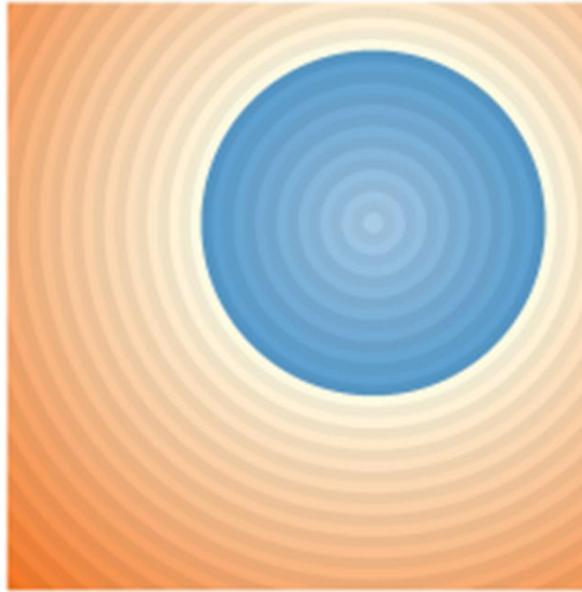


$$f(x, y, z) = \sqrt{x^2 + y^2 + z^2} - 1$$

Training a neural network
to represent the geometry



Voxel - Signed distance field/function



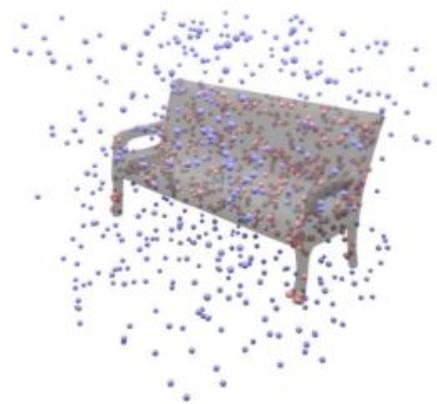
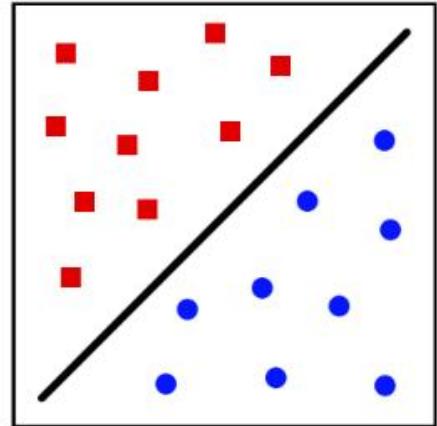
$$f : \mathbb{R}^3 \rightarrow \mathbb{R}, f(x) = \begin{cases} d, & \text{if } x \text{ is outside} \\ -d, & \text{otherwise} \end{cases}$$

Key Idea:

- ▶ Do not represent 3D shape explicitly
- ▶ Instead, consider surface **implicitly** as **decision boundary** of a non-linear classifier:

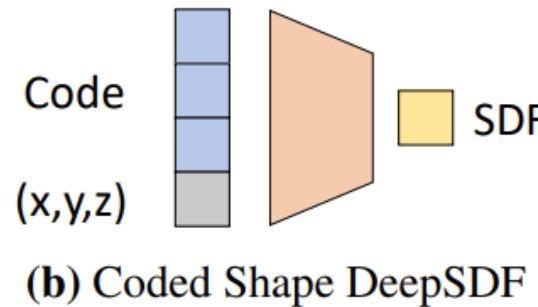
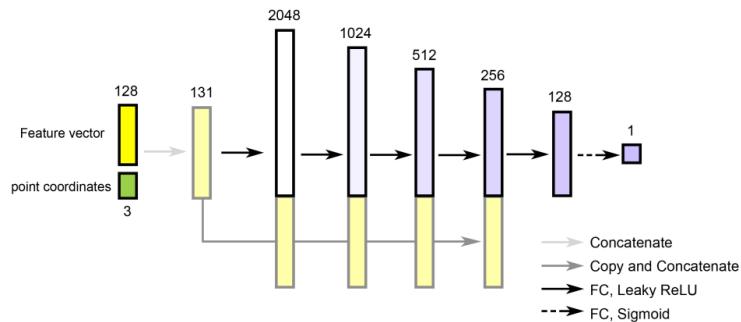
$$f_{\theta} : \mathbb{R}^3 \times \mathcal{X} \rightarrow [0, 1]$$

↑ ↑ ↑
3D Location Condition
(eg, Image) Occupancy
Probability



Neural Fields

CVPR19: Three contemporary works proposed almost the same idea

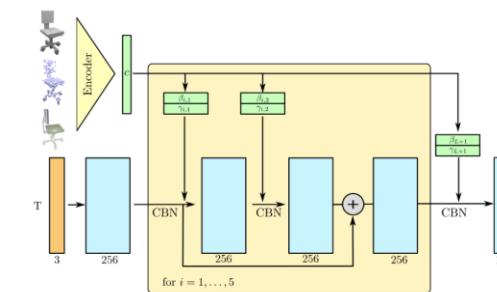


Learning Implicit Fields for Generative Shape Modeling, (IMNET)

$$\mathcal{F}(p) = \begin{cases} 0 & \text{if point } p \text{ is outside the shape,} \\ 1 & \text{otherwise.} \end{cases}$$

DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation

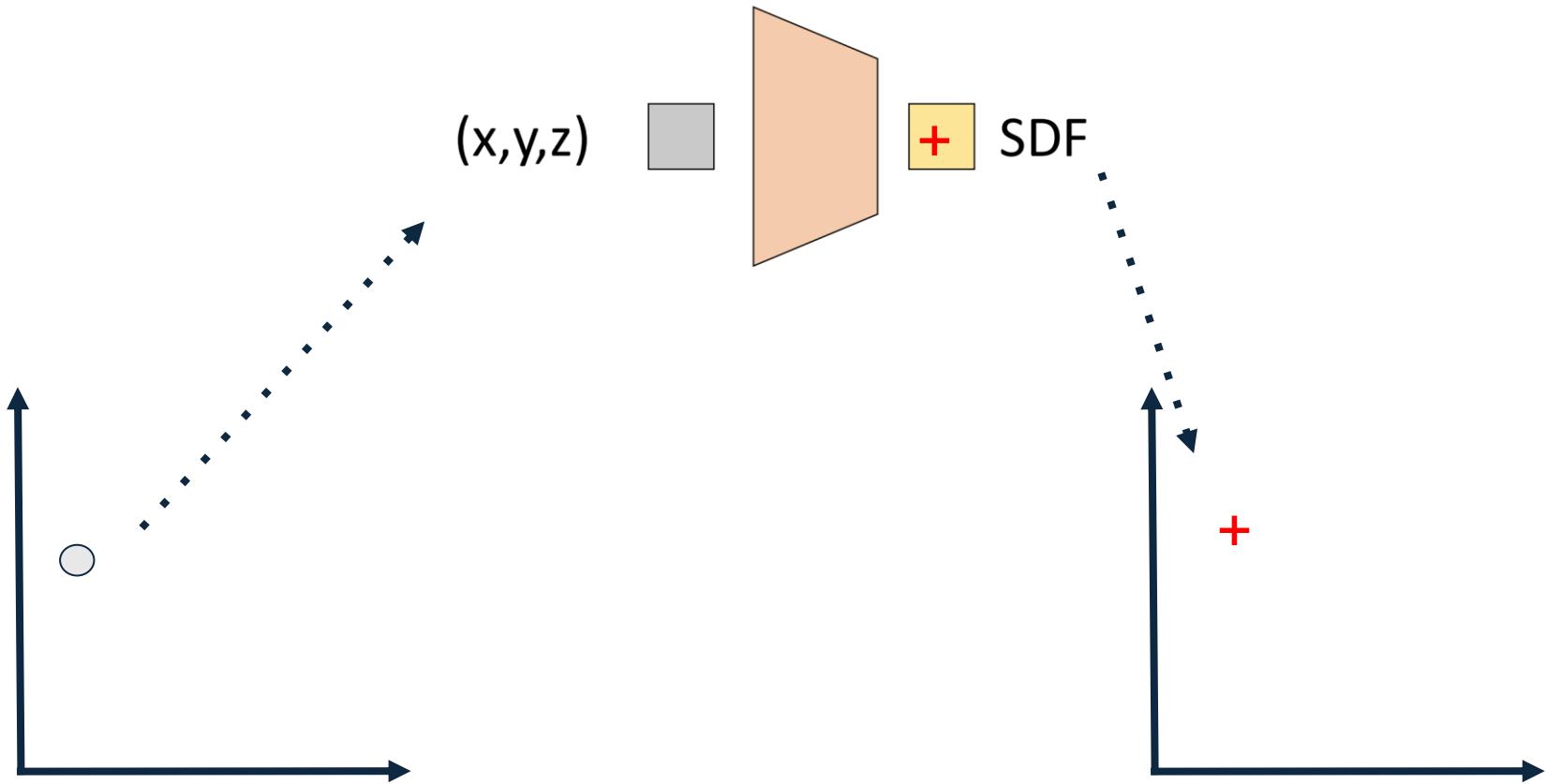
$$f_{\theta}(\mathbf{x}) \approx SDF(\mathbf{x})$$



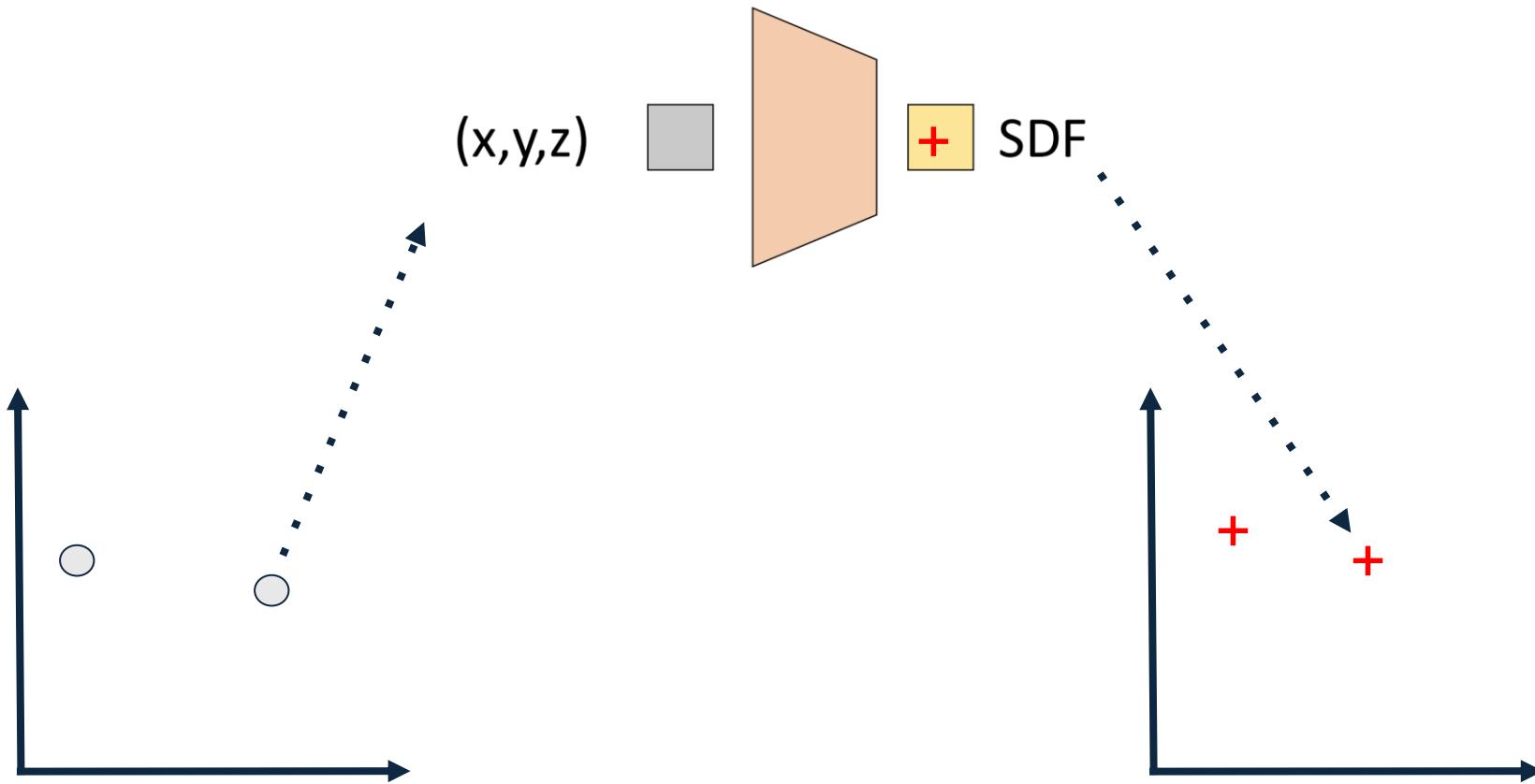
Occupancy Networks:
Learning 3D Reconstruction
in Function Space.

$$f_{\theta} : \mathbb{R}^3 \times \mathcal{X} \rightarrow [0, 1]$$

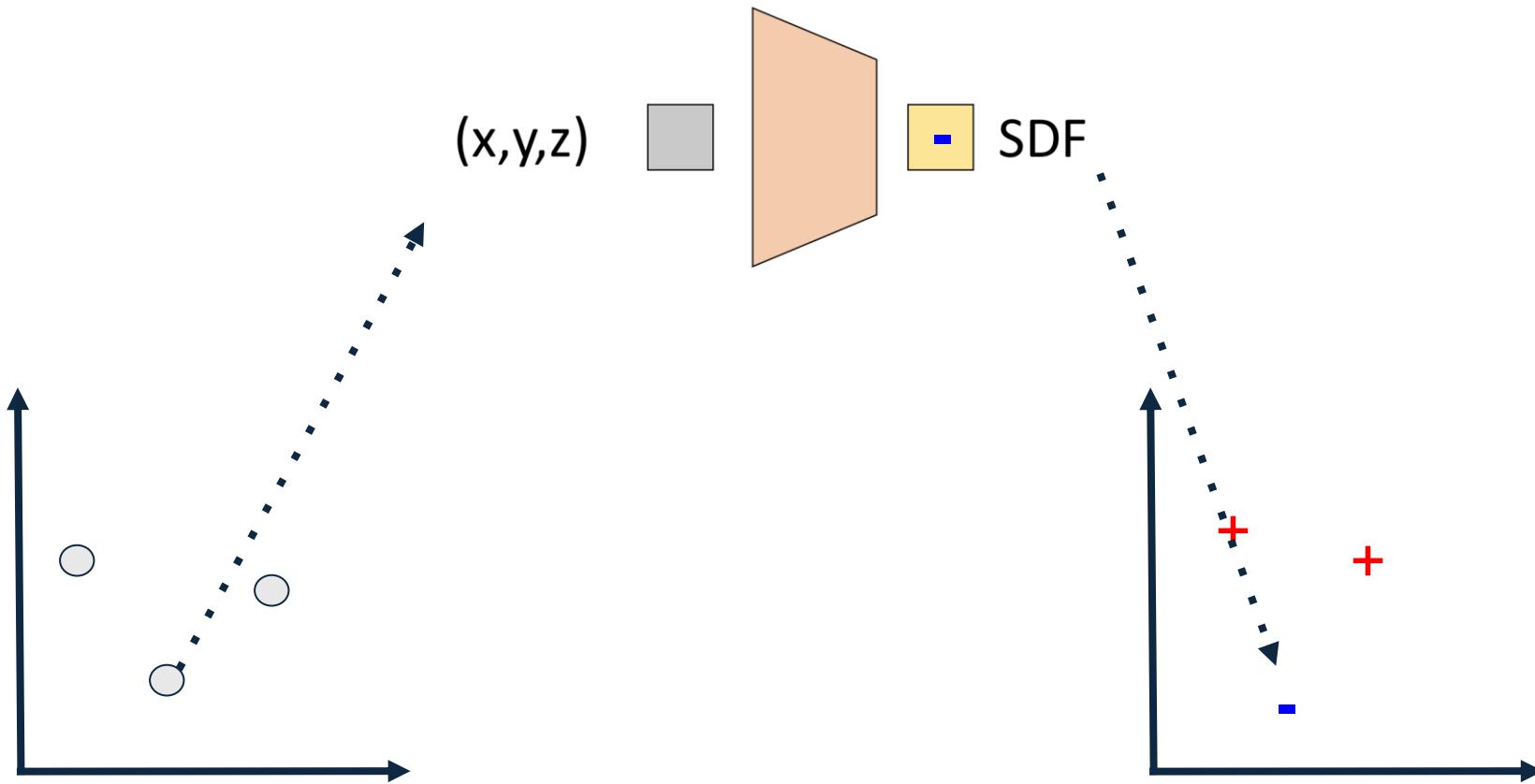
Deep SDF



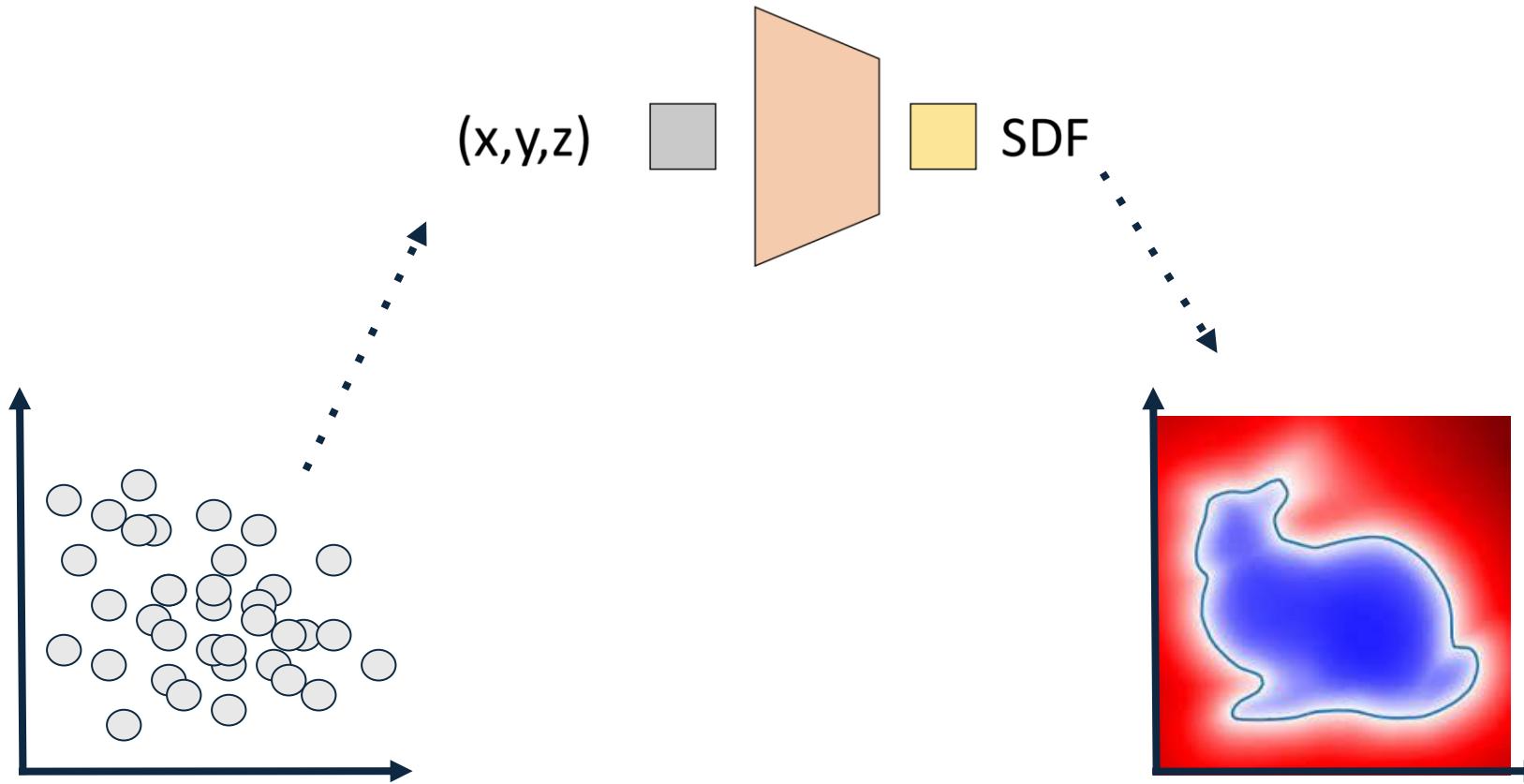
Deep SDF



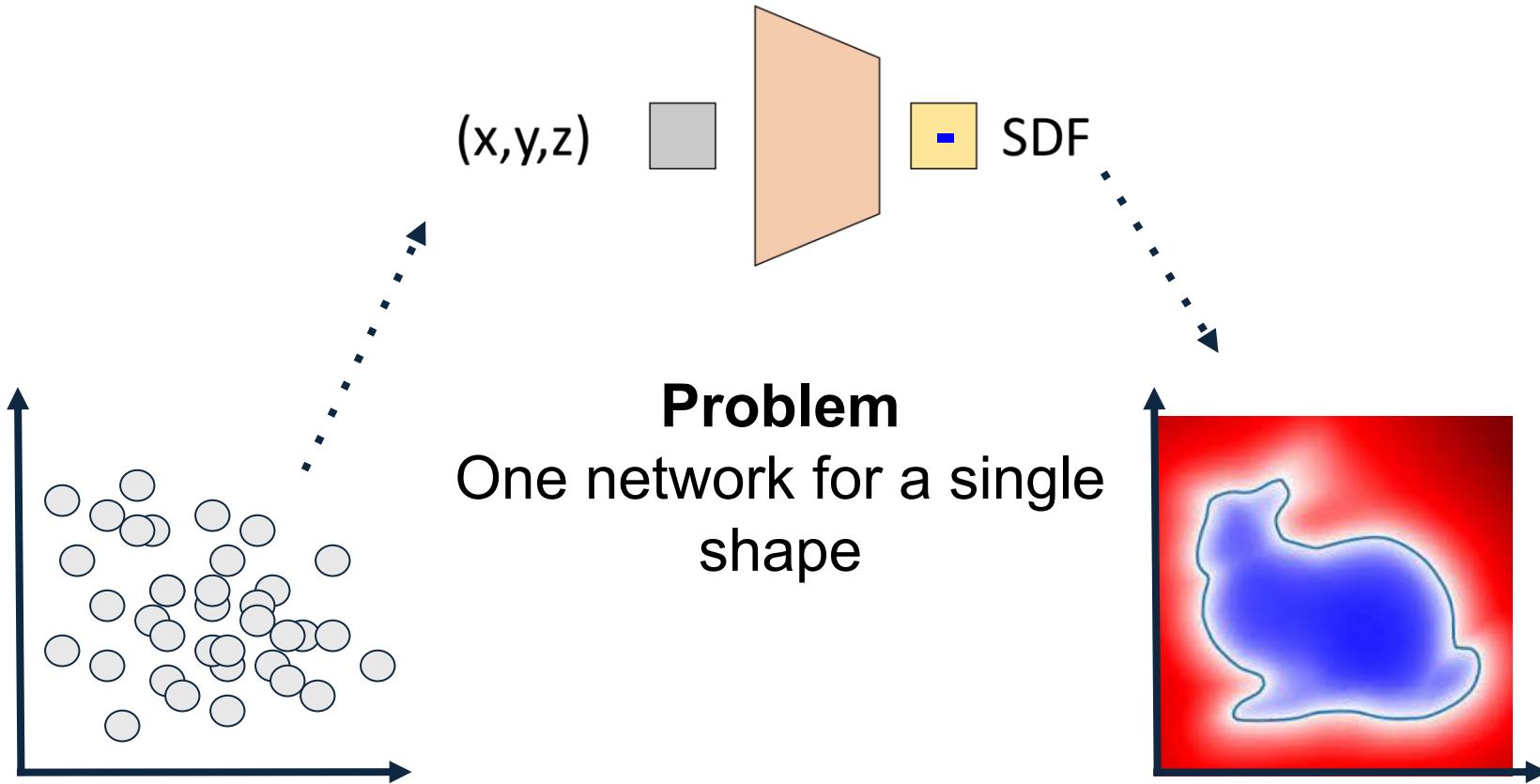
Deep SDF



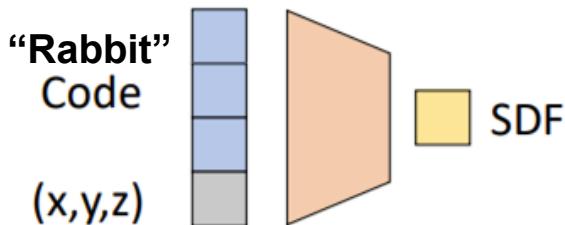
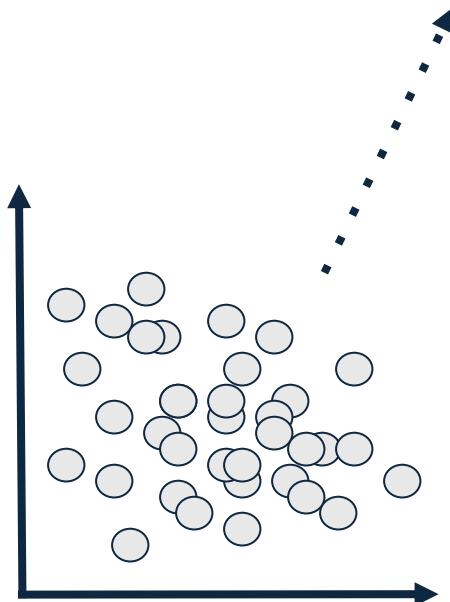
Deep SDF



Deep SDF

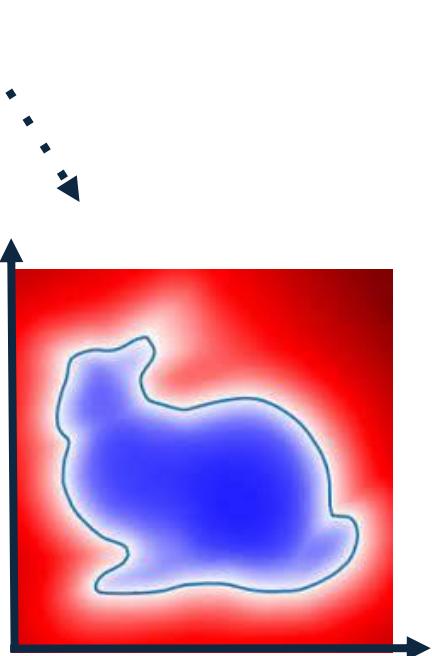


Deep SDF

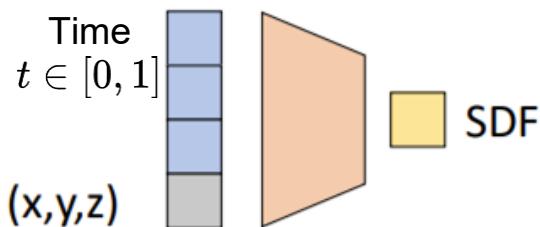
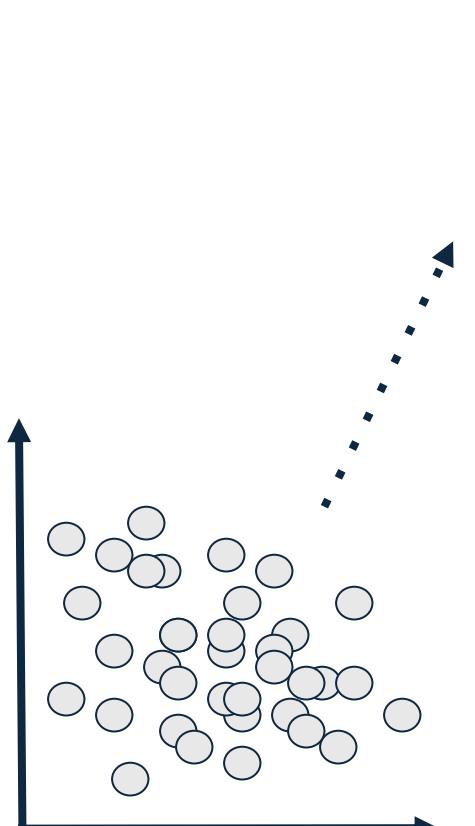


Problem
One network for a single
shape

Solution
Condition the input



Deep SDF

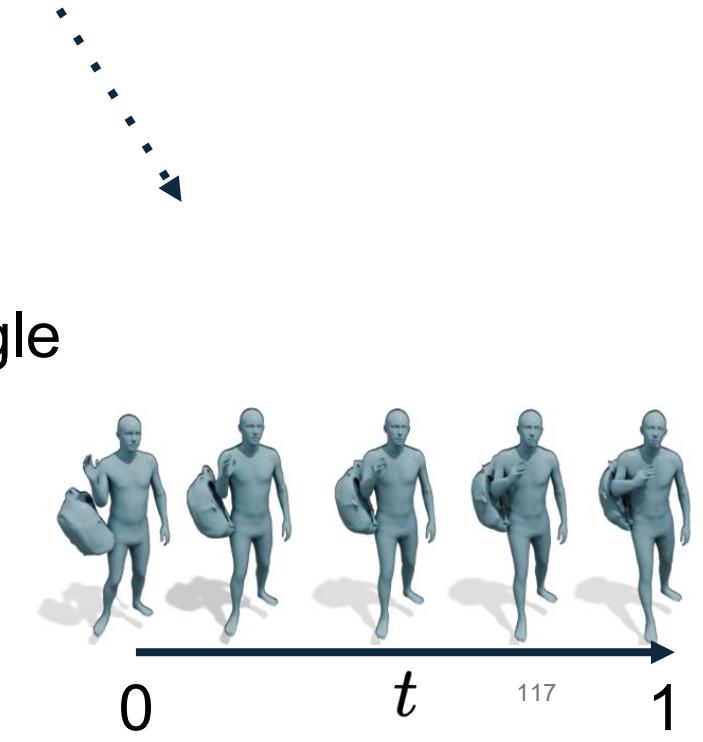


Problem

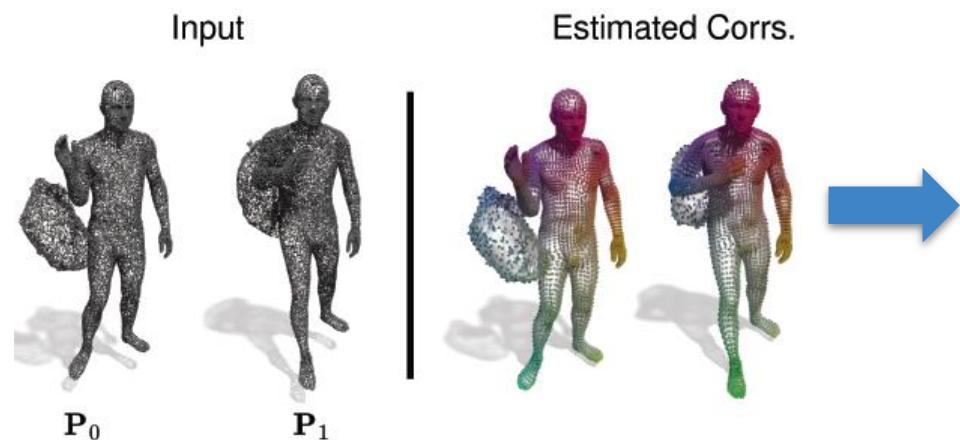
One network for a single shape

Solution

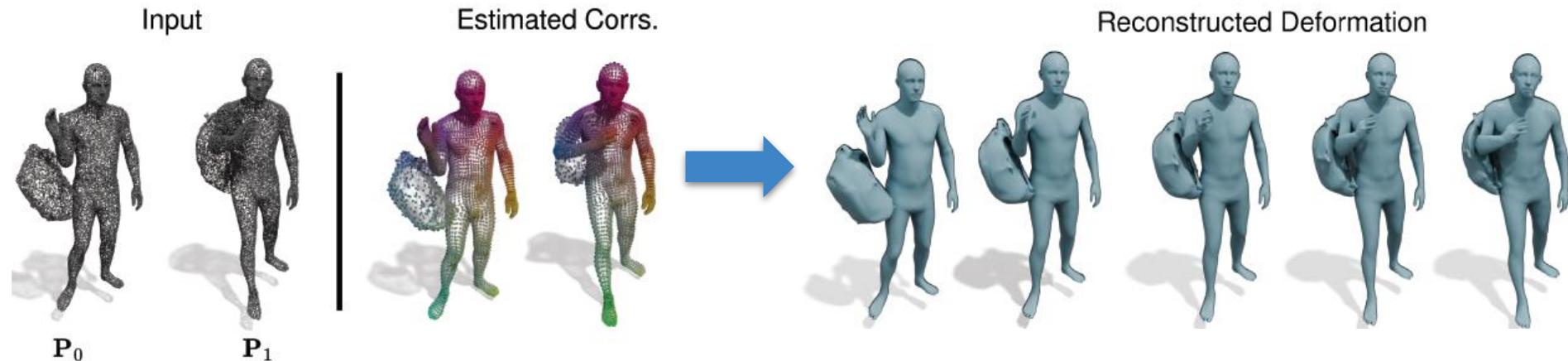
Condition the input



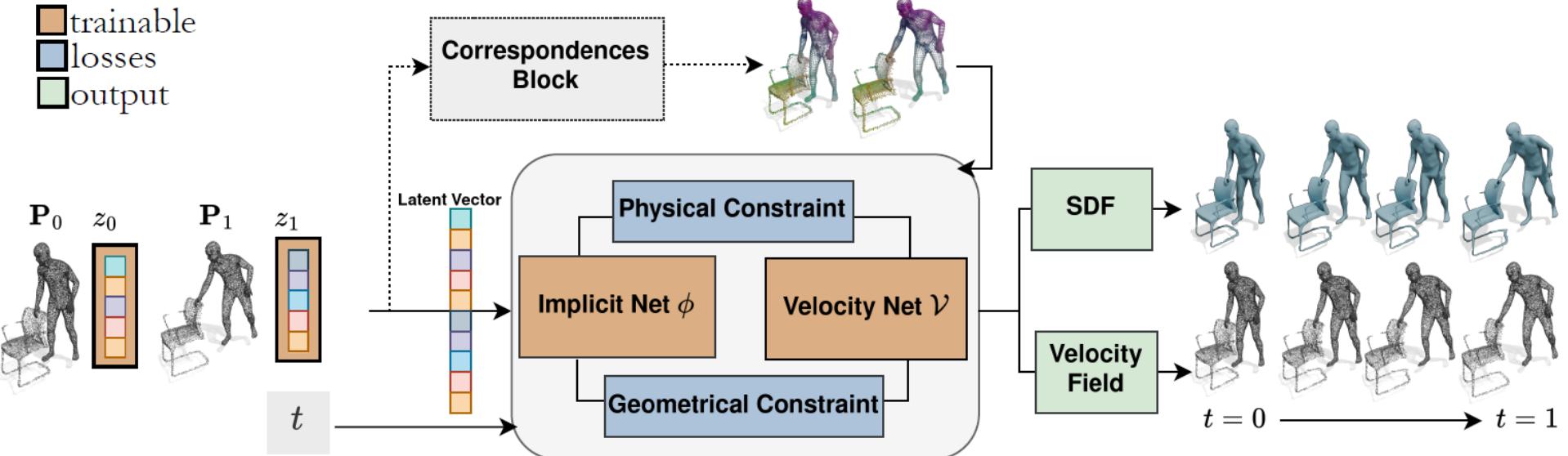
Goal: 3D to 4D reconstruction



Goal: 3D to 4D reconstruction



4Deform: Neural Surface Deformation for Robust Shape Interpolation



Input images



Input images



Input images



Input images



Novel View Synthesis



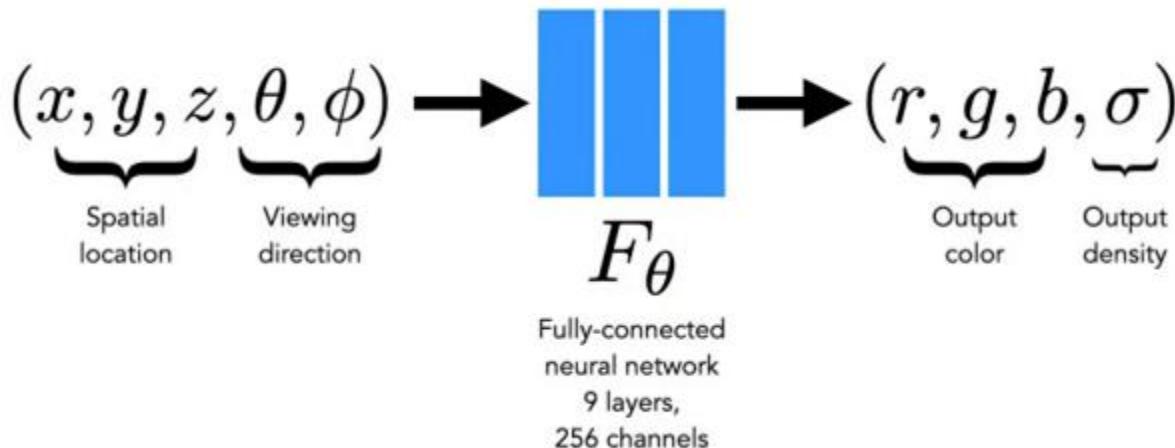
Inputs: sparsely sampled images of scene

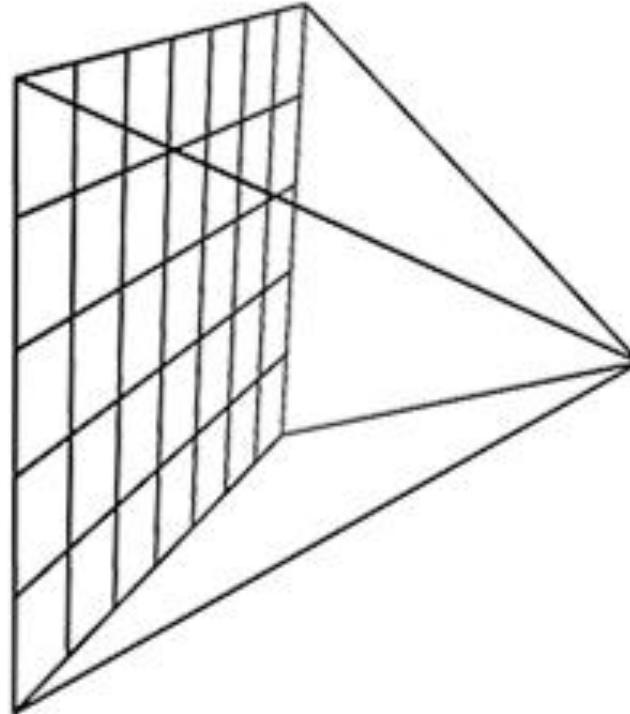


Outputs: *new views of same scene*

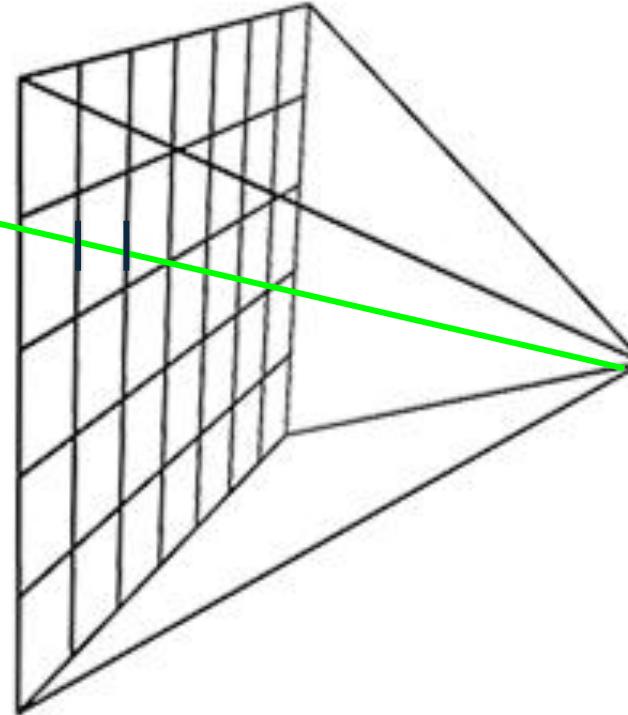
- ▶ **Task:** Given a set of images of a scene, render image from novel viewpoint
- ▶ Slide credits: Ben Mildenhall and Jon Barron

NeRF: Representing Scenes as Neural Radiance Fields

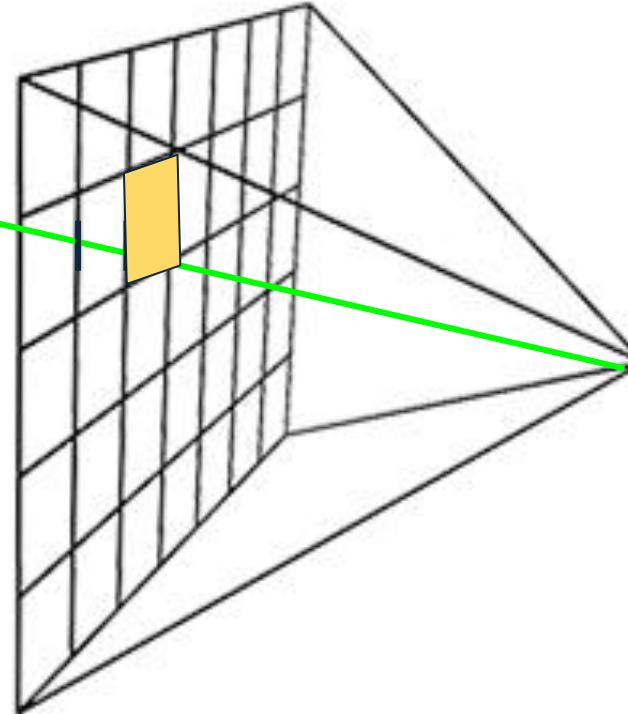




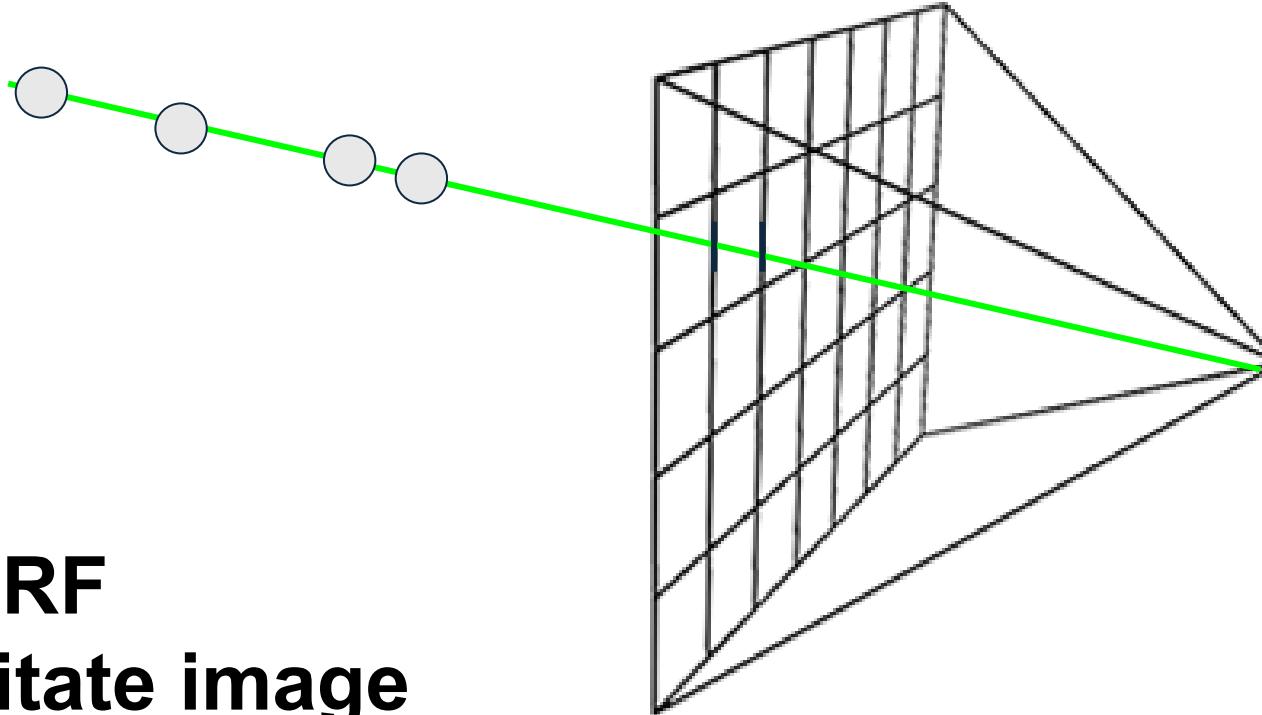
**Standard
image
formation**



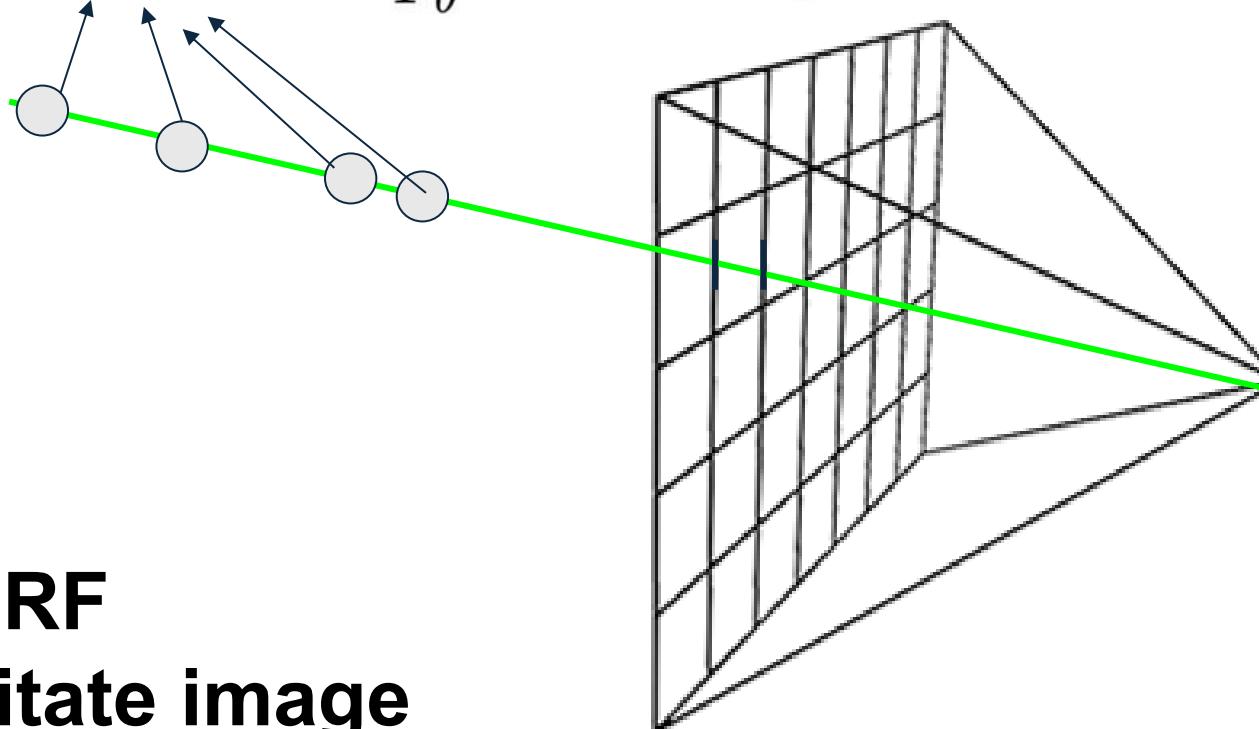
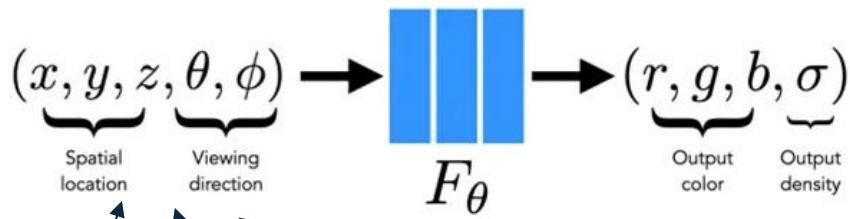
**Standard
image
formation**



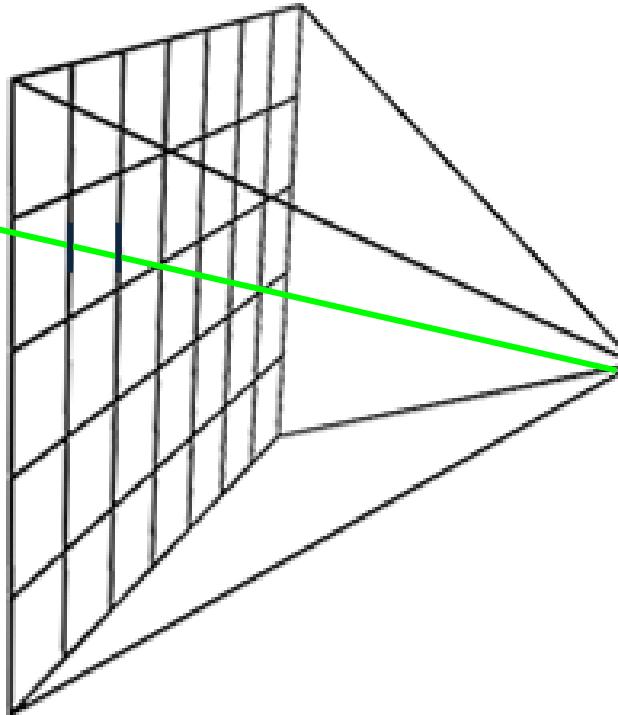
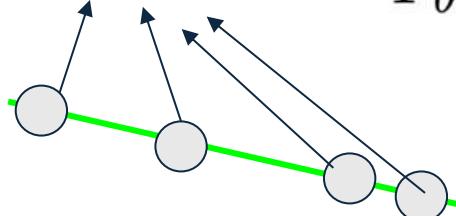
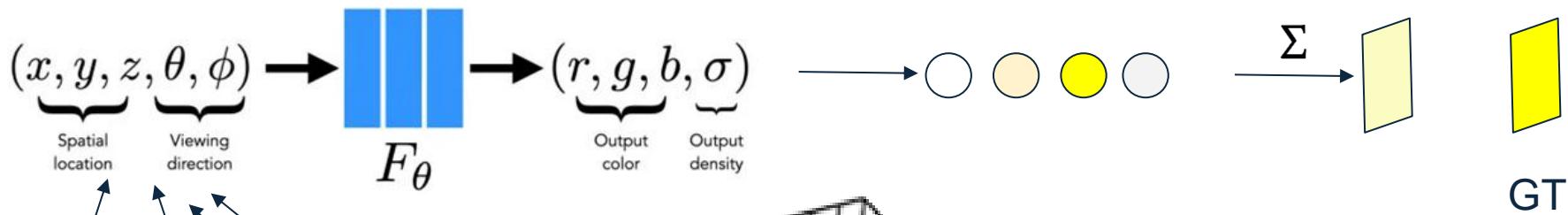
**Standard
image
formation**



NeRF
imitate image
formation

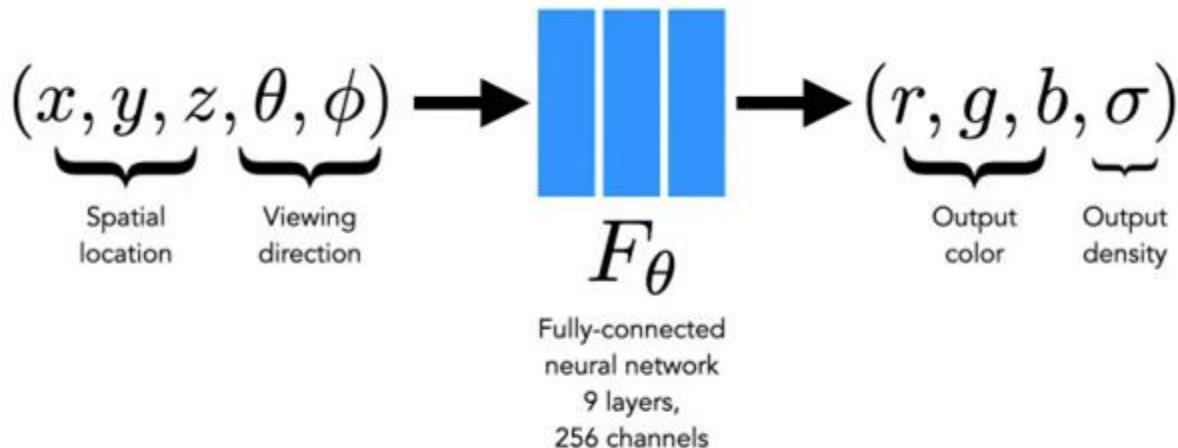


NeRF
imitate image
formation



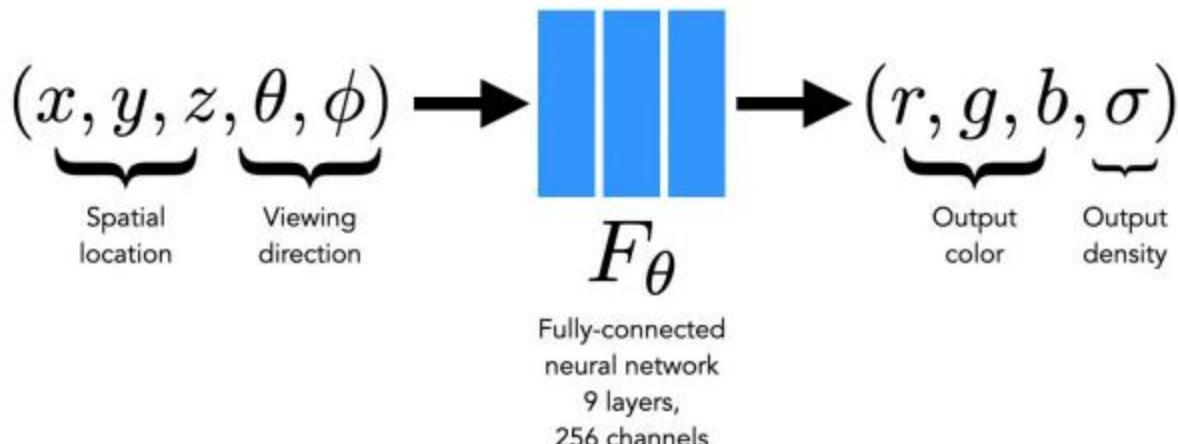
NeRF imitate image formation

NeRF: Representing Scenes as Neural Radiance Fields



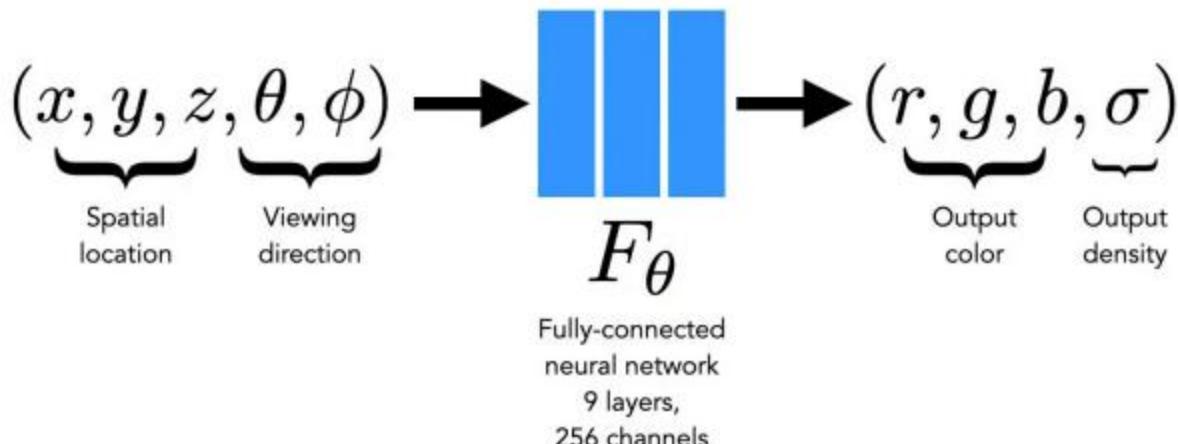
- Vanilla **ReLU MLP** that maps from **location/view direction to color/density**

NeRF: Representing Scenes as Neural Radiance Fields



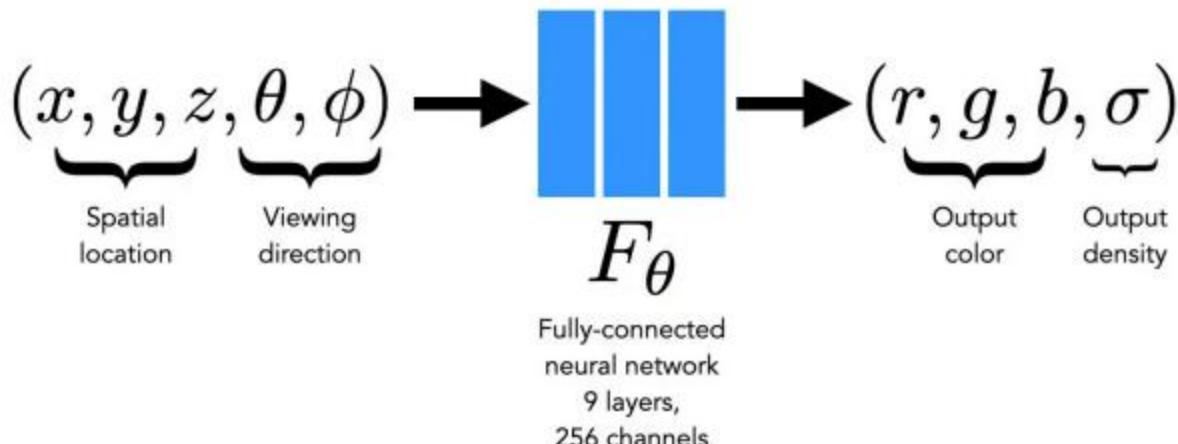
- ▶ Vanilla **ReLU MLP** that maps from **location/view direction to color/density**
- ▶ **Density** σ describes how solid/transparent a 3D point is (can model, e.g., fog)

NeRF: Representing Scenes as Neural Radiance Fields



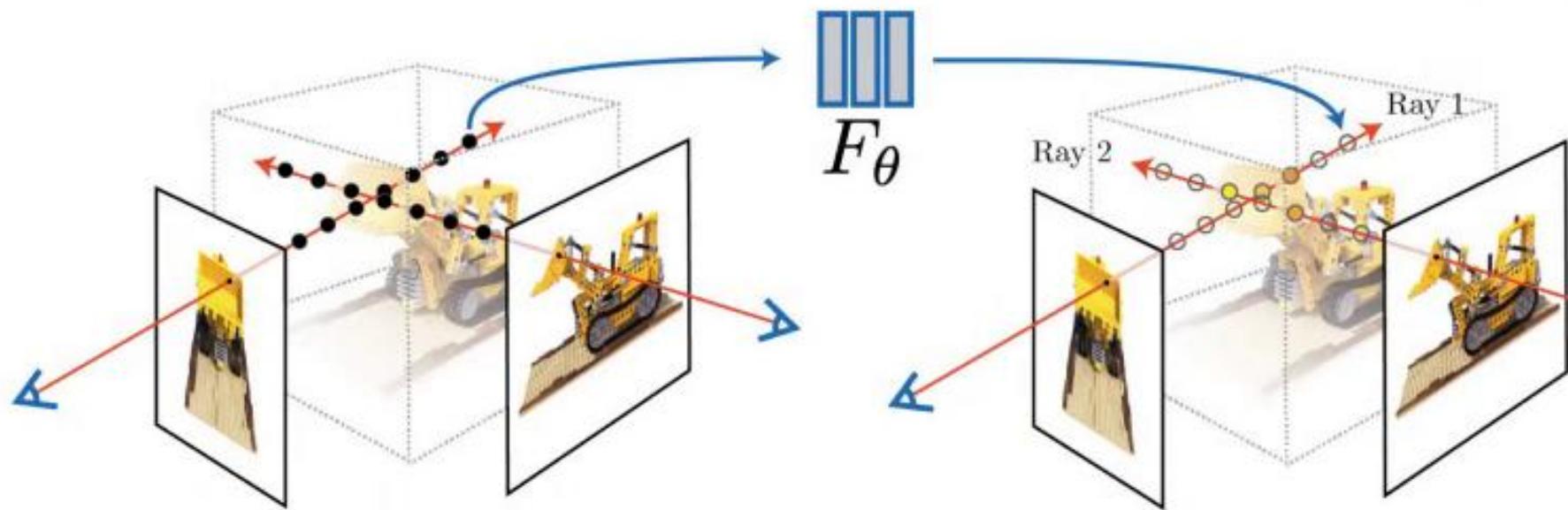
- ▶ Vanilla **ReLU MLP** that maps from **location/view direction to color/density**
- ▶ **Density** σ describes how solid/transparent a 3D point is (can model, e.g., fog)
- ▶ Conditioning on view direction allows for modeling **view-dependent effects**

NeRF: Representing Scenes as Neural Radiance Fields



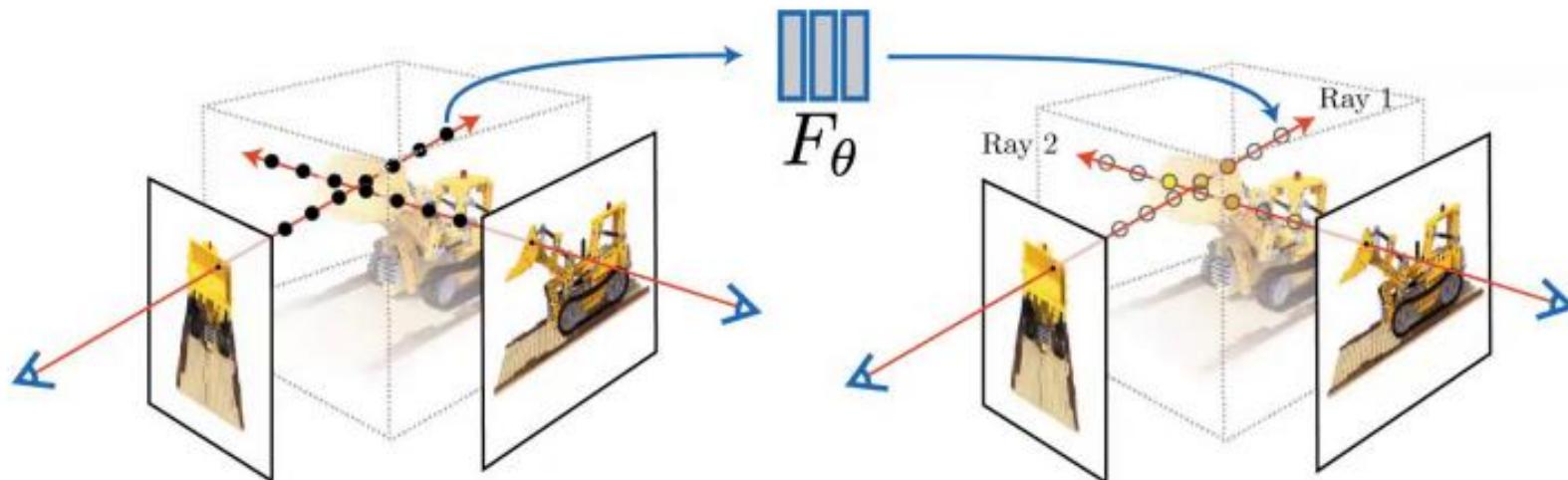
- ▶ Vanilla **ReLU MLP** that maps from **location/view direction to color/density**
- ▶ **Density** σ describes how solid/transparent a 3D point is (can model, e.g., fog)
- ▶ Conditioning on view direction allows for modeling **view-dependent effects**
- ▶ In practice, the view direction is input as a normalized 3D vector \mathbf{d} , not (θ, ϕ)

Volume Rendering



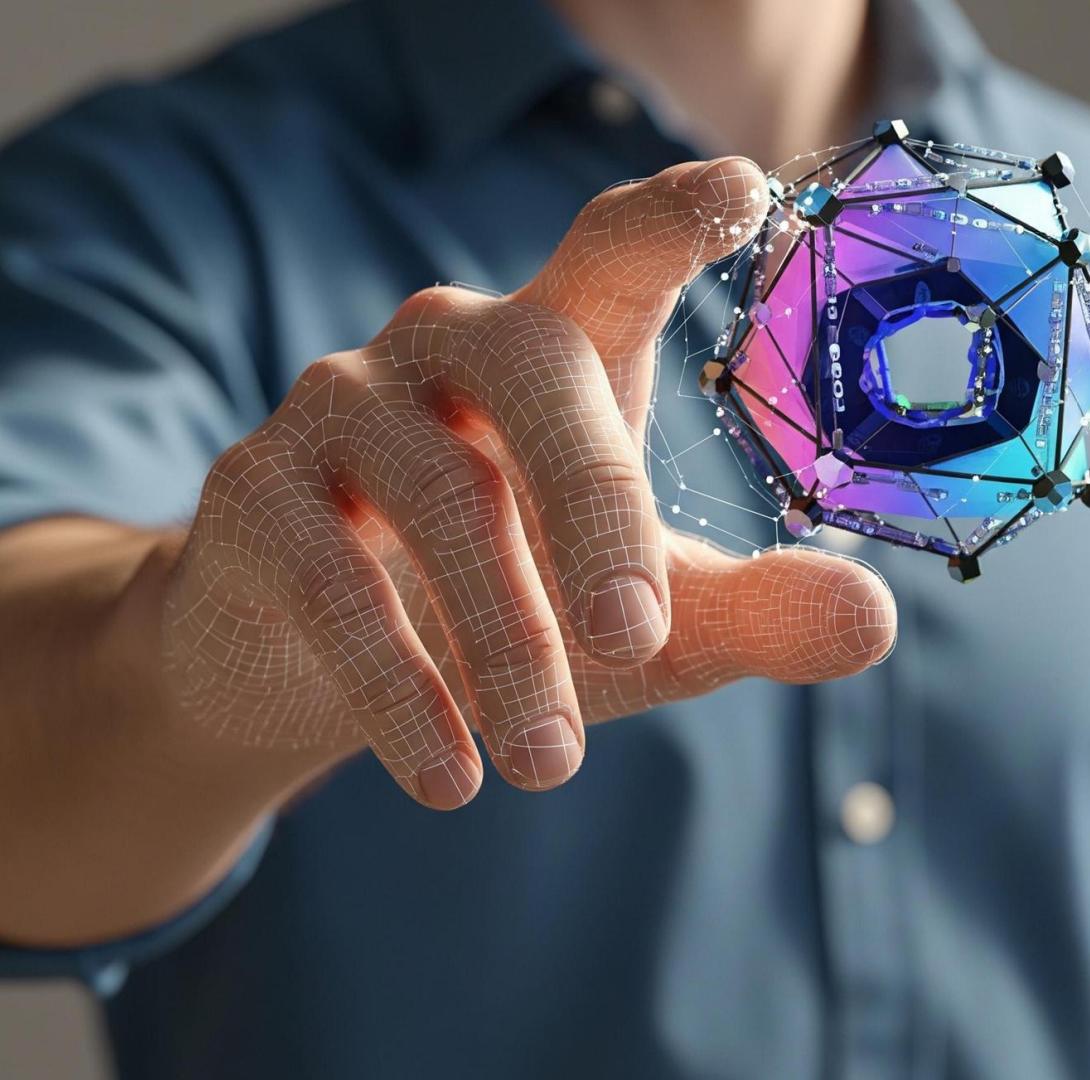
- ▶ **Volume rendering** works very similar to traditional **ray tracing** in graphics
- ▶ Shoot ray through the scene, sample points along ray, query radiance field to obtain color/density, apply alpha composition to obtain pixel color

NeRF Training



$$\min_{\theta} \sum_i || \text{render}_i(F_\theta) - I_i ||^2$$

- Shoot ray, render ray to pixel, minimize **reconstruction error** via backpropagation



Geometric Deep Learning for Virtual Humans

Intro & Geometric DL

Riccardo Marin



20th November 2025