OS: Humanoid Robots

3D Scene Representations

Kiran Varanasi

IM Fakultät Informatik und Medien

HTWK

Overview

- Motivation for 3D Representations in Robotics
- 3D Representations for Visualization
- Levels of robot autonomy
- 3D Representations for Robotics

Humanoid Robotics Applications

What is needed to build a general purpose humanoid robot?

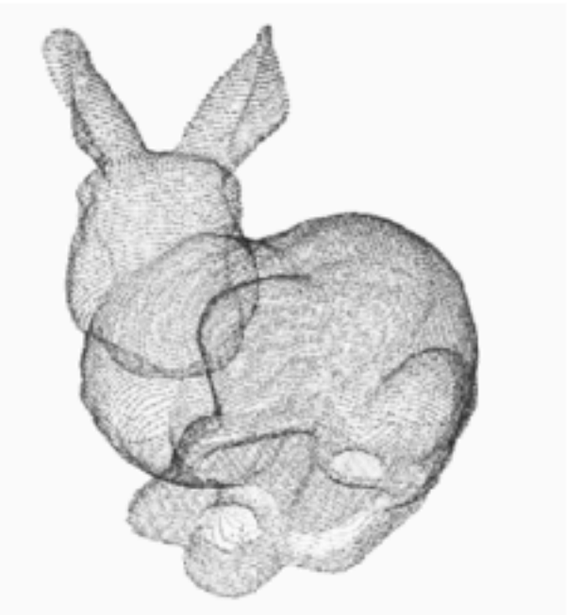
- An adaptable world model: that makes physically based predictions about the objects in the world

- A continuous update of the robot telemetry in relation to this world model: geometric positioning, relative velocities with respect to objects, operating forces and dynamics

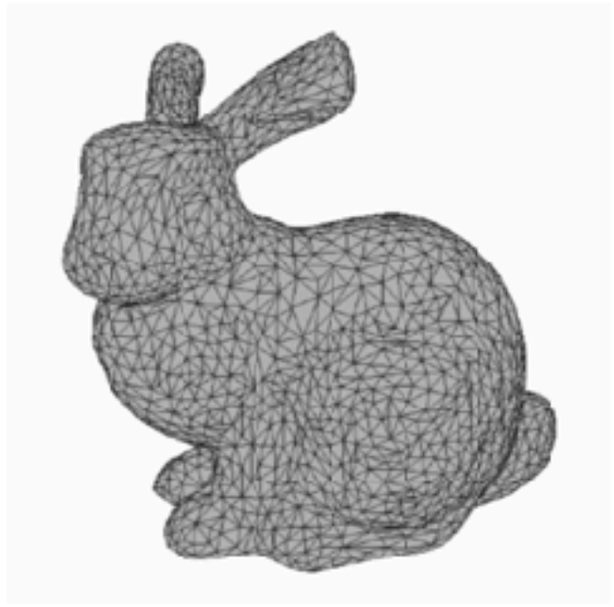
- Cognitive skills for planning: path planning, obstacle avoidance, calculation of affordance, modelling human expectations with respect to various tasks

- Why are 3D representations needed? For consistent processing while the robot is actively moving. For making physically based predictions about the state of the world.

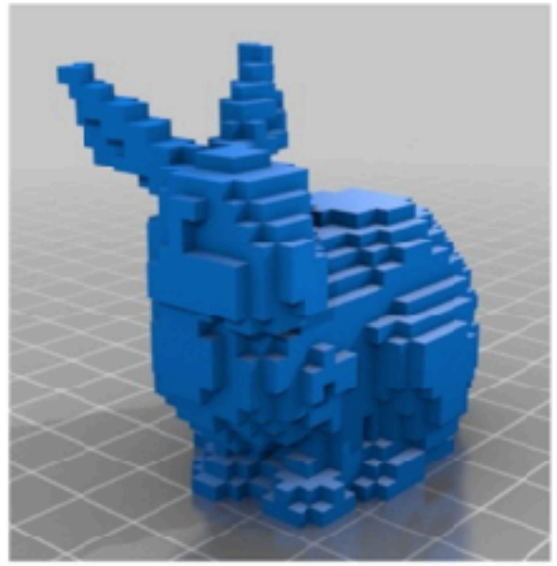
3D Scene representations for visualisation



Point cloud



3D Mesh



Voxels

Projected 2D renderings

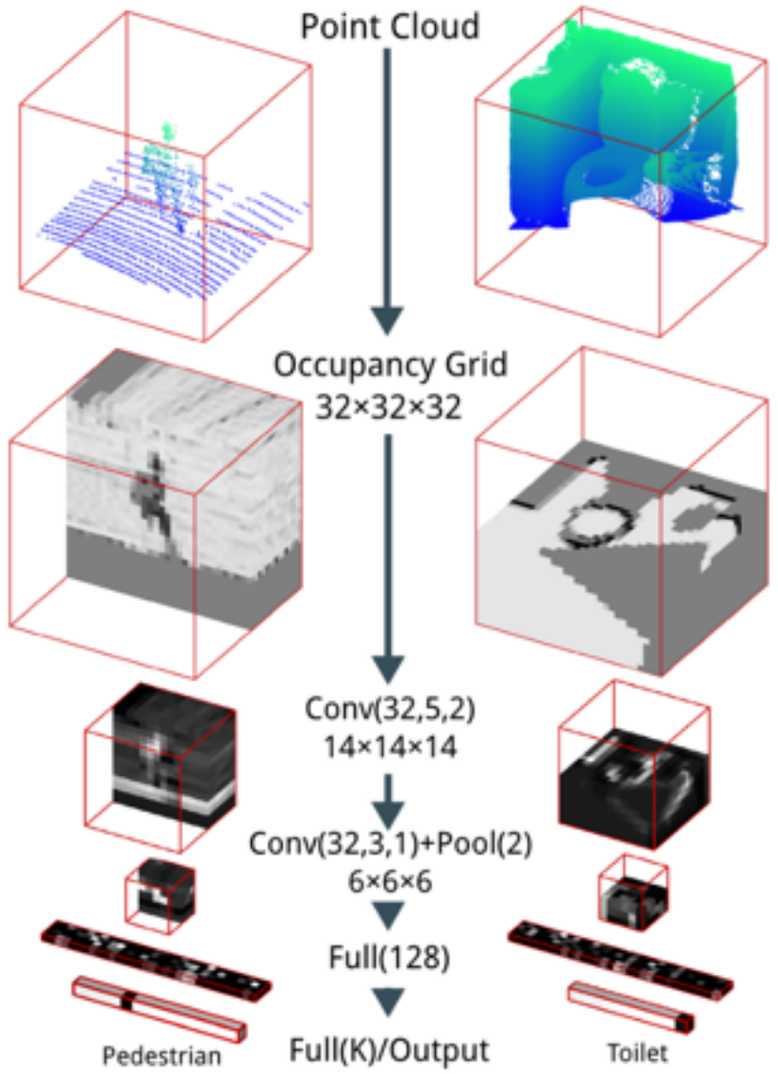
Voxels

- Voxels: Discrete 3D-grid of cells (3D-Pixels): this can be Occupancy, Colour, Point Density etc.

- **Advantages:** 3D-convolutions are easy to define. CNNs can be trained with similar architectures as for 2D images.

- **Disadvantages:** Very memory intensive. Difficult for sparse or large 3D scenes.

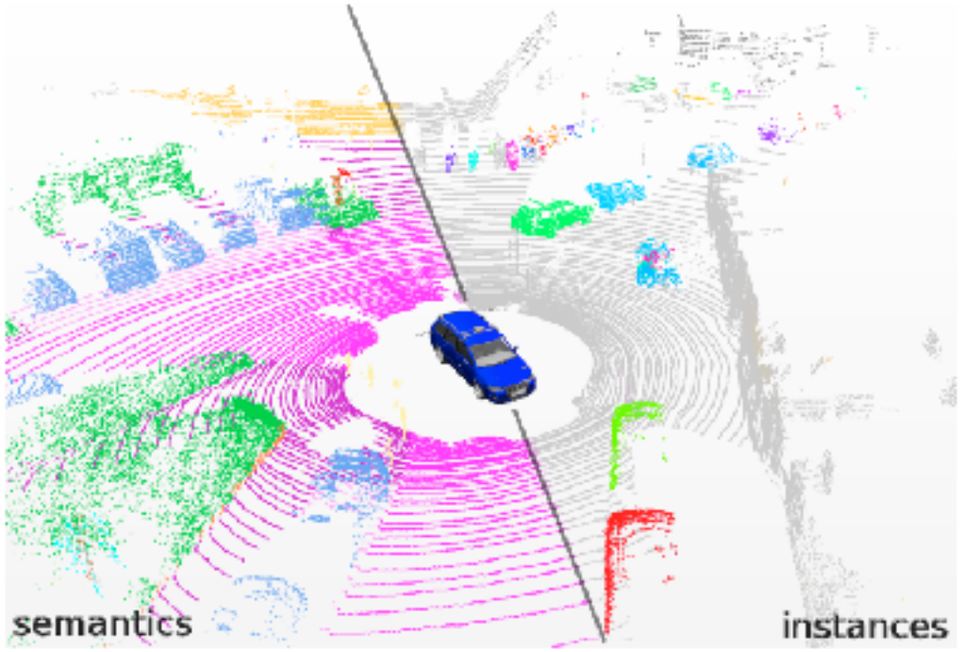
- Suitable for supervised learning with CNNs.



Vox Net [Maturana & Scherer 2015]
https://dimatura.net/publications/voxnet_maturana_scherer_iros15.pdf

Point Clouds

- Point Clouds: Unordered sets of 3D-points. Can be just XYZ-coordinates, or can contain colour or normal information.



- **Advantages**: Memory-efficient, contain fine geometric details

- **Disadvantages**: Unstructured. Require special network architectures for training

- Largely used in the areas of autonomous driving (KITTI), robotics and in 3D-reconstruction.

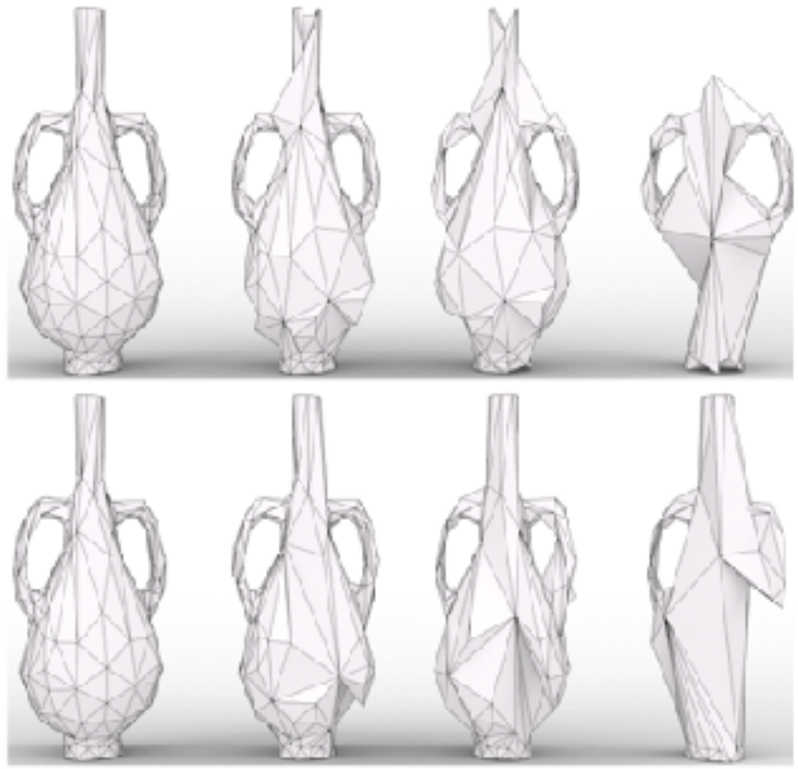
3D-Meshes

- Discrete geometric models with vertices and polygonal surfaces: the surface of a 3D geometric form is represented

- **Advantages**: Industry standard in CAD, 3D-games etc. Compact, suitable for rendering

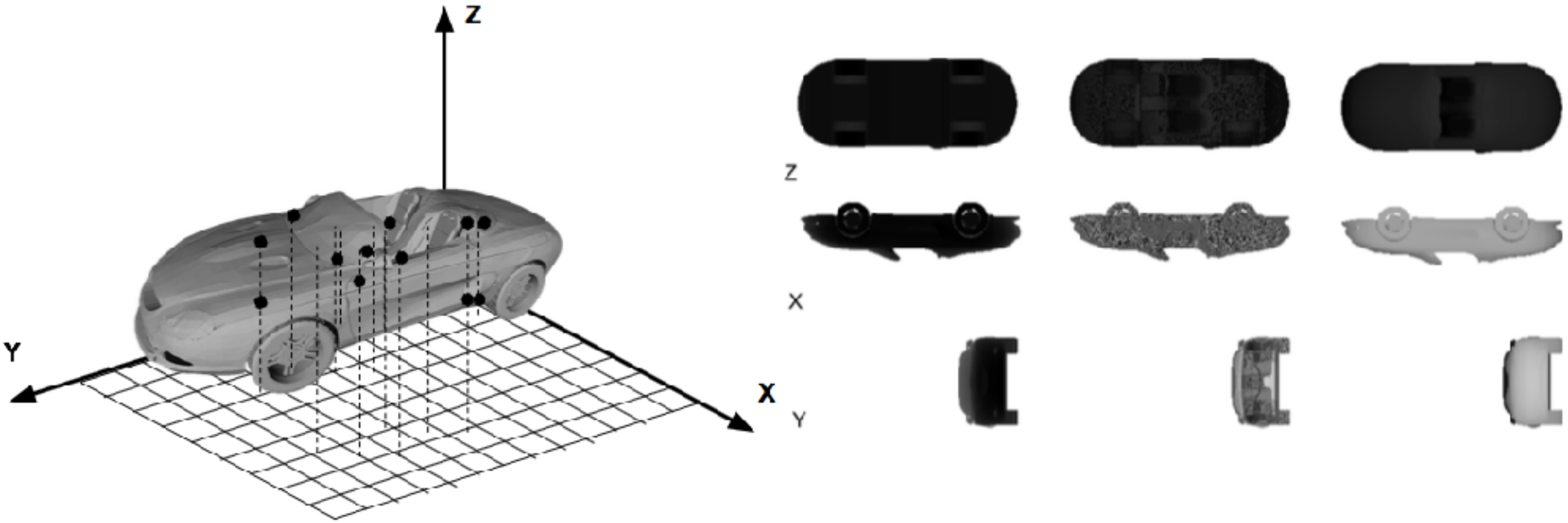
- **Disadvantages**: Topological changes are very difficult to model. Not easy with rendering volumetric effects.

- Graph-neural-networks can be used for defining convolutions over a graph or mesh neighbourhood. This can generate 3D-shapes.



Mesh-CNN: Mesh-Simplification durch Mesh-Pooling [Hanocka et al. 2019] https://arxiv.org/pdf/1809.05910

Triplane / Multiplane depth map representations



"Learning 3D shapes as multi-layered height maps using 2D convolutional neural networks" Sarkar, Hampiholi, Varanasi, Stricker ECCV 2018

Neural Radiance Fields (NERFs)

- Implicit neural representations: Latent-Embedding through weights of the network. This can represent continuous functions (3D-shapes and light transport).

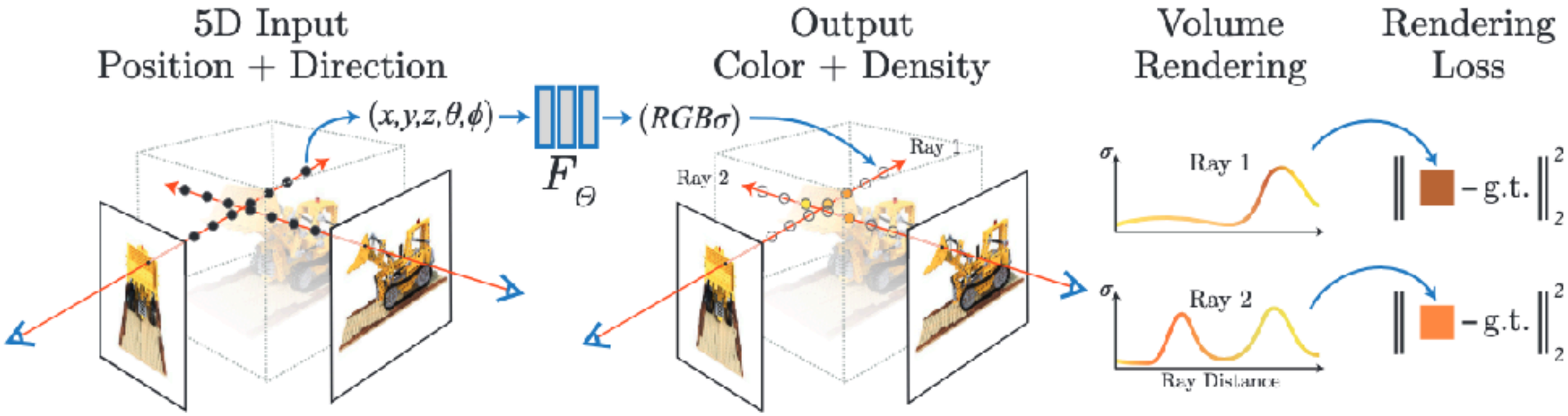
- NERF: 5D coordinates (x,y,z,θ,ϕ) → Point density + Colour.Can be trained through a multi layer perceptron.

- **Advantages**: Very good quality, continuous functions, compact.

- **Disadvantages**: Lengthy and complex training and rendering

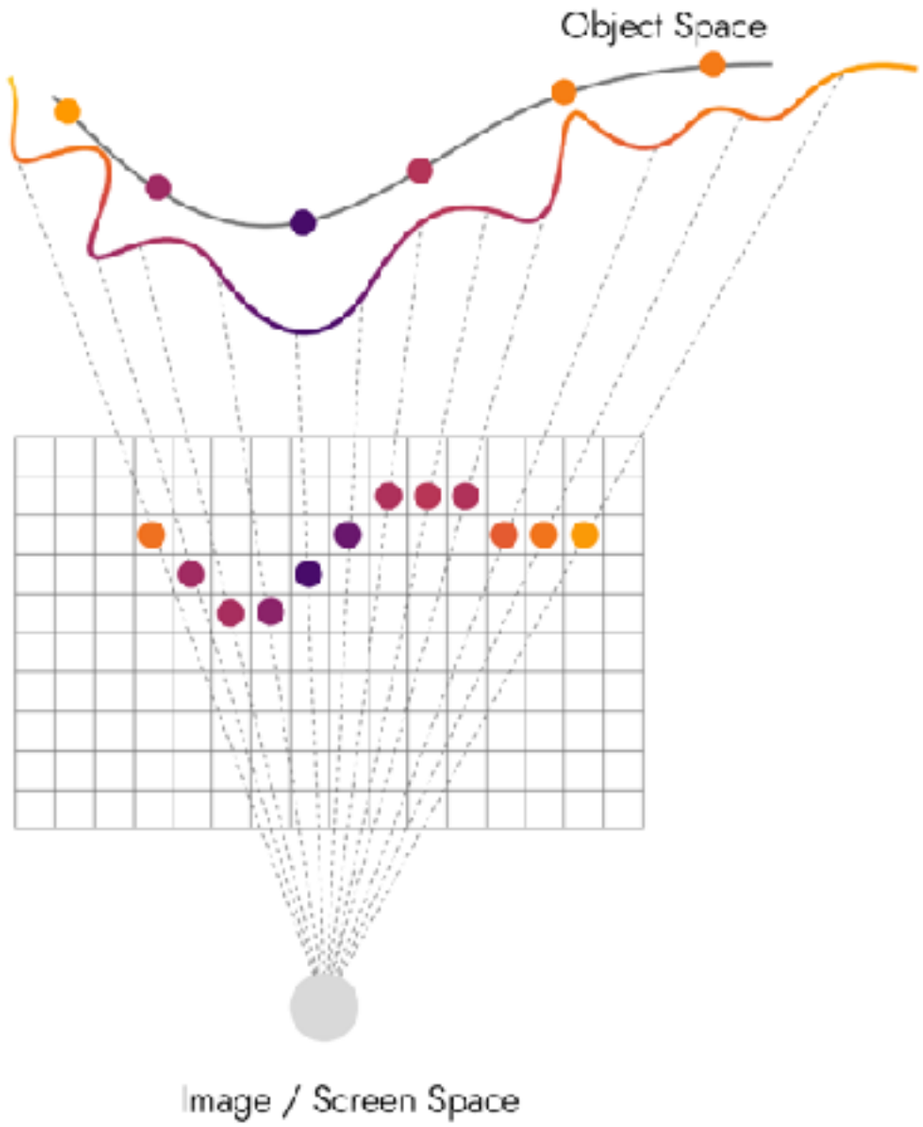
- Suitable for relighting, photo realistic rendering and high quality visualisations.

Neural Radiance Fields

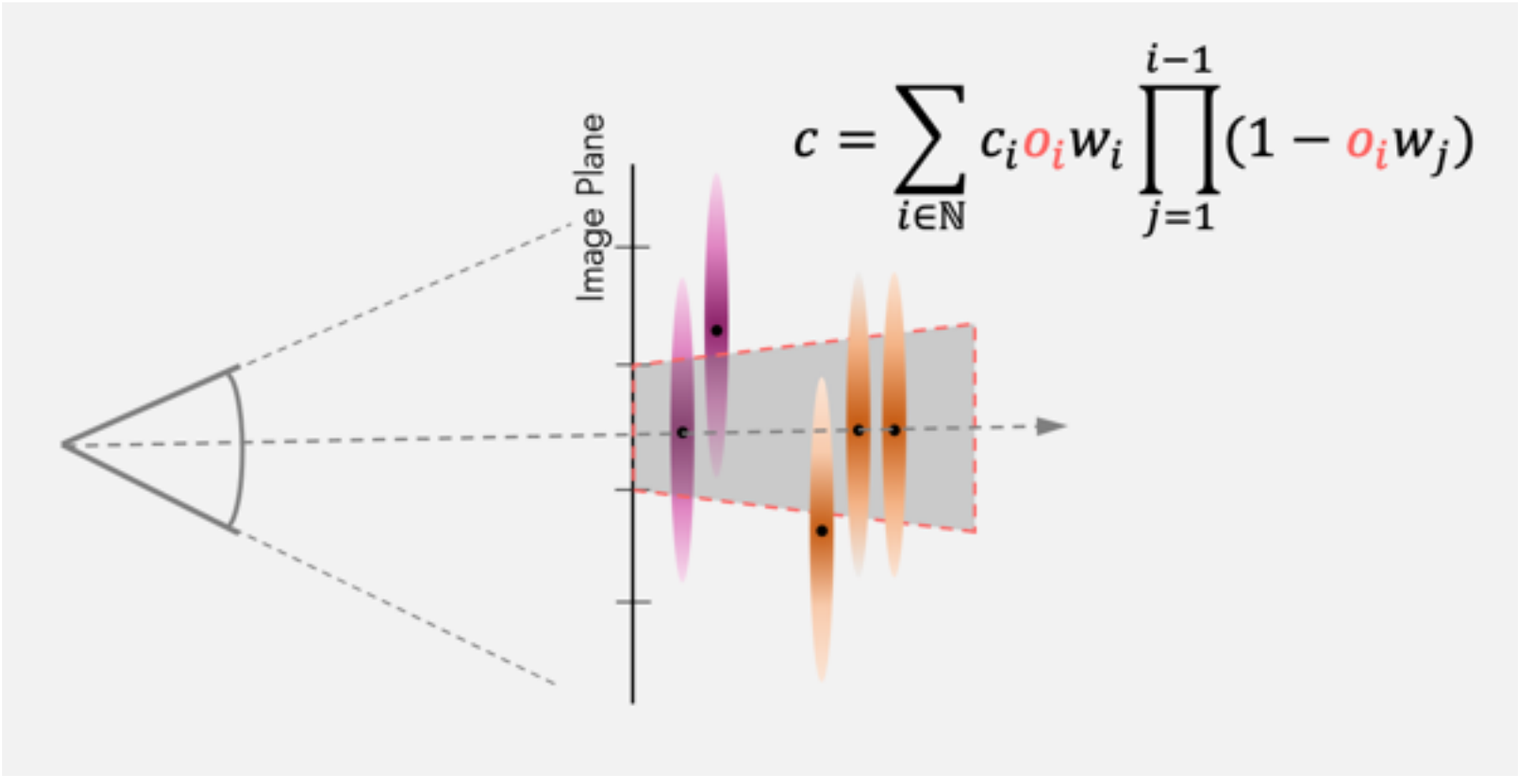


[NeRF: Representing scenes as neural radiance fields for view synthesis] ***Mildenhall et al.*** *ECCV 2020*

Splatting: Interpolated rendering from point clouds

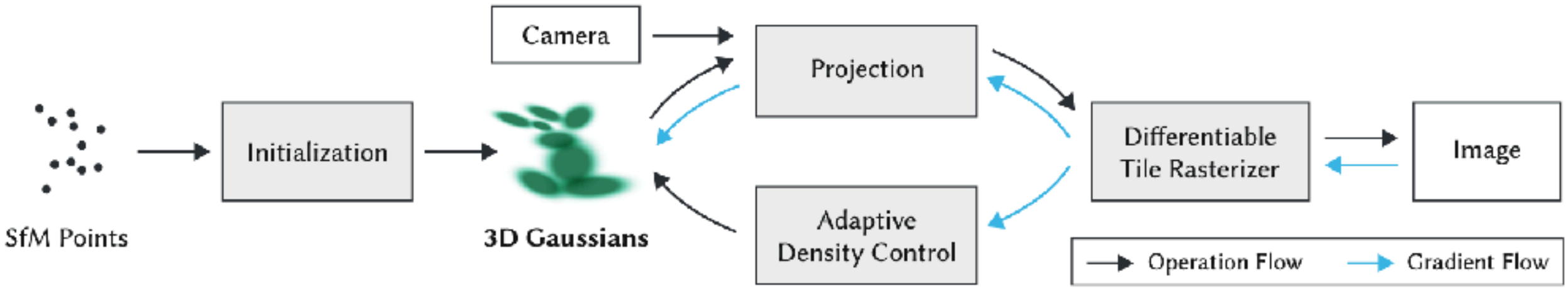


3D Gaussian Splatting: Rendering of 3D “Blobs” (3D Gaussian Ellipsoids)



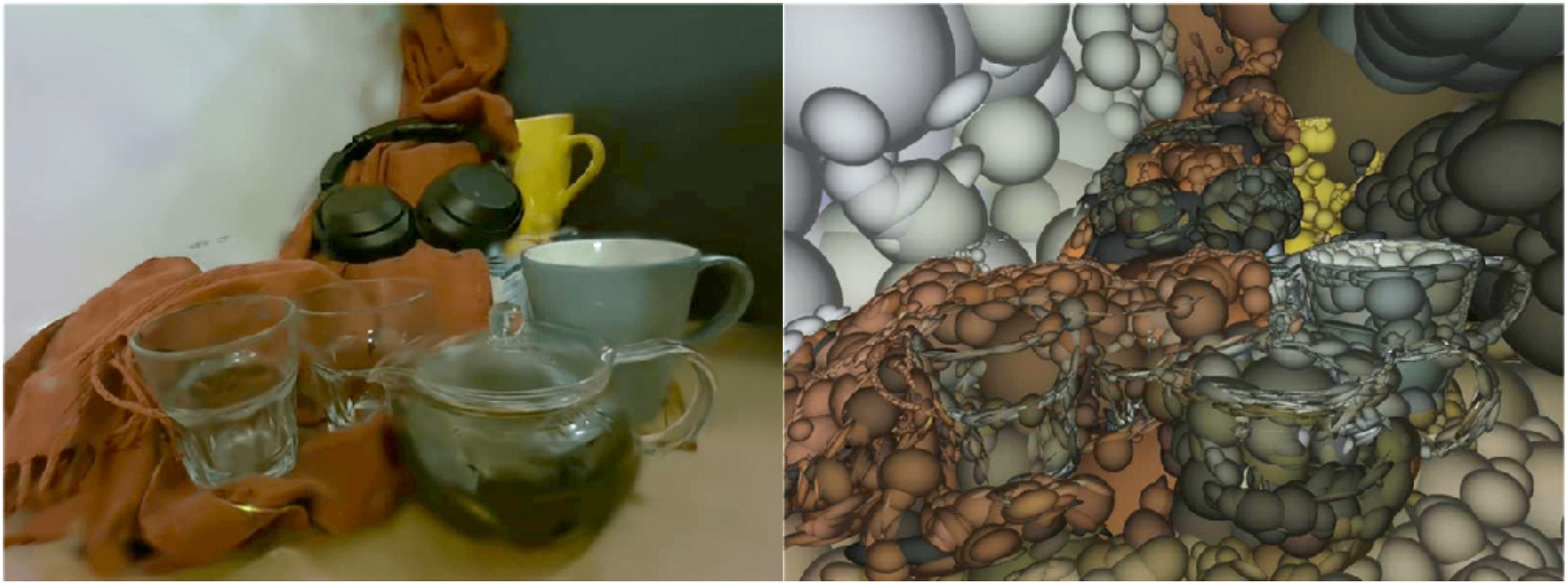
[3D Gaussian Splatting for real time radiance field rendering]
Kerbl et al. *SIGGRAPH 2023*

3D Gaussian Splatting: Workflow



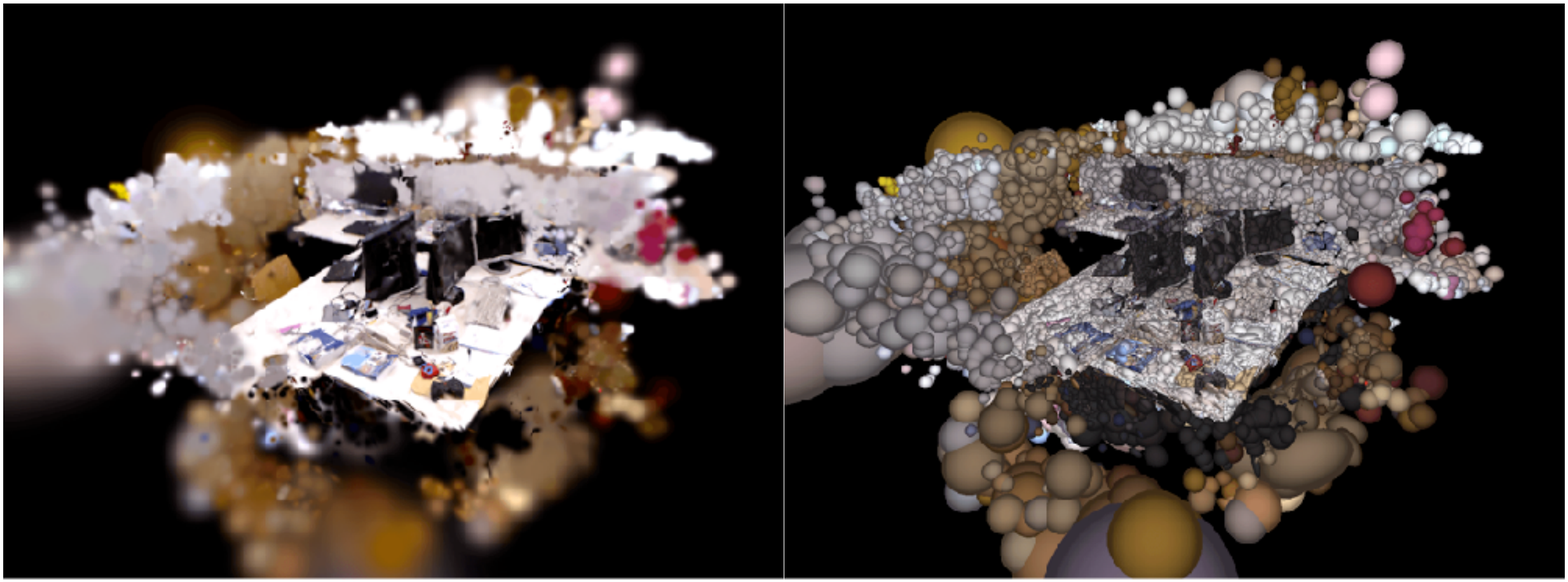
[3D Gaussian Splatting for real time radiance field rendering]
Kerbl et al. *SIGGRAPH 2023*

Visualizations of 3D Gaussians



[Gaussian Splatting SLAM] **Matsuki et al.** *CVPR 2024*

Visualizations of 3D Gaussians



[Gaussian Splatting SLAM] ***Matsuki et al.*** *CVPR 2024*

3D-Gaussian-Splatting for AI-Models

- Description: Anisotropic Gaussian functions (Mean and Variance) with 3D position, degree of transparency, colour / shading.

- **Advantages**: Very efficient, high quality and real-time rendering

- **Disadvantages**: Similar to point clouds, require more memory (this is being addressed through compression mechanisms) than pure implicit representations.

- AI-Models: Suitable for Transformer-based representations (LLMs, VLMs etc.). Can be trained with cross-modal-Attention. Suitable for real-time applications (VR/AR/Robotics).

Teleoperation of Humanoid Robots

- Unitree Embodied Avatar

Level of autonomy in Robotic Applications

1. Direct Teleoperation (Master-Slave Control)
2. Secured Teleoperation
3. Shared Autonomy: Recognition of user intent
4. Task level autonomy under supervision
5. Collaborative autonomy: Team work along with humans

(1) Direct Teleoperation

- In dieser Phase ist der Roboter eine physische Erweiterung des Menschen. Er besitzt keinerlei Autonomie. Jede Millimeterbewegung der menschlichen Hand in der VR wird vom Roboter in Echtzeit gespiegelt.

- **Required Robot capabilities:** *Kommunikation with small latency.* The capability, to transfer the required data in a few milliseconds.

    - **Proprioception:** Highly precise sensor telemetry, with exact calculation of positions of the joints.

- **Cognitive Functions**: *Signal Processing*: Conversion of digital commands into the corresponding motor actuations.

- **Security Mechanisms:** An elementary stopping logic (Kill-Switch) when there is a detection of communication loss.

(2) Secured Teleoperation

- The user still controls the robot, but it behaves as an “intelligent defence shield” to prevent harmful actions. The robot knows its own surroundings, so that any mistake by the teleoperator (e.g, driving the robot through a wall) is prevented.

- **Required robot capabilities:**

    - **Recognition of immediate surroundings:** LIDAR or ultra sound sensors are used, along with automatic detection of immediate obstacles

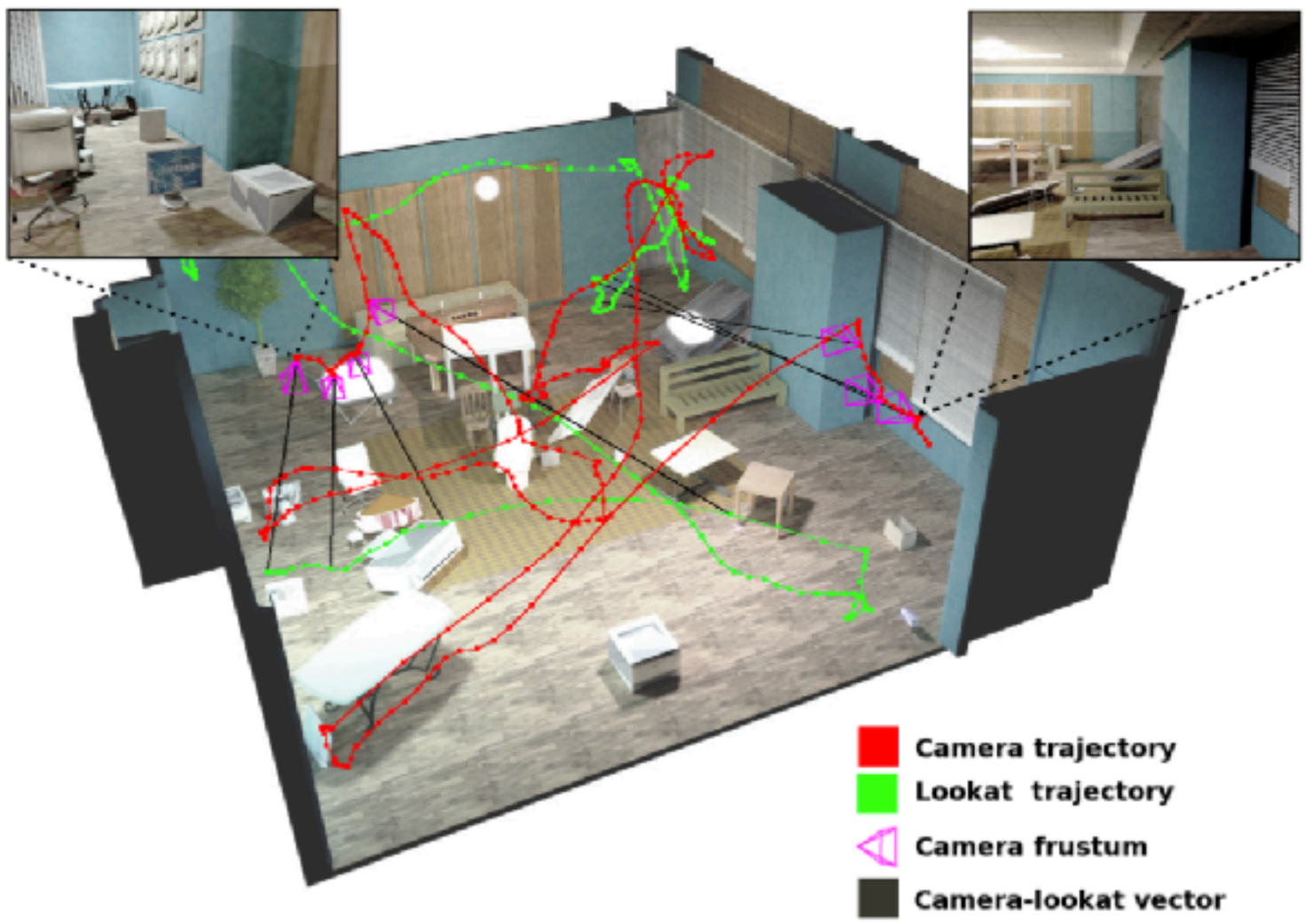
    - **Virtual Barriers:** The creation of “Software Barriers”, that the robot would not cross physically.

- **Cognitive Functions:**

    - **Spatial Perception:** Creation of a 3D map of the immediate surroundings, even if not very precise

- **Reactive Planning:** The capability to override a human command when it clearly leads towards a threatening collision.

3D-Mapping of the environment (SLAM)



SceneNet RGB-D: https://www.imperial.ac.uk/media/imperial-college/research-centres-and-groups/dyson-robotics-lab/jmccormac_etal_iccv2017.pdf

SLAM through 3D Gaussian Splatting

[Gaussian Splatting SLAM] ***Matsuki et al.*** *CVPR 2024*

SLAM through 3D Gaussian Splatting: SplaTAM

[SplaTAM: Splat, track and map 3D Gaussians for dense RGB-D SLAM] ***Kreta et al.*** *CVPR 2024*

Automatic depth estimation from RGB-video and SLAM

[Depth Anything 3: Recovering the Visual Space from Any Views] ***Lin et al. (ByteDance-Seed)*** *Arxiv 2025*

(3) Shared Autonomy: Recognition of human intent

- Here, the robot begins to predict the intention of the human teleoperator. For example, if the operator is moving the hand of the robot towards a cup, it will prepare itself for a perfect gripping posture for the cup's handle.

- **Required robot capabilities:**

  - **Object recognition**: Identification of specific objects (cup, handle, door etc) through computer vision.

  - **Intelligent manipulation**: Estimation of the optimal gripping based on the predicted object weight, geometric form, behaviour etc.

- **Cognitive Functions:**

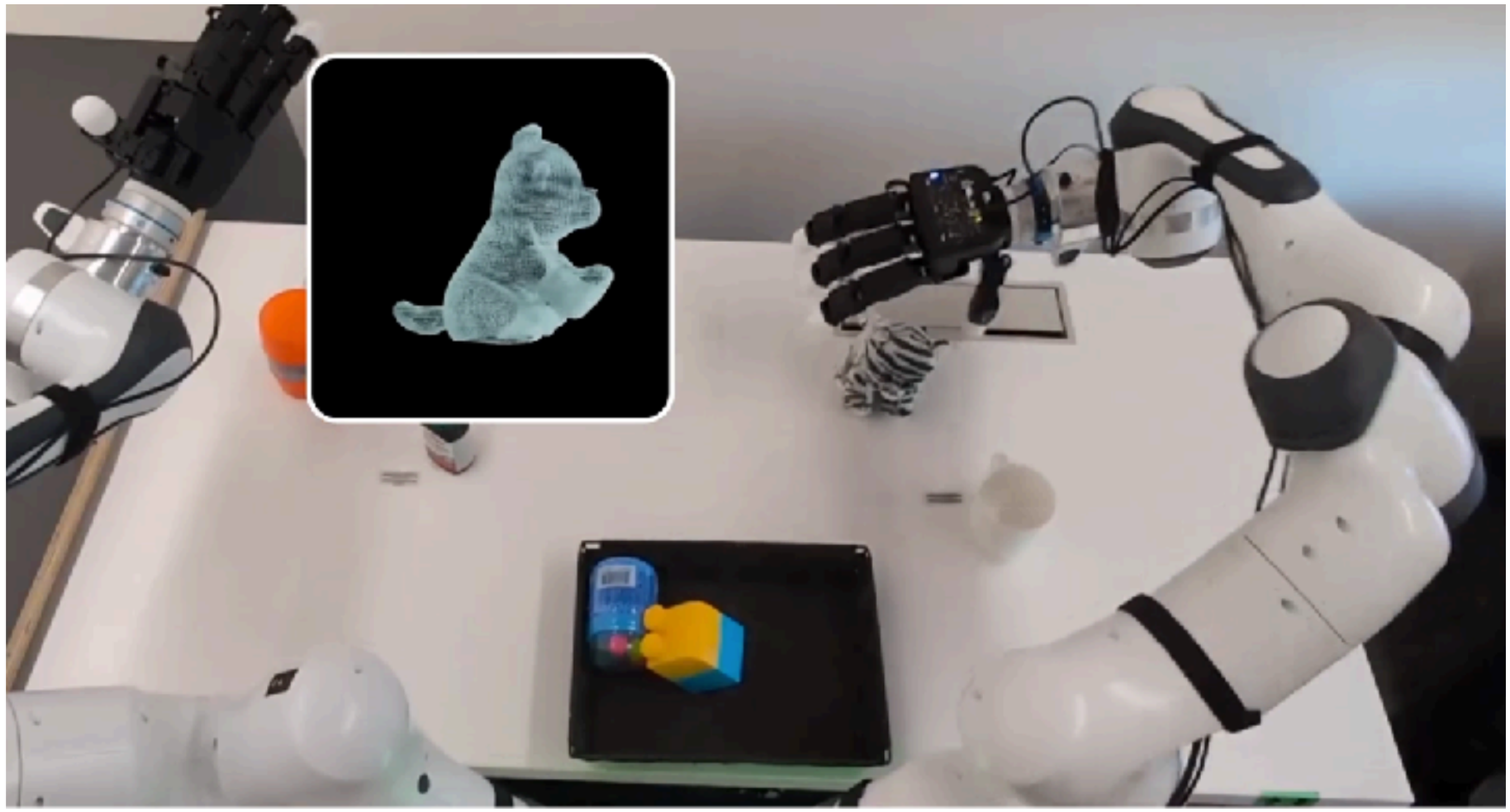
  - **Intent recognition**: Prediction of the physical properties of the objects intended to by the operator through probabilistic models (e.g, Bayesian Inference).

  - **Multi-sensory fusion:** Combination of visual and haptic data (manipulation/touch sensors) for the confirmation of a successful action.

3D scene segmentation of the mapped environment

[Group any Gaussians via 3D aware memory bank] **Lyu et al.**
Arxiv 2024

Segment-Anything-3D

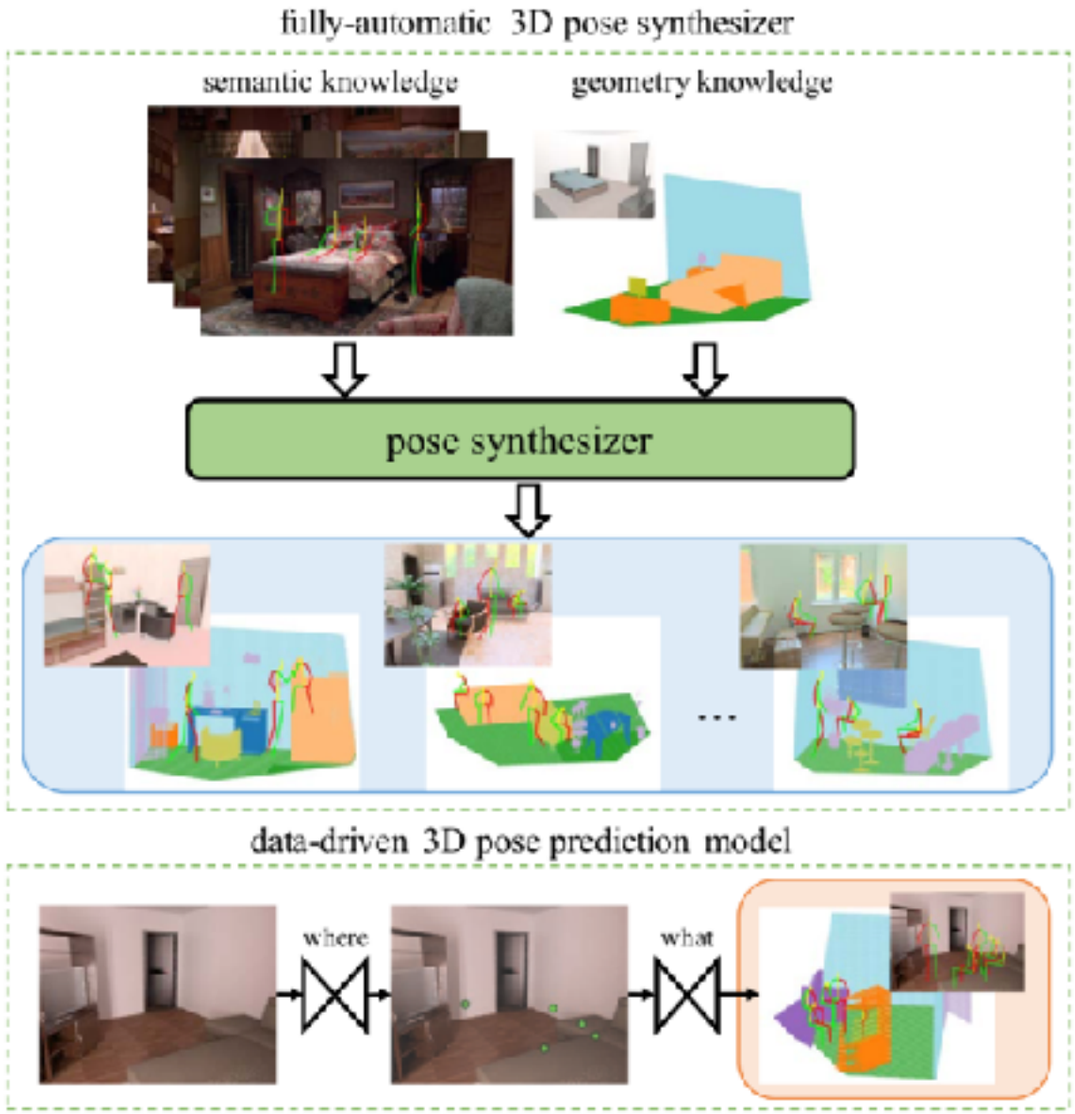


Affordance: The behaviour of humans with respect to objects in a 3D scene

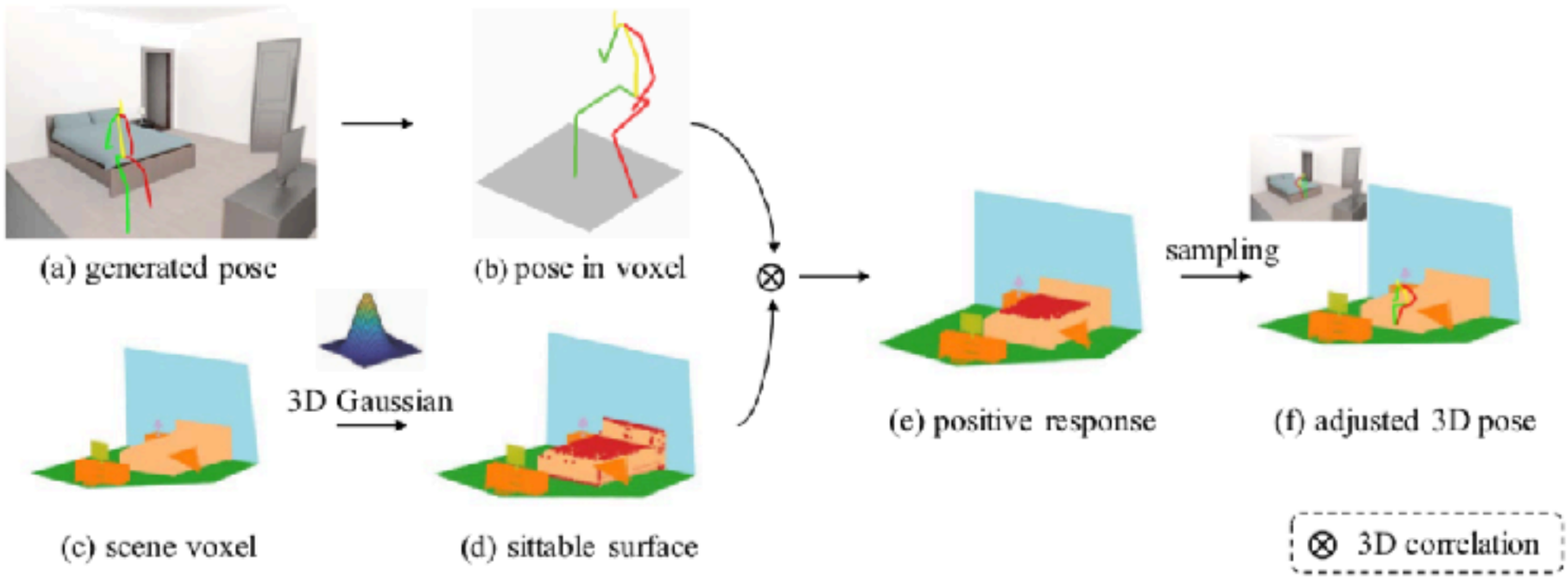
- What human body poses are afforded by a certain object in the scene: a chair, a bed, a machine etc.

- The 3D geometric representation and semantics of the scene need to be computed along with the possibilities of human interaction.

- An extensive dataset with 3D affordances is required to model such interactions.

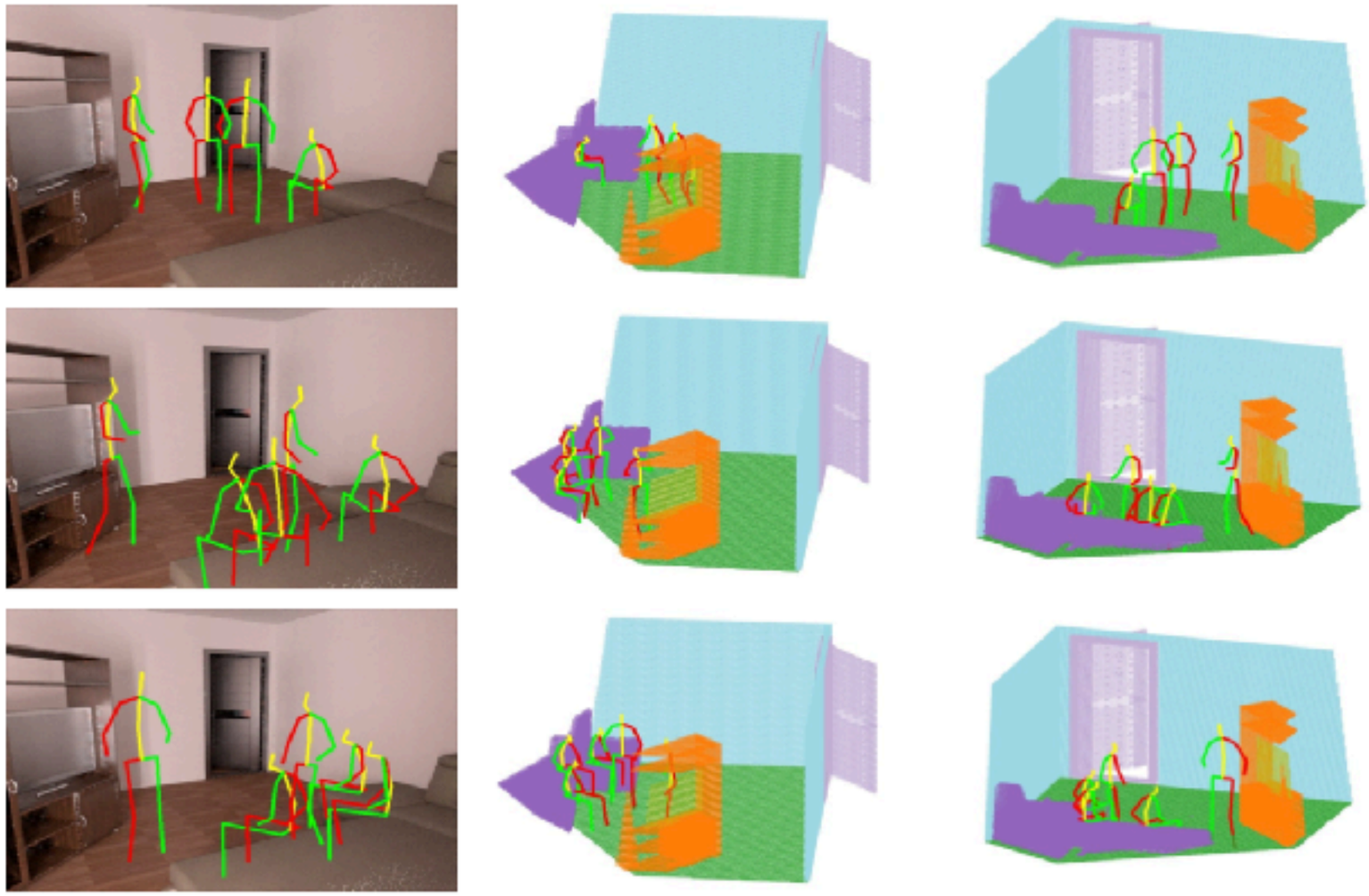


Affordance adjustment



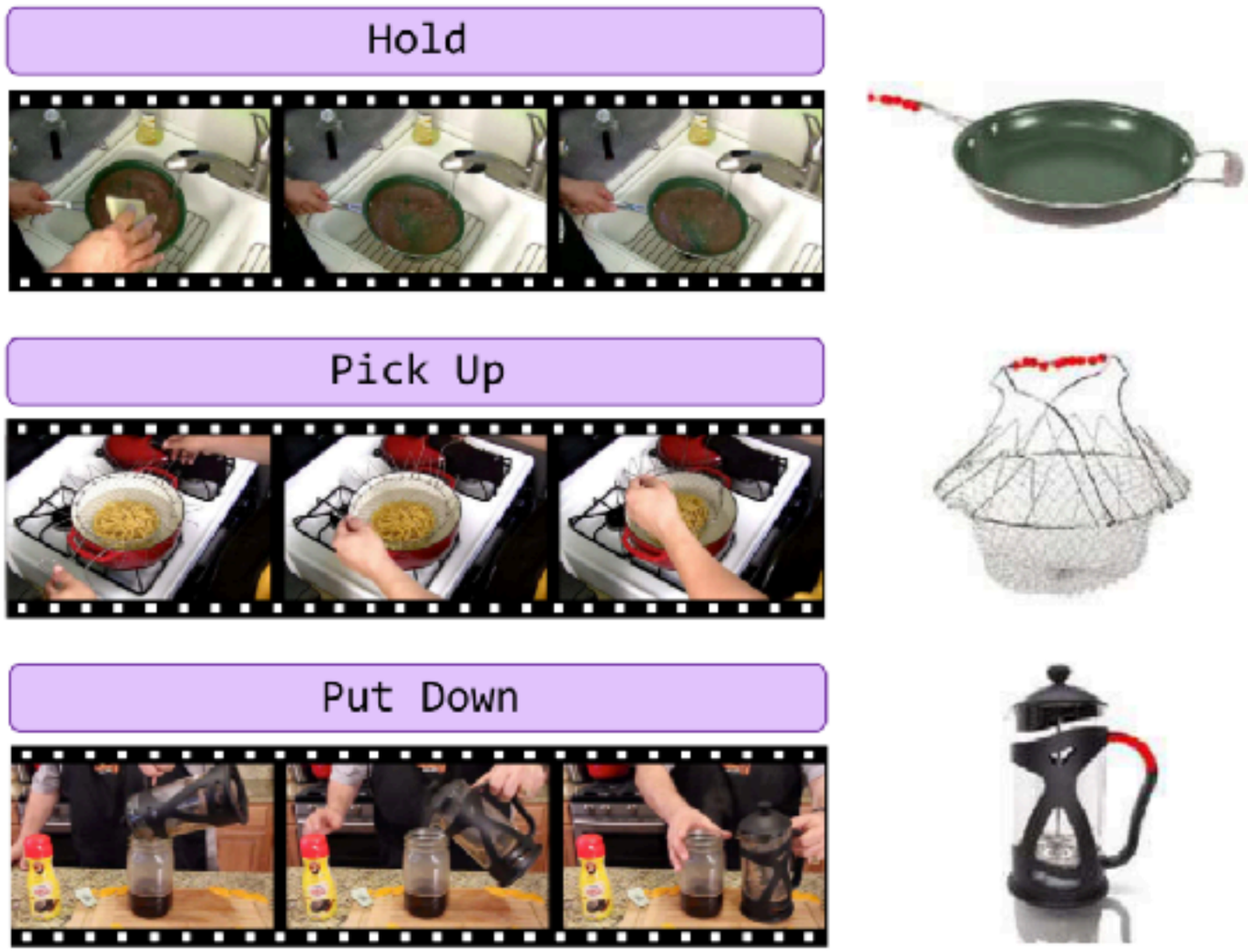
"Putting Humans in a Scene: Learning Affordance in 3D Indoor Environments" Li et al. CVPR 2019
http://openaccess.thecvf.com/content_CVPR_2019/papers/Li_Putting_Humans_in_a_Scene_Learning_Affordance_in_3D_Indoor_CVPR_2019_paper.pdf

Prediction of human activities within a 3D space



"Putting Humans in a Scene: Learning Affordance in 3D Indoor Environments" Li et al. CVPR 2019
http://openaccess.thecvf.com/content_CVPR_2019/papers/Li_Putting_Humans_in_a_Scene_Learning_Affordance_in_3D_Indoor_CVPR_2019_paper.pdf

Prediction of human interactions and manipulations with an object



“Demo2Vec: Reasoning Object Affordances from Online Videos” Fang et al. CVPR 2018
http://openaccess.thecvf.com/content_cvpr_2018/papers/Fang_Demo2Vec_Reasoning_Object_CVPR_2018_paper.pdf

(4) Task-level autonomy

- The Human stops directly operating the robot and sends complex commands instead. e.g, Instead of operating the robot arm, the human can point towards a door from the streamed visuals in a VR-device, and say „open this“. This task must be executed autonomously by the robot.

- **Required robot capabilities:**
  - **Semantic Mapping:** Not only understand the geometric structure of objects, but also their semantic components (e.g, that the door has a handle)
  - **Motion Planning:** Ability to generate complex articulated motions, and autonomously avoid dynamic obstacles while executing that motion

- **Cognitive Functions:**
  - **Task Division and Planning:**The user command (“open the door”) must be divided into multiple sub-tasks that can be autonomously planned and executed
  - **Error correction:** The capability to recognise automatically that a sub-task is not completed successfully (e.g, to grip the handle of a door) and autonomously correct this without asking the human operator.

Dataset collection by human teleoperation of the humanoid root

- Humanoide Roboter können aus der Ferne ferngesteuert werden, und diese Daten werden gesammelt, um autonome Aufgaben zu erlernen.

(5) Collaborative autonomy: Cognitive team-work

- The robot acts as an peer co-worker. It understands the context of the mission. It can autonomously detect the weakness or stress in a human co-worker and take over specific tasks by itself to reduce the work load.

- **Required robot capabilities:**

    - **Contextual understanding:** Understanding the overall context of a mission (e.g, „We are here to rescue a person, not to clean up the area“).

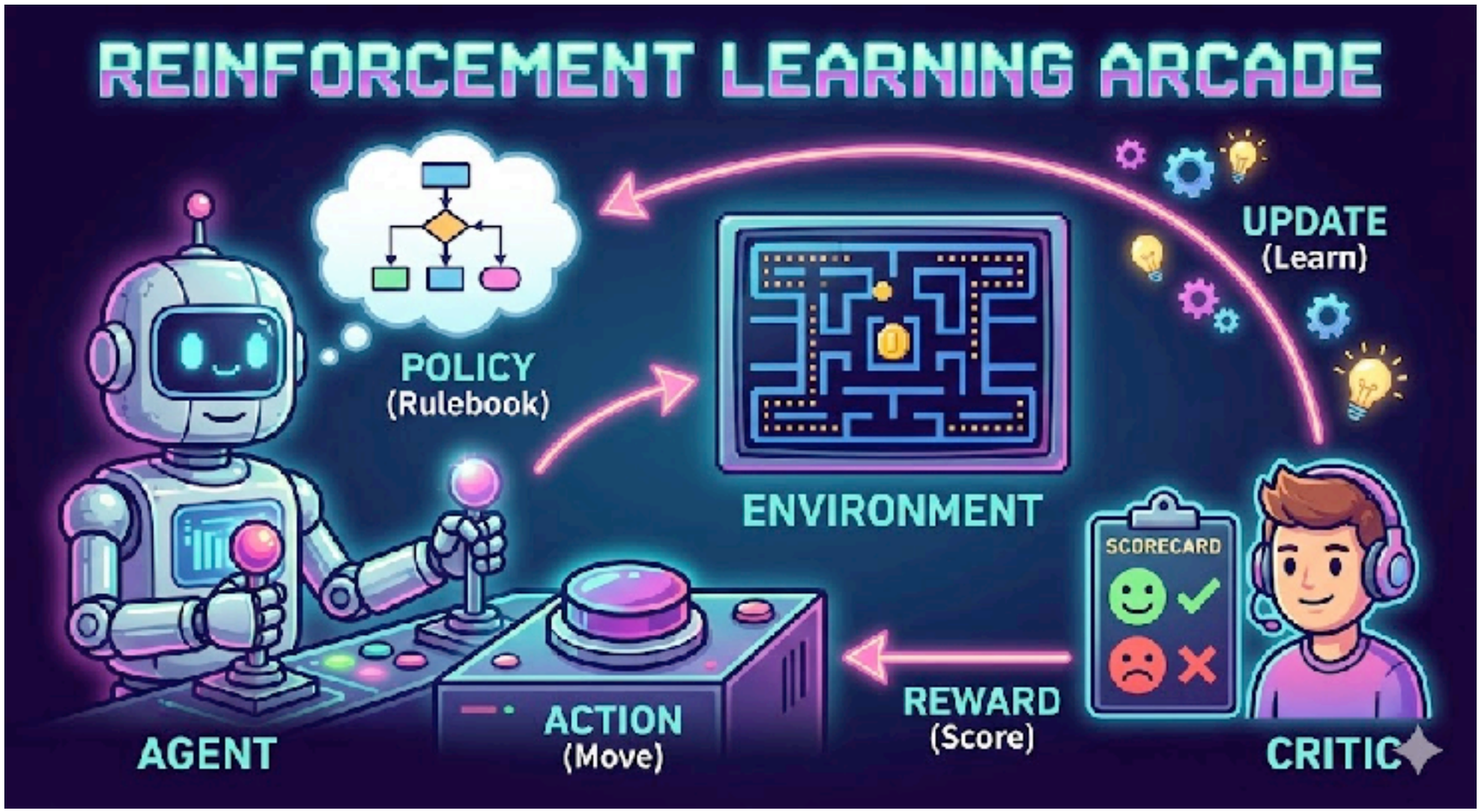
    - **Adaptive learning:** Maintaining a model of human co-worker and learning their preferences and proclivities over time through collaboration.

- **Cognitive Functions:**

    - **Theory of Mind:** The robot has a complete mental model of the human co-worker (Knowledge, goals and limits).

    - **Proactive Thinking:** Anticipation of future requirements and the execution of tasks, in order to reduce the cognitive load of human co-workers.

Reinforcement Learning



- For the learning of autonomous actions, the robot requires an interactive simulated environment that has all the relevant physical laws appropriately simulated: World Model

World Model

- **Passive world model:** Watch videos, predict future frames. Diffusion based generative models are a good start.

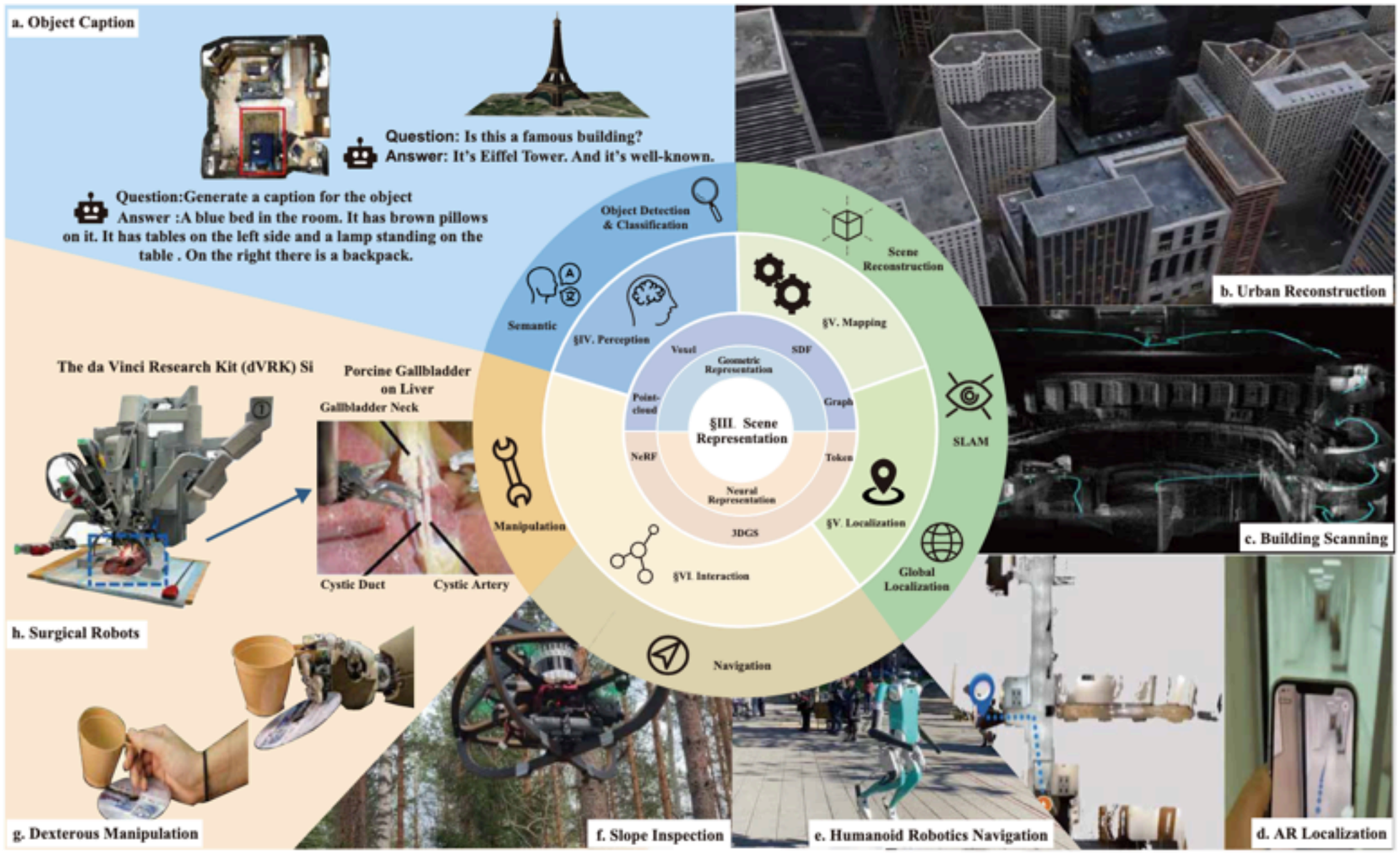
- **Aktive, interaktive world model:** enable interactive predictions, so that the agent can learn on the basis of a complex chain of examples scenarios.

- Detailled realistic **3D-modeling of the environment:** can be conducted through 3D Gaussian Splatting, SLAM and other 3D scanning methods

- Realistic **simulation of physics**: can be trained through diffusion based generative models, or causal predictive models (JEPA etc).

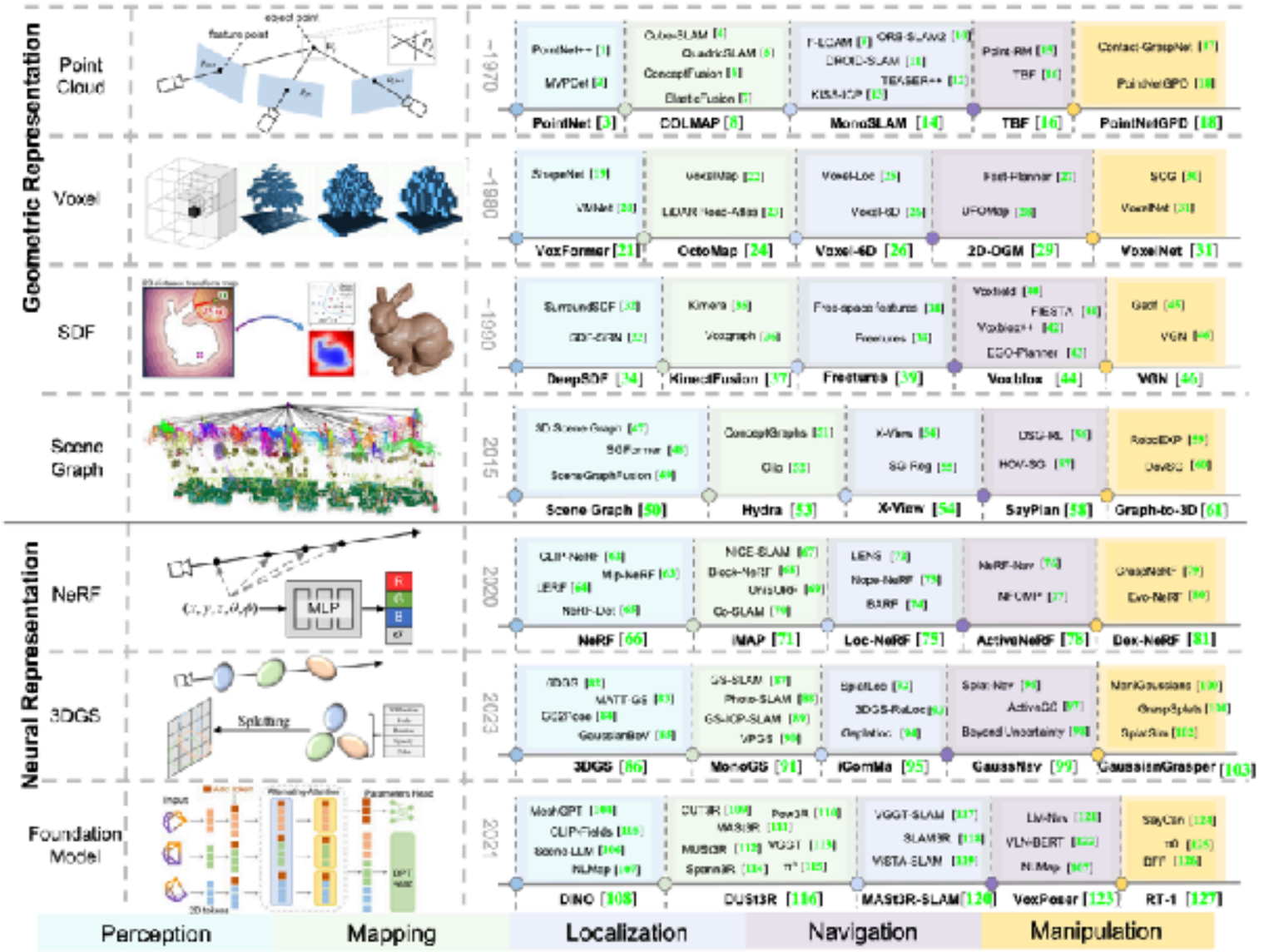
- **Vision-Language-Action-M (VLA)**: Extension of visual-language generative models, where the robot actions are tokenized and represented in the same latent space. This is learned on an appropriate multi-model dataset.

3D scene representations in robotics



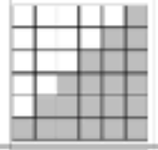
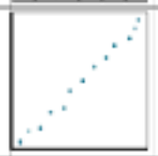
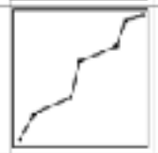
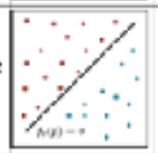
"What is the best 3D Scene representation for robotics? From geometric to foundation models"
Deng et al. IEEE Transactions in Robotics 2025

3D-Scene representation: Perception, Mapping, Localization, Navigation, Manipulation



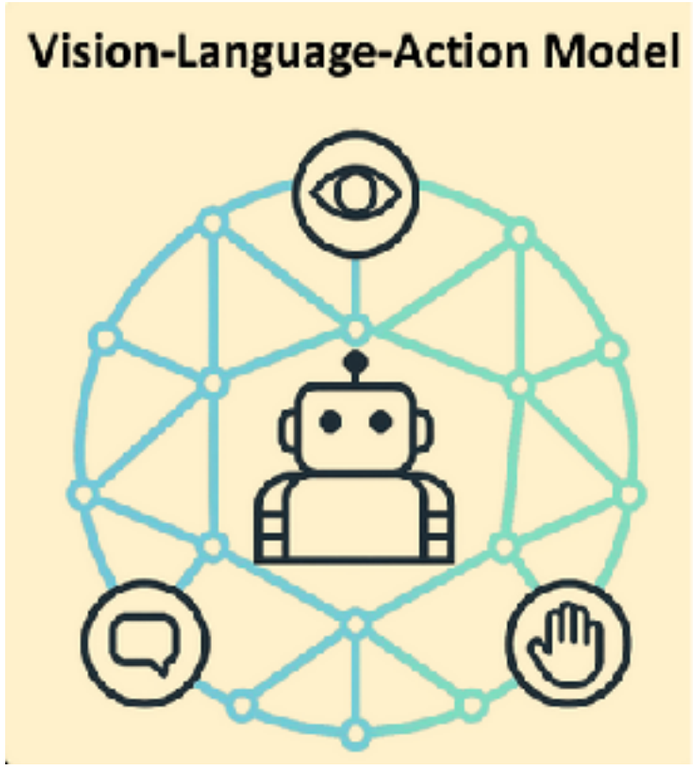
"What is the best 3D Scene representation for robotics? From geometric to foundation models"
Deng et al. IEEE Transactions in Robotics 2025

Types of 3D-Scene Representations: Advantages and Disadvantages

Category	Data form	Continuous & Differentiable	Memory Efficiency	Photorealism	Flexibility	Geometric representation capability
Voxel		Discrete & Non-Differentiable	+++	+	++	++
Point Cloud		Discrete & Non-Differentiable	++	+++	+++	+++
Mesh		Discrete & Non-Differentiable	+++	++	+	++++
Scene Graph		Discrete & Non-Differentiable	++++	+	++++	+
Neural Radiance Fields (NeRF)		Continuous & Differentiable	+++	++++	+++	+++
3D Gaussian Splatting(3DGS)		Continuous & Differentiable	+	++++	+++	+++
Tokenizer		Continuous & Differentiable	++++	++++	+++	+++

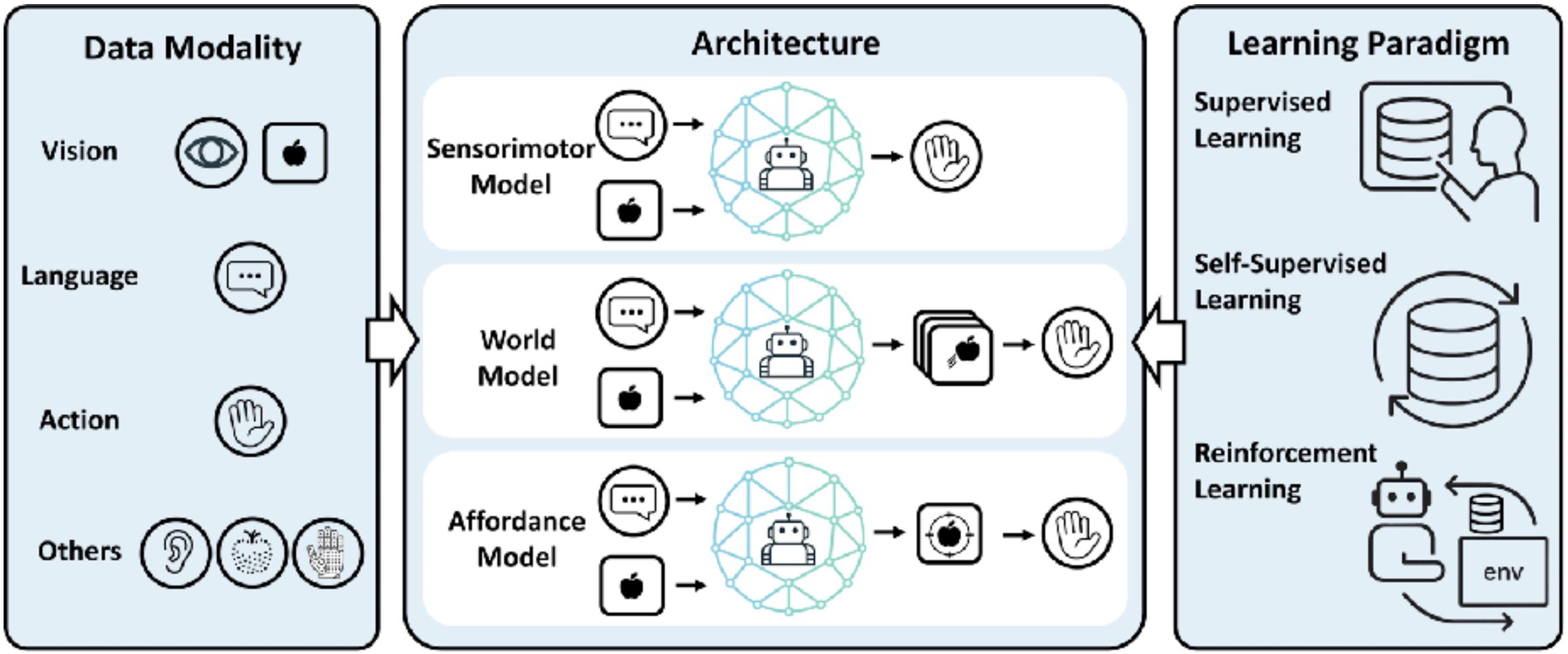
“What is the best 3D Scene representation for robotics? From geometric to foundation models”
Deng et al. IEEE Transactions in Robotics 2025

VLA-Models



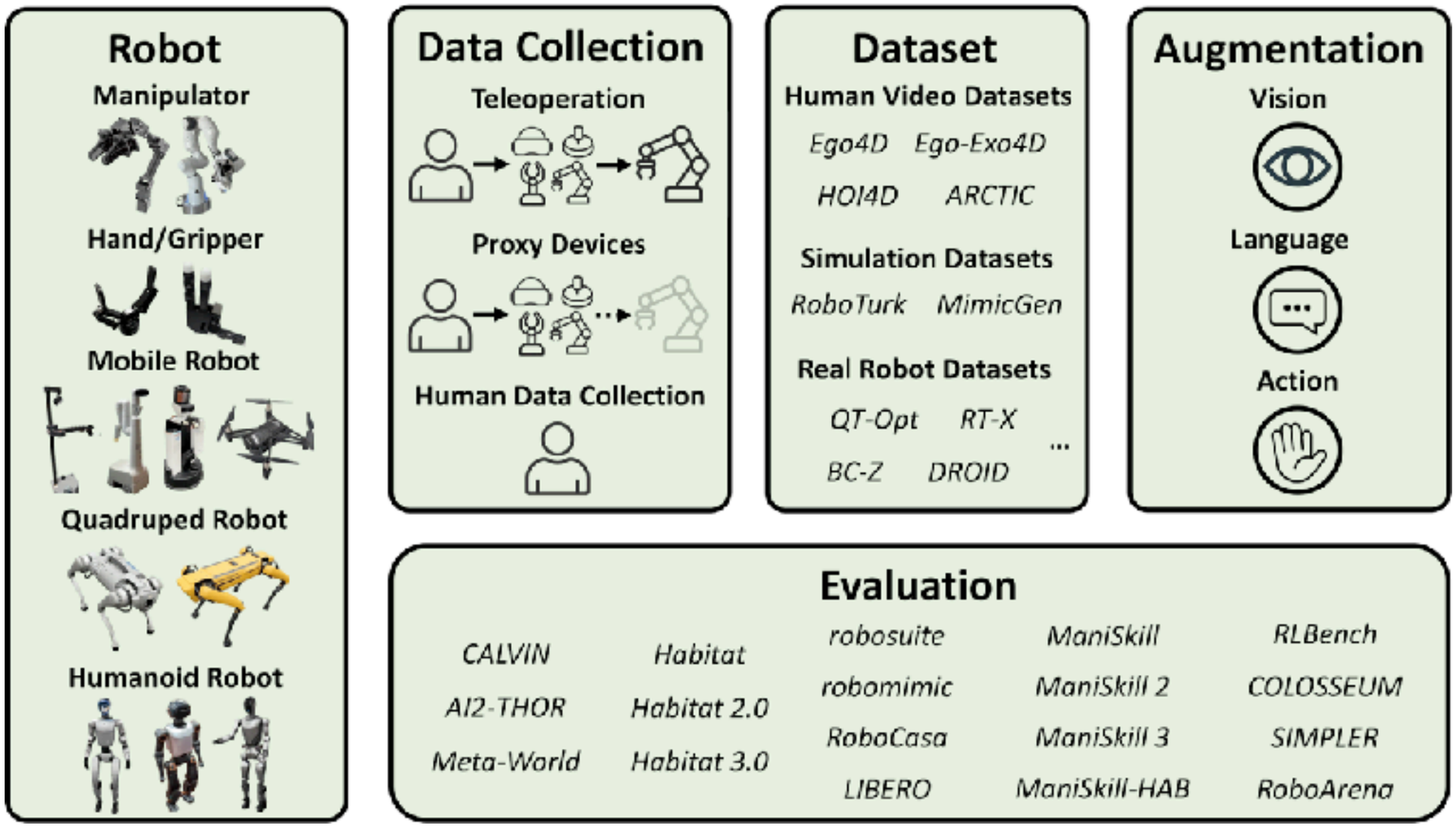
- Uniform Tokenisation of language commands, visual/sensor streams and robot actions.
- Coarse-to-fine learning of robot skills
- A possible foundation model for robotics (like LLMs for language processing)
- Generative model: May have certain limitations with making physically based predictions.

VLA-Models: Architecture



"Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications" Kawaharazuka et al. 2025 (https://vla-survey.github.io/)

VLA-Models: Datasets



"Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications" Kawaharazuka et al. 2025 (https://vla-survey.github.io/)

Questions