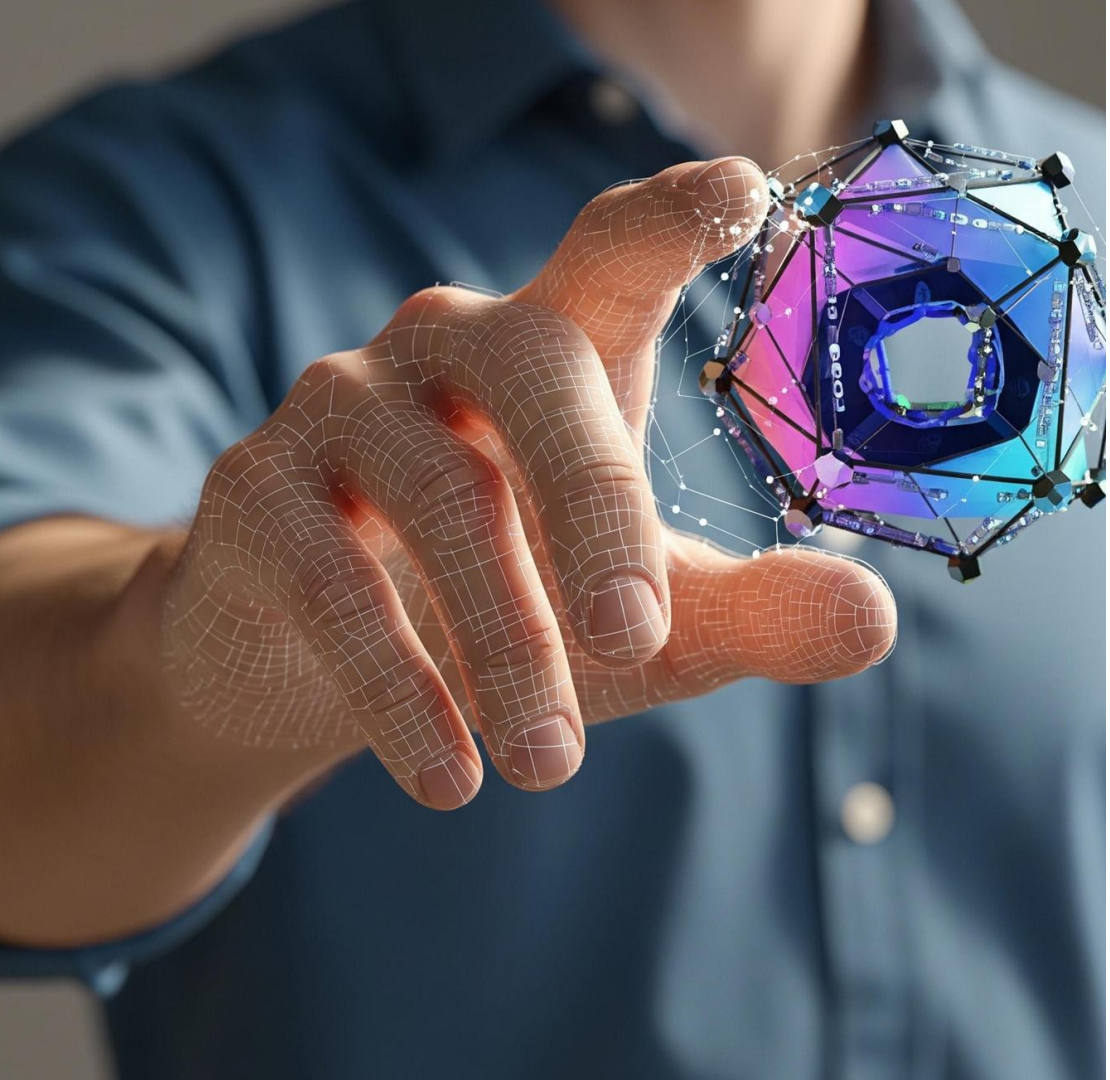


Geometric Deep Learning for Virtual Humans

Intro & Geometric DL

Riccardo Marin

20th November 2025

About me

Verona
PhD Computer Science

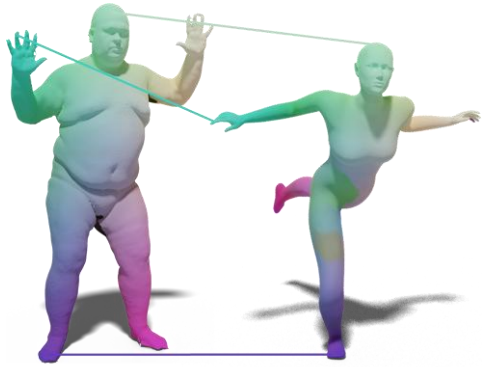
Paris
École Polytechnique
(Visiting)

Rome
Sapienza

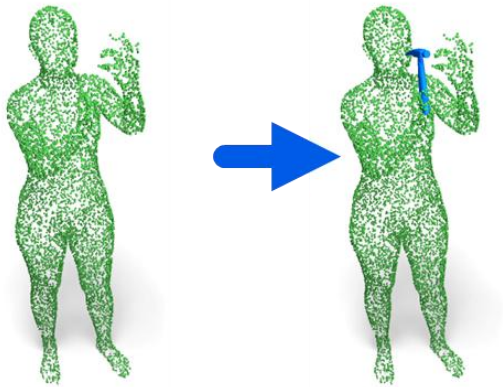
Tuebingen
AI Center

Munich
TUM/MCML

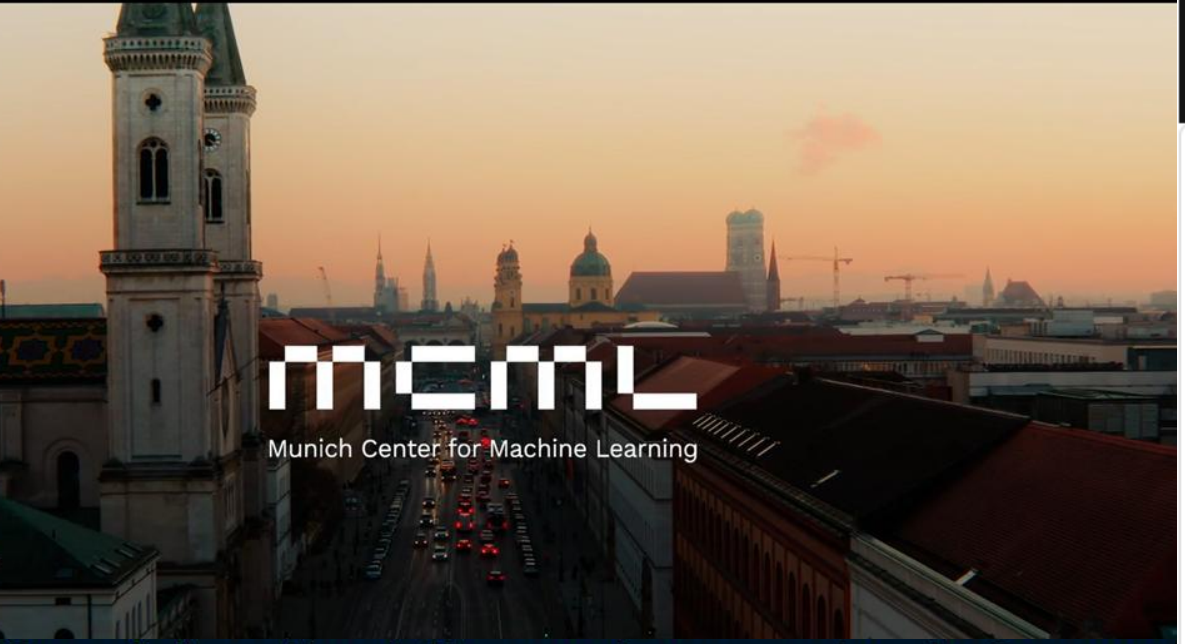
3D Matching/Shape Analysis



Virtual Humans



https://riccardomarin.github.io/



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

LMU LUDWIG-MAXIMILIANS UNIVERSITÄT MÜNCHEN

TUM

≈ 80
Principal Investigators

460+
Junior Members

2210+
Publications

Opportunities

News

For Junior Members

For Research X-Change

For PhD Applicants

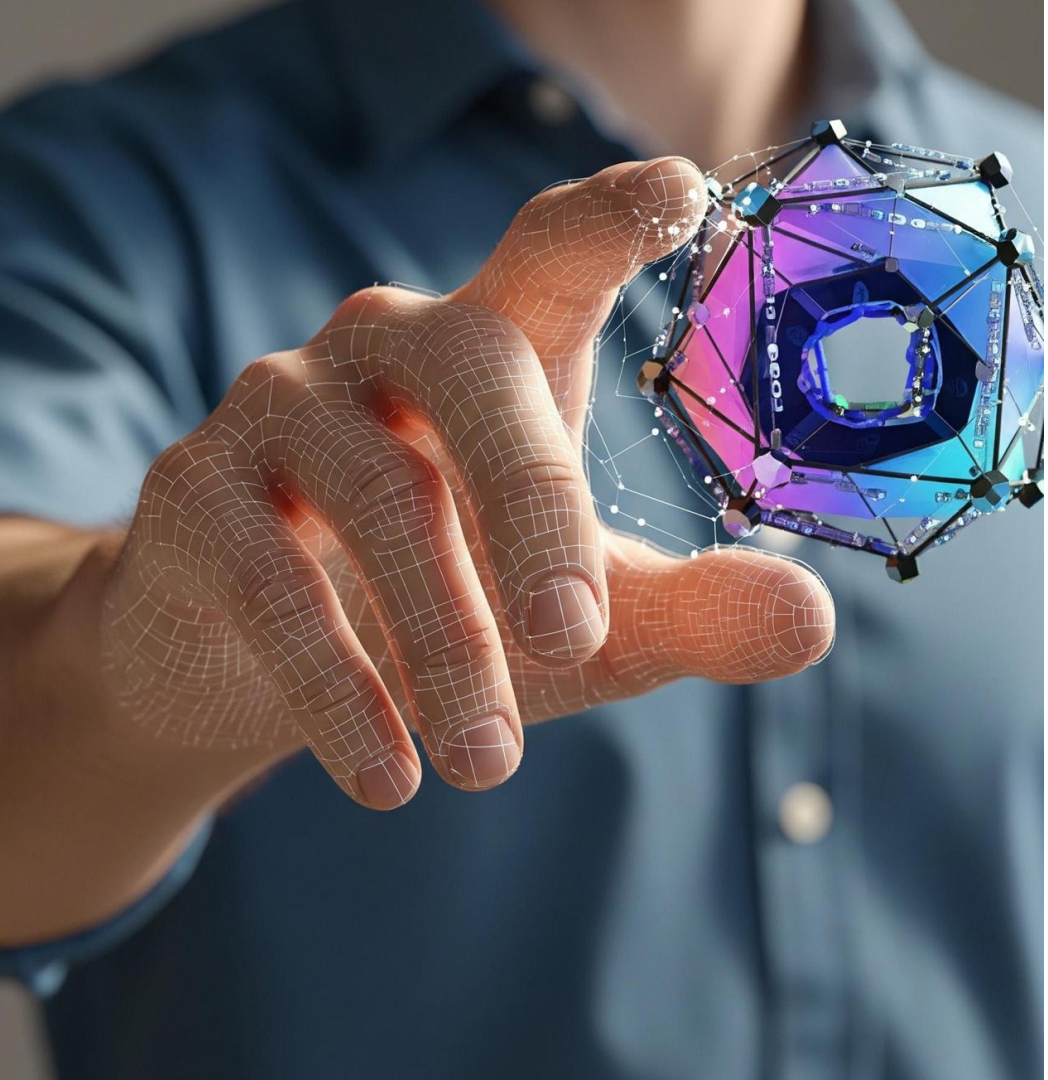
For Schools

For Potential Partners

For ML Consulting Services

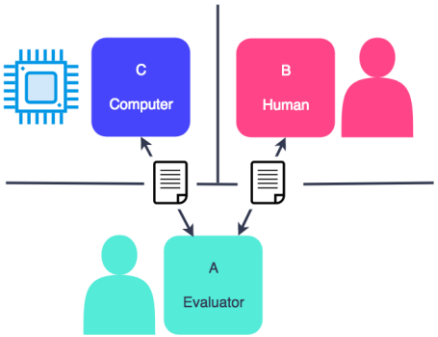
For Job Candidates

Feel free to reach out!

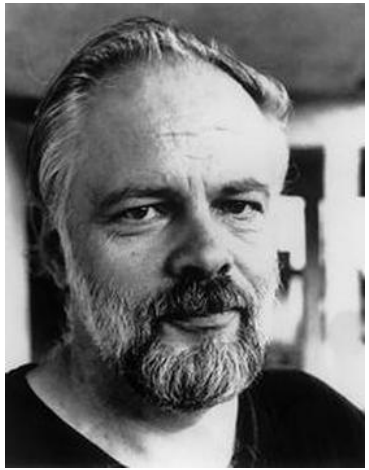


Motivation

Alan Turing



Philip K. Dick



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

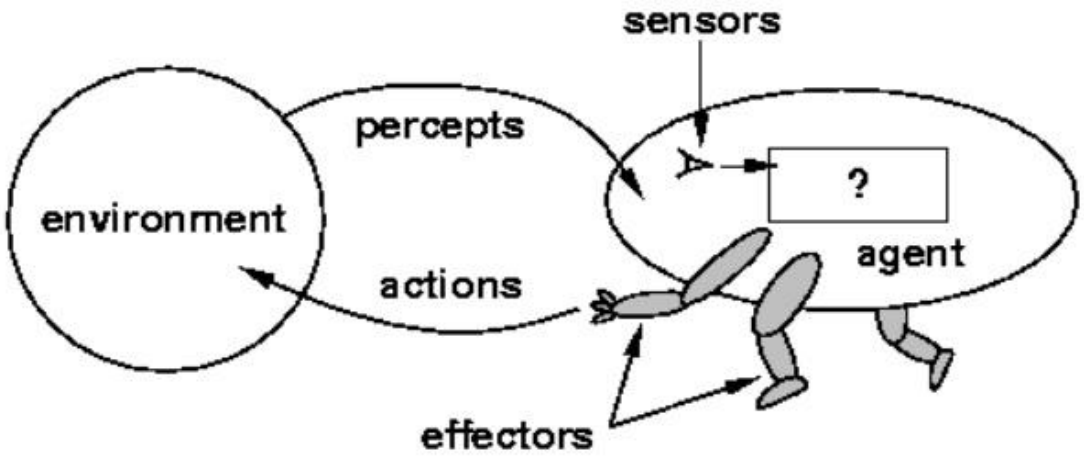
Steve Wozniak

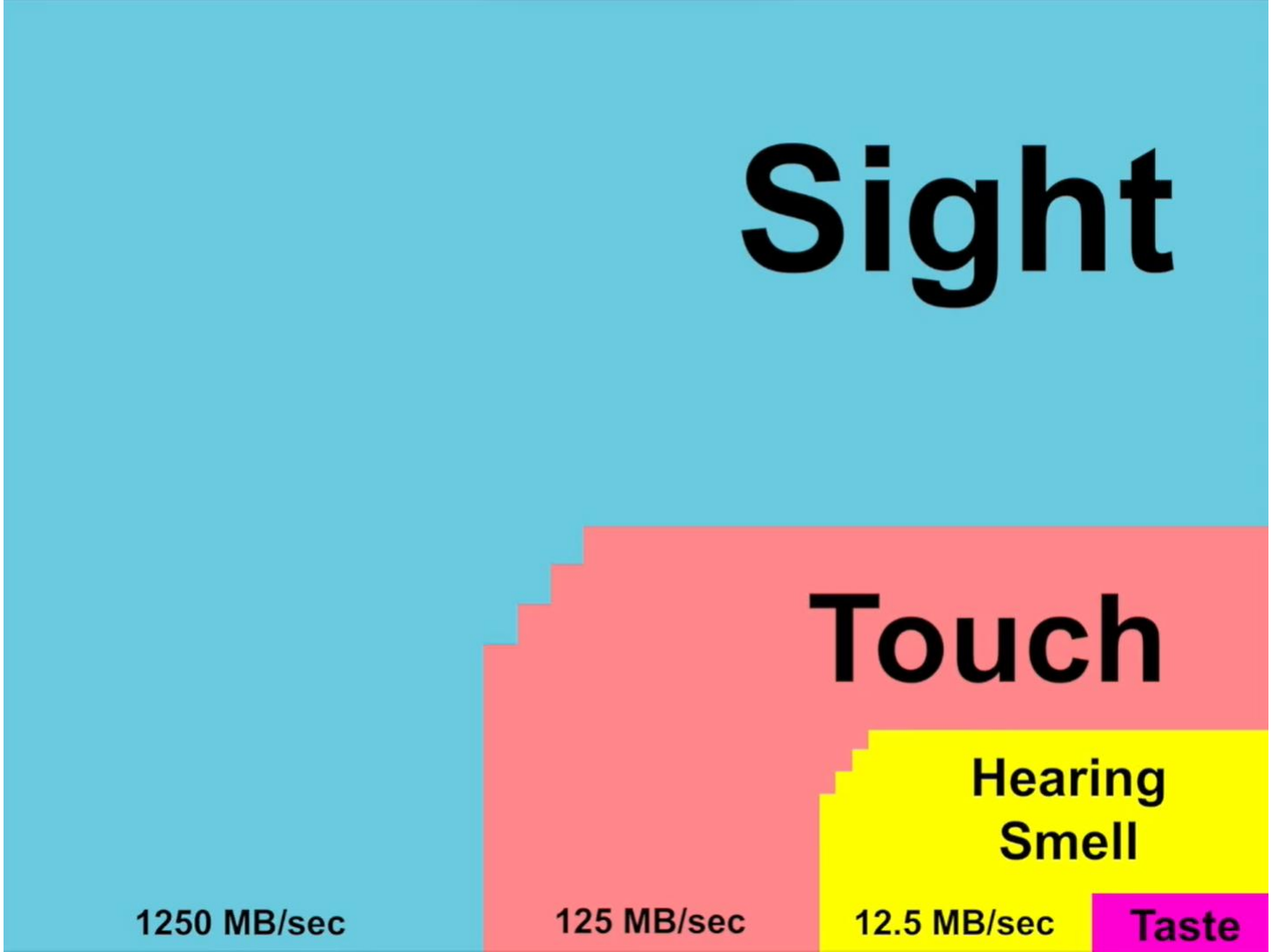
Artificial intelligence

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]





The start of research in CV

- Project MAC (est. 1963) – CSAIL today

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
PROJECT MAC

Artificial Intelligence Group
Vision Memo. No. 100.

July 7, 1966

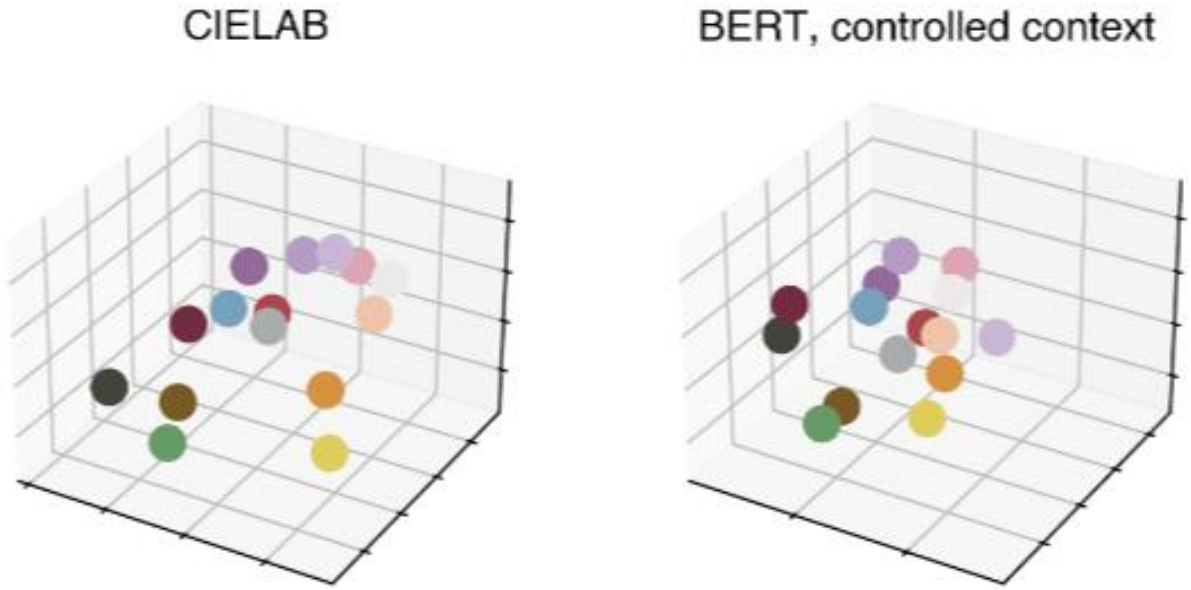
THE SUMMER VISION PROJECT

Seymour Papert

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

About LLM perception – Yes…

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

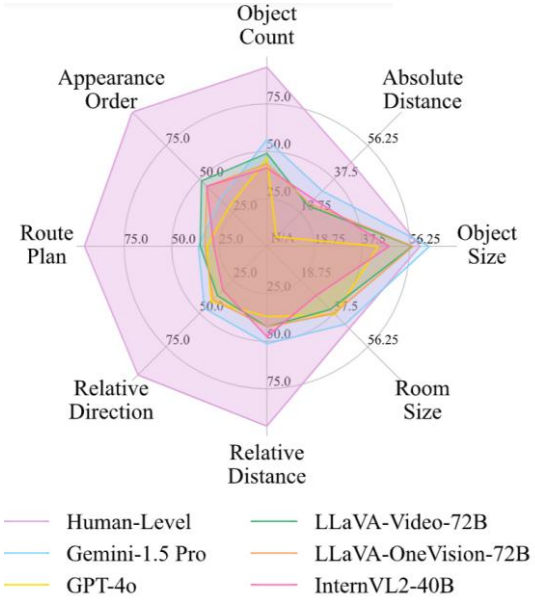


About LLM perception – … But

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-Fei[3] Saining Xie[1]
[1]New York University [2]Yale University [3]Stanford 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

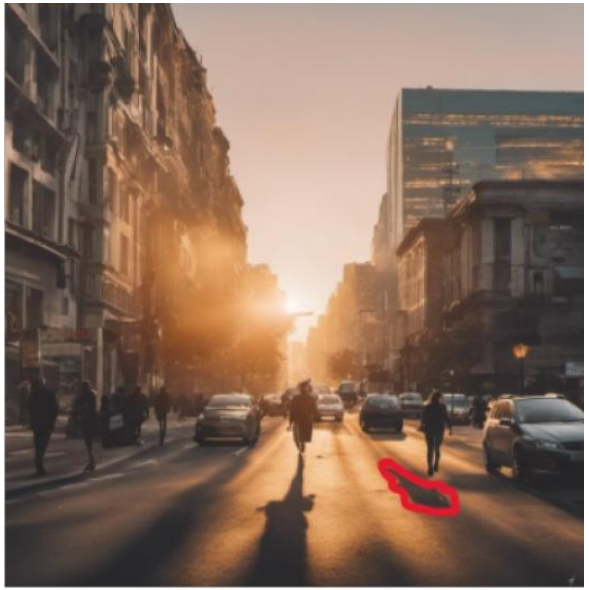
About GenAI – Yes…

https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/

About GenAI – …But

Shadows Don't Lie and Lines Can't Bend!
Generative Models don't know Projective Geometry...for now

Generated Image



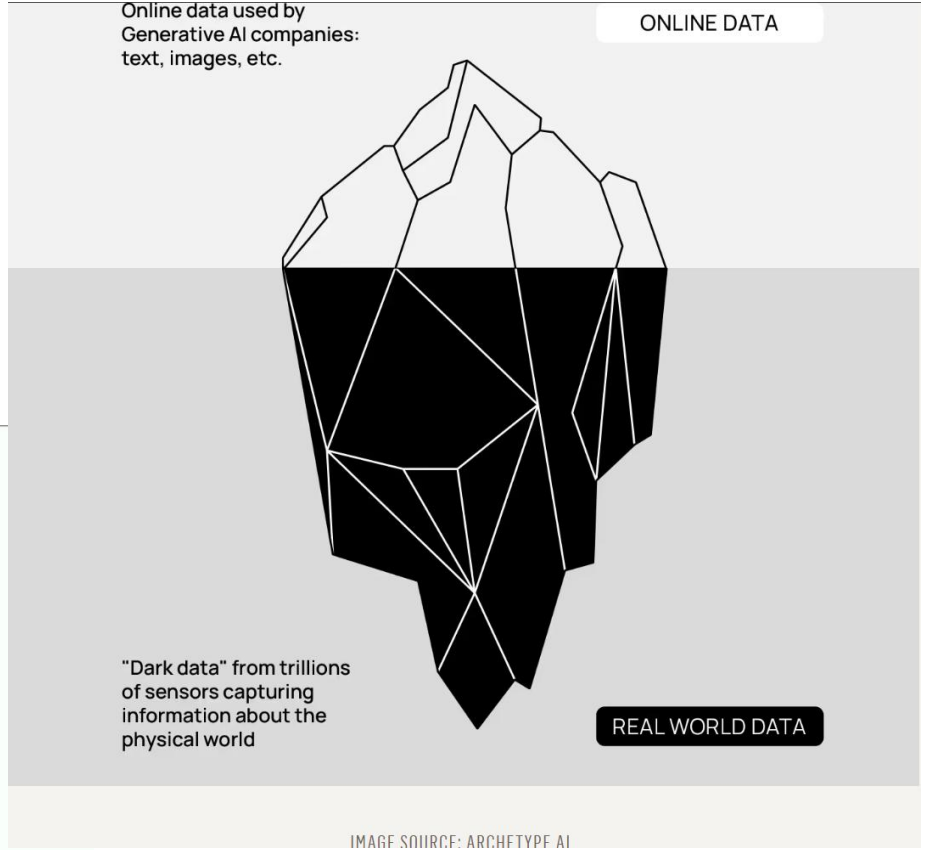
Shadow Errors

Vanishing Point Errors

Deal with data bottleneck

Ilya Sustkever, NeurIPS 2024

Vs.



Pre-training as we know it will end

Compute is growing:

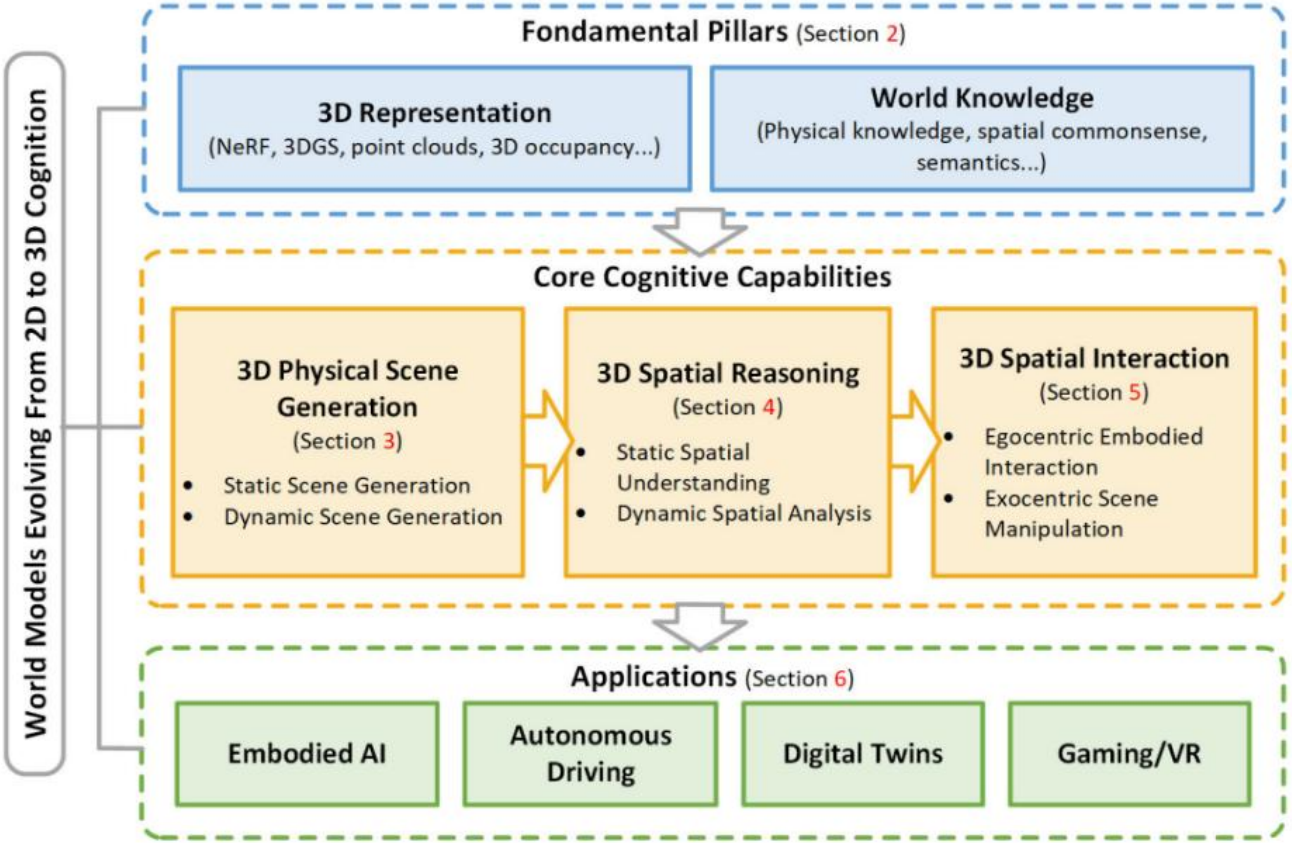
- Better hardware
- Better algorithms
- Larger clusters

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



Dr. Fei-Fei Li

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

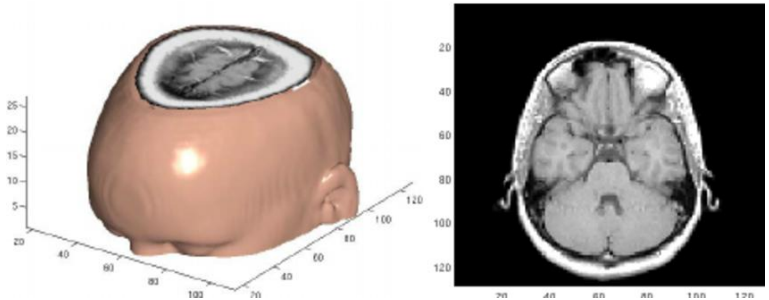
Entertainment

Architecture

Fashion

Digital Fabrication

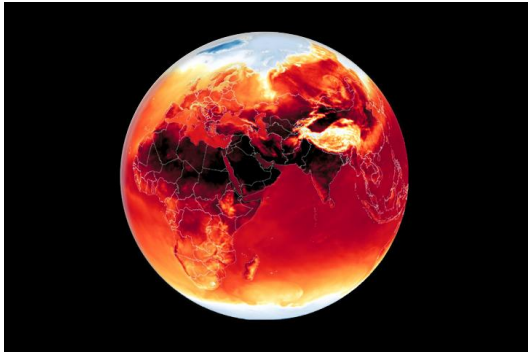
Medical Imaging



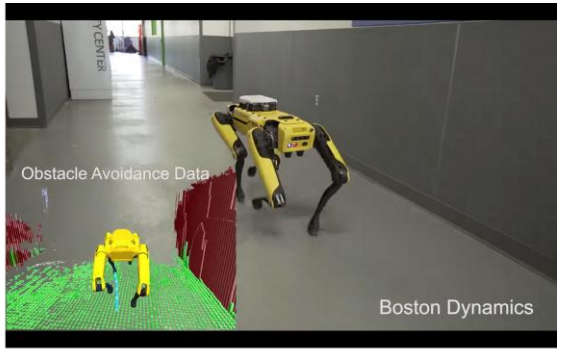
Archeology

And more…

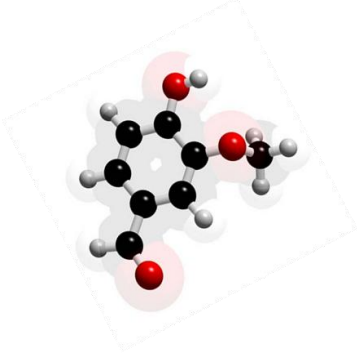
Geology/Climate



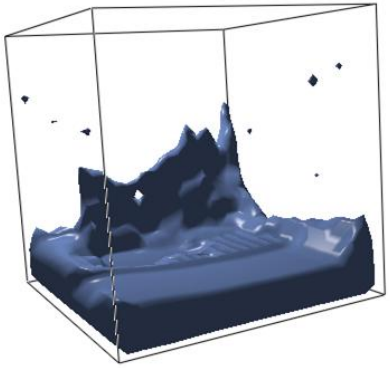
Robotics



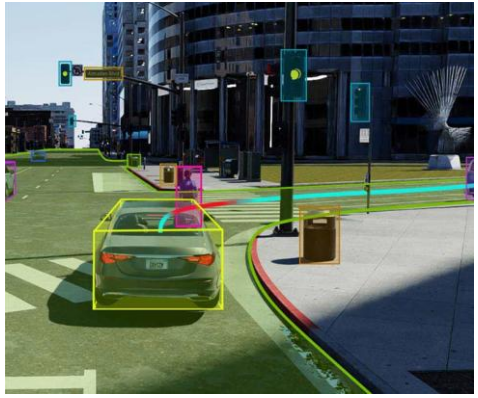
Biology



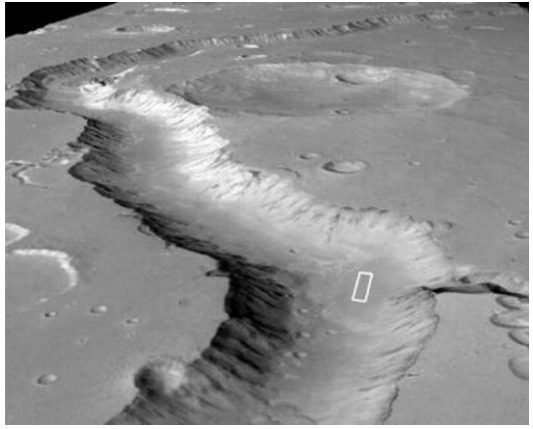
Physics/Simulations



Autonomous driving



Space Exploration



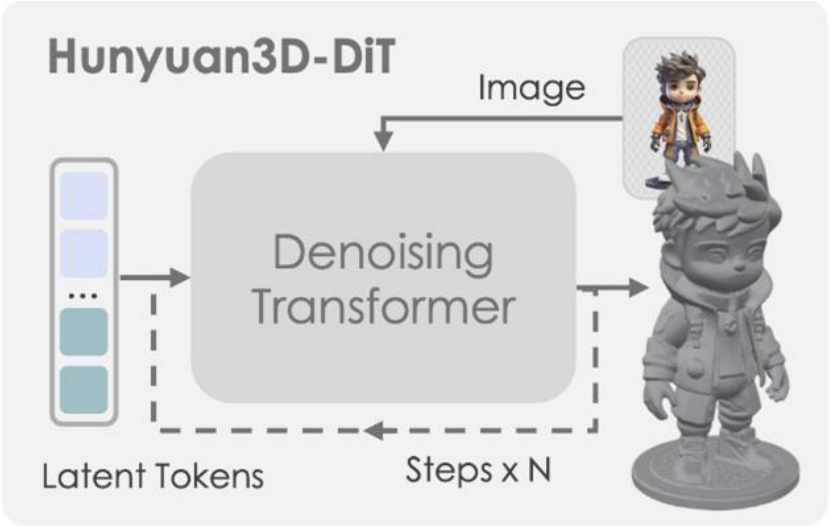
…

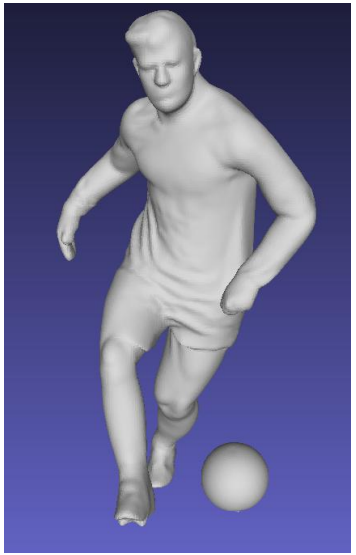
3D data are scaling

How to handle them?

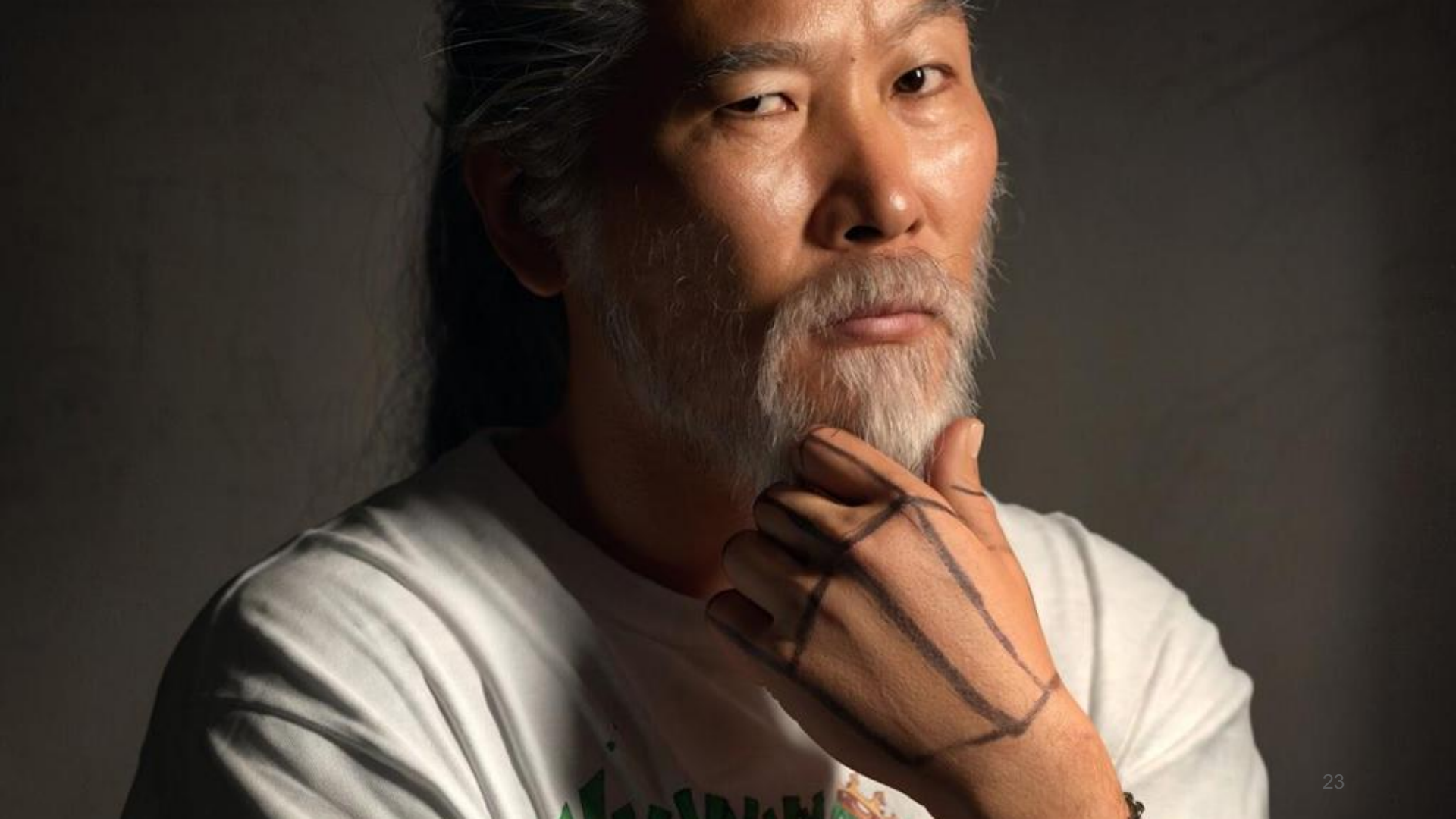
3D/4D Generative AI as a new data source

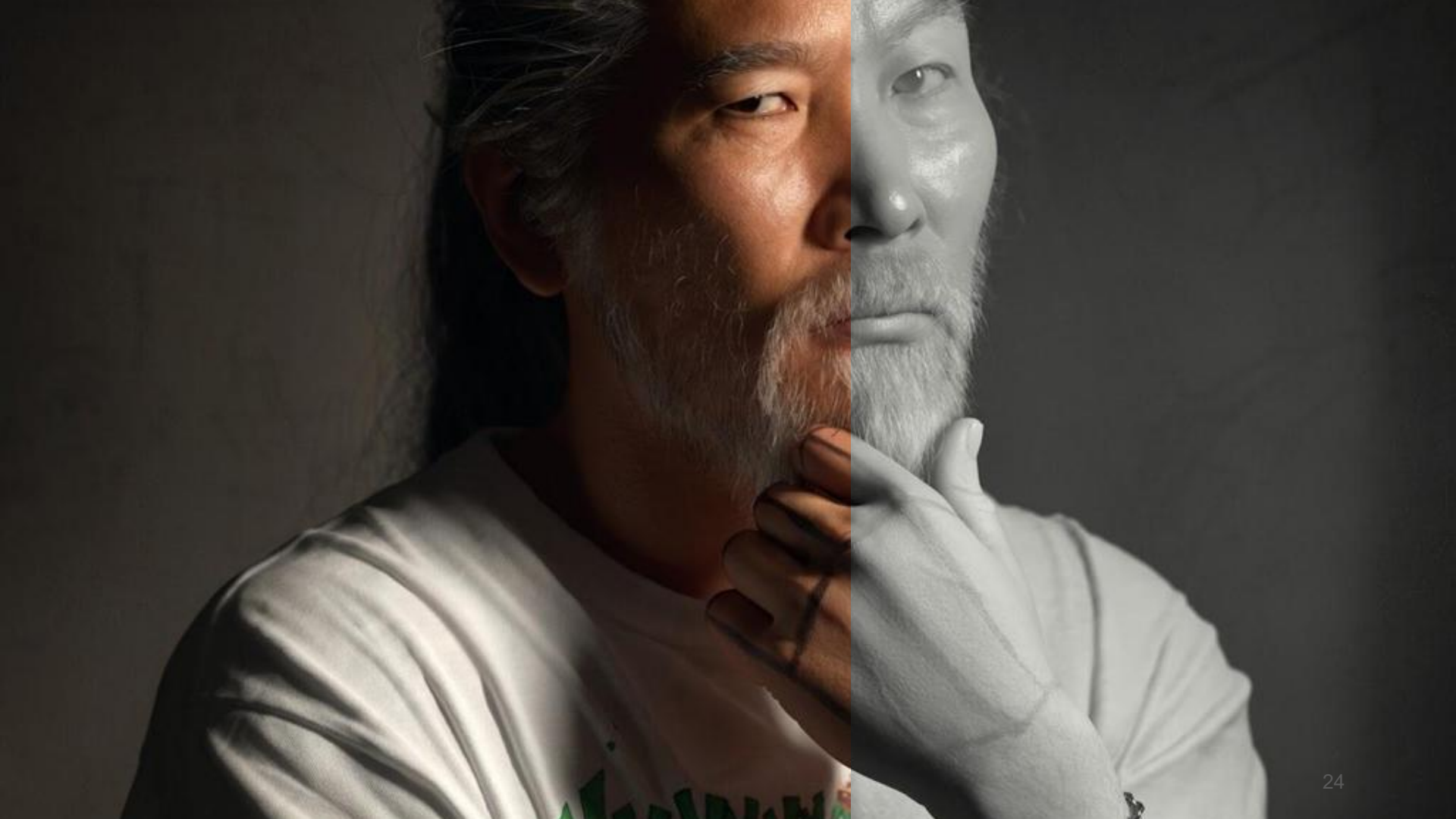
https://huggingface.co/spaces/tencent/Hunyuan3D-2





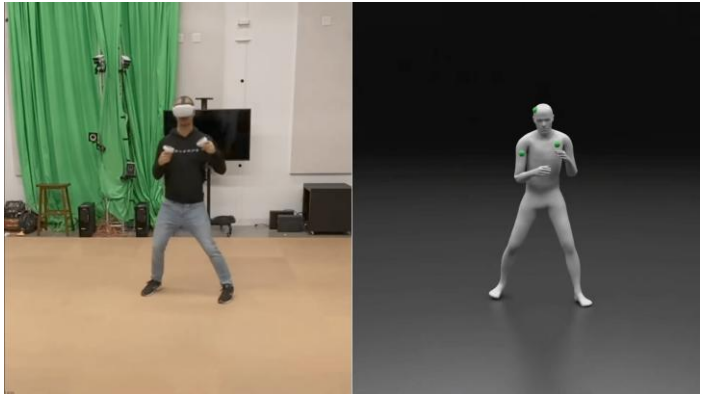
Disparate sources of data call for tools to relate them!



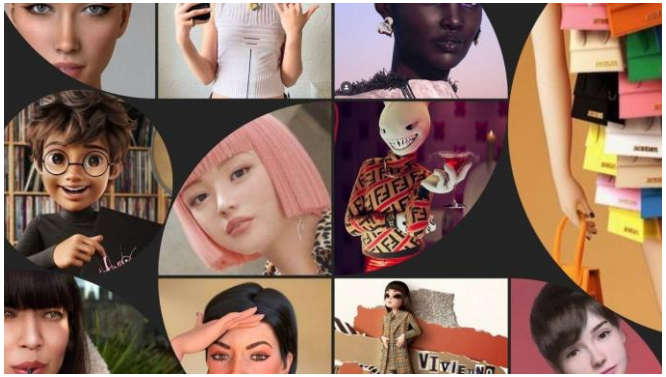


https://www.artstation.com/ianspriggs

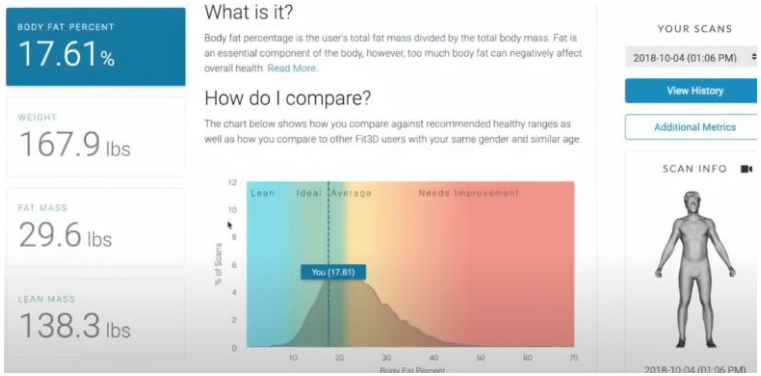
Virtual Humans as enablers of Spatial AI



VR/AR



Marketing
(Virtual Influencers)



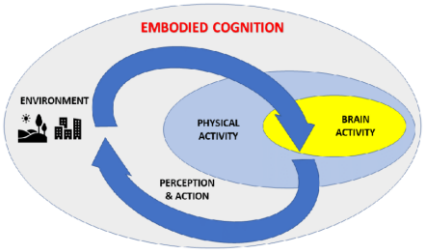
Health, Medics, Training

Assistance and Therapy

Does intelligence need a body?

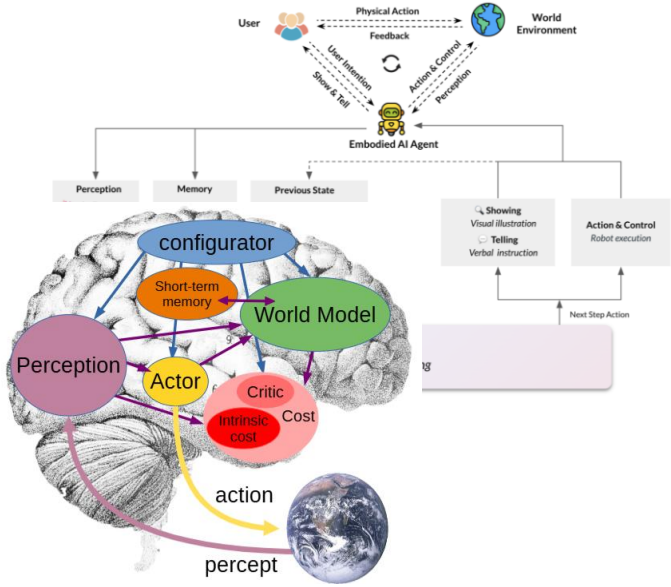
Descartes: No

https://pmc.ncbi.nlm.nih.gov/articles/PMC3512413/



Psycologists:
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

PhD Internships

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

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

Resarch Groups

MPI-IS Tuebingen Director

MAX-PLANCK-INSTITUT FÜR INTELLIGENTE SYSTEME

UNIVERSITY OF AMSTERDAM

BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

ETH zürich

Teaching

Seminar on Digital Humans

ETH zürich Course Catalogue

MSc Course

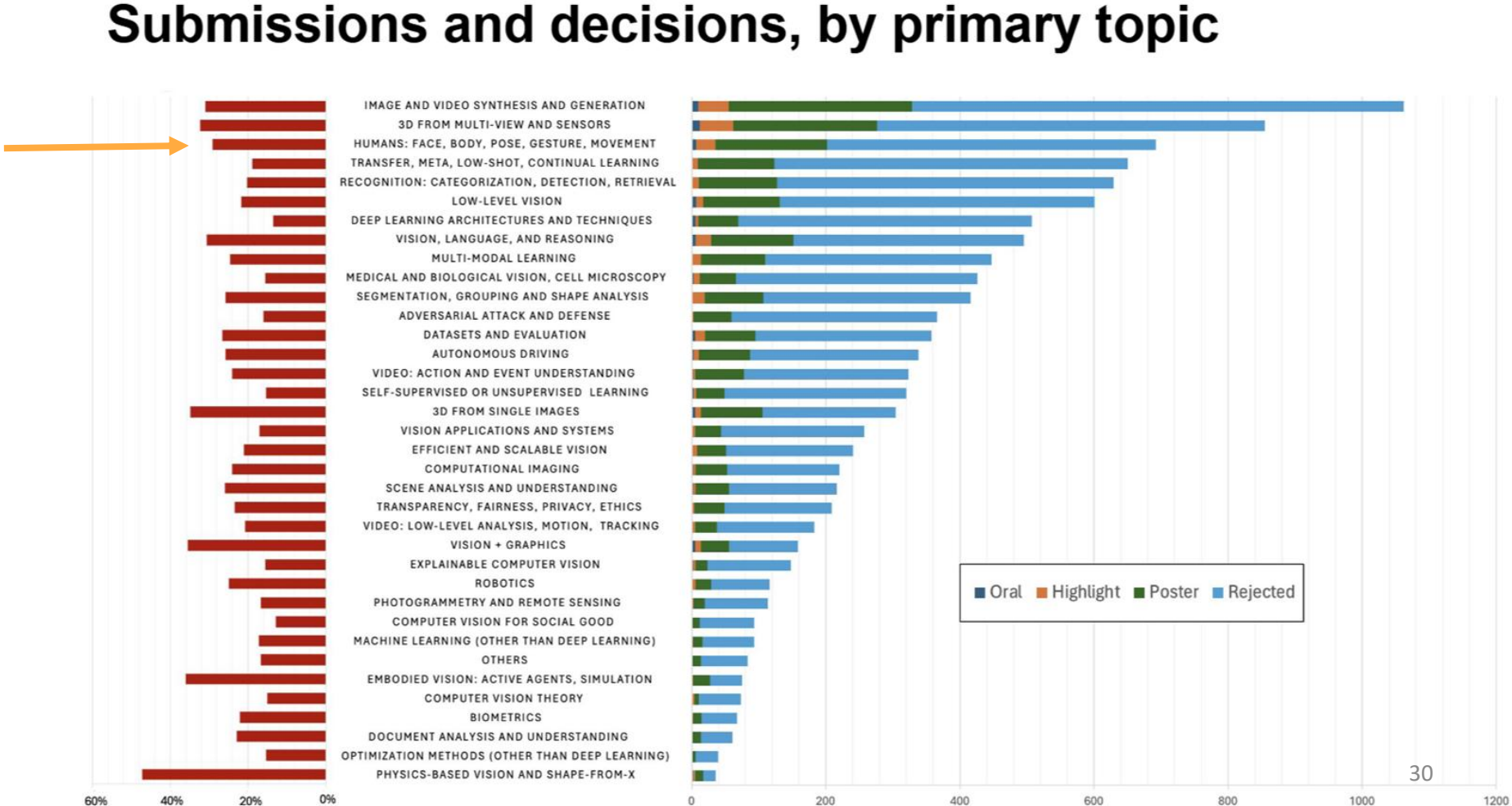
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Tutorials, Workshops

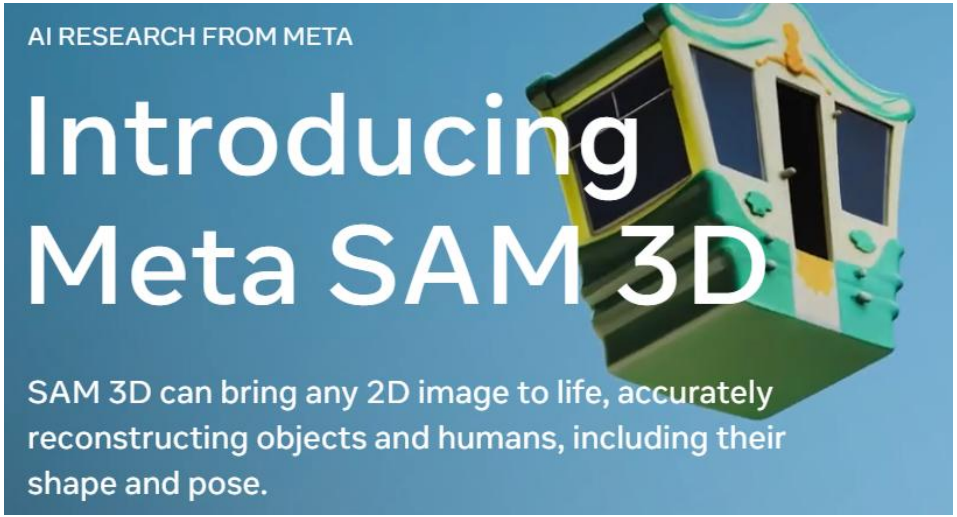
CVPR VIRTUAL JUNE 19-25

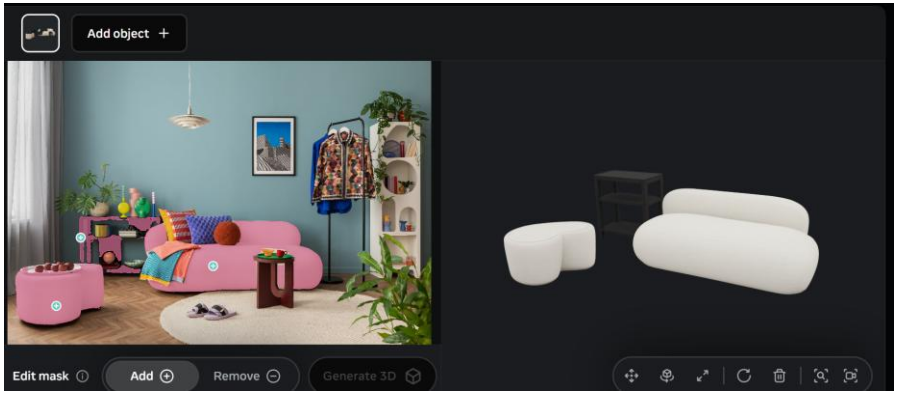
ECCV TEL AVIV 2022

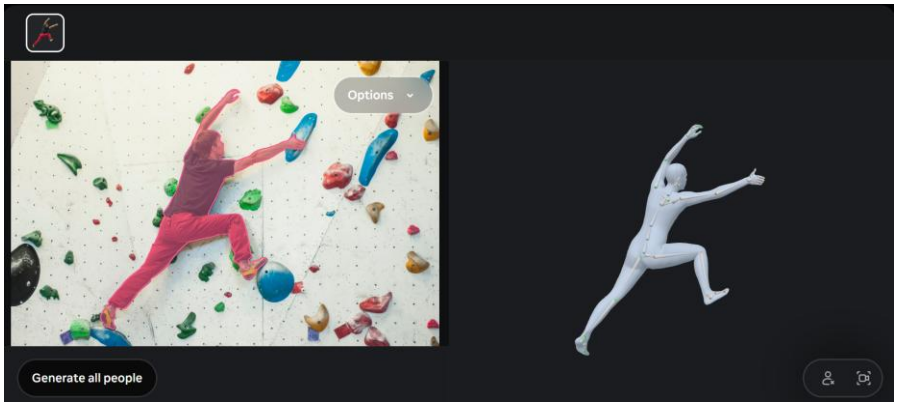
CVPR24 Statistics



BREAKING NEWS: last night Meta released SAM 3D Models







https://aidemos.meta.com/segment-anything/gallery

Structure

Today

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

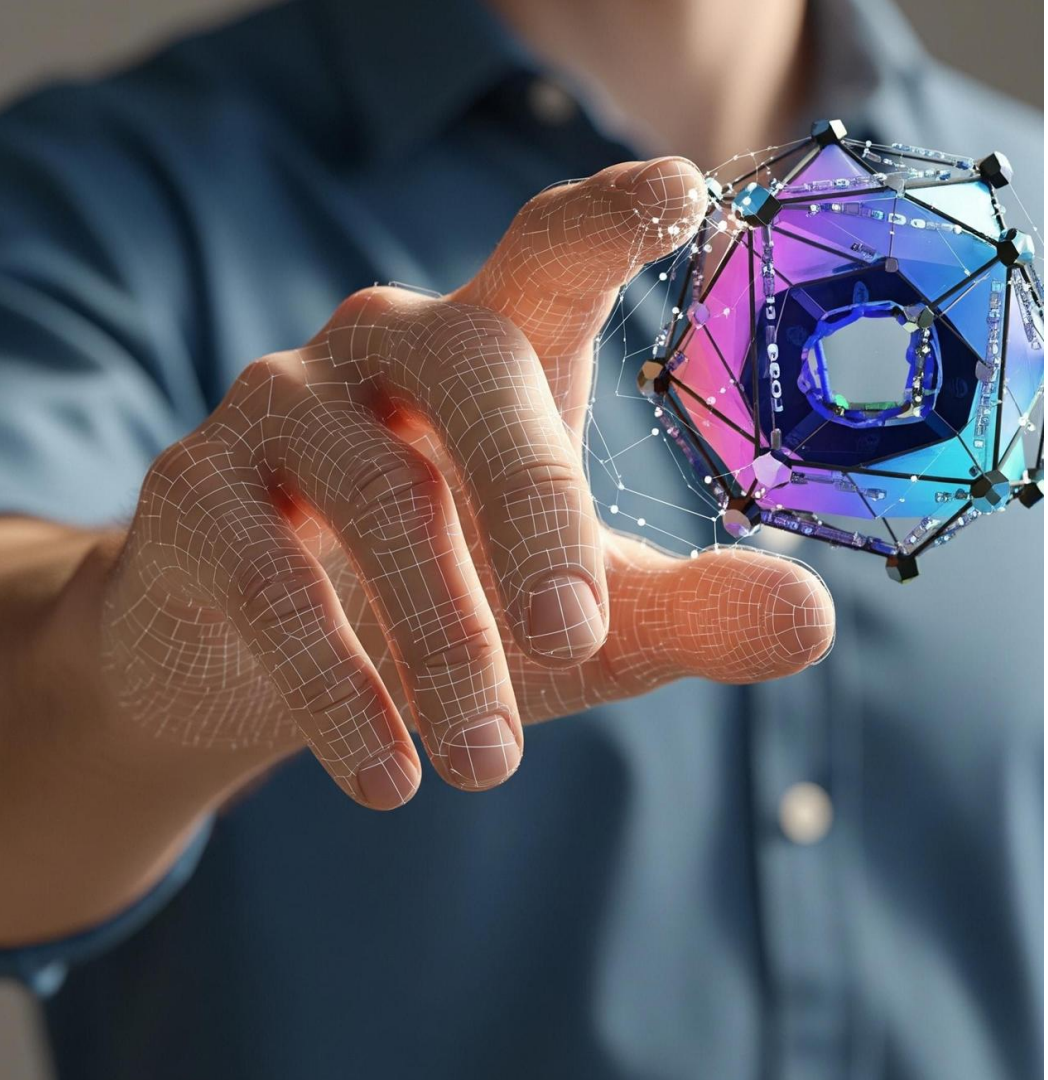
+ hands-on code Lab

Tomorrow

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

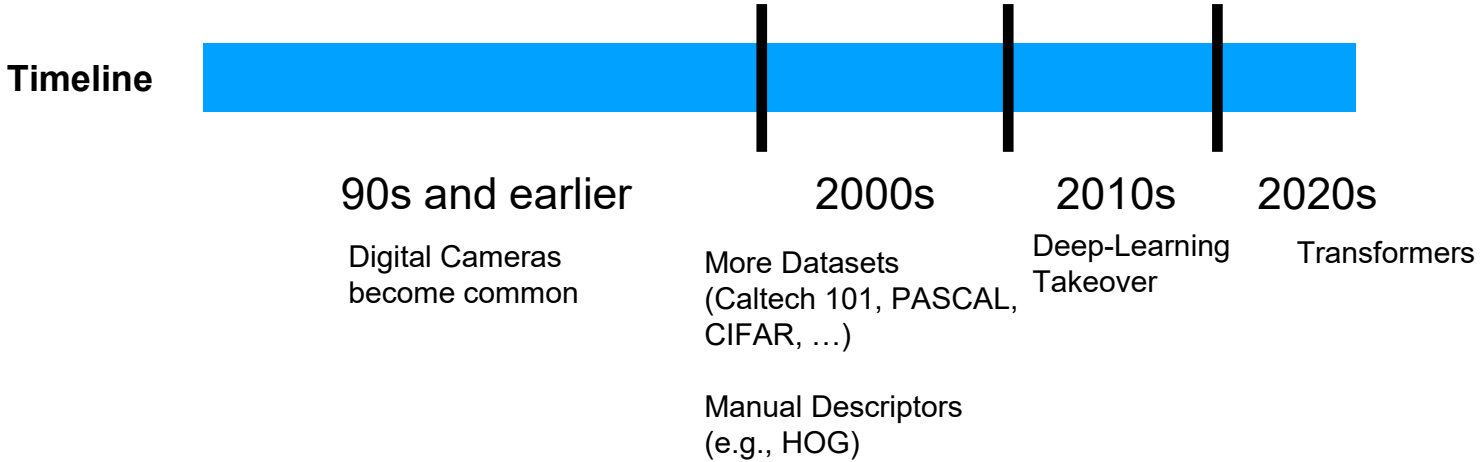
+ hands-on code Lab

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

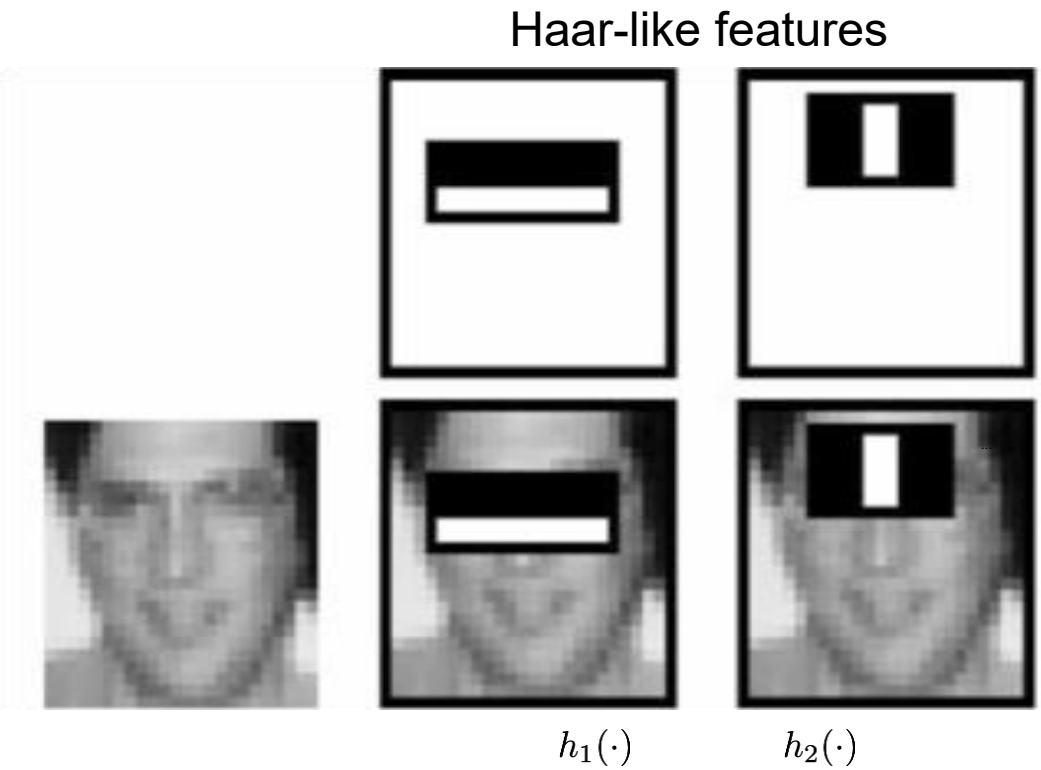


Processing Geometrical Data

Progress: Now and then



Pre-Learning era: Handcrafted features

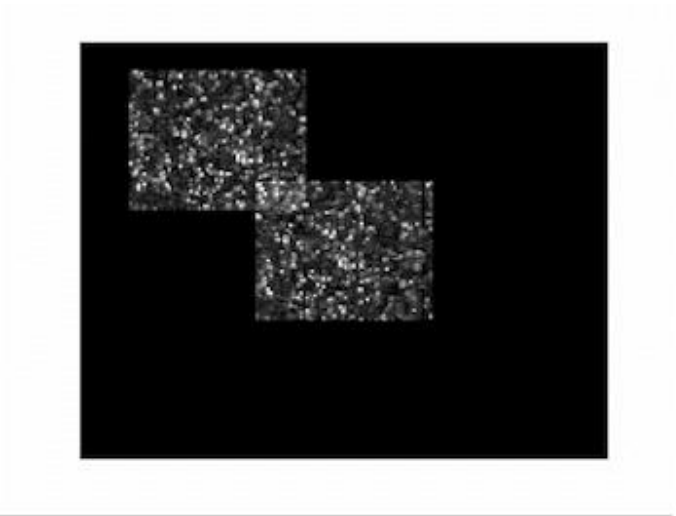


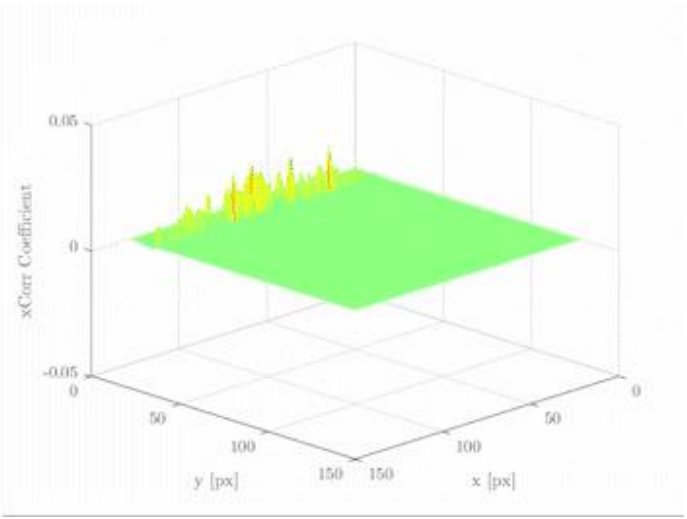
Viola and Jones. Rapid object detection using a boosted cascade of simple features. CVPR 2001.

Pre-Learning era: Handcrafted features

- Zero-normalised cross-correlation (ZNCC):

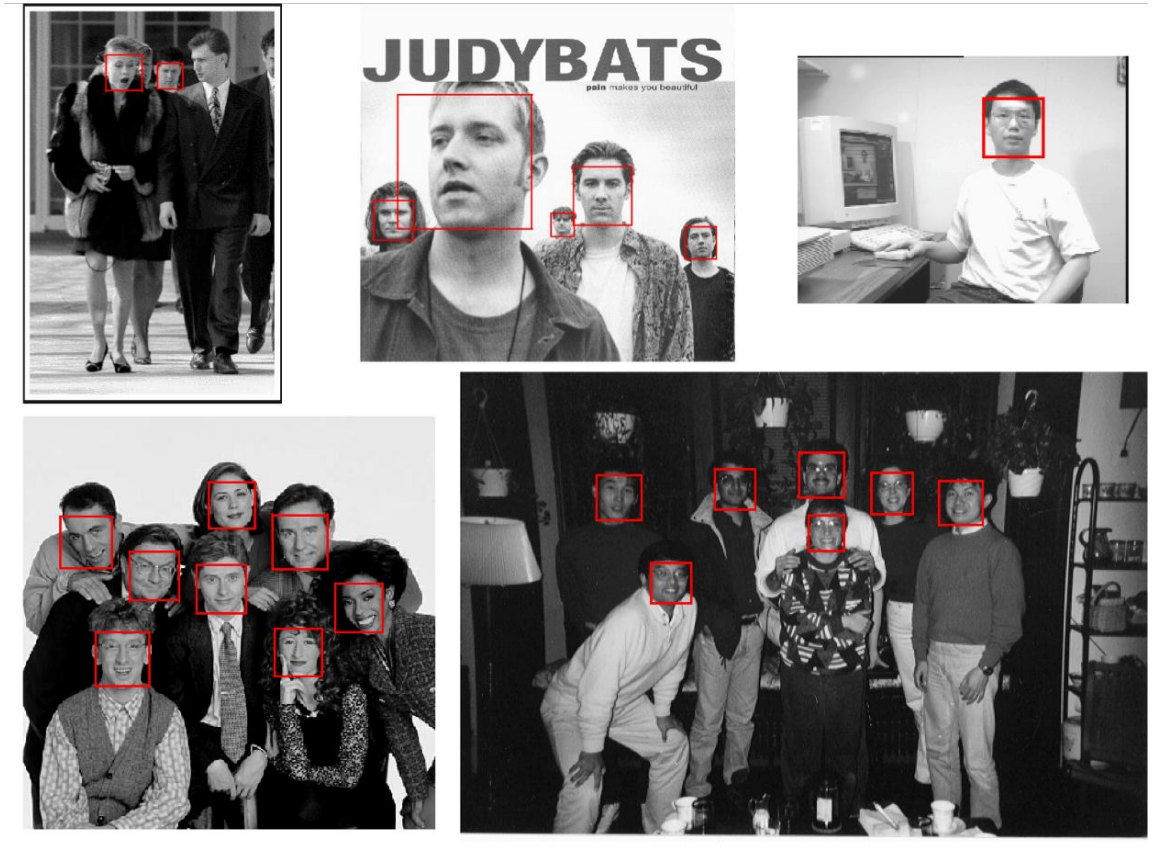
$$d(I_{(x_0,y_0)}, T) = \frac{1}{n} \sum_{x,y} \frac{1}{\sigma_I \sigma_T} (I_{(x_0,y_0)}(x,y) - \mu_I)(T(x,y) - \mu_T)$$





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

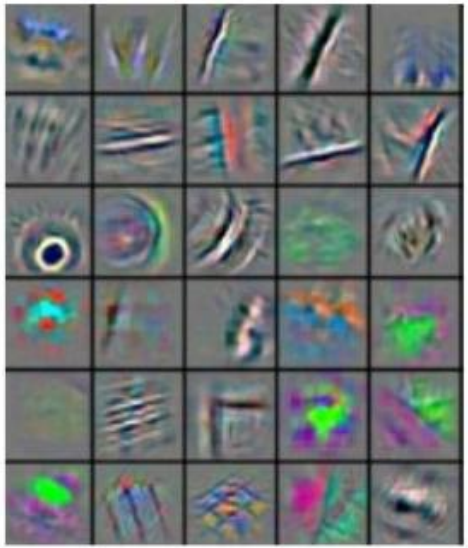
Pre-Learning era: Handcrafted features

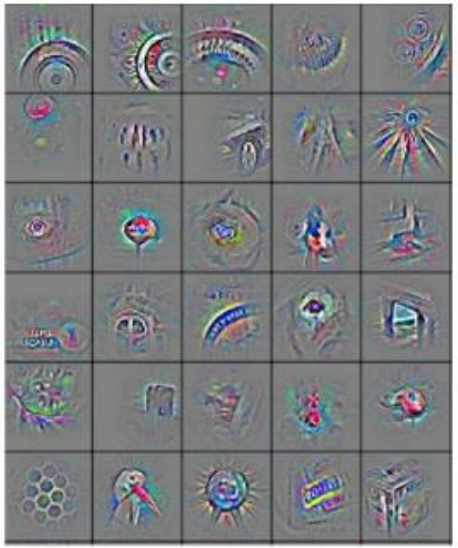


Viola and Jones. Rapid object detection using a boosted cascade of simple features. CVPR 2001.

Deep Learning:

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

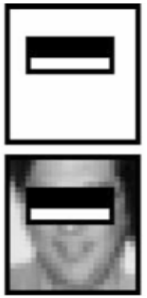




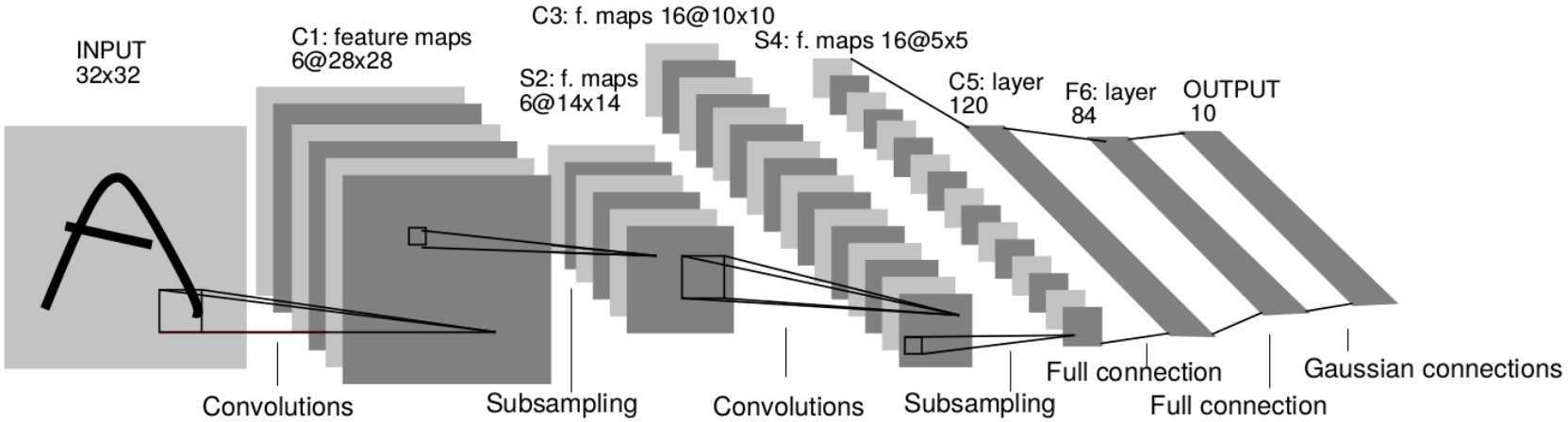
Slide credit E. *Rodolà*

CNN:

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

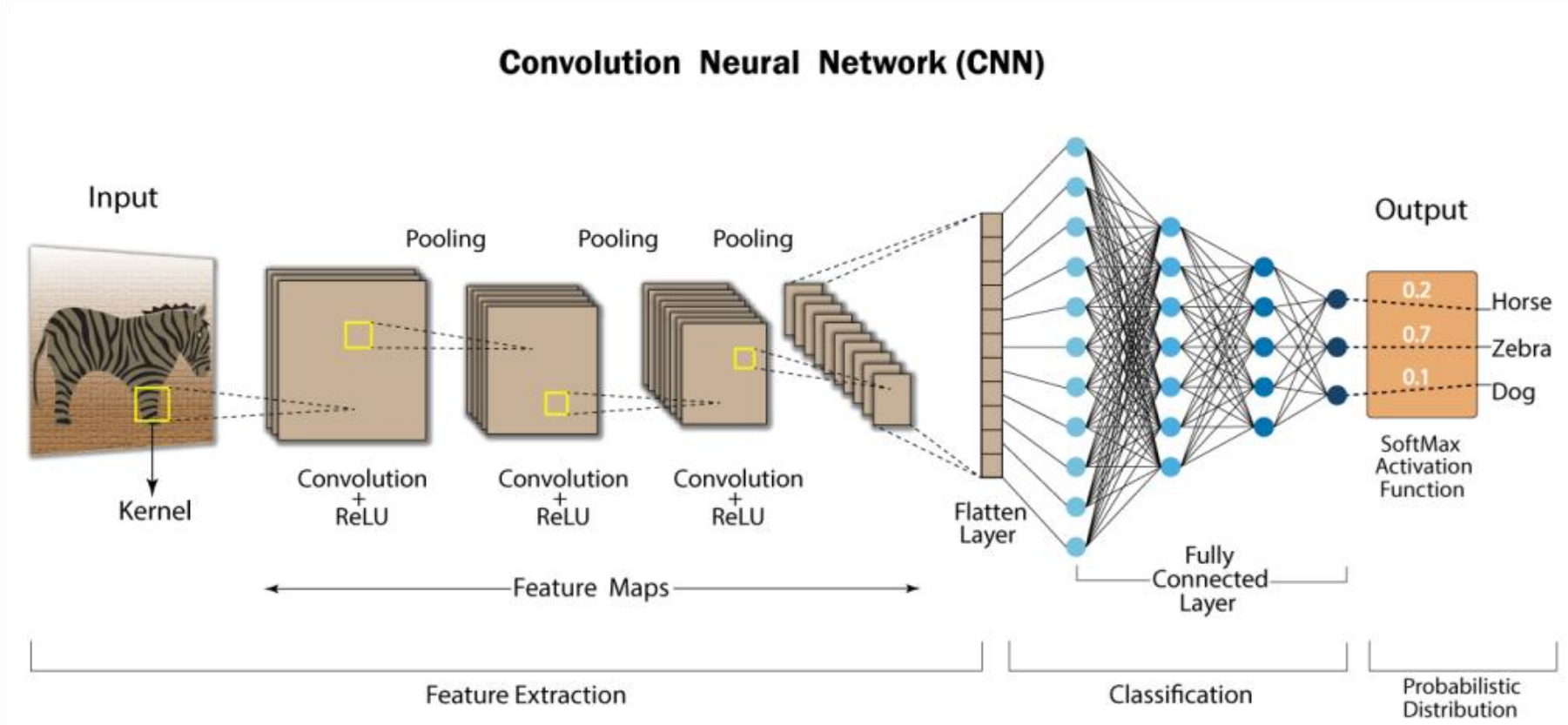


CNNs directly exploit this fact as a prior.



Slide credit E. *Rodolà*

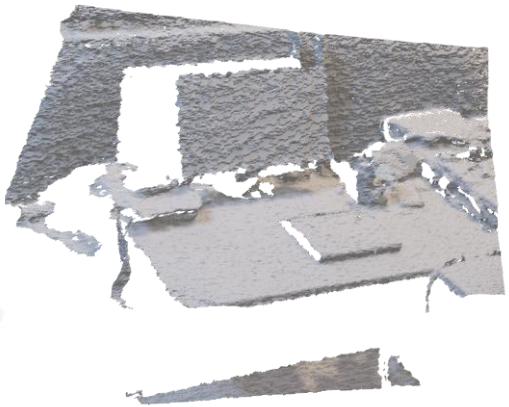
Convolutional network



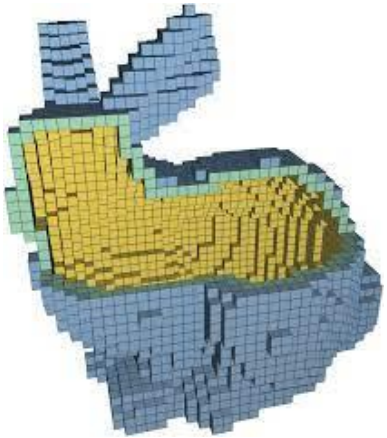
Convolution relies on the underlying structure

Geometrical data have different structure

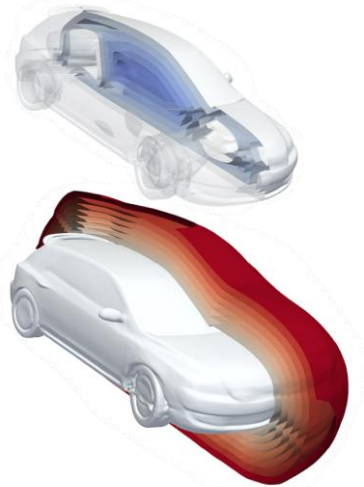
Point clouds



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

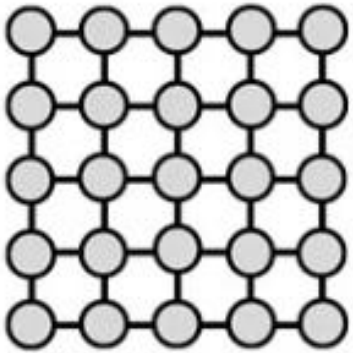
Slide credit E. *Rodolà*

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

https://arxiv.org/abs/2104.13478

The roles of a representation

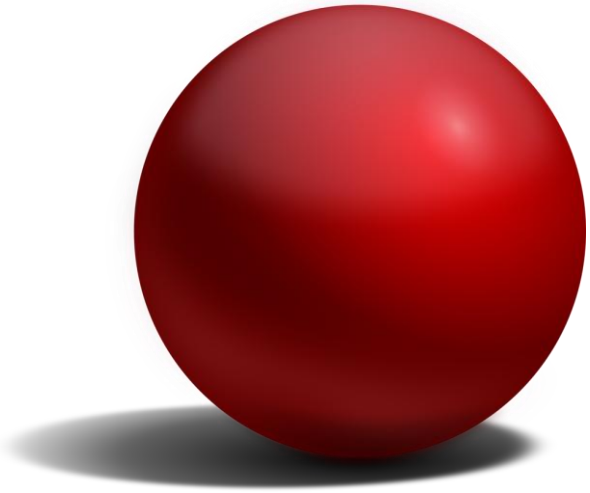
Convey the geometry

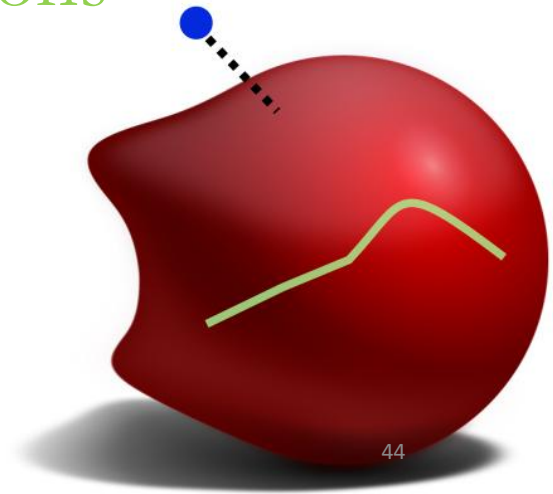
Support computations

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

https://ieee-dataport.org/documents/uwa-3d-multiview-activity-ii-dataset

Multiple RGBs

- We can infer spatial relations

- Can be processed by (CNNs)

- Often requires calibrated multi-view cameras

https://ieee-dataport.org/documents/uwa-3d-multiview-activity-ii-dataset

Point Cloud

- Geometry is where points are

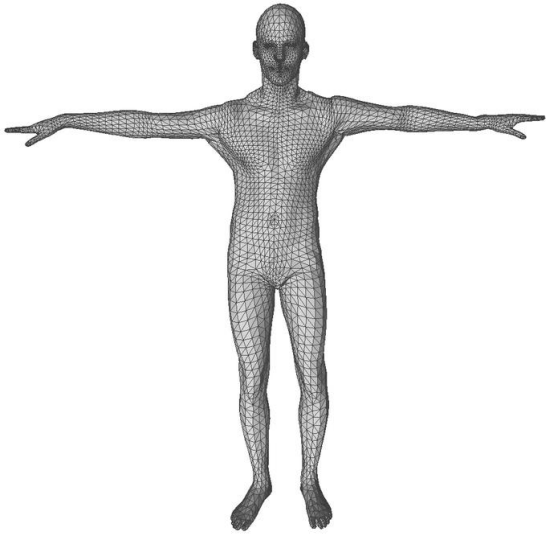
- Computations?
Adding or subtracting points, spatial distance, …

- What happen in the middle?

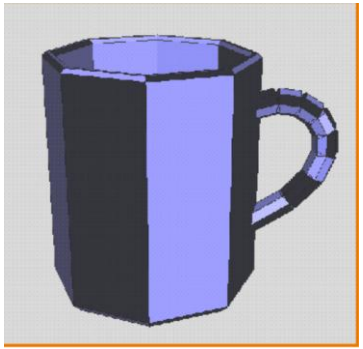
- Output of different pipelines

Human Body Reconstruction from Image Sequences, Remondino et al, 2001

(Triangular) Mesh



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



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

- Geometry is a set of piecewise flat surfaces

- Points + Connectivity

- Great theoretical properties

- Not everything is easy
(e.g., spatial query, remove geometry)

Implicit Representations (E.g., Voxelization)

- Geometry as a function in a 3D grid

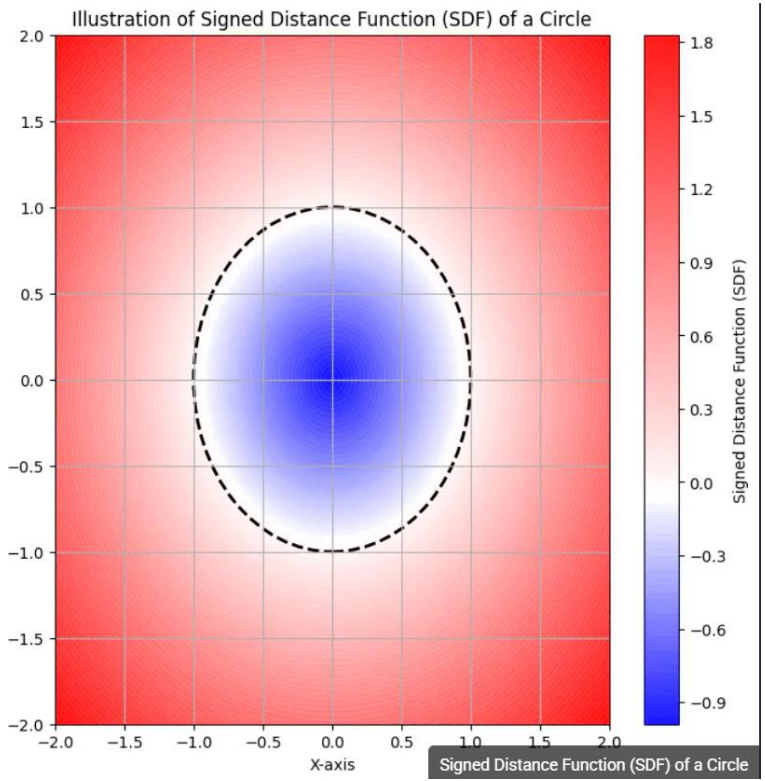
- Like an image, but in 3D

- Great for learning, (e.g, CNNs), model objects with different toplogies

- Expensive, not that easy to visualize

https://www.deviantart.com/guifeur/art/Male-Human-Model-in-Voxel-885885832

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 toplogies

- Expensive, not that easy to visualize

https://sadjadabedi.ir/post/demystifying-signed-distance-functions/

To sum up, choose your representations wisely!

Do I need to edit it?

How will I visualize it?

Intrinsic or spatial queries?

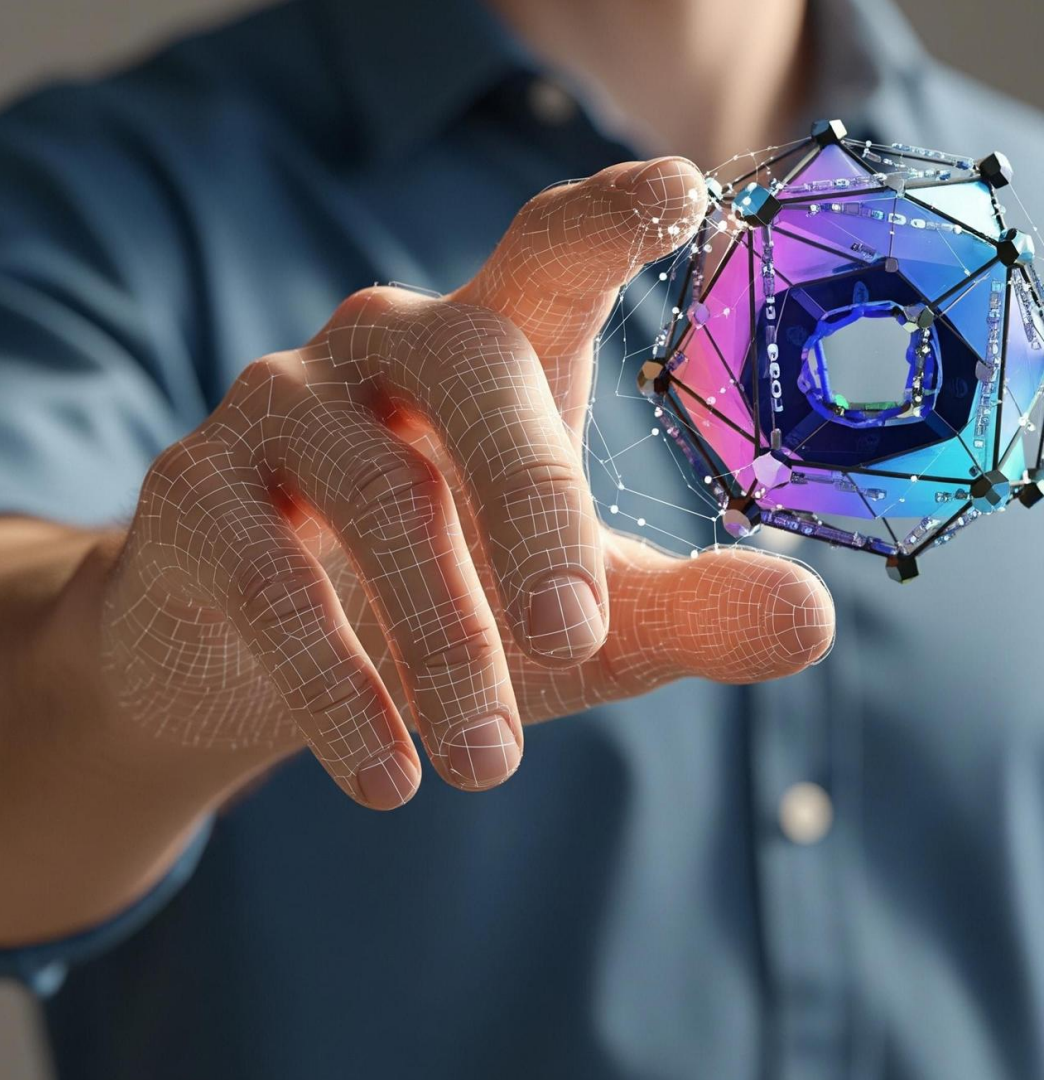
Invariance to 3D embedding?

Can I combine, extend, learn representations?

…

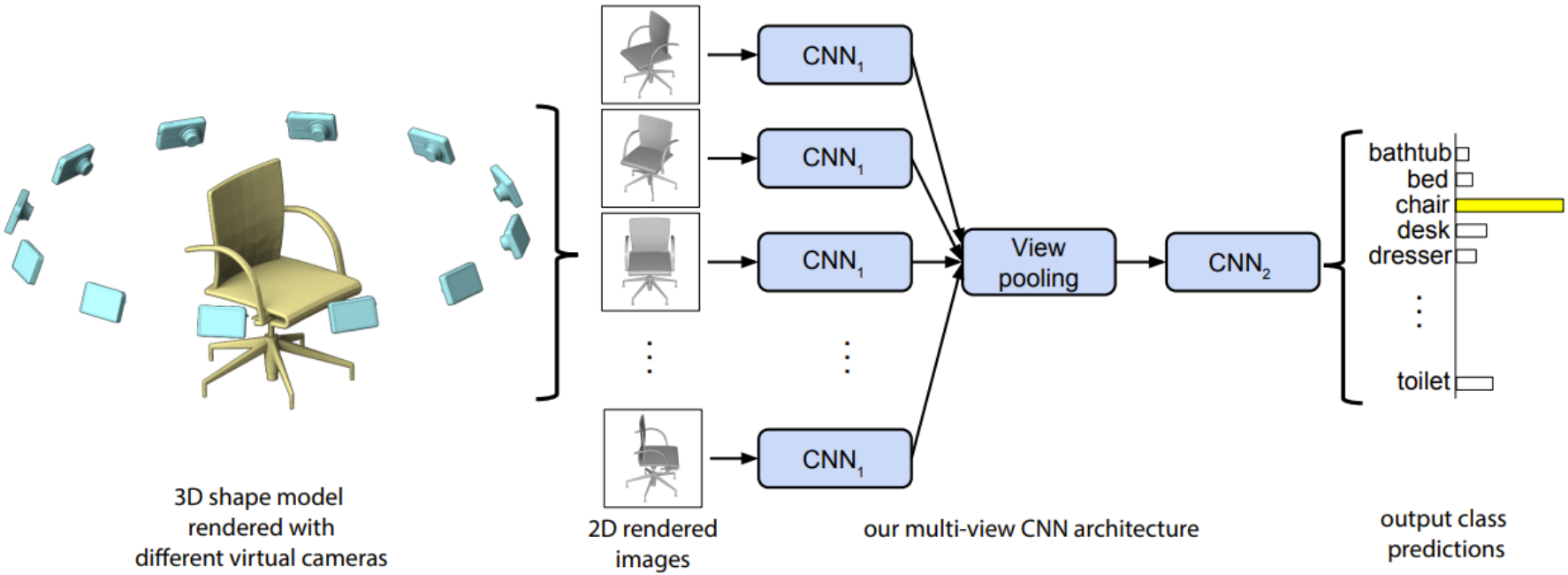
To sum up, choose your representations wisely!

What about learning?



Different architectures for different representations

Multi-view CNN



Su et al 2015

Multi-view CNN - Saliency

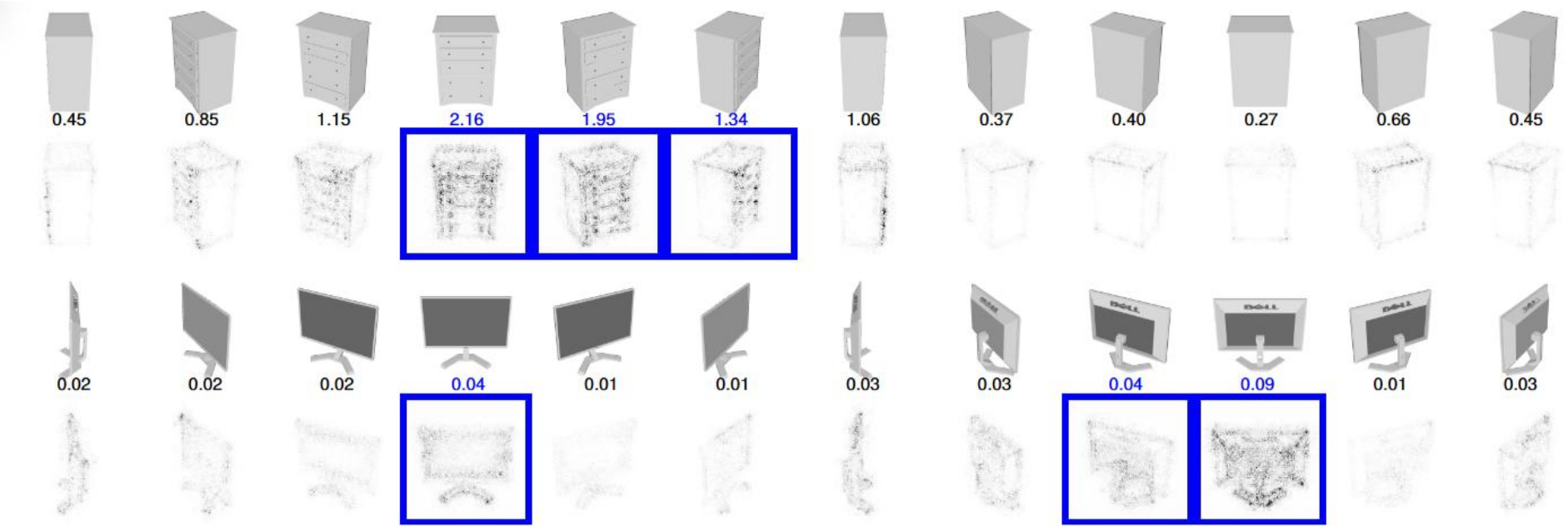
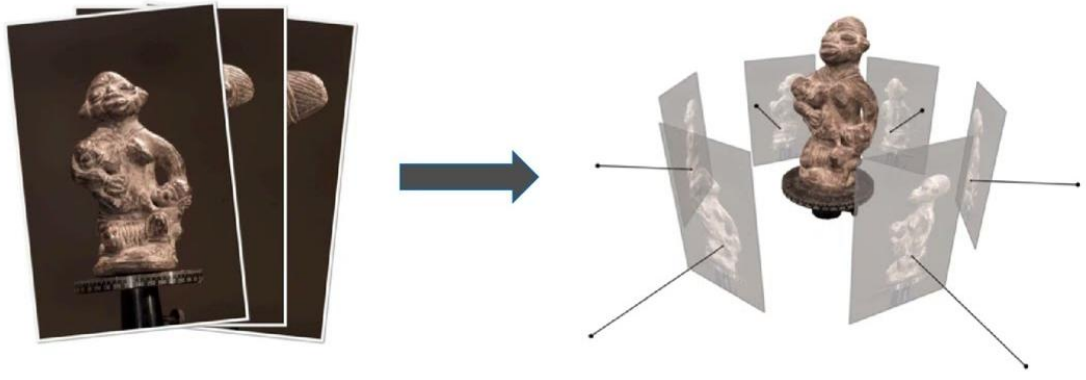
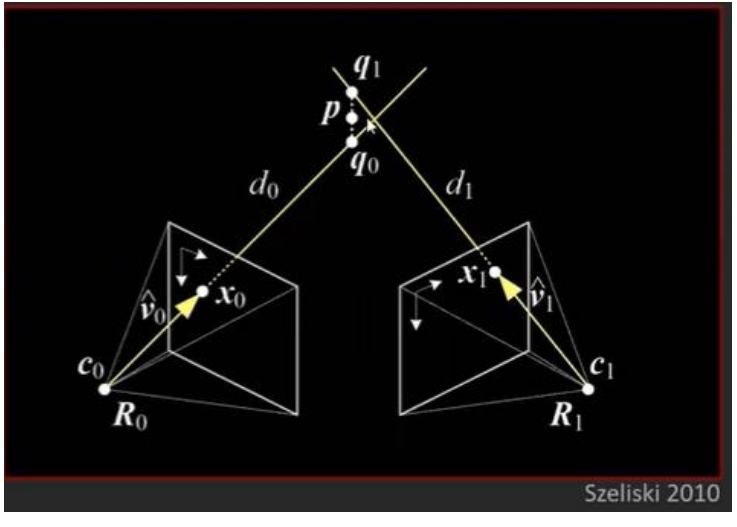


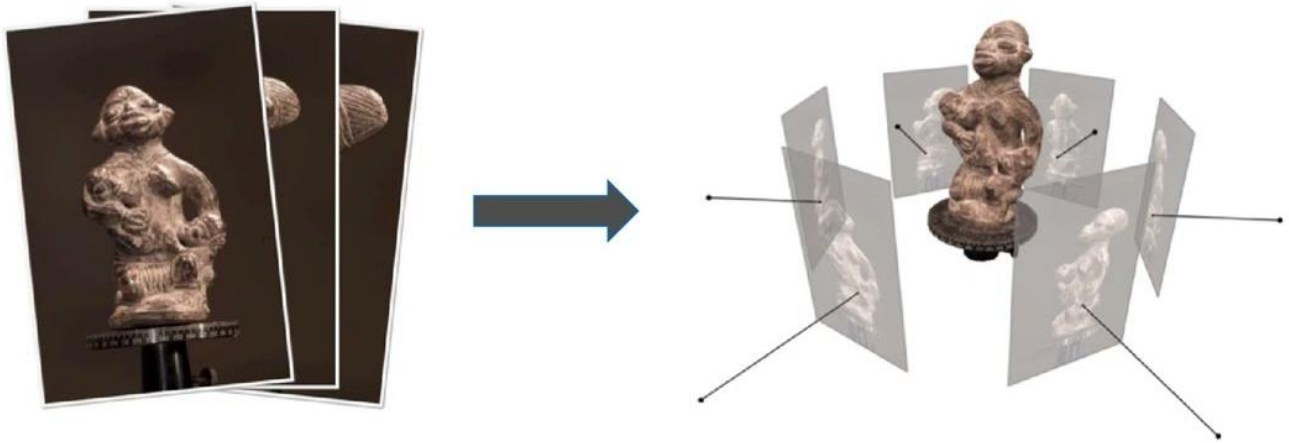
Figure 3. Top three views with the highest saliency are highlighted in blue and the relative magnitude of gradient energy for each view is shown on top. The saliency maps are computed by back-propagating the gradients of the class score onto the images via the view-pooling layer. Notice that the handles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

Multiple RGBs are the foundation for 3D reconstruction



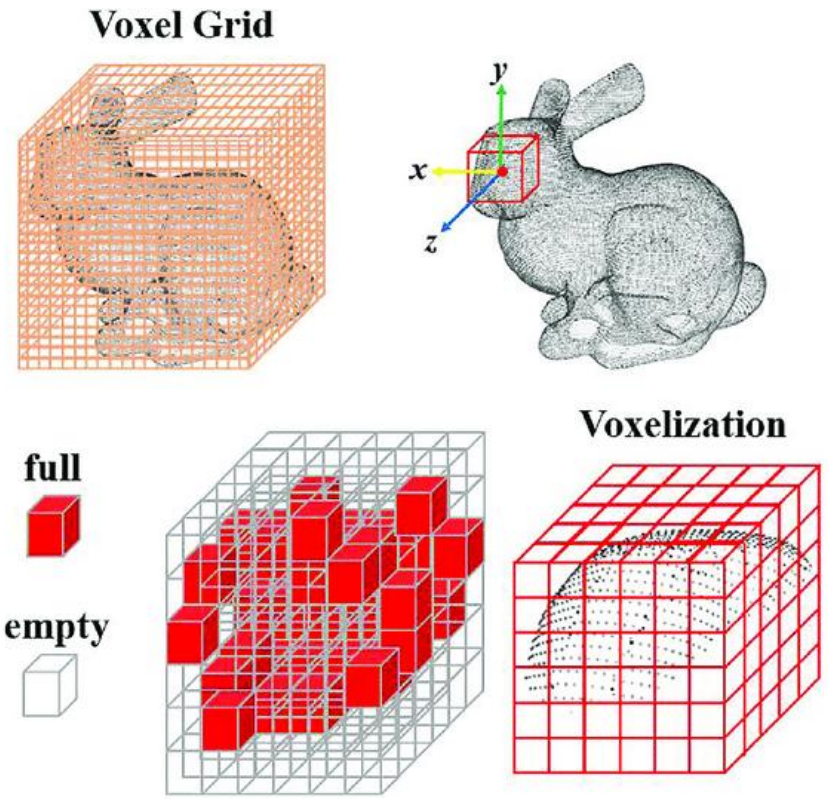


Multiple RGBs are the foundation for 3D reconstruction



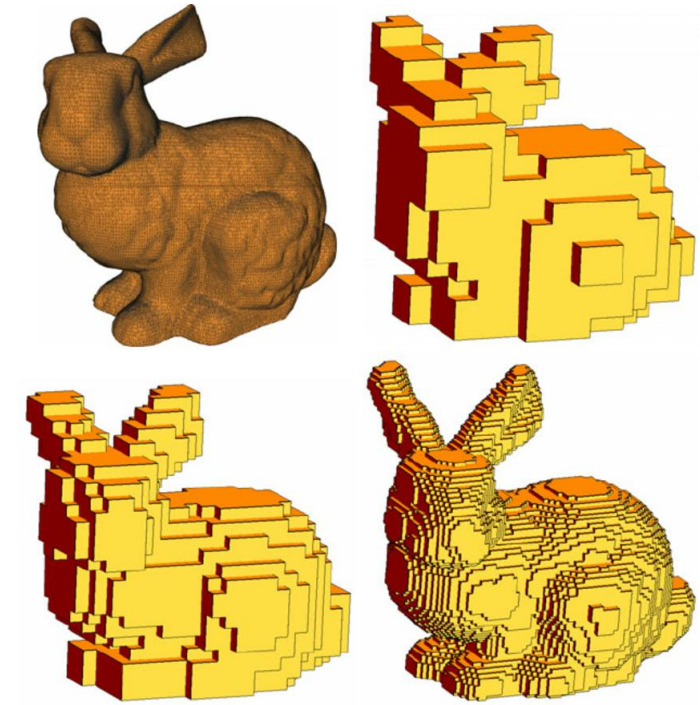
Why learning from images
when I can incorporate an explicit
structural bias (3D reconstruction)?

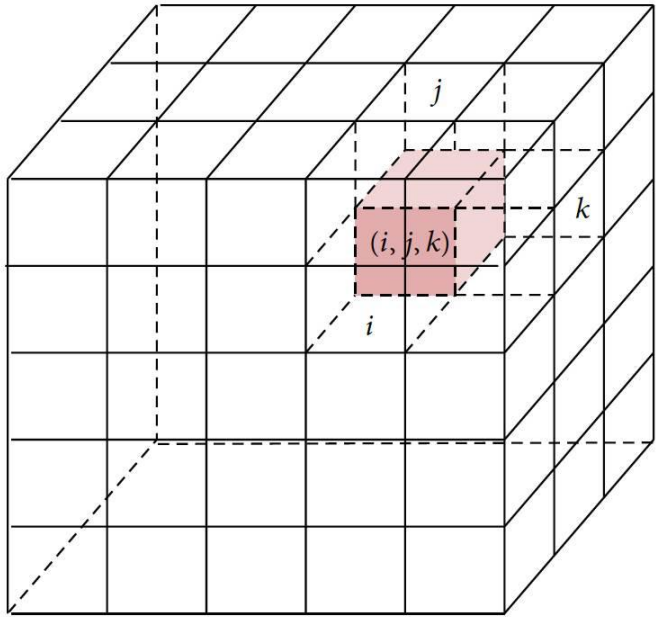
Converting 3D data to a fixed grid



https://www.researchgate.net/figure/Drawing-of-the-LVS-descriptor-with-cubic-volume-definition-voxelization-and-voxel-label_fig3_328129223

Voxel - Occupancy

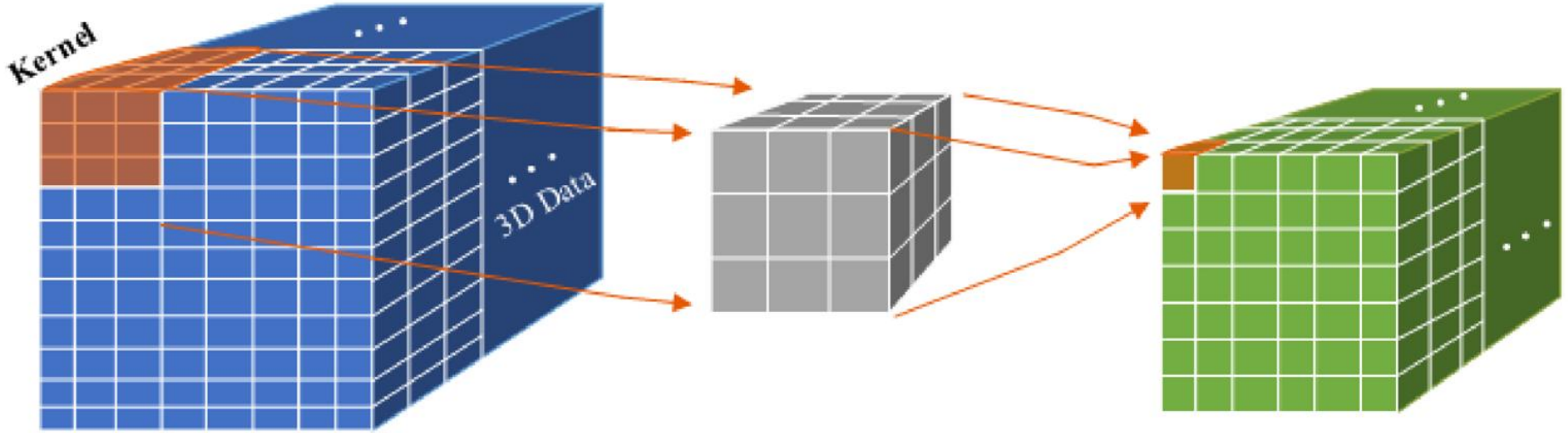




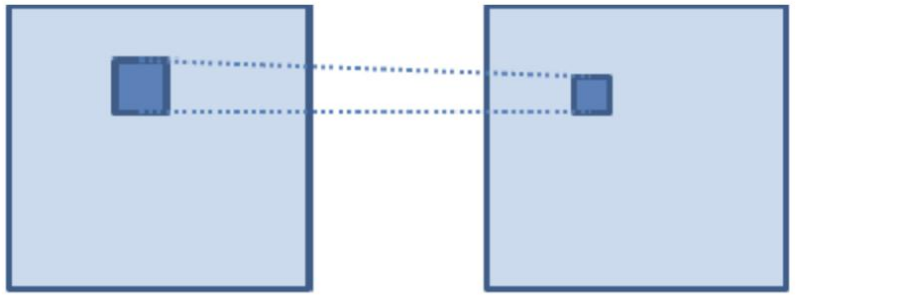
A grid (N x N x N)

https://www.researchgate.net/publication/221016806/figure/fig1/AS:305420410343424@1449829308153/Stanford-Bunny-and-its-covers-for-different-grid-sizes-Top-left-original-object-n.png
https://discourse.vtk.org/uploads/default/original/2X/3/34fe8488730c7da4f3e3ccd8d44bdd4d7e34693b.jpeg

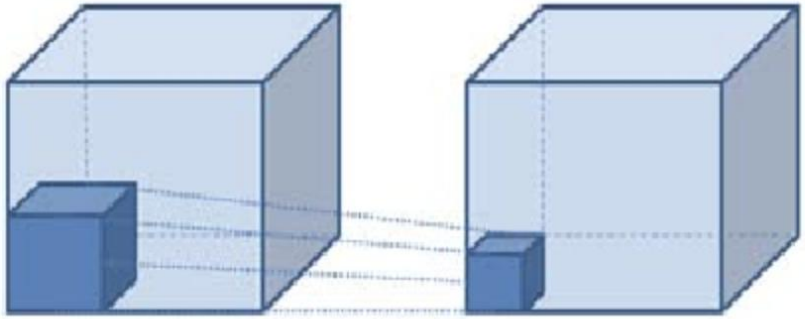
3D Convolutional network



Voxel - 3D Convolution



(*a*) 2D convolution



(*b*) 3D convolution

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

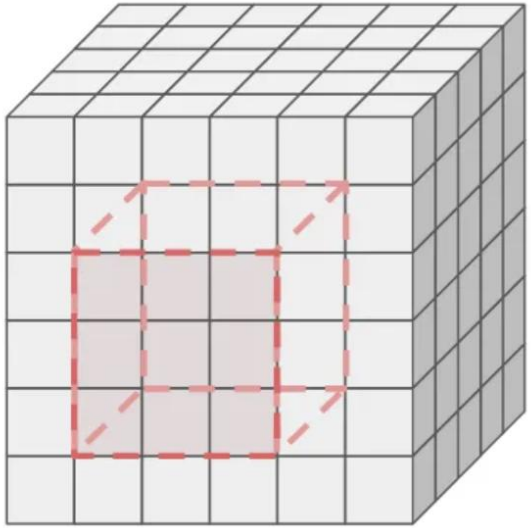


Fig. 7: Example with a 3×3×3 kernel sliding across the voxel region

https://paperswithcode.com/method/3d-convolution

https://medium.com/@tvial_77168/convolutional-neural-networks-in-a-3d-world-30d66b304bfd

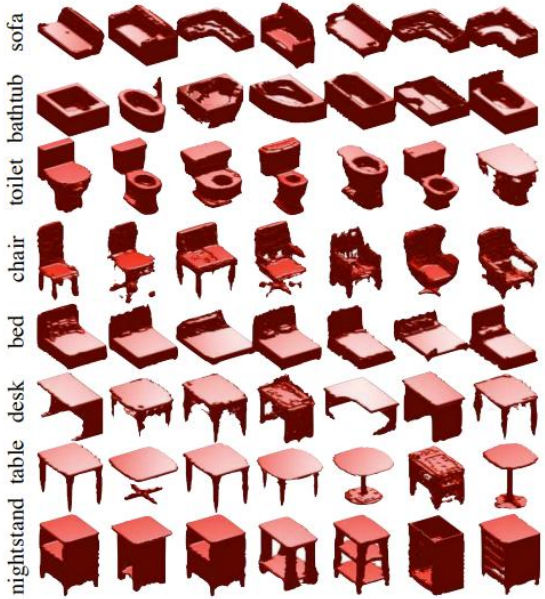
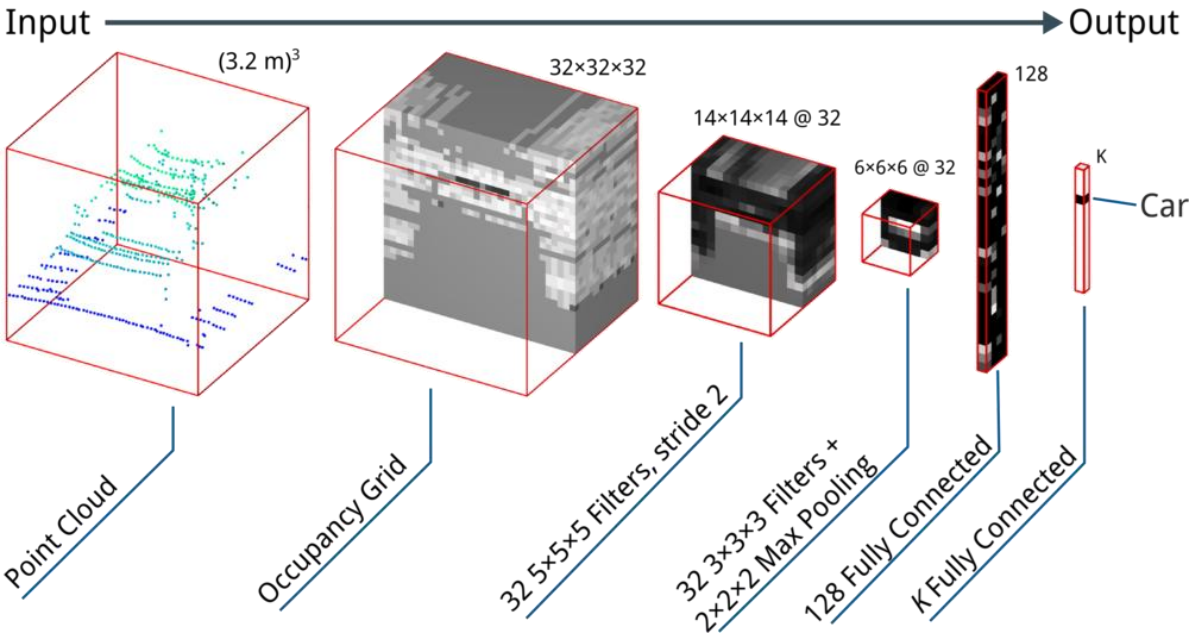


Figure 6: **Shape Sampling.** Example shapes generated by sampling our 3D ShapeNets for some categories.

ShapeNet 3D

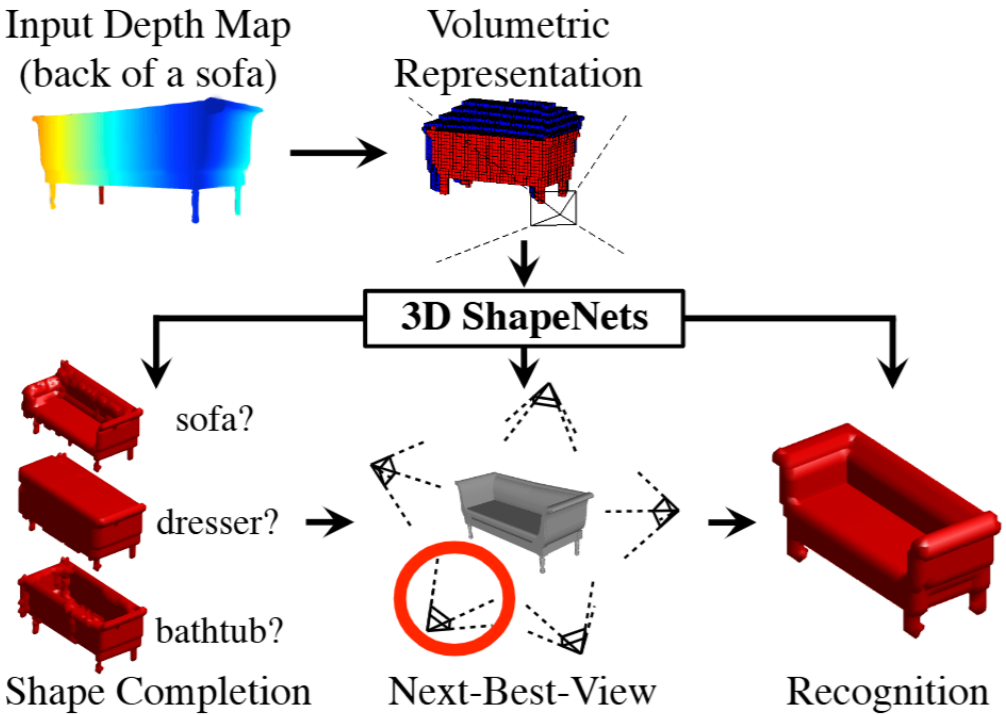
CVPR
June 2015



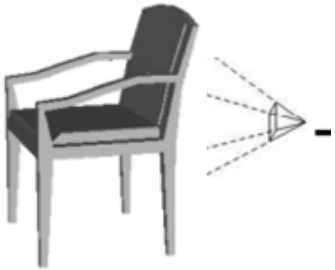
VoxNet

IROS
September 2015

ShapeNet 3D Training (classification&completion)

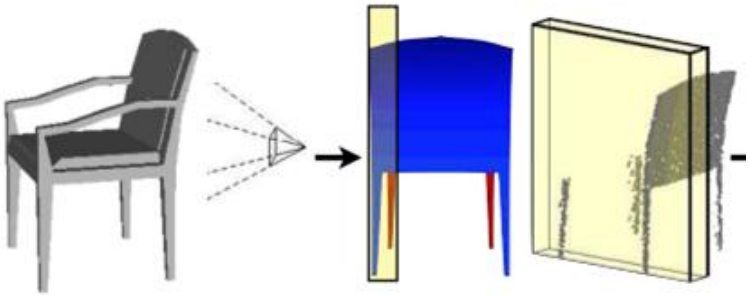


ShapeNet 3D Training (classification&completion)



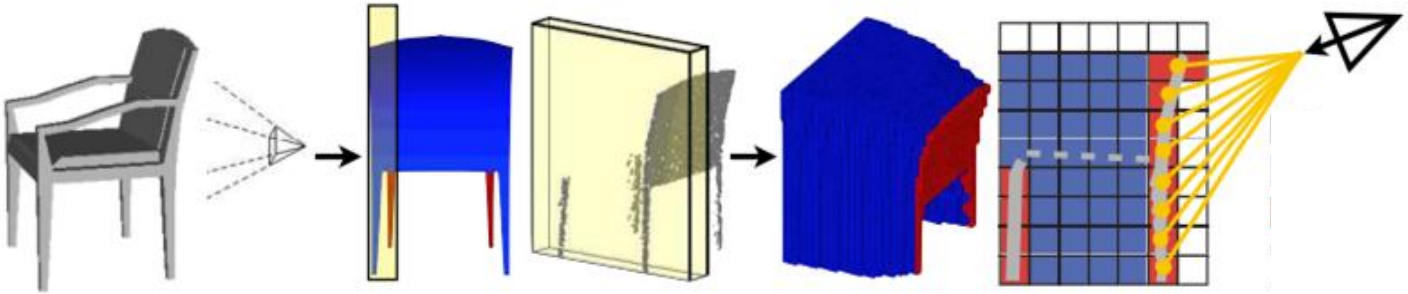
(1) object

ShapeNet 3D Training (classification&completion)



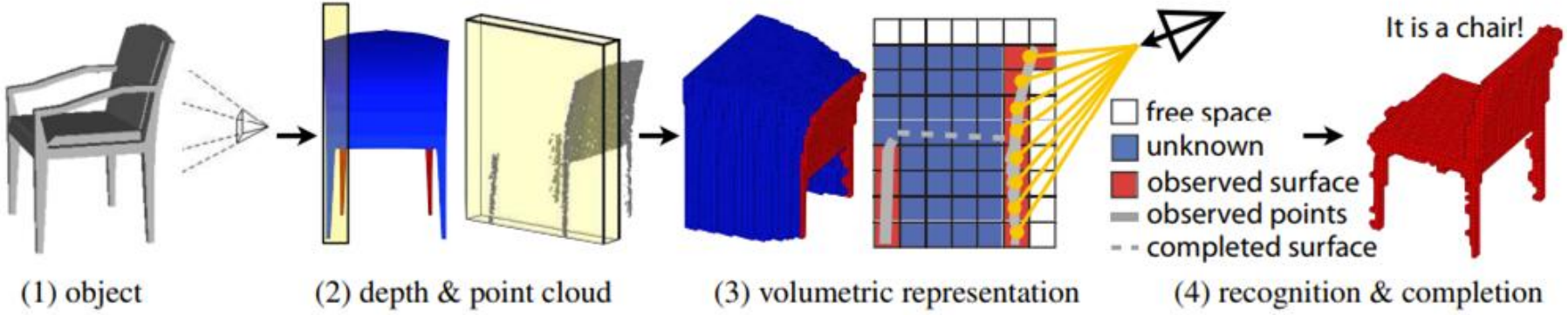
(1) object (2) depth & point cloud

ShapeNet 3D Training (classification&completion)



(1) object (2) depth & point cloud (3) volumetric representation

ShapeNet 3D Training (classification&completion)



Completion

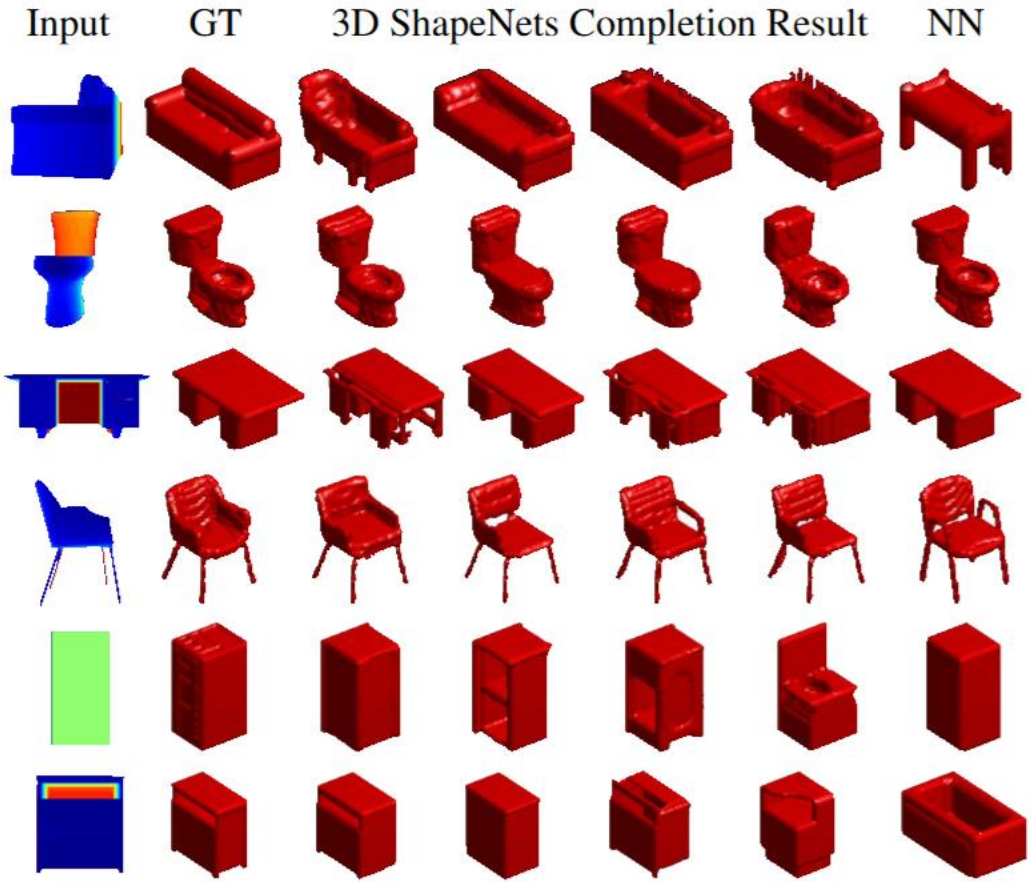
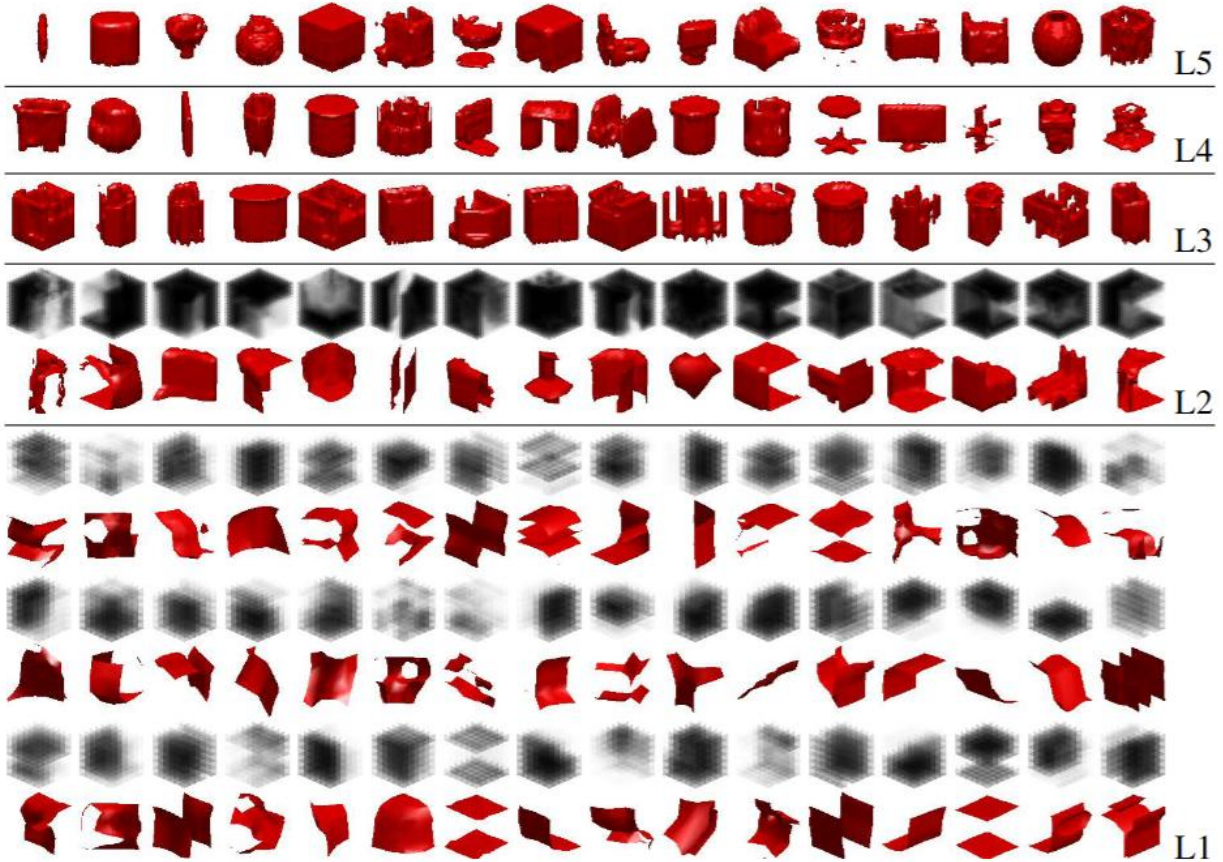


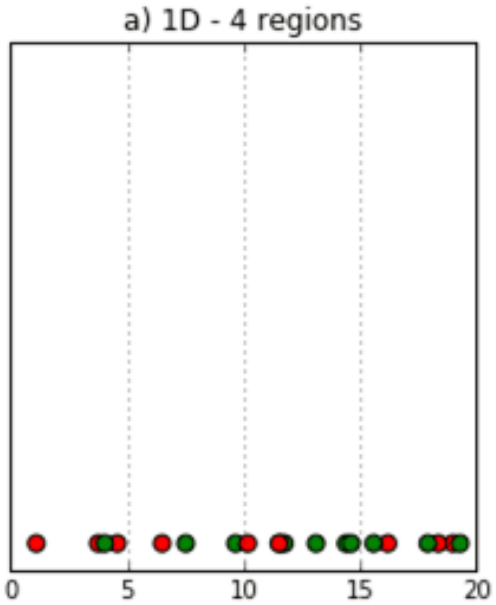
Figure 8: **Shape Completion.** From left to right: input depth map from a single view, ground truth shape, shape completion result (4 cols), nearest neighbor result (1 col).

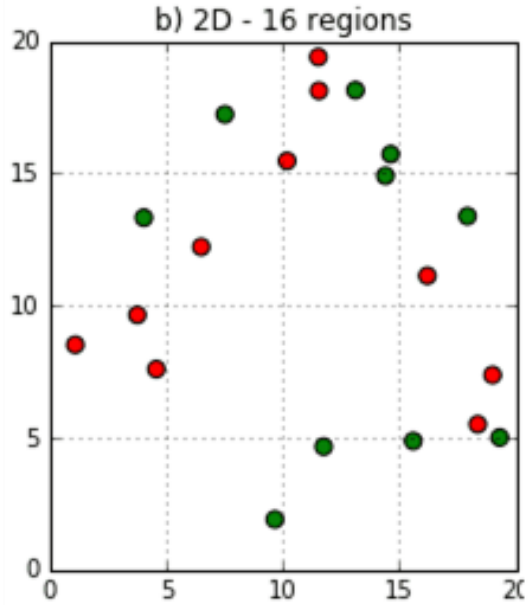
Interpretability study

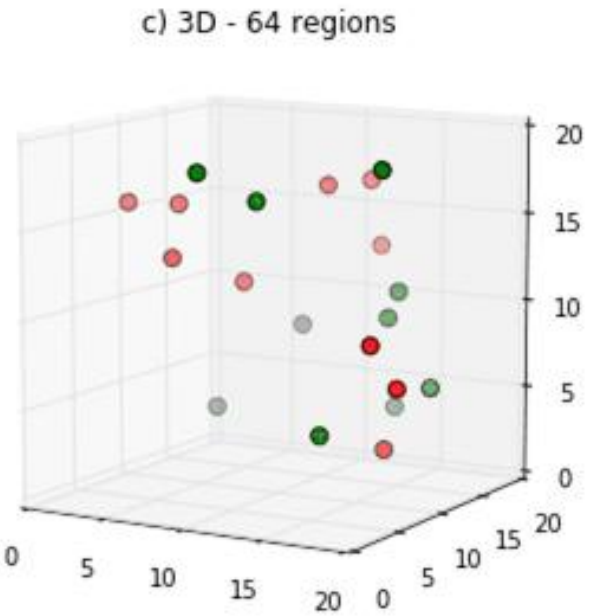


(b) Data-driven visualization: For each neuron, we average the top 100 training examples with highest responses (>0.99) and crop the volume inside the receptive field. The averaged result is visualized by transparency in 3D (Gray) and by the average surface obtained from zero-crossing (Red). 3D ShapeNets 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

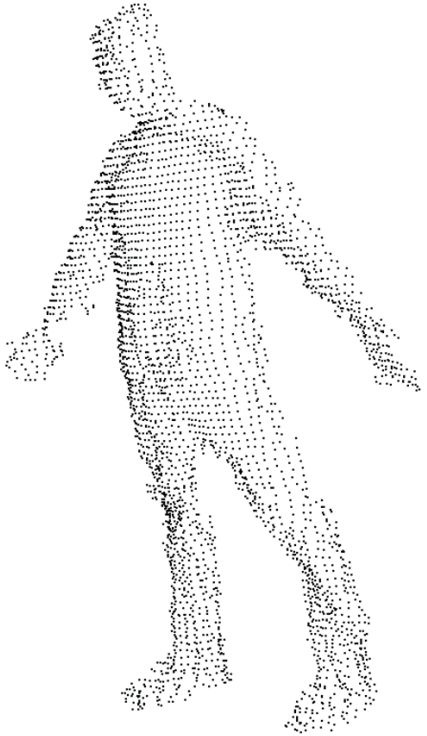






https://www.youtube.com/watch?v=vfL6uJYFrp4&t=2683s

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

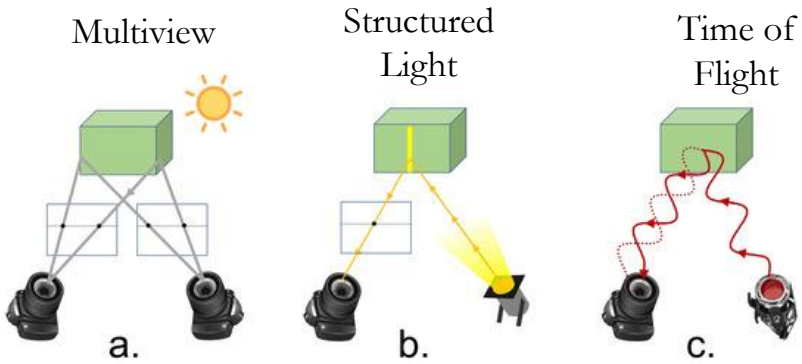
1.1	- 2	0.1
1.5	- 1	0.4
0.1	1.1	0.7
…	…	…

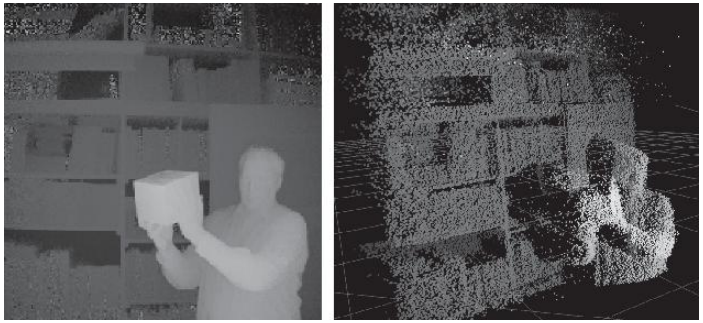
=

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)

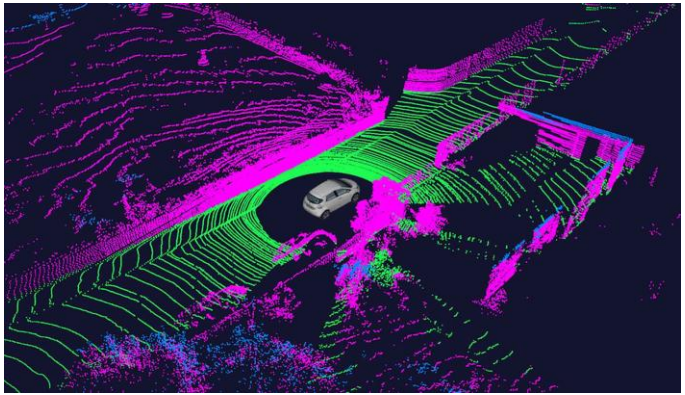




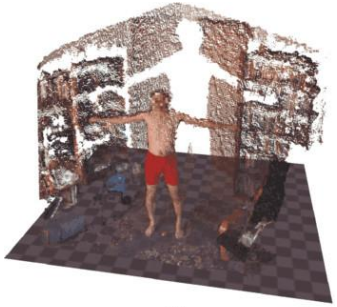
Depth Image

Unprojected 3D Pointcloud

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/Simulation/manual/author/set-up-sensors/configure-a-tof.html

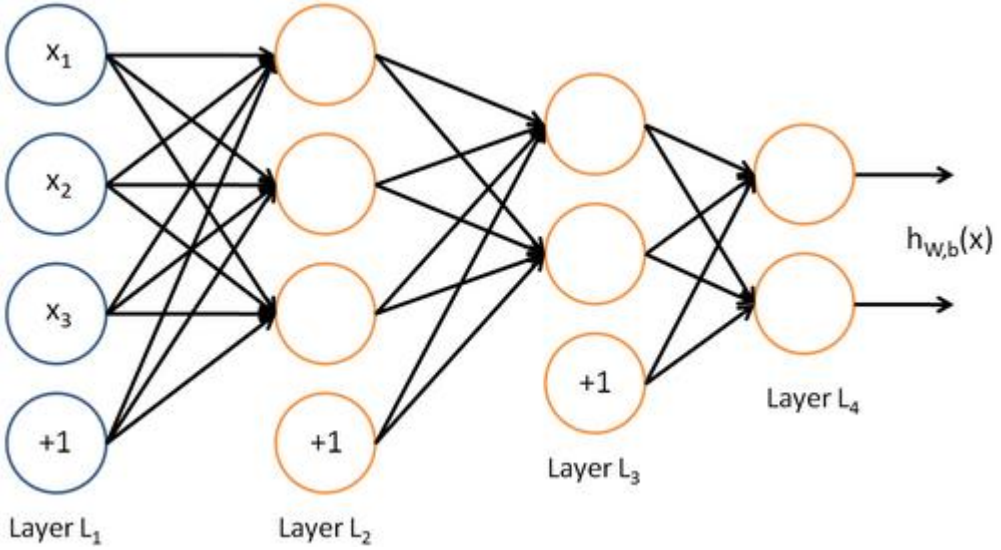
Standard MLP

Features

Cholesterol Level

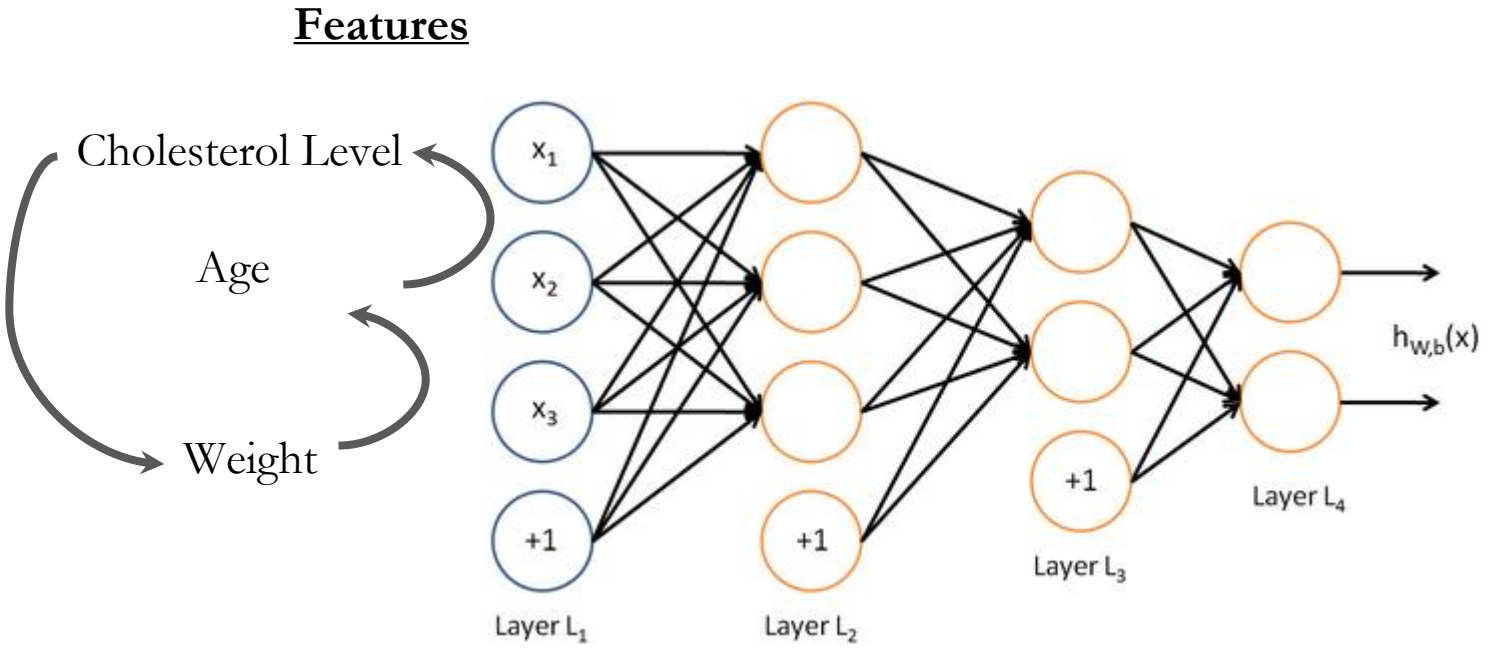
Age

Weight

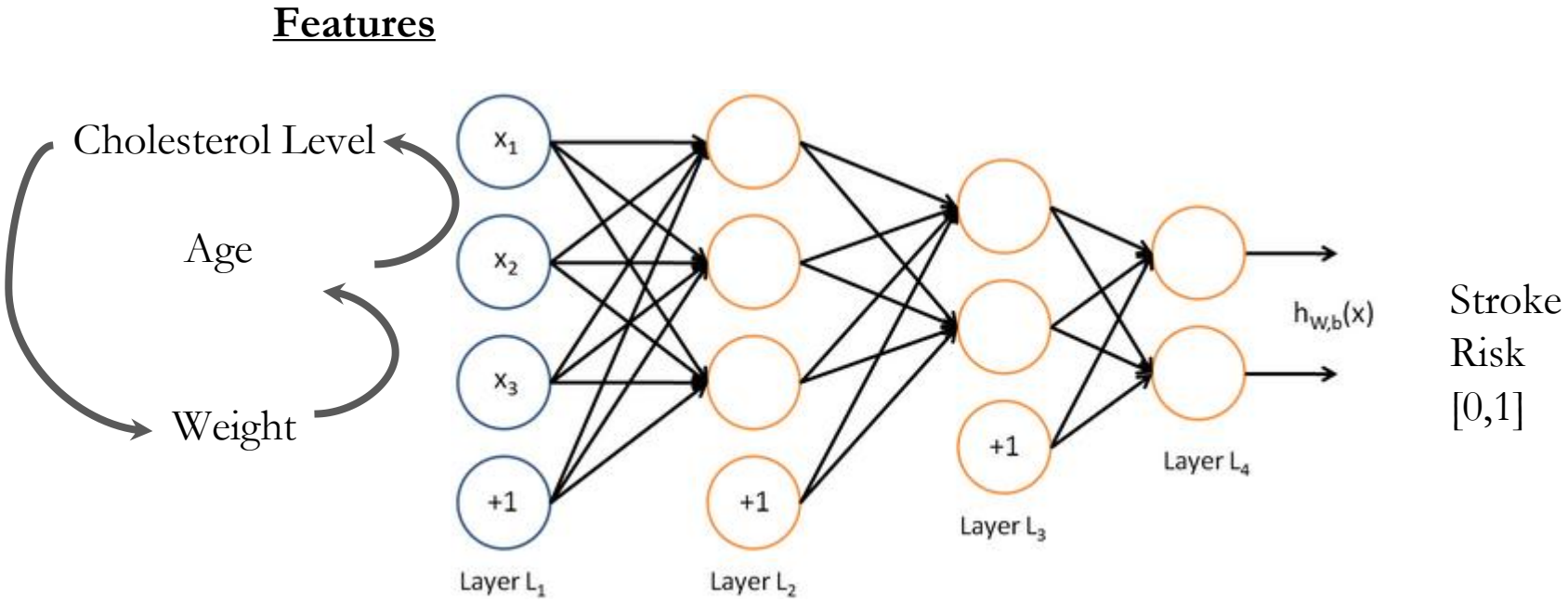


Stroke Risk [0,1]

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 Leonidas J. Guibas
Stanford University

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

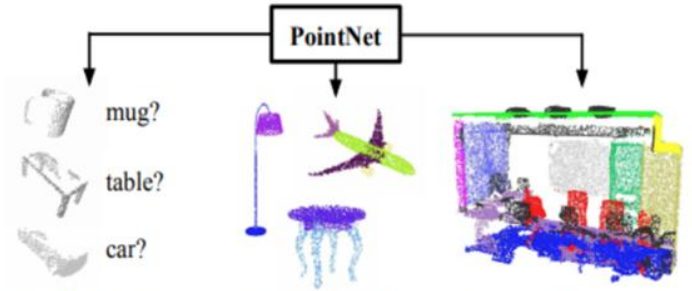
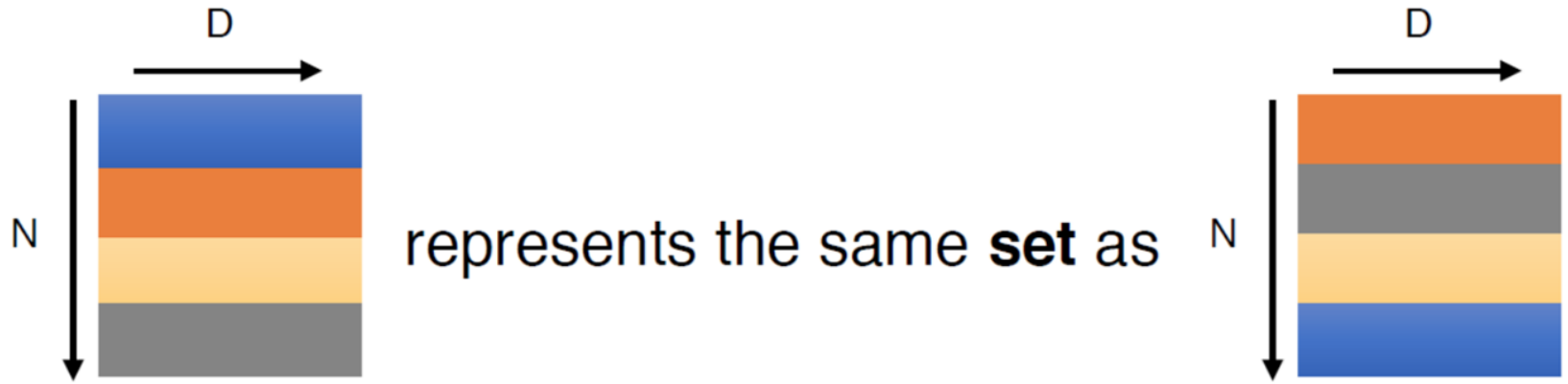


Figure 1. **Applications of PointNet.** We propose a novel deep net architecture that consumes raw point cloud (set of points) without voxelization or rendering. It is a unified architecture that learns both global and local point features, providing a simple, efficient

PointNET properties:

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



"PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation" *Qi et al., 2016*

Permutation invariance

Examples of symmetric functions:

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

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

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

$$f(x_1, x_2, \cdots, x_n) = x_1 + x_2 + \cdots + x_n\}$$

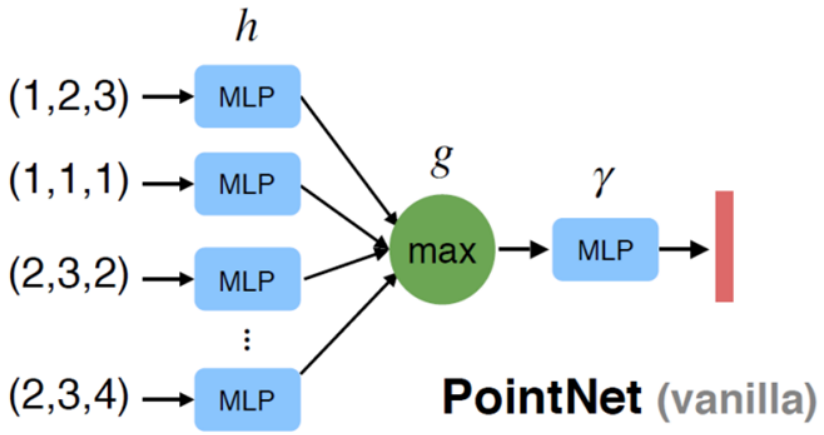
$$\cdots$$

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

Basic PointNet

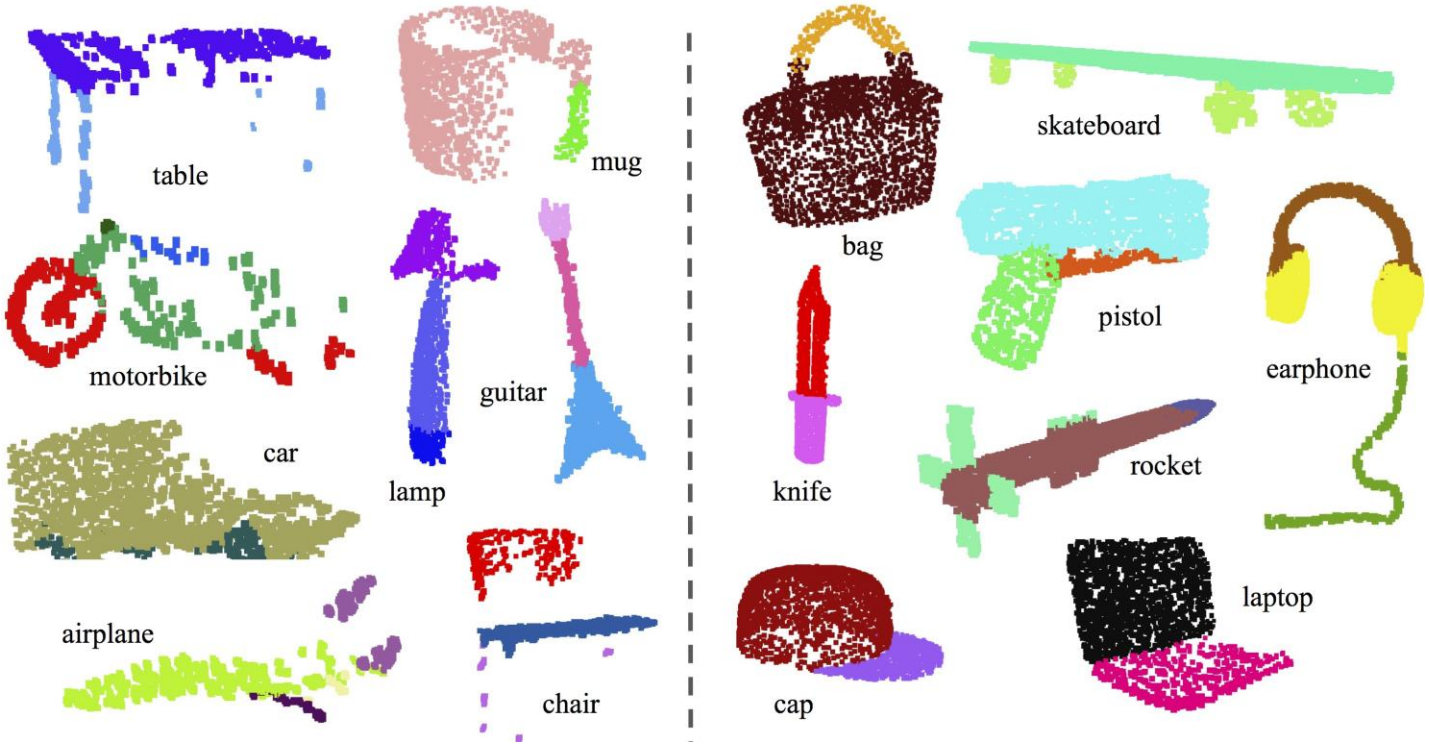
g = max pooling

f, γ = Multi-Layer percepton



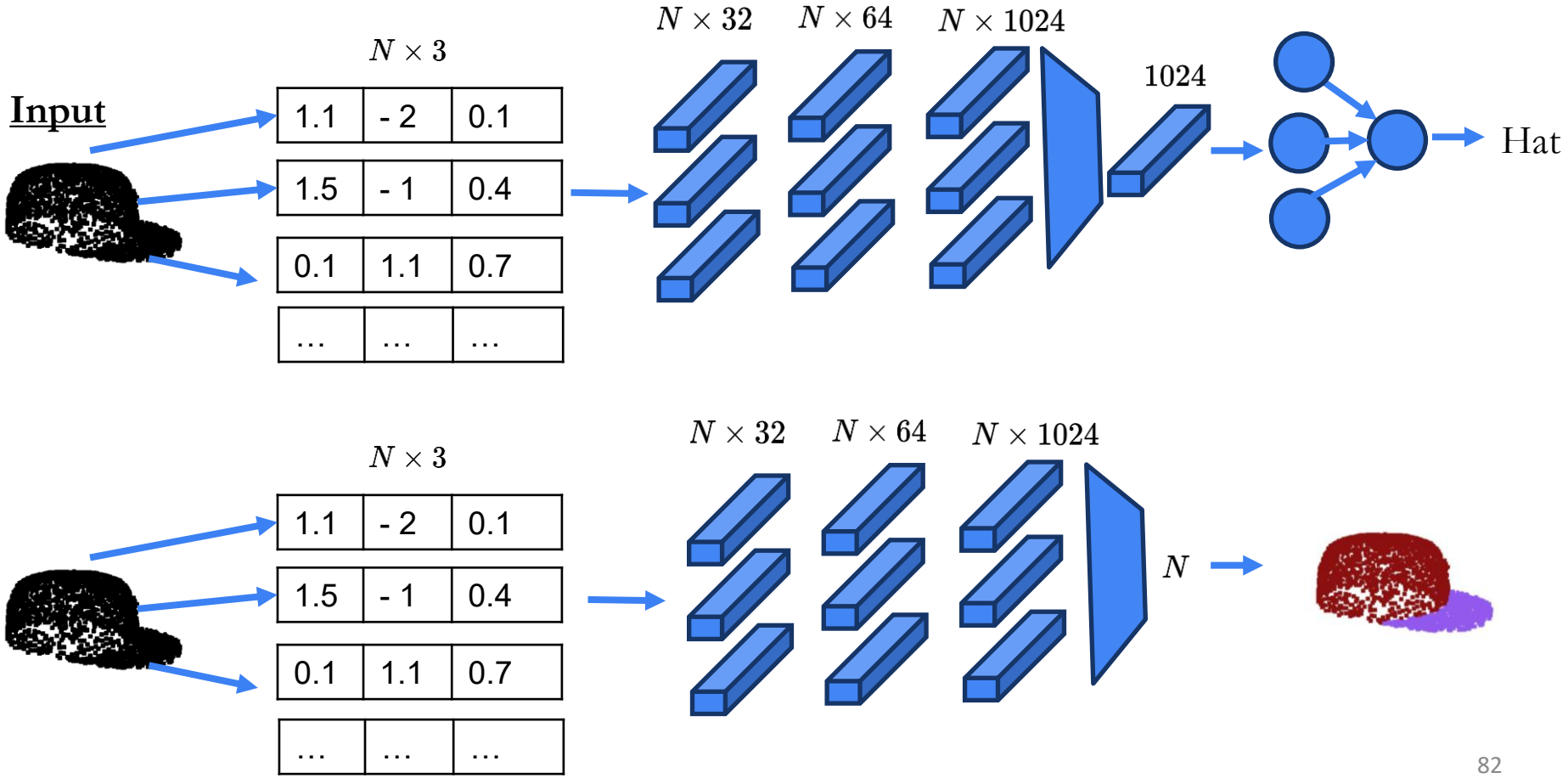
"PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation" *Qi et al., 2016*

Point cloud classification and segmentation

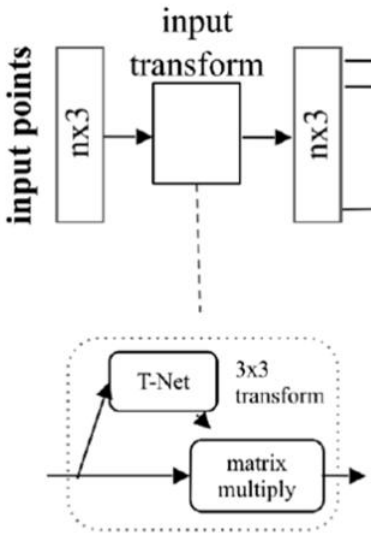


Classification = one label per point cloud
Segmentation = one label per point

Point cloud classification and segmentation

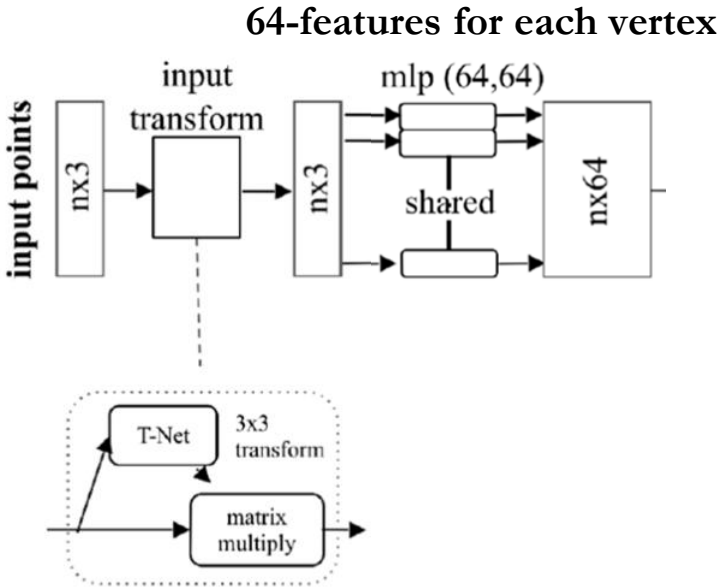


Point cloud classification

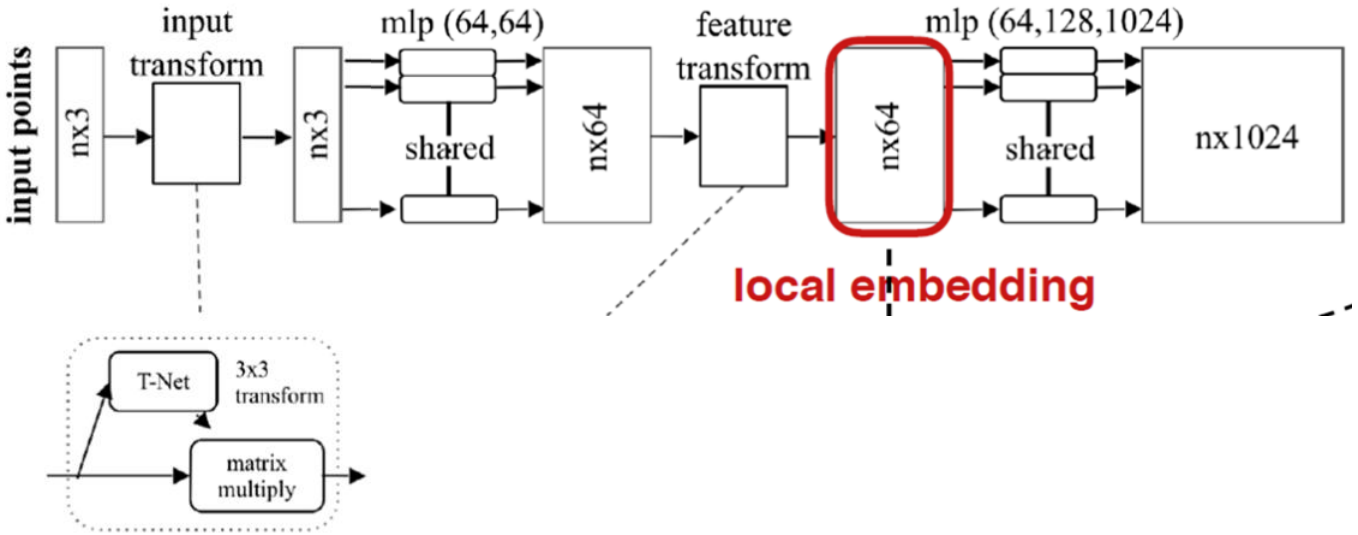


The network learns a canonicalization

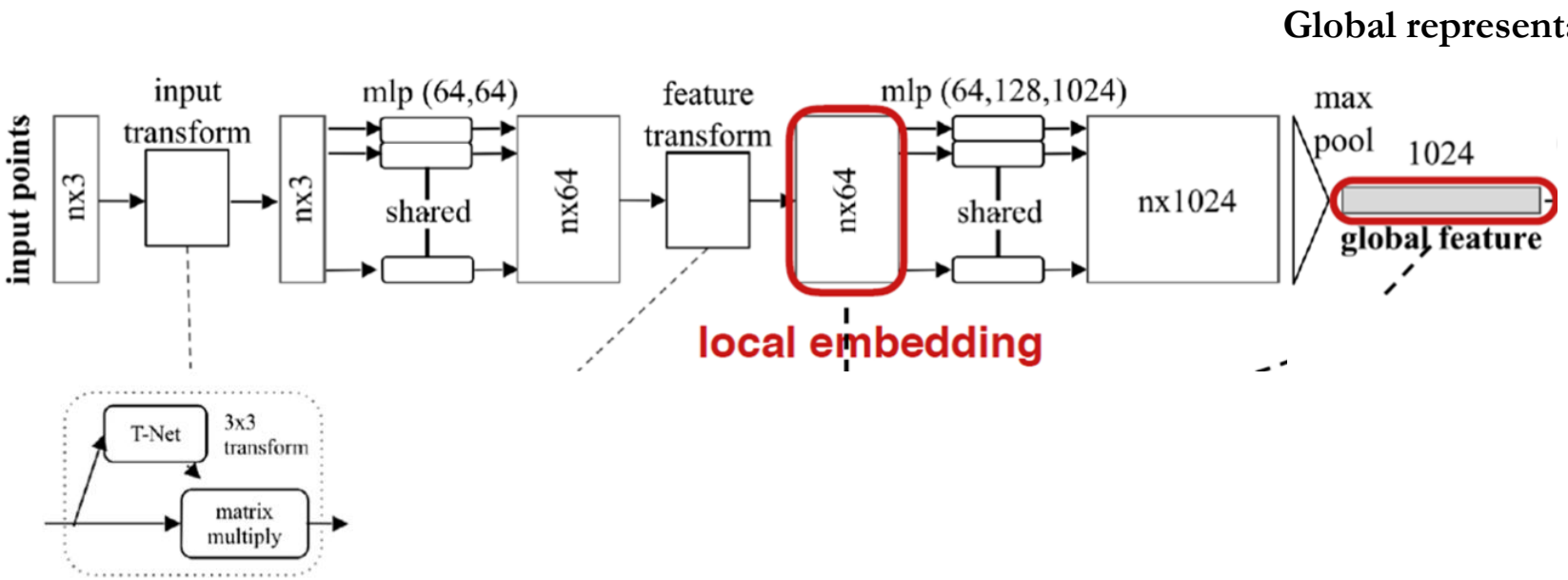
Point cloud classification



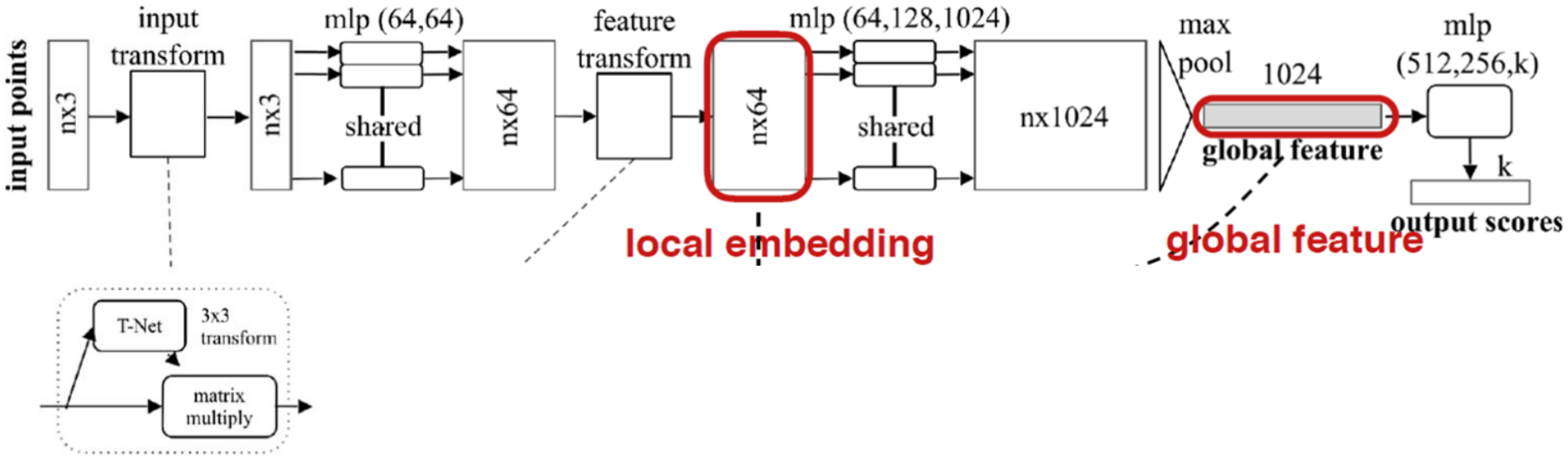
Point cloud classification



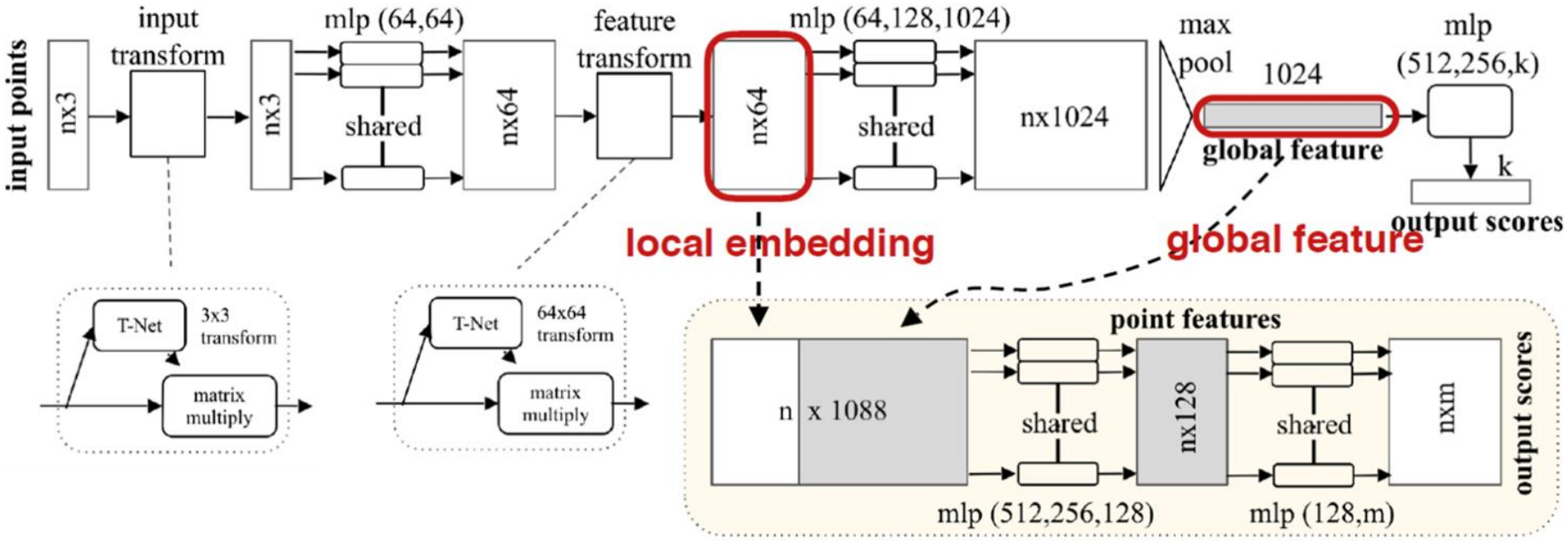
Point cloud classification



Point cloud classification



Point cloud classification and Segmentation



Point cloud Segmentation

Important points for global features:

Original sets of points

Points represented in the global feature

Which ones are these points?

"PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation" *Qi et al., 2016*

Sailency Study

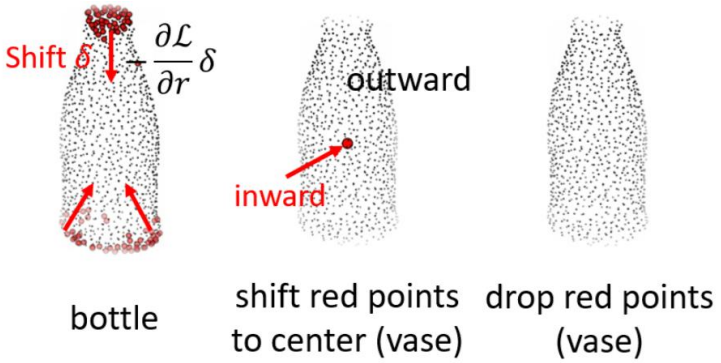


Figure 2. Approximate point dropping with point shifting toward the point-cloud center.

1. Given an input point cloud $\mathbf{P}$ and the associated class c, first of all we compute the center of $\mathbf{P}$ as the median of the three individual coordinates of the points:

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

2. For each point, we compute the vector that connect the center $\mathbf{p}_m$ to it:

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

3. We cast $\mathbf{P}$ 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.

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

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

5. We construct the point-wise saliency map as:

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

6. 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:

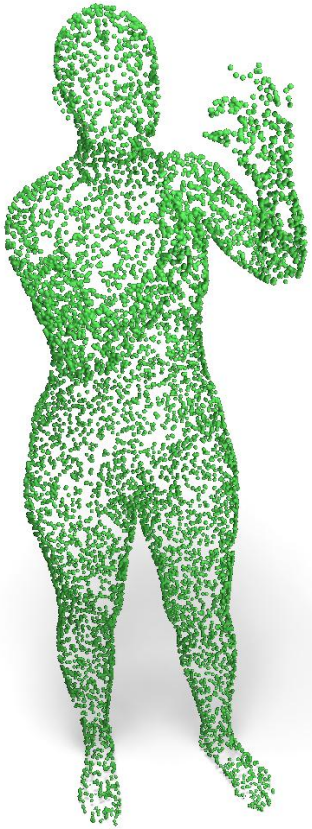
$$\tilde{\mathbf{p}}_j = \mathbf{p}_j - 0.05\mathbf{r}_i \quad (16)$$

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

PointCloud Saliency Maps, Zheng et al., ICCV19

Input

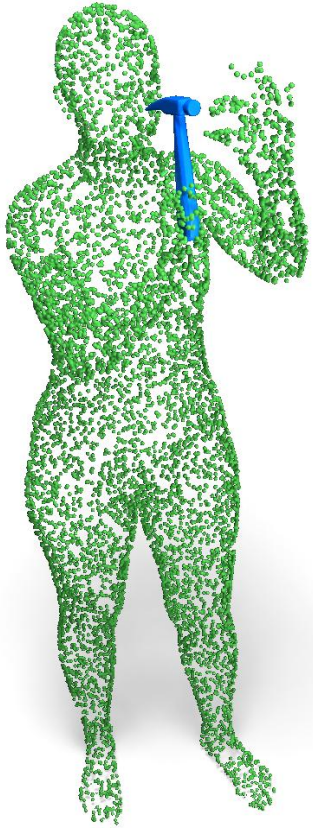
Point Cloud



Object pop-up

Output

Posed object



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




Ilya A. Petrov[1,2] Riccardo Marin[1,2] Julian Chibane[1,3] Gerard Pons-Moll[1,2,3]

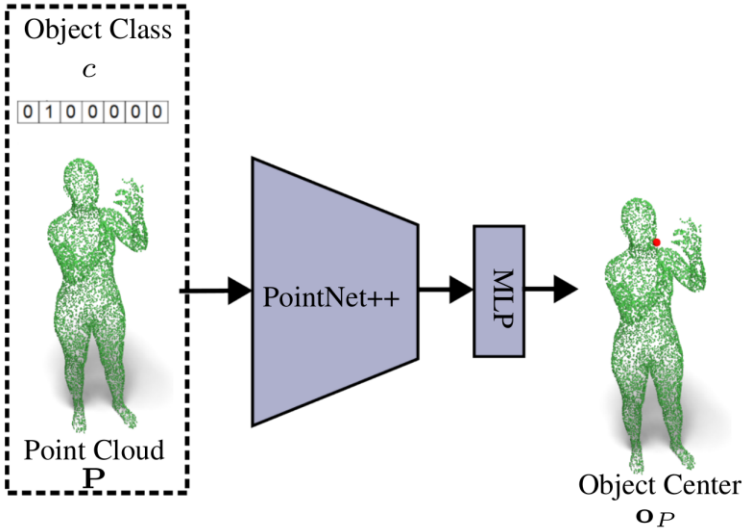
[1]University of Tübingen; [2]Tübingen AI Center;
[3]Max Planck Institute for Informatics, Saarland Informatics Campus

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

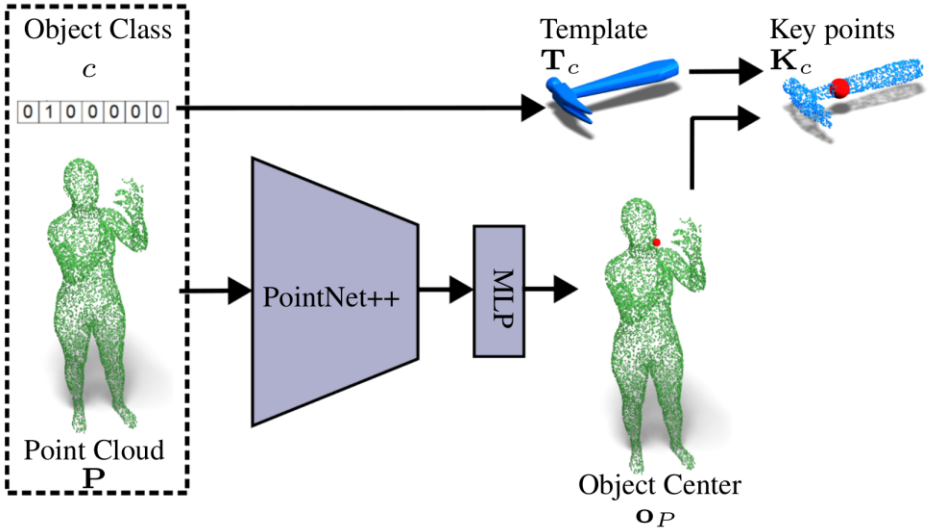
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Includes Audio

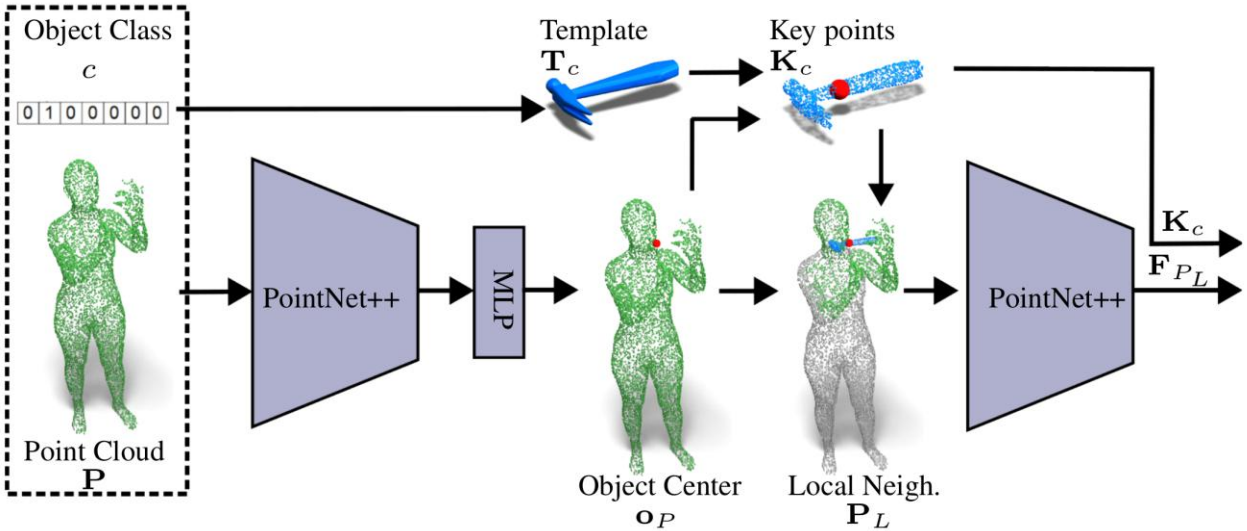
Object pop-up: whole body features and center prediction



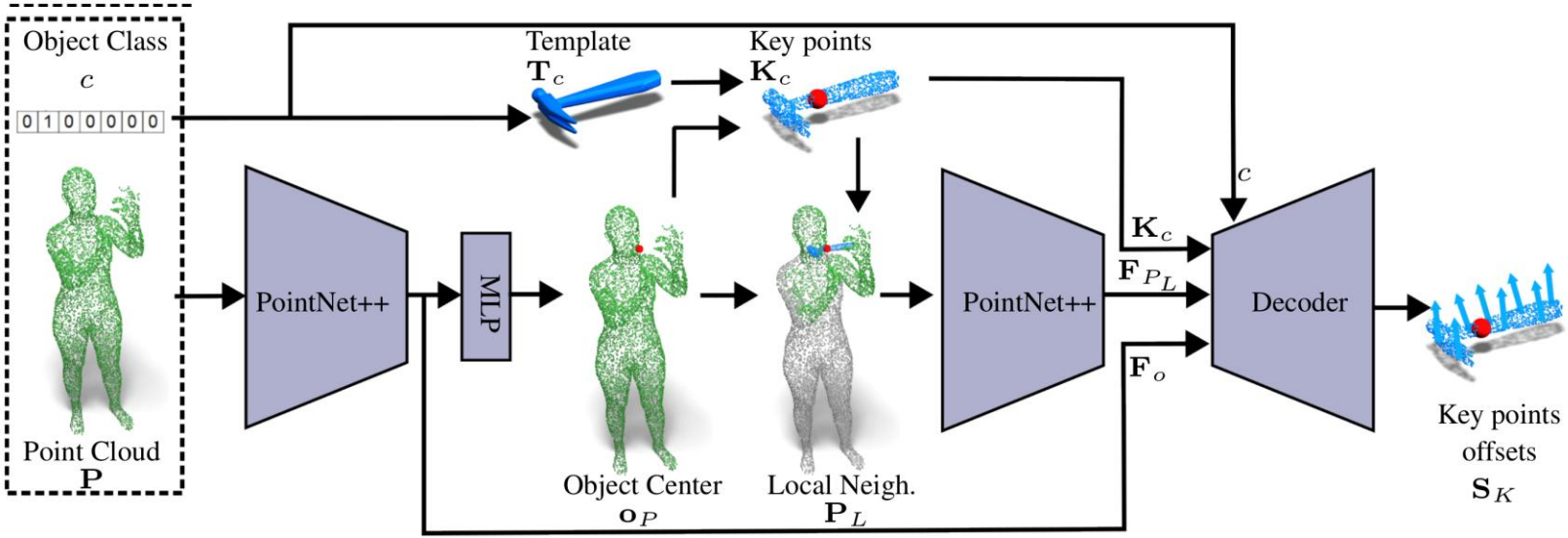
Object pop-up: object keypoints



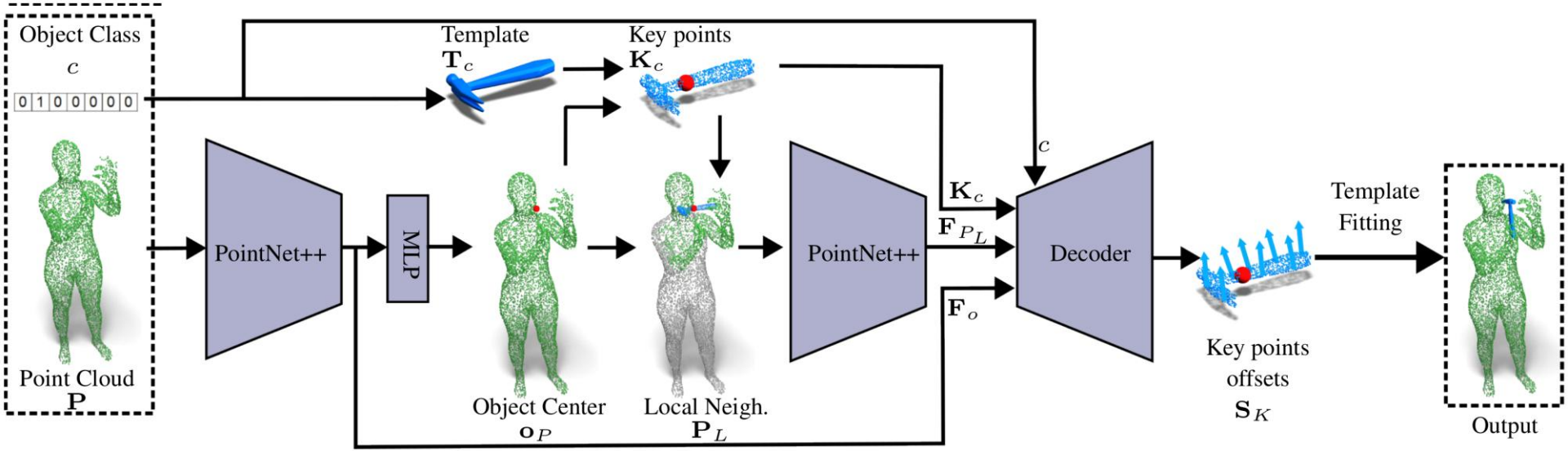
Object pop-up: local features



Object pop-up: per-point offset prediction



Object pop-up: overview




Input

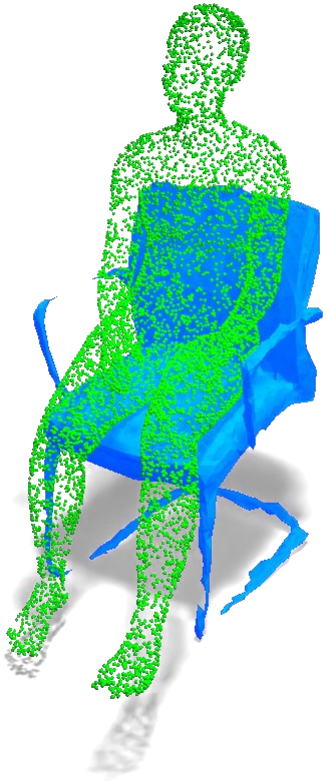

Input

Can we classify an object of interaction?

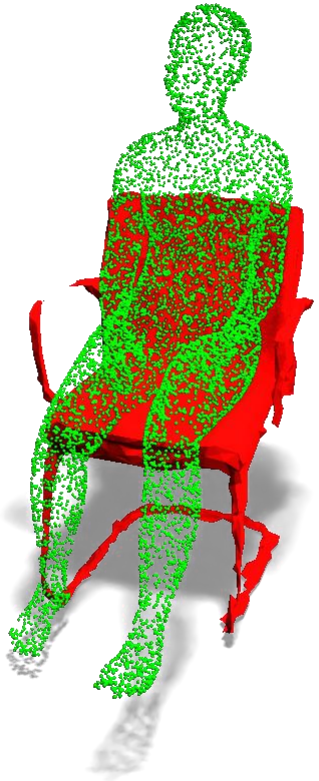
Yes

Input

Prediction



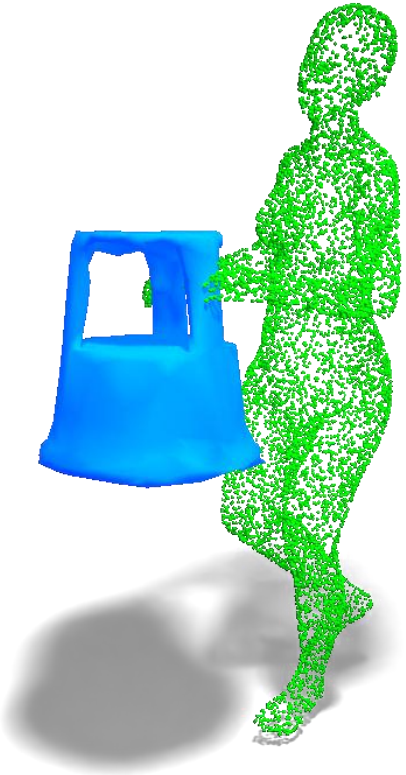
Ground-truth



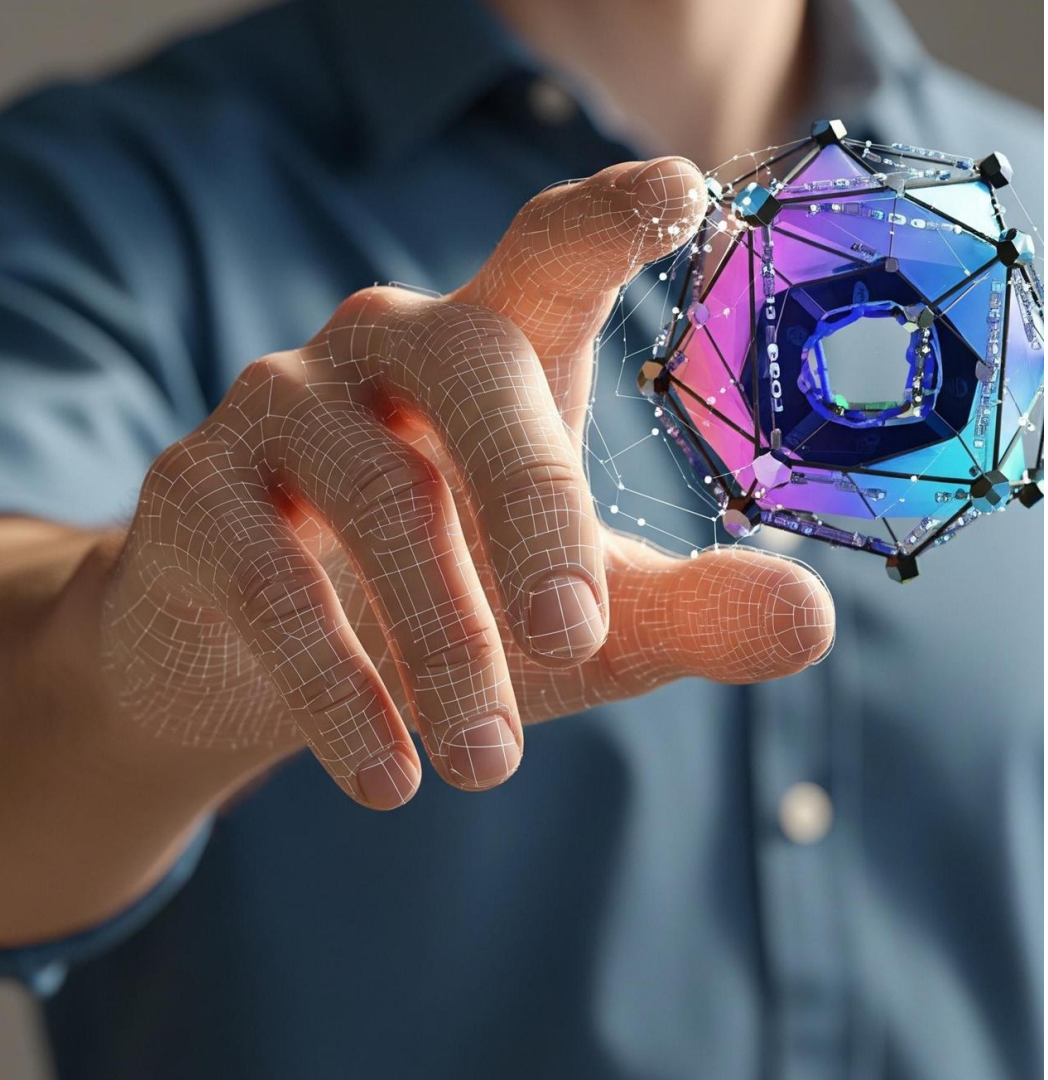
But the task is inherently ambiguous

Input

Prediction



Ground-truth

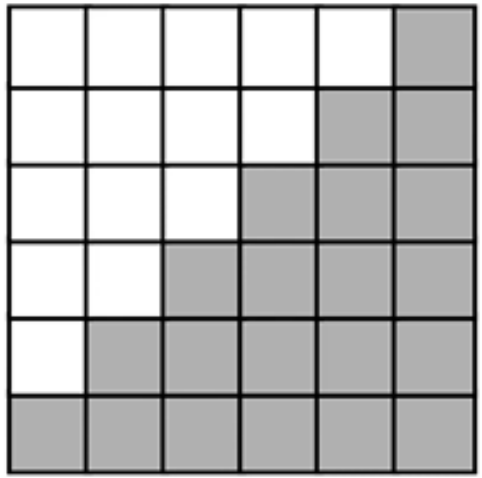


Neural Field
Geometry as networks

Voxels:

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

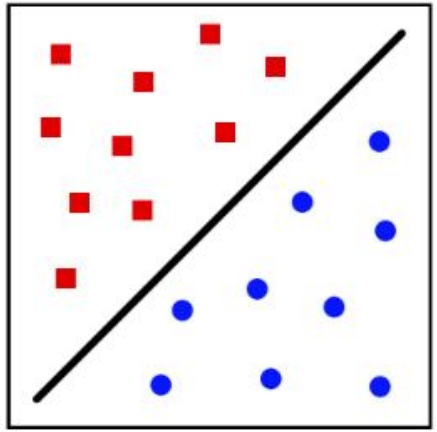
[Maturana et al., IROS 2015]



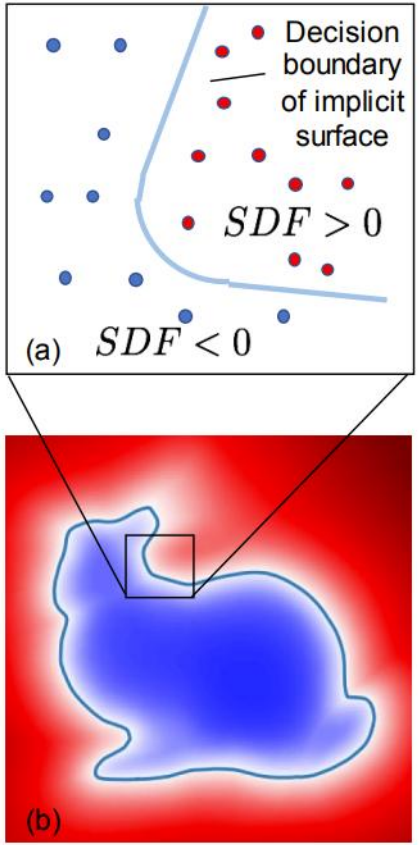
Slides from: Andreas Gaiger, Computer Vision Course

Key Idea:

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

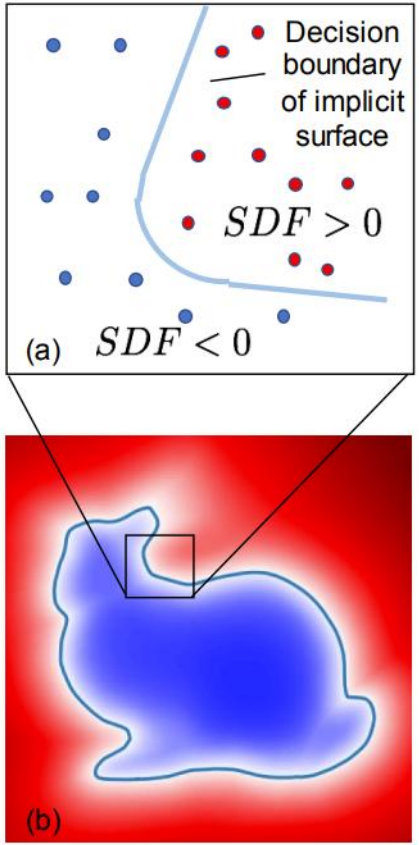


Continous SDF



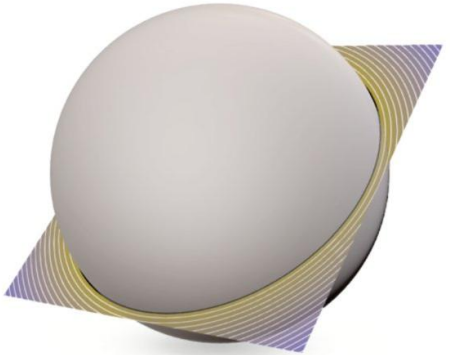
Continous SDF

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



Problem
analytical form is rare

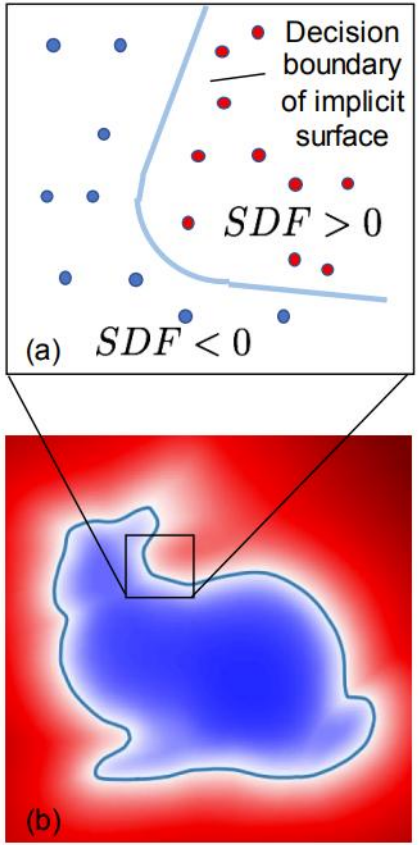
Idea:
Use a parametric function



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

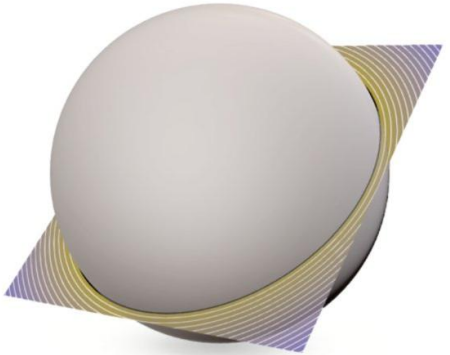
Continous SDF

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



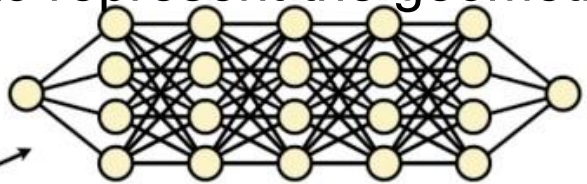
Problem
analytical form is rare

Idea:
Use a parametric function

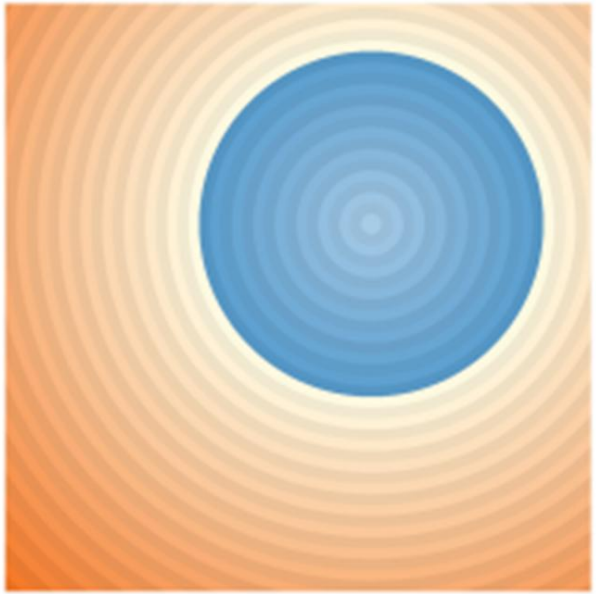


$$f(x, y, z) = \sqrt{x^2 + y^2 + z^2} - 1$$

Training a neural network
to represent the geometry



Voxel - Signed distance field/function



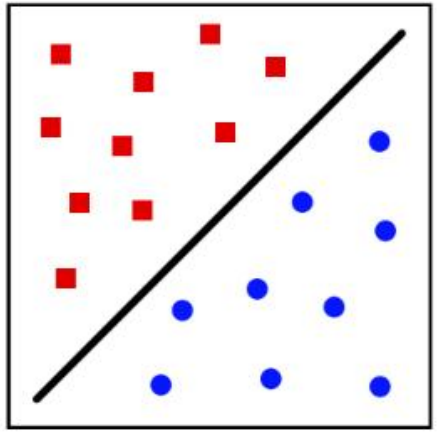
$$f : \mathbb{R}^3 \to \mathbb{R}, f(x) = \begin{cases} d, & \text{if } x \text{ is outside} \\ -d, & \text{otherwise} \end{cases}$$

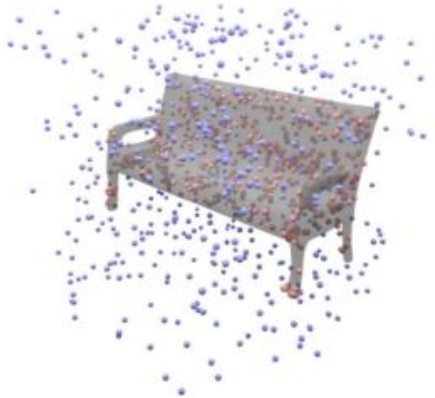
Key Idea:

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

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

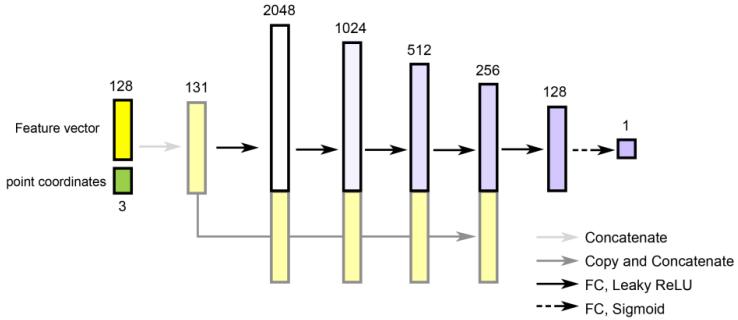
3D Location | Condition (eg, Image) | Occupancy Probability





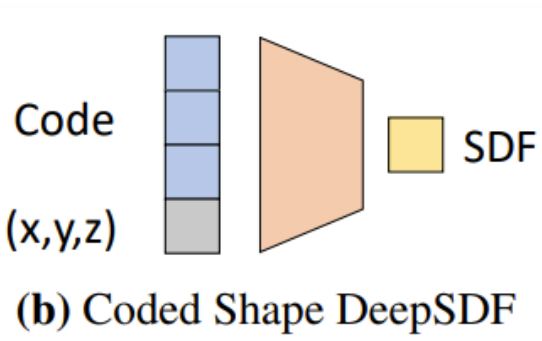
Neural Fields

CVPR19: Three contemporary works proposed almost the same idea



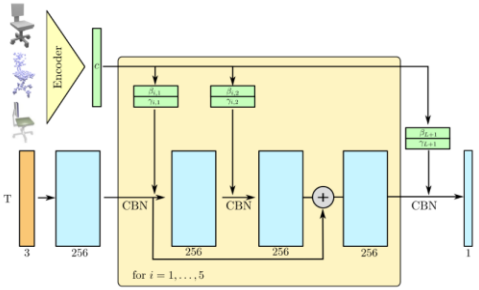
Learning Implicit Fields for Generative Shape Modeling, (IMNET)

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



DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation

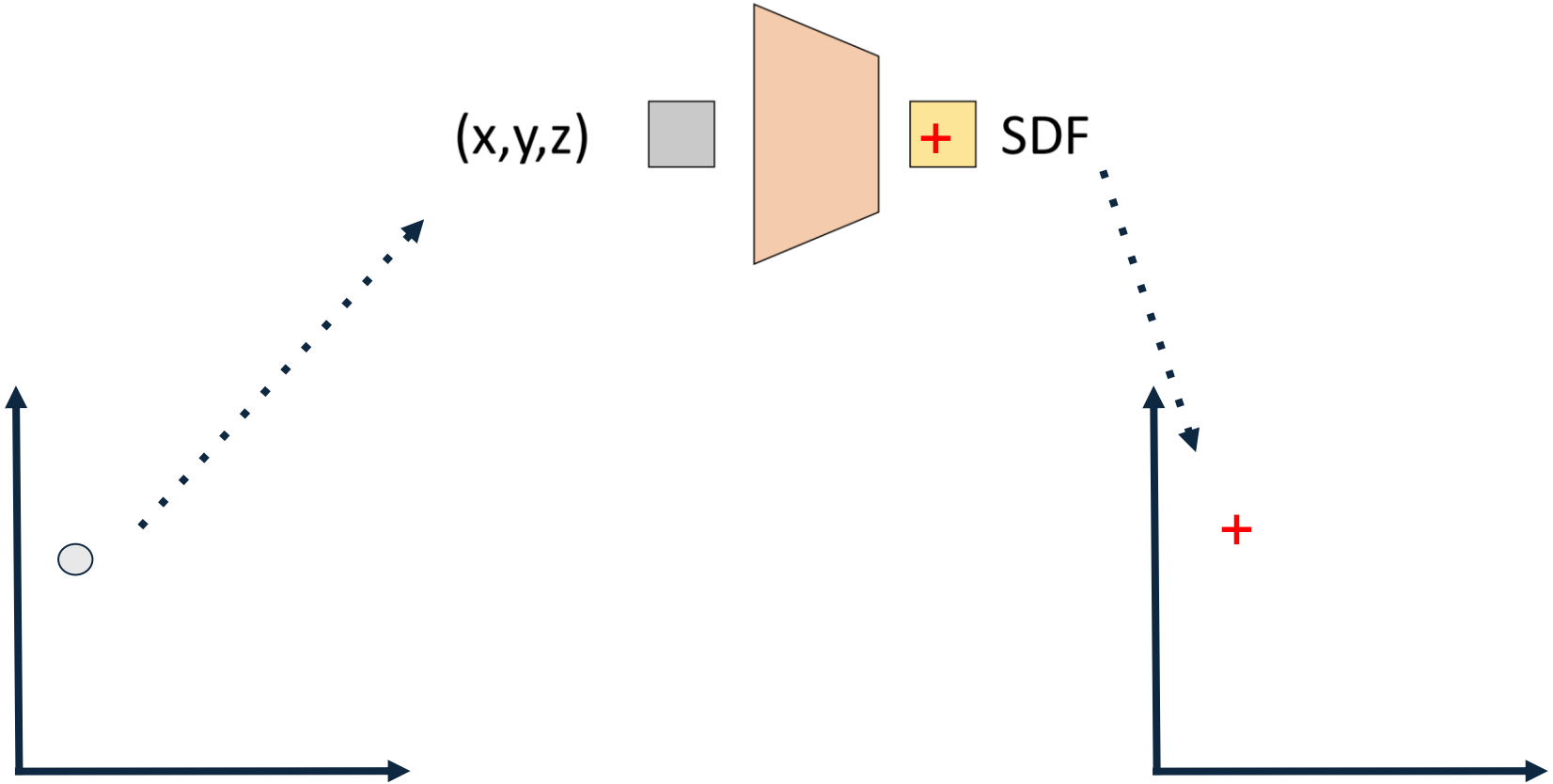
$$f_\theta(\boldsymbol{x}) \approx SDF(\boldsymbol{x})$$



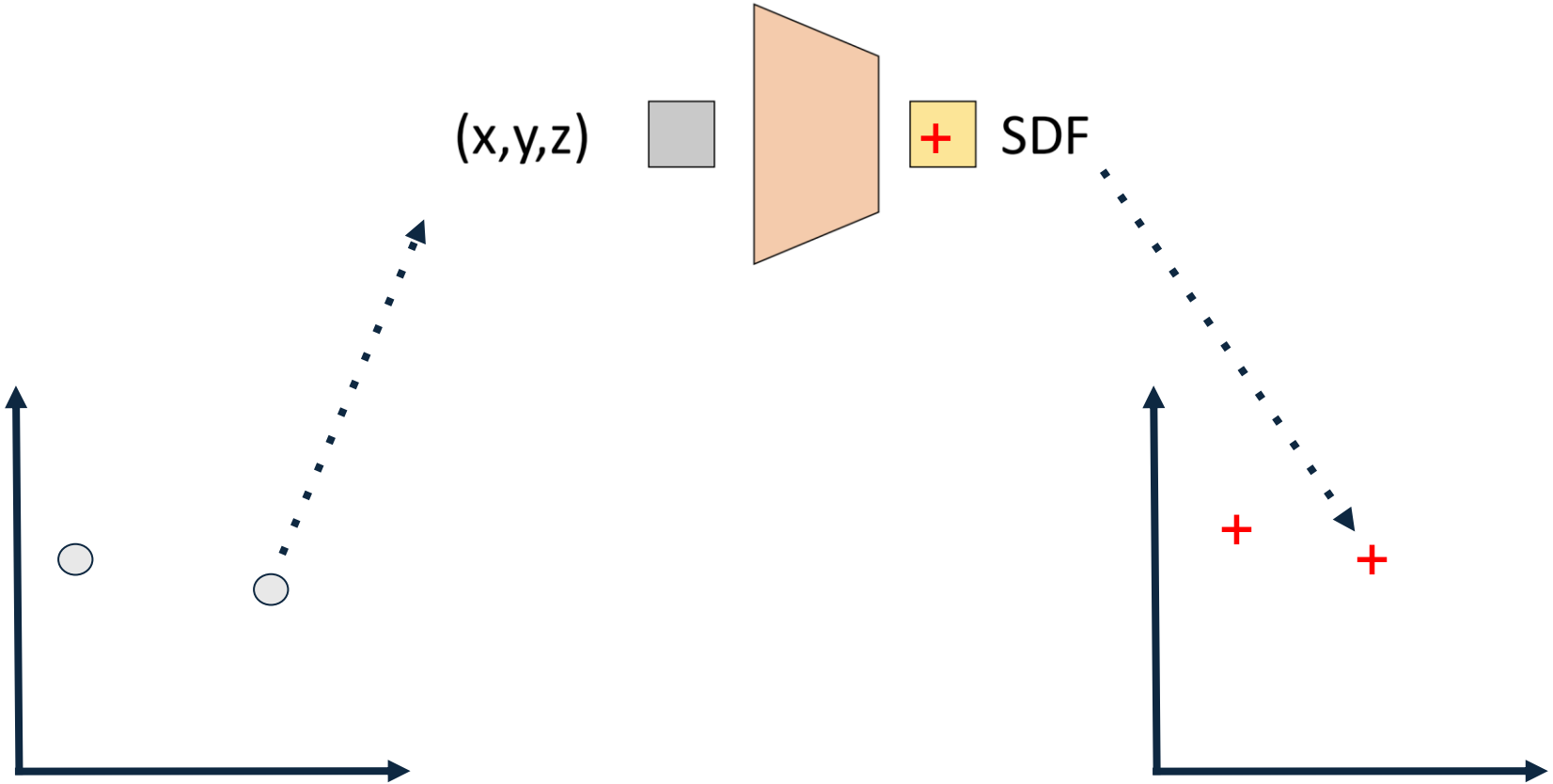
Occupancy Networks: Learning 3D Reconstruction in Function Space.

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

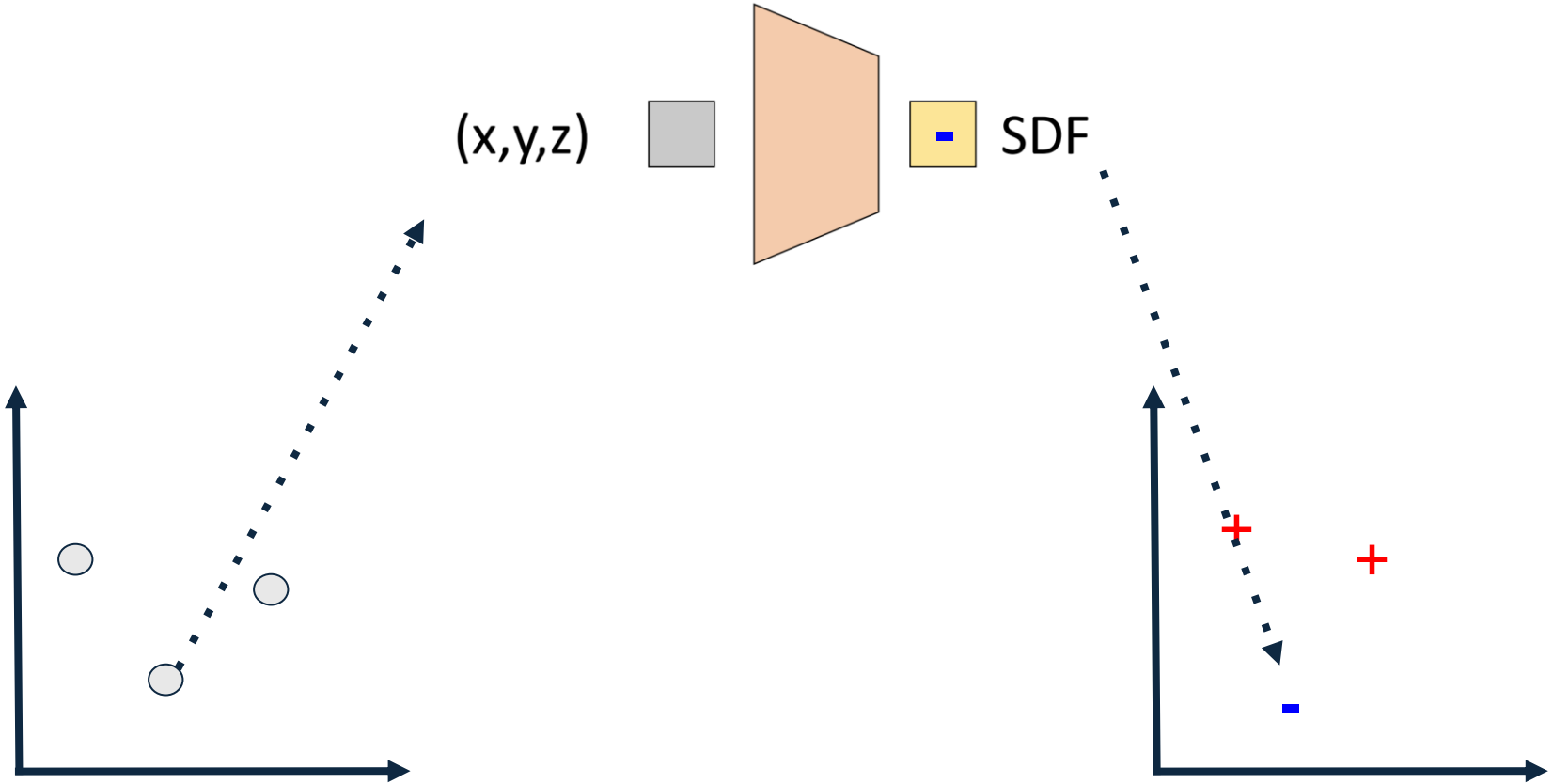
Deep SDF



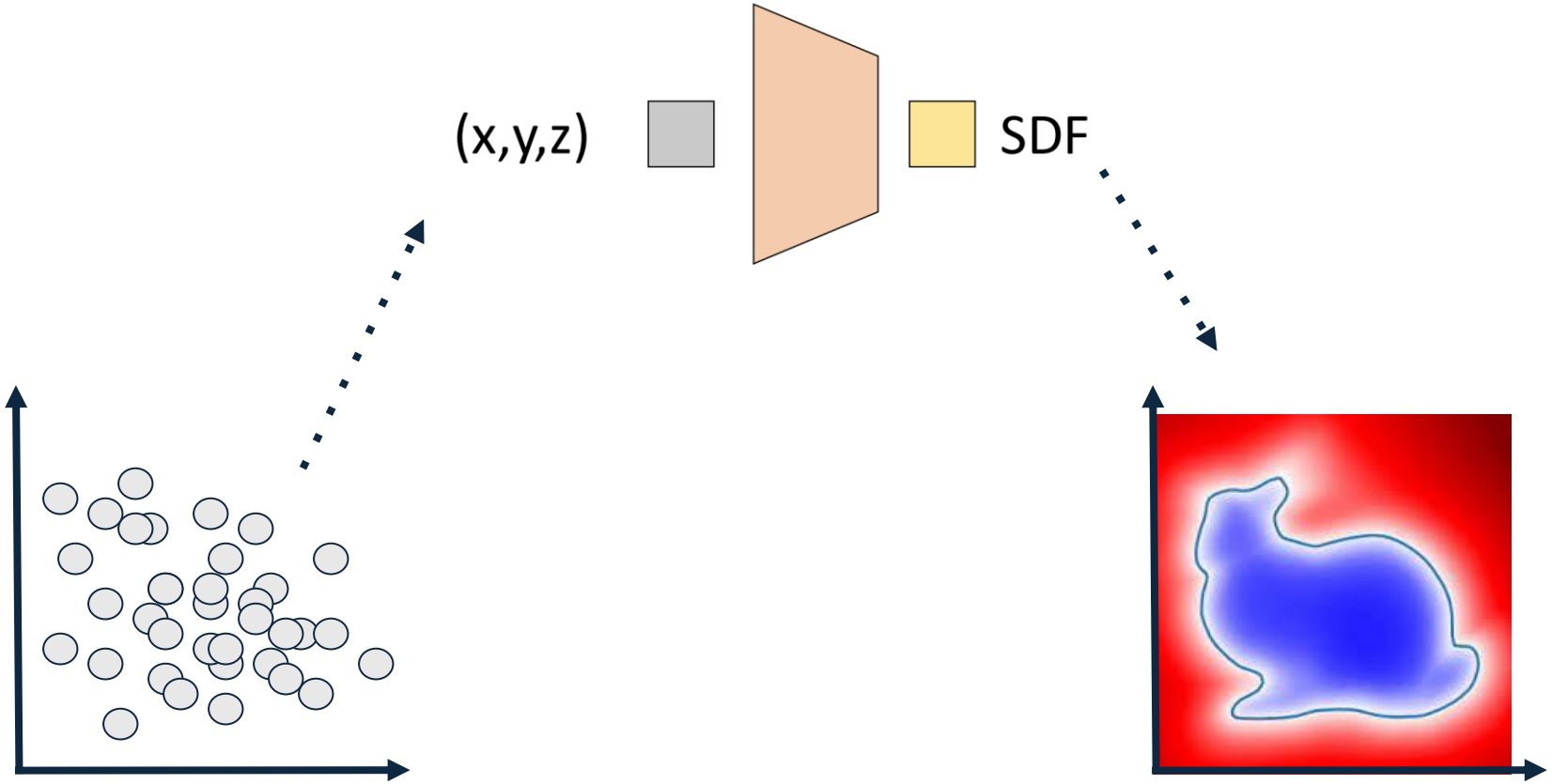
Deep SDF



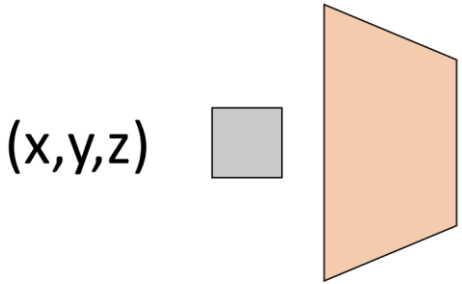
Deep SDF

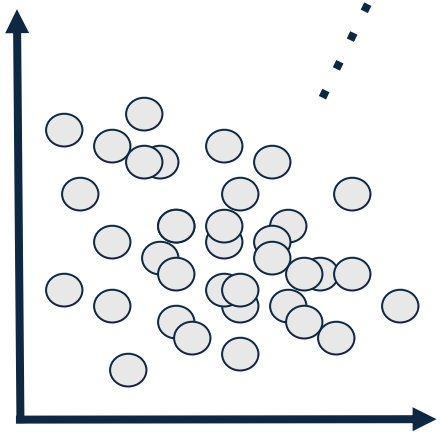


Deep SDF



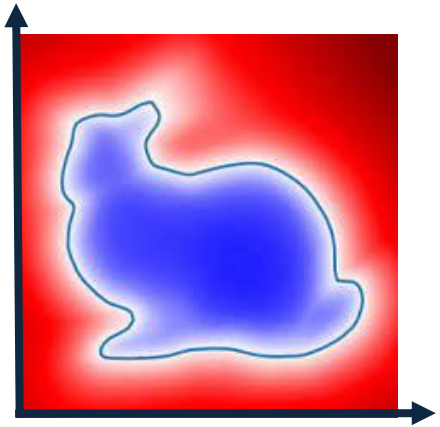
Deep SDF



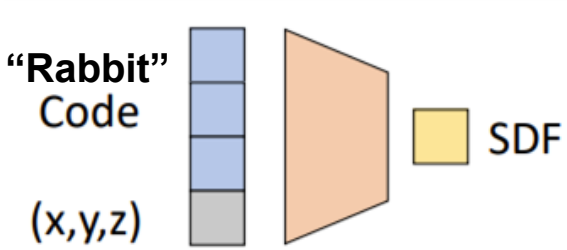


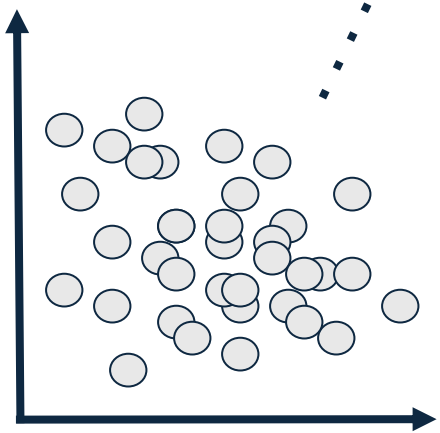
Problem

One network for a single shape



Deep SDF



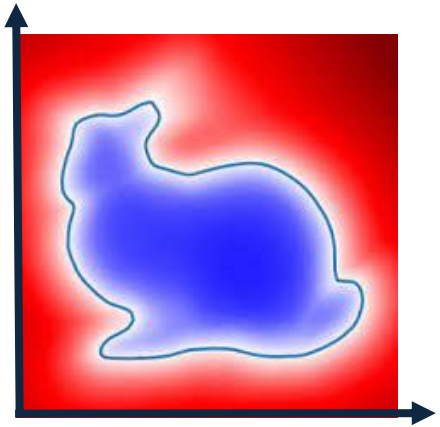


Problem

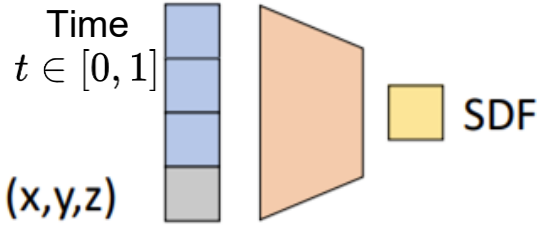
One network for a single shape

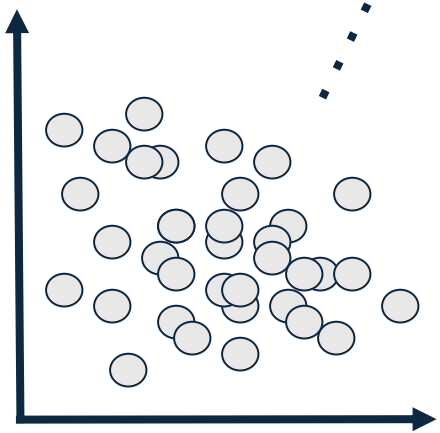
Solution

Condition the input



Deep SDF



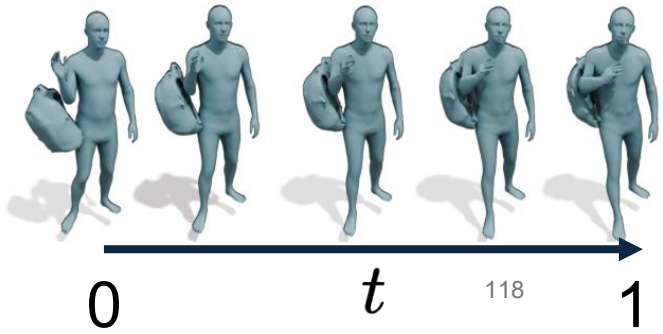


Problem

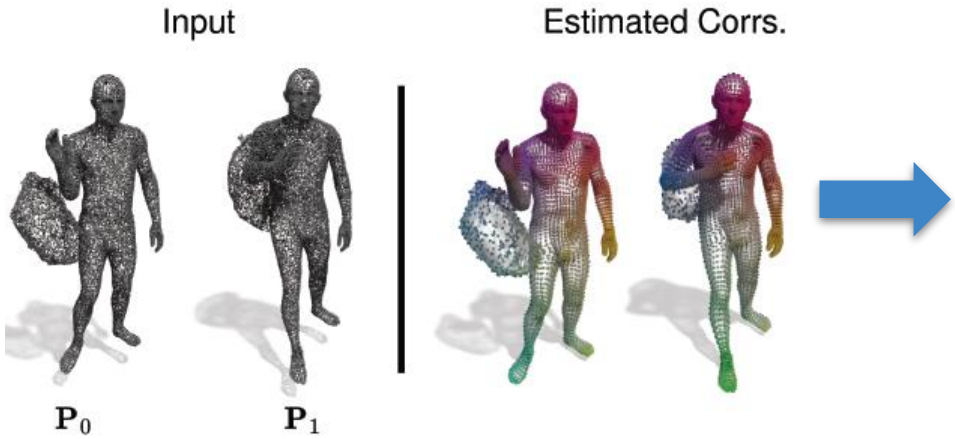
One network for a single shape

Solution

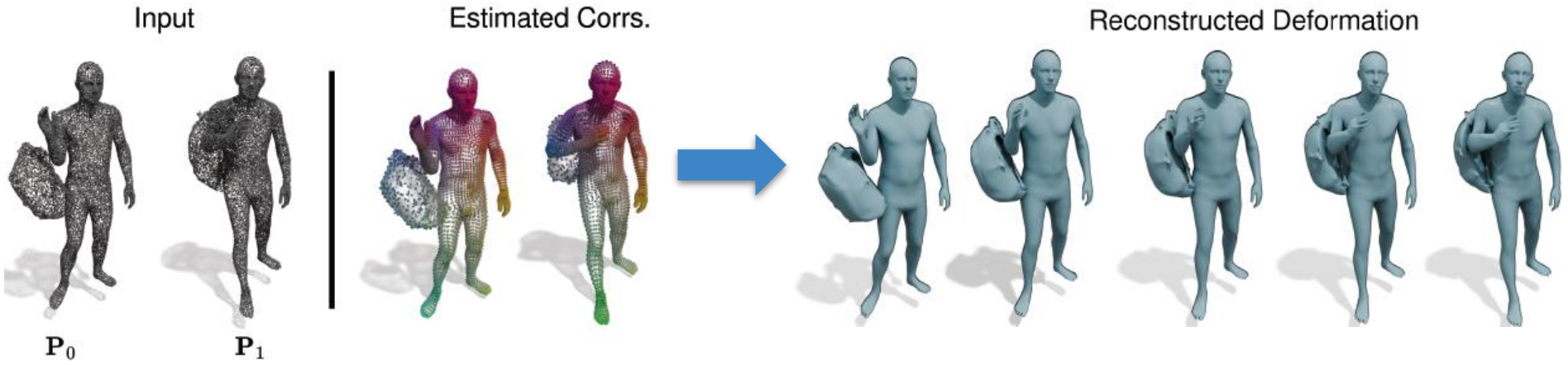
Condition the input



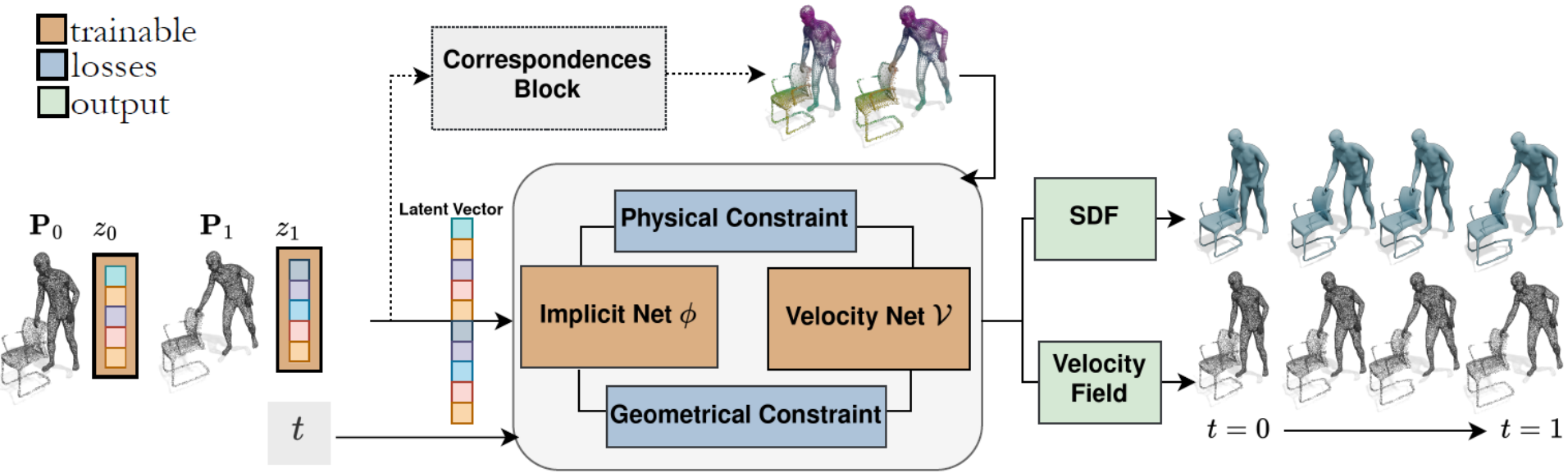
Goal: 3D to 4D reconstruction



Goal: 3D to 4D reconstruction



4Deform: Neural Surface Deformation for Robust Shape Interpolation



Input images

Input images





Input images

Input images

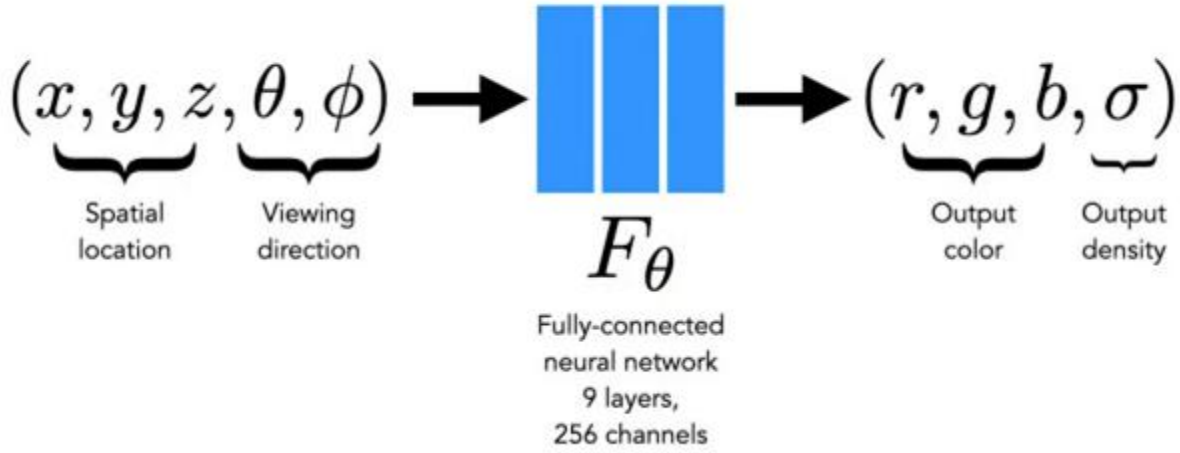
Novel View Synthesis

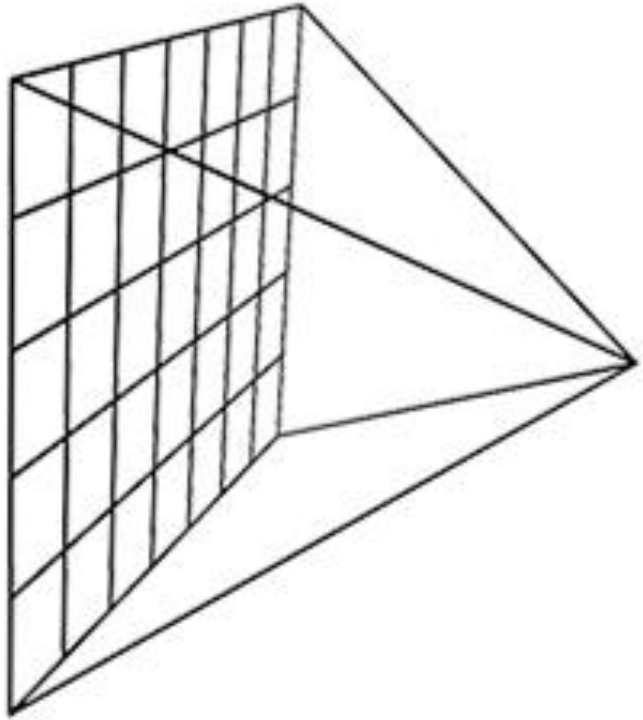
Inputs: sparsely sampled images of scene

Outputs: *new* views of same scene

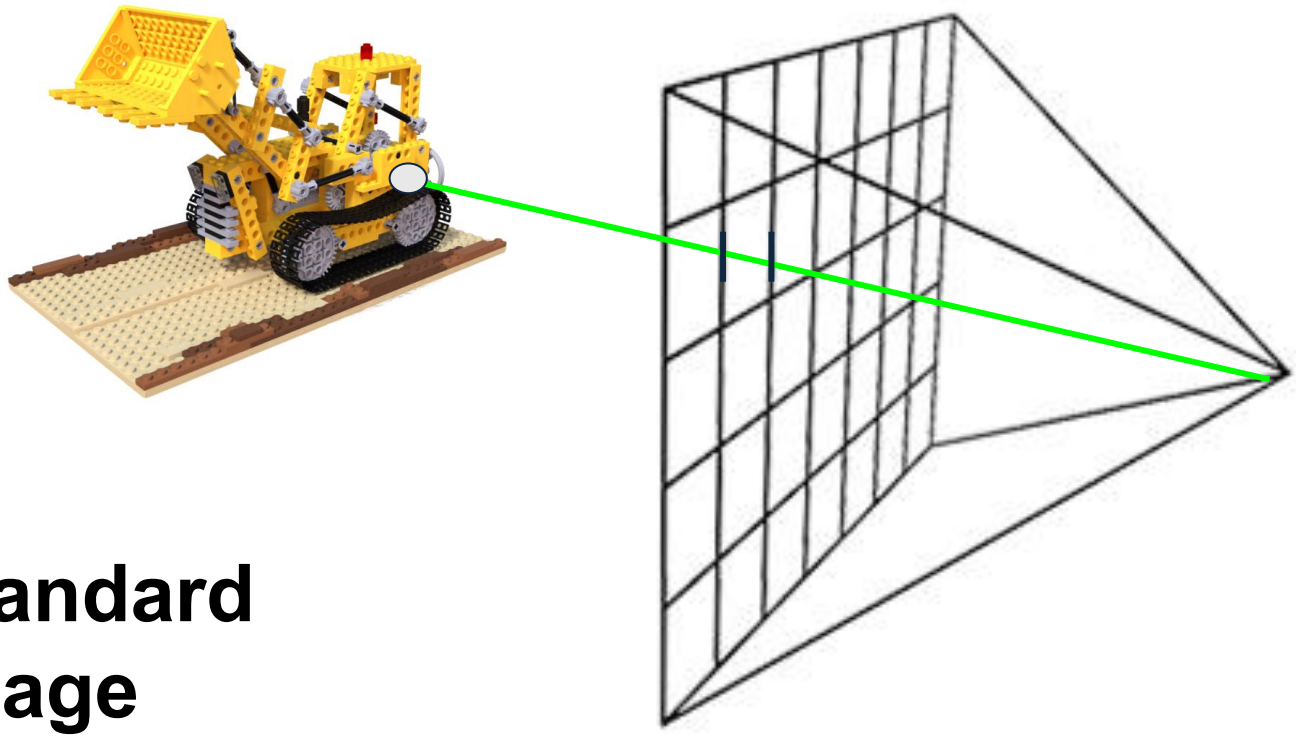
- **Task:** Given a set of images of a scene, render image from novel viewpoint
- Slide credits: Ben Mildenhall and Jon Barron

NeRF: Representing Scenes as Neural Radiance Fields

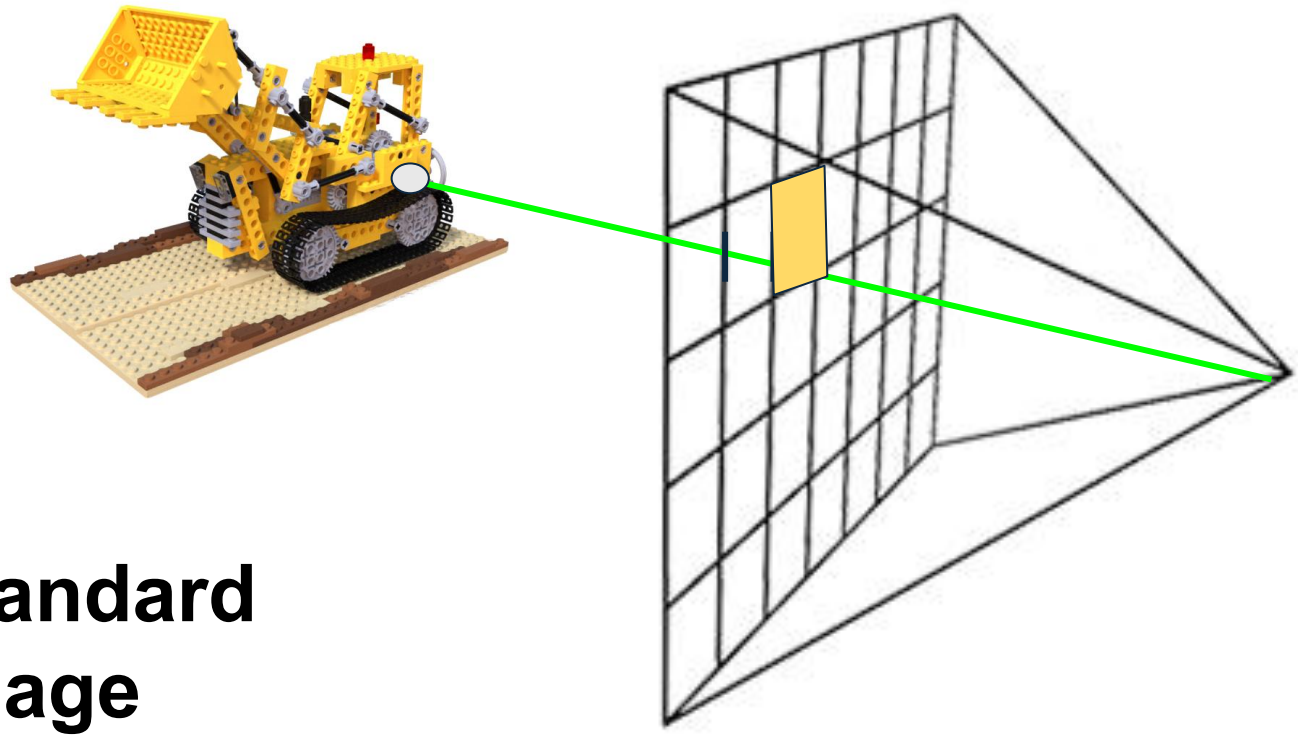




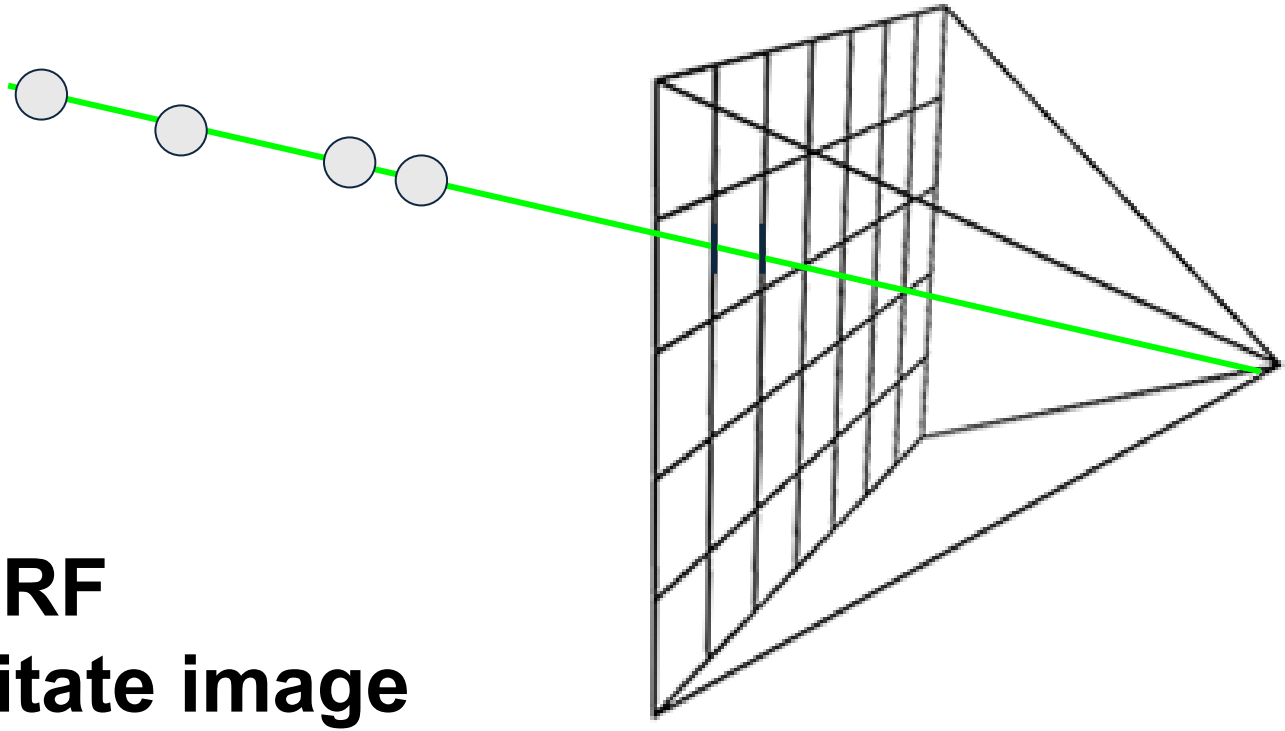
Standard image formation



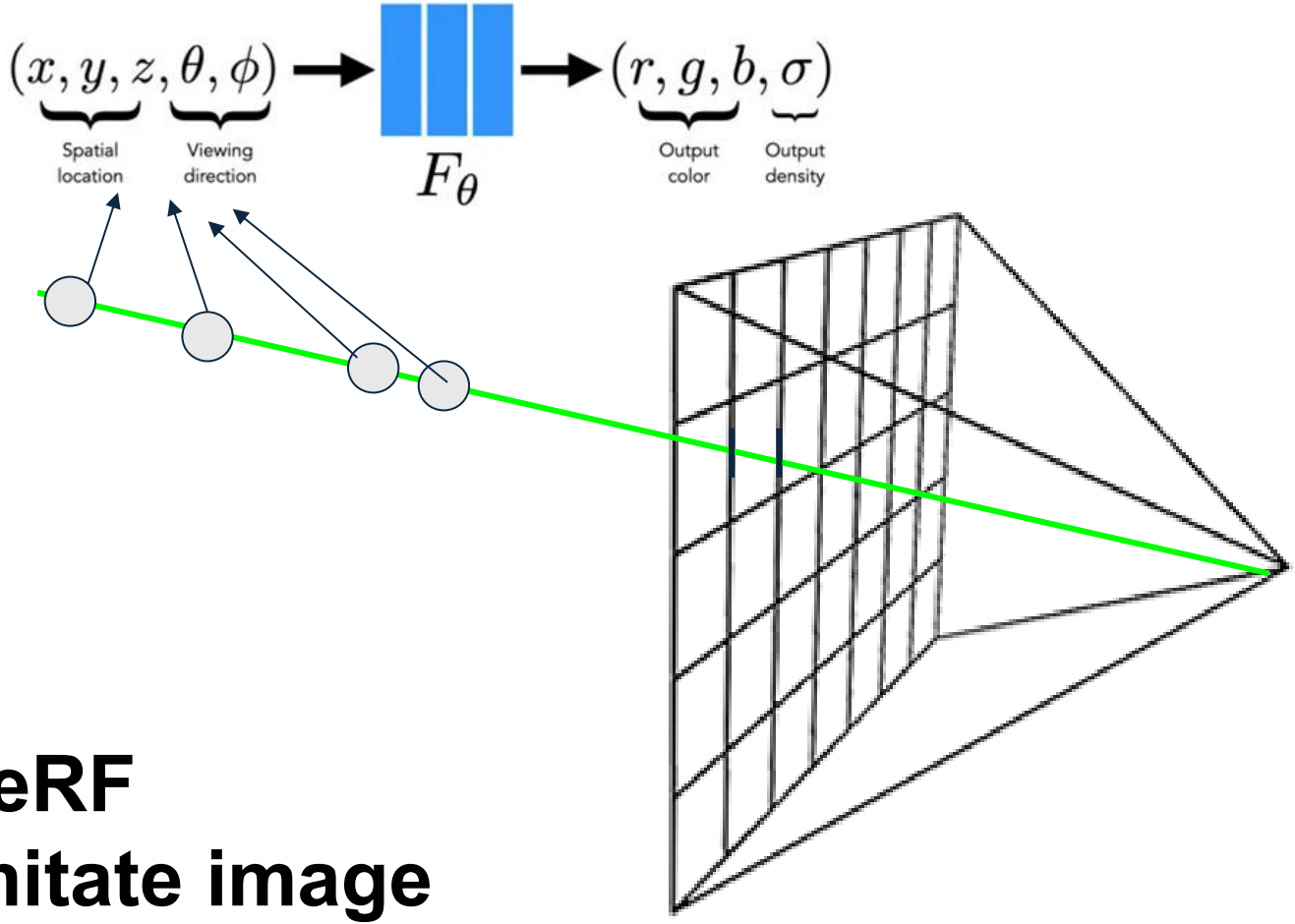
Standard image formation



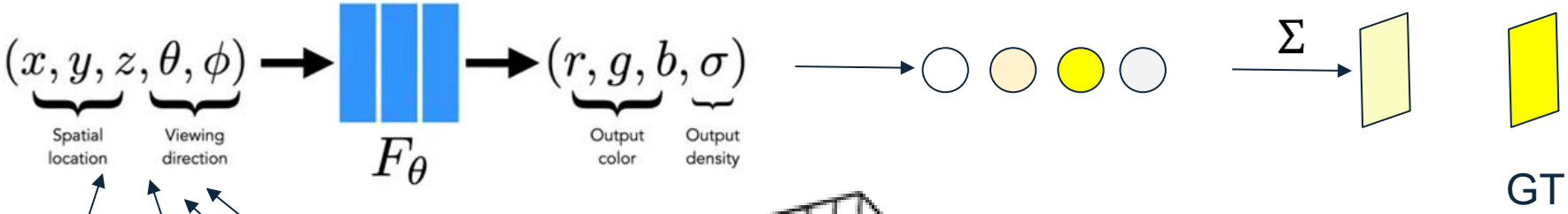
Standard image formation

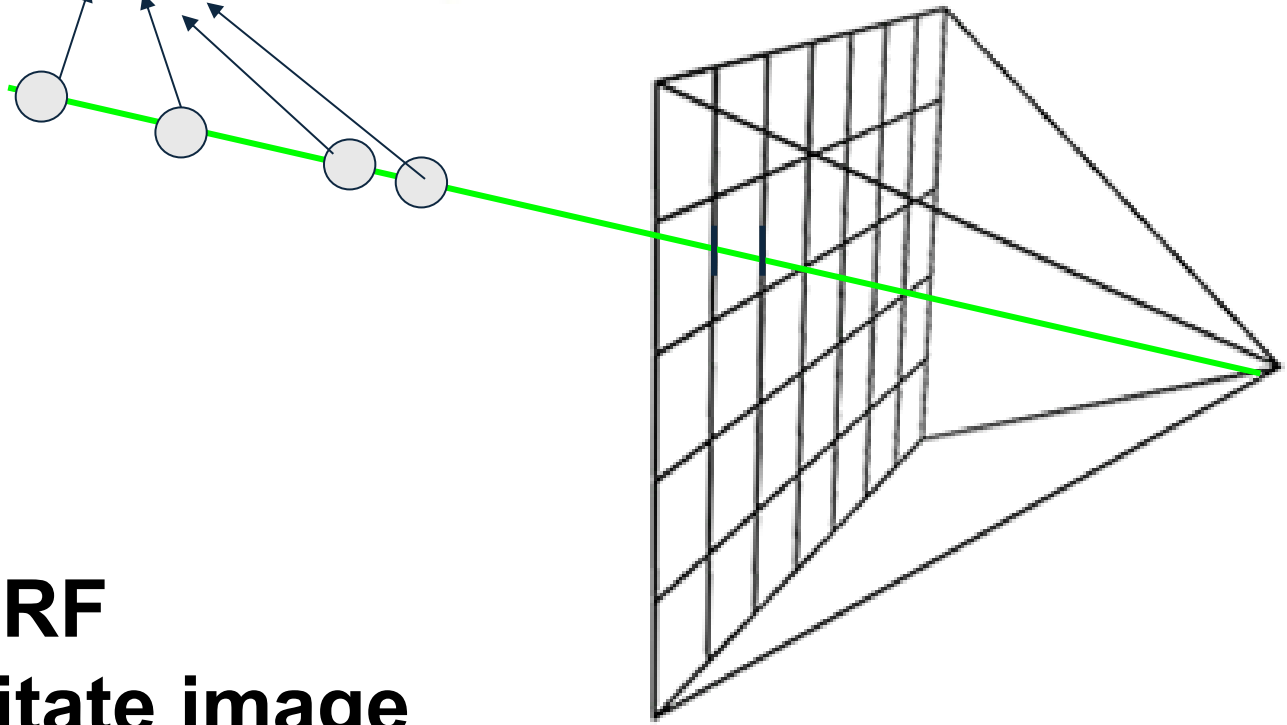


NeRF
imitate image formation



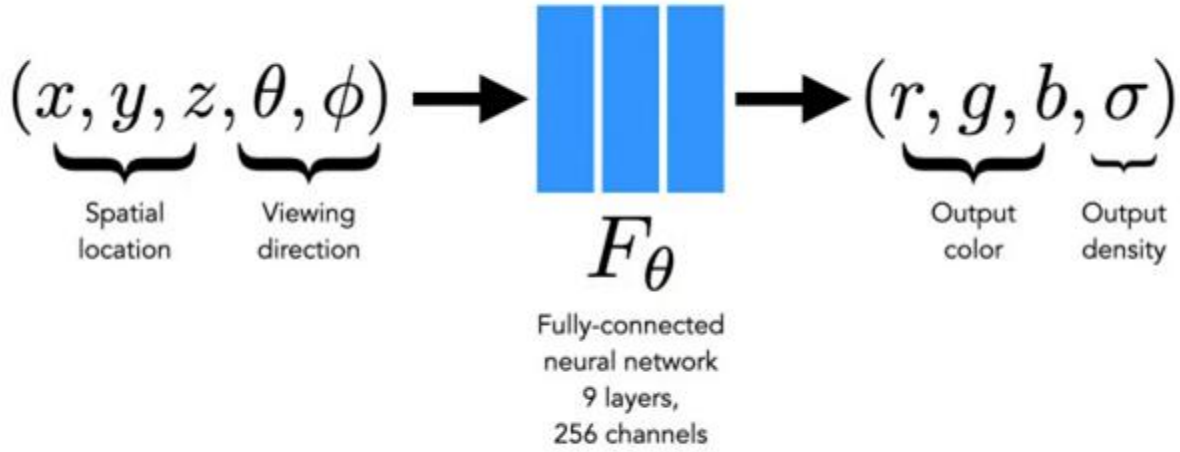
NeRF imitate image formation





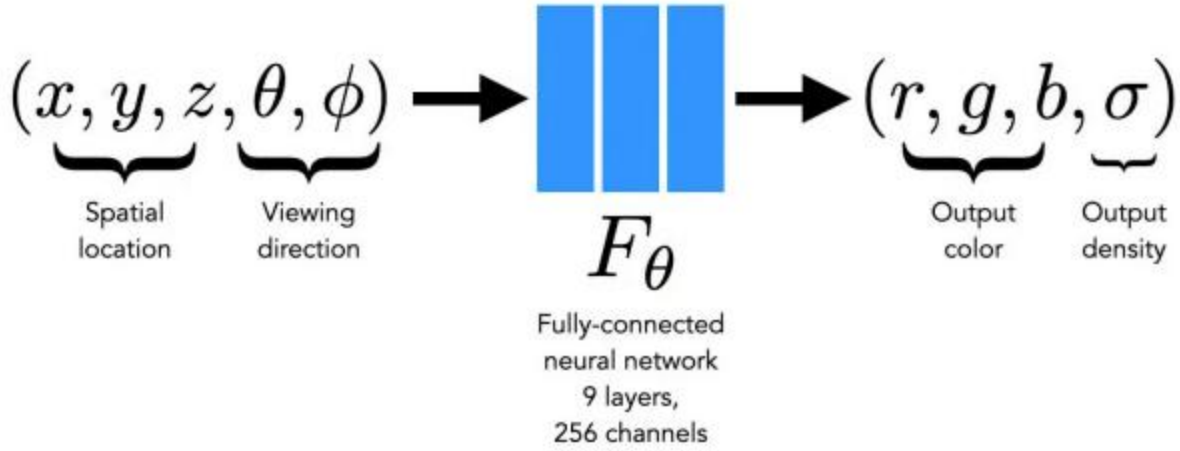
NeRF
imitate image formation

NeRF: Representing Scenes as Neural Radiance Fields



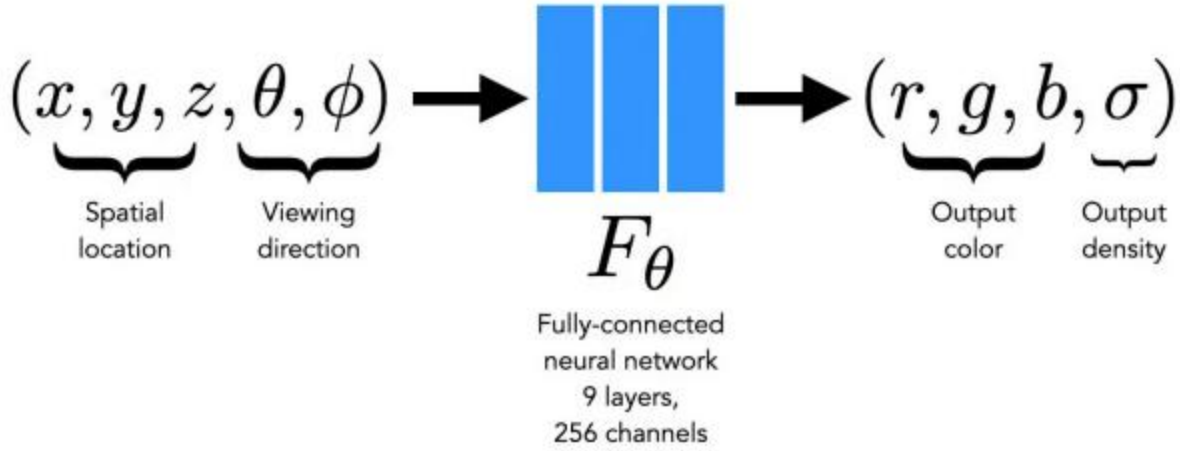
- Vanilla **ReLU MLP** that maps from **location/view direction to color/density**

NeRF: Representing Scenes as Neural Radiance Fields



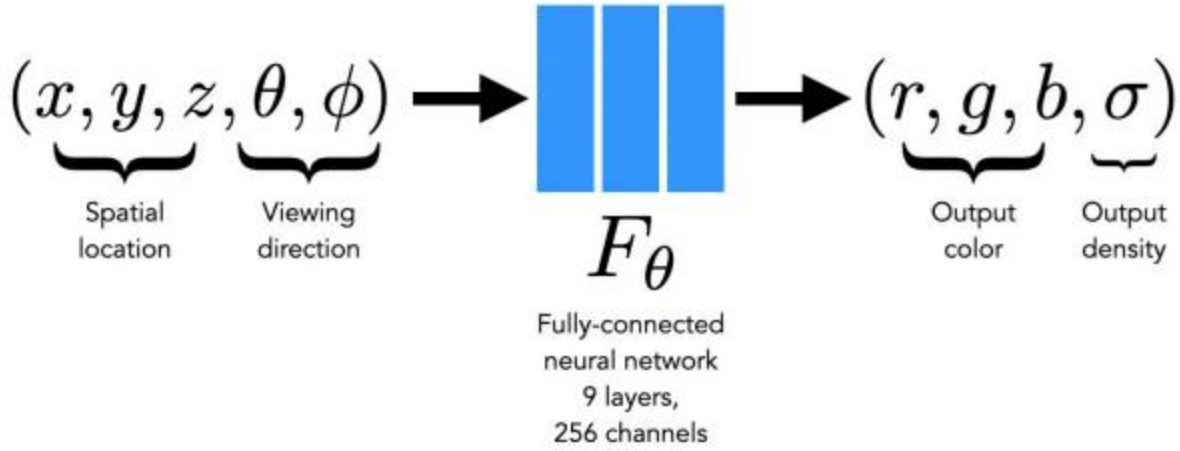
- Vanilla **ReLU MLP** that maps from **location/view direction to color/density**
- **Density** σ describes how solid/transparent a 3D point is (can model, e.g., fog)

NeRF: Representing Scenes as Neural Radiance Fields



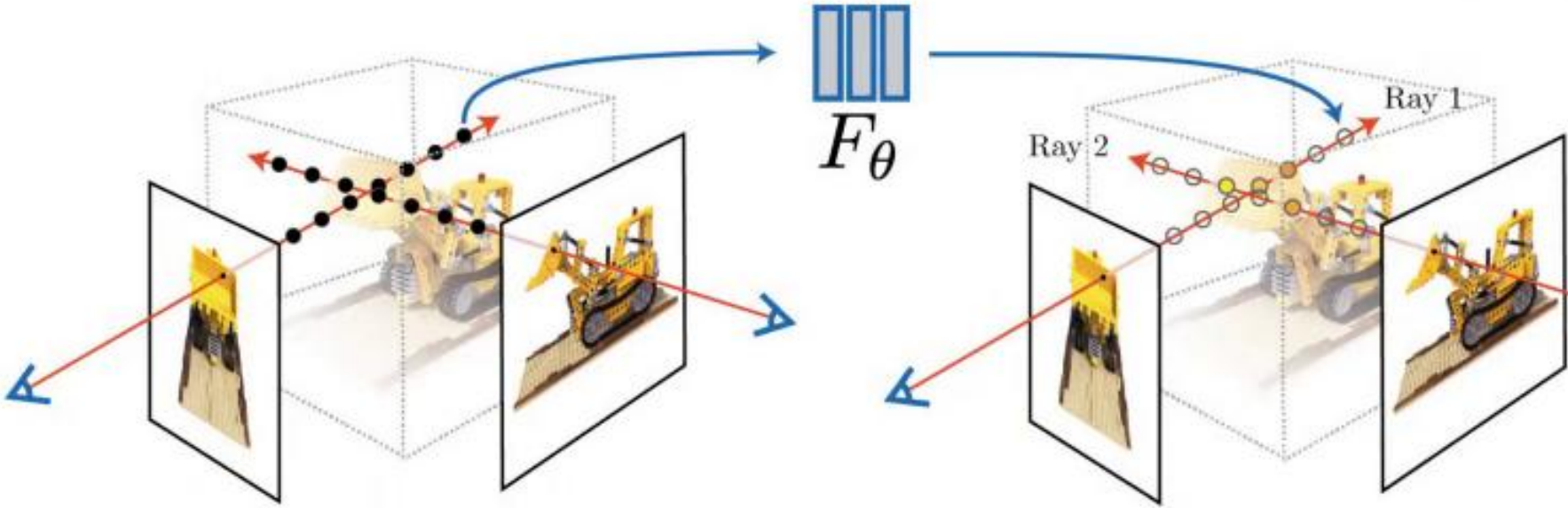
- Vanilla **ReLU MLP** that maps from **location/view direction to color/density**
- **Density** σ describes how solid/transparent a 3D point is (can model, e.g., fog)
- Conditioning on view direction allows for modeling **view-dependent effects**

NeRF: Representing Scenes as Neural Radiance Fields



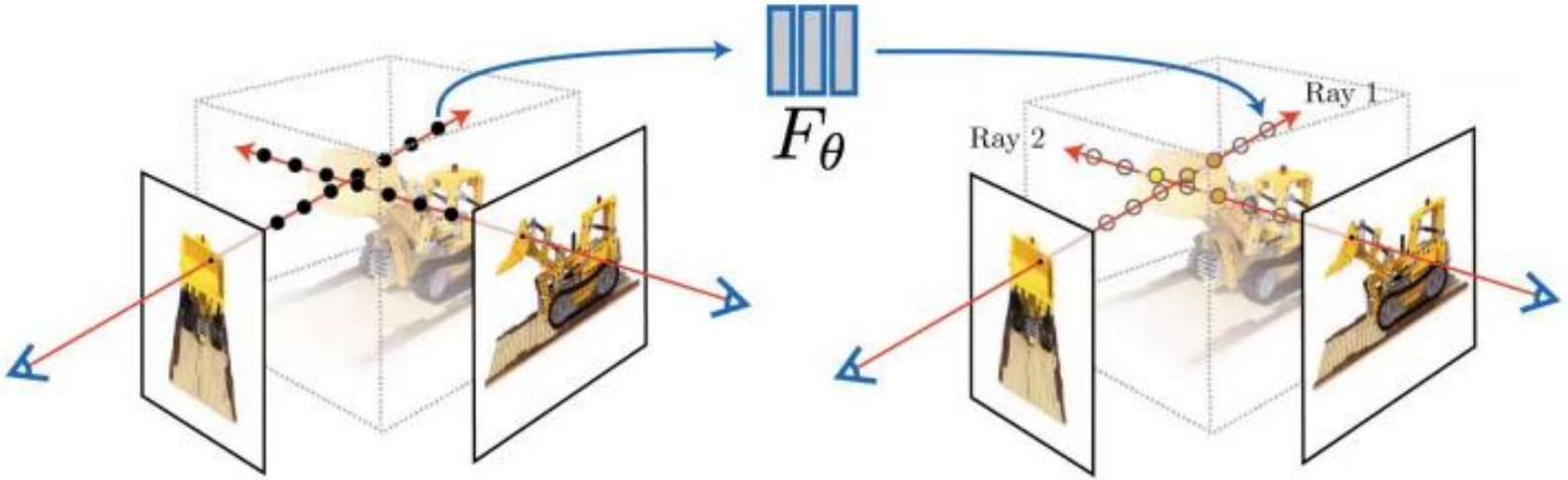
- Vanilla **ReLU MLP** that maps from **location/view direction to color/density**
- **Density** σ describes how solid/transparent a 3D point is (can model, e.g., fog)
- Conditioning on view direction allows for modeling **view-dependent effects**
- In practice, the view direction is input as a normalized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

Volume Rendering



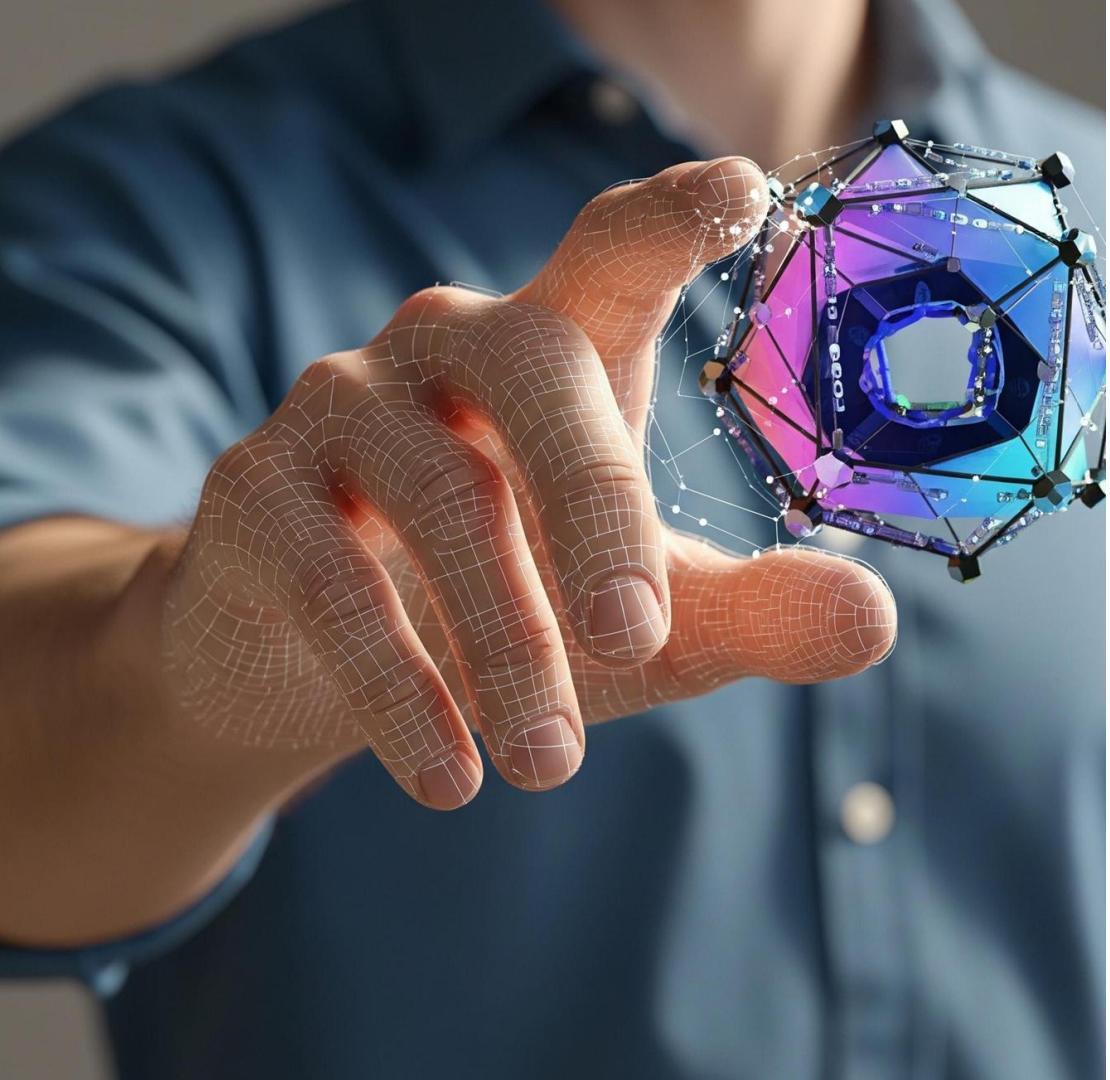
- **Volume rendering** works very similar to traditional **ray tracing** in graphics
- Shoot ray through the scene, sample points along ray, query radiance field to obtain color/density, apply alpha composition to obtain pixel color

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