Hands on AI based 3D Vision Summer Semester 26

Lecture 1 – AI 3D Vision- Introduction

Prof. Dr.-Ing. Gerard Pons-Moll

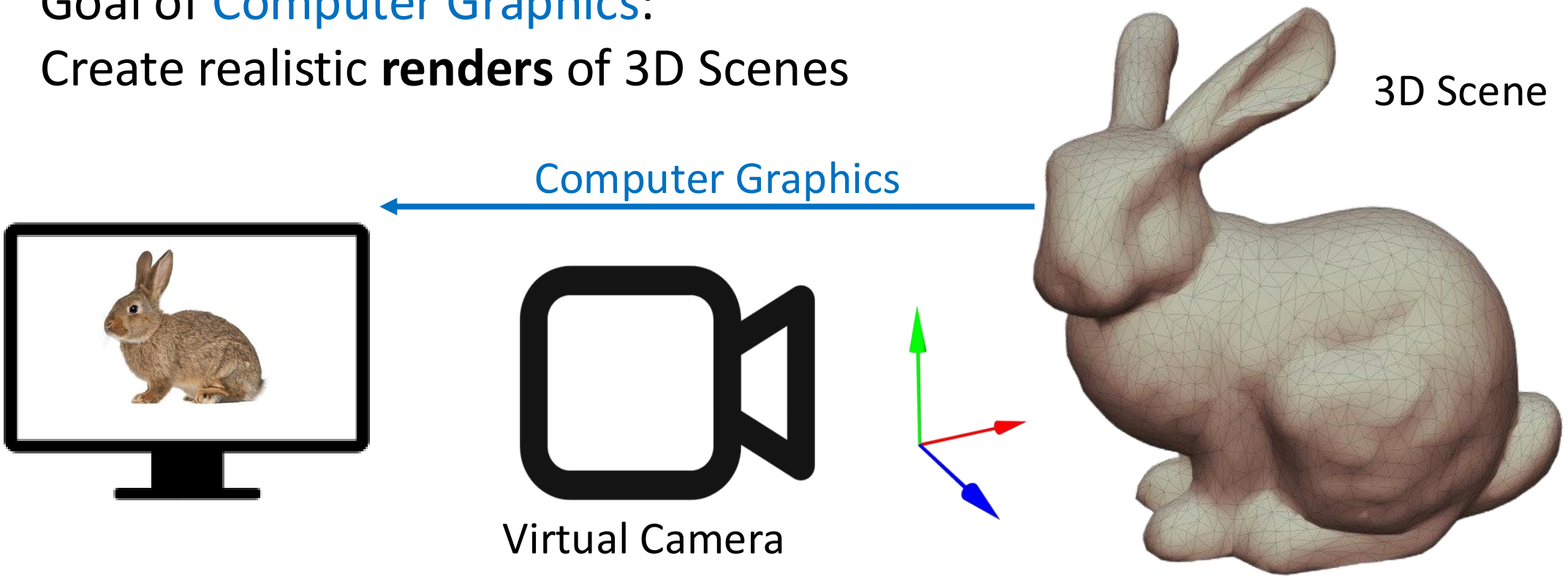
University of Tübingen / MPI-Informatics

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

3D Computer Vision vs Computer Graphics

3D Computer Vision vs Computer Graphics

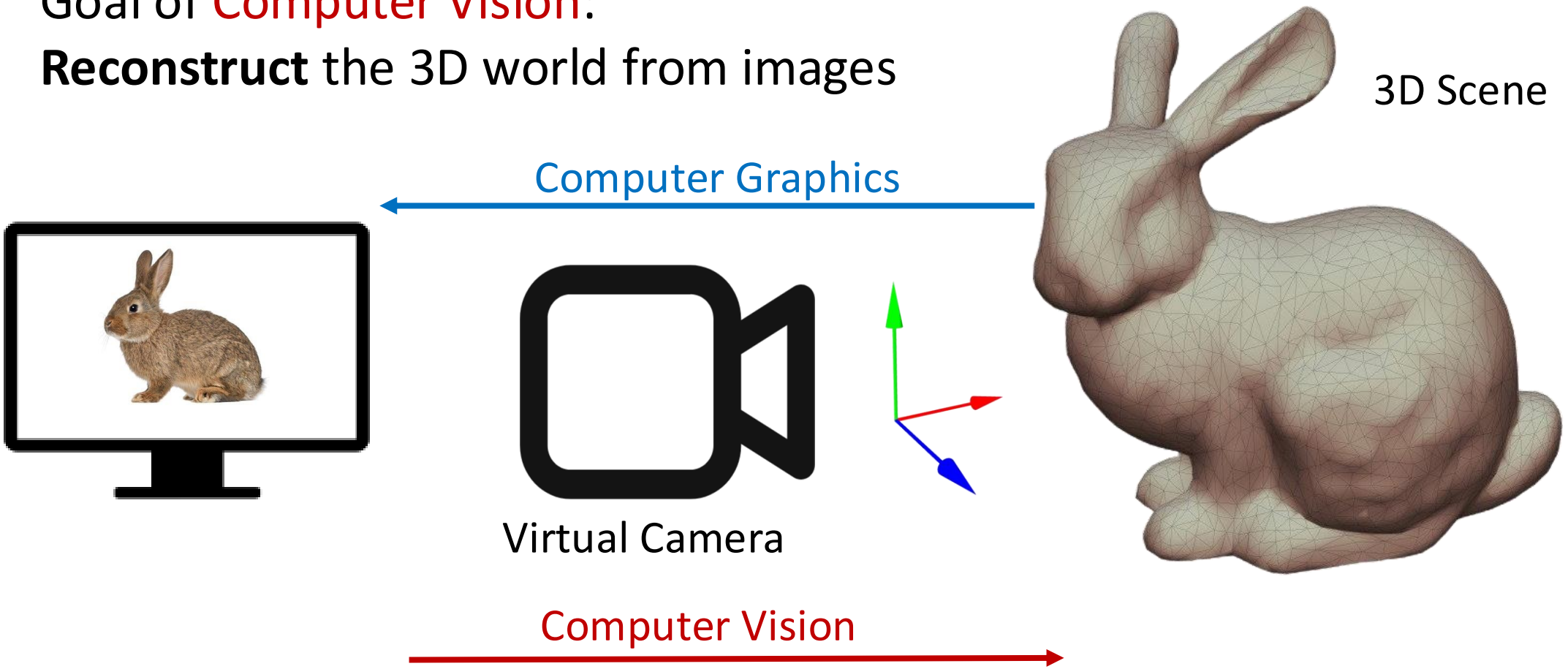
Goal of Computer Graphics:
Create realistic **renders** of 3D Scenes



3D Computer Vision vs Computer Graphics

Goal of Computer Vision:

Reconstruct the 3D world from images



3D Computer Vision vs Computer Graphics

Why is 3D vision important?

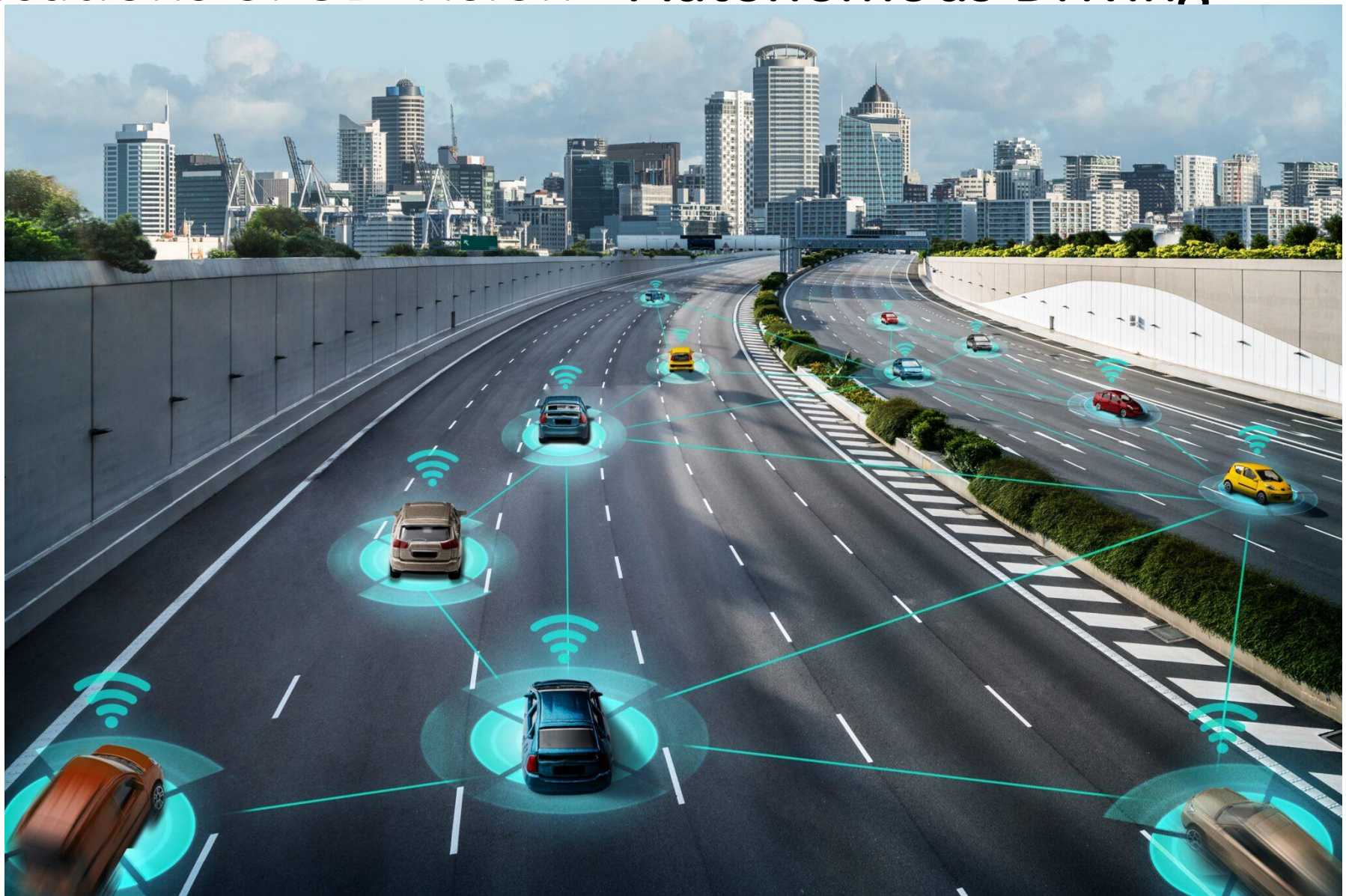
Applications of 3D Vision - **Entertainment Industry**

Star Wars The Mandalorian: Luke Skywalker Behind the Scenes | Disney+

Applications of 3D Vision - **Entertainment Industry**

How They Made Me Look 23 in Gemini Man

Applications of 3D Vision - **Autonomous Driving**



Applications of 3D Vision - **Autonomous Driving**

How Tesla's Driver Monitoring System Works

Applications of 3D Vision - **Autonomous Driving**

[A self-driving car from Waymo]

[Block-Nerf], Tancik et al. CVPR 2024

Applications of 3D Vision - **Virtual Reality**

Half-Life: Alyx

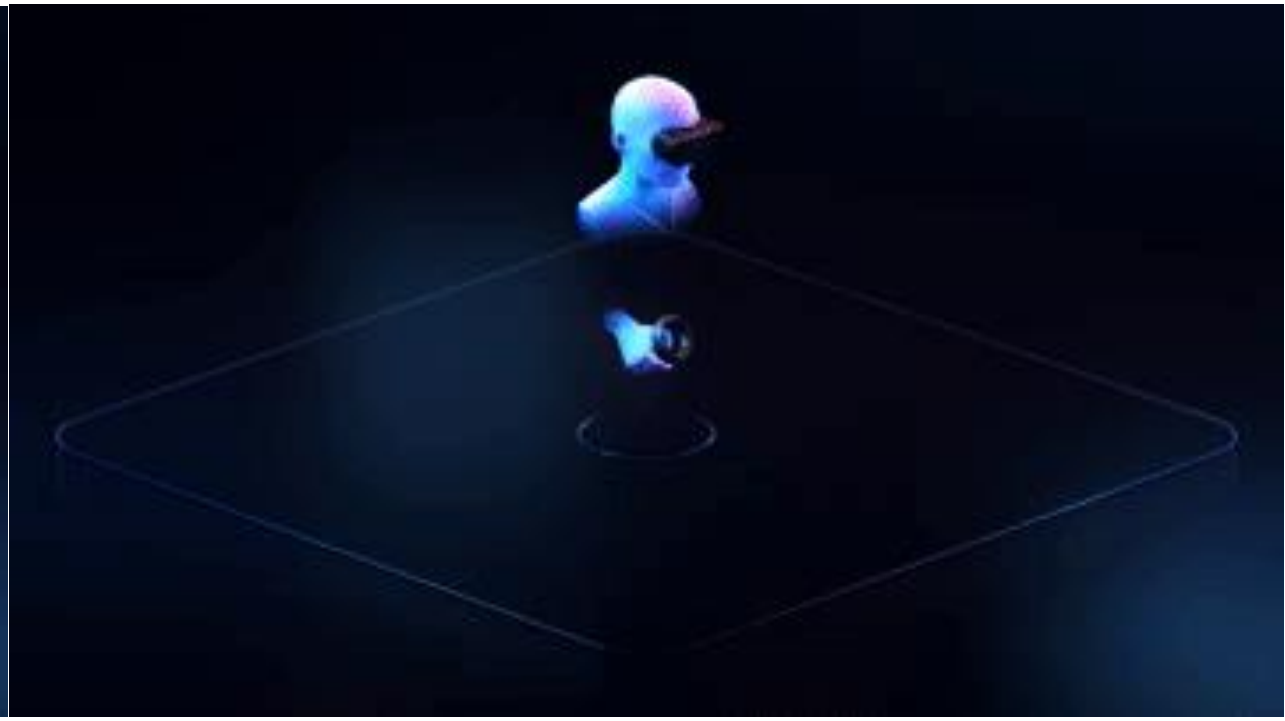
Applications of 3D Vision - **Virtual Reality**

Many sensors are involved

[Apple Vision Pro](). RGB cameras, LiDAR, TrueDepth cameras, IR cameras

Applications of 3D Vision - **Virtual Reality**

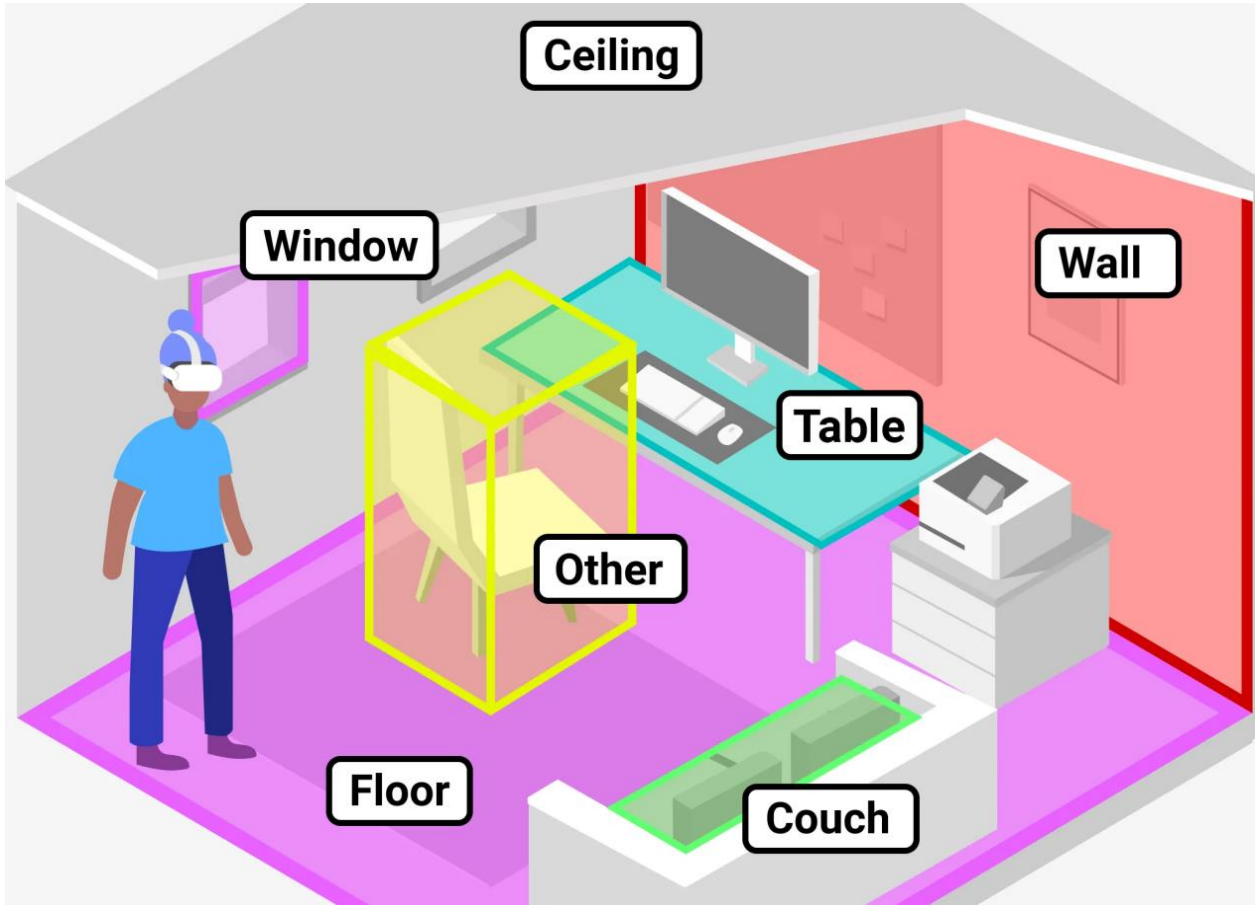
Track user position



[Oculus Rift setup]

Applications of 3D Vision - **Virtual Reality**

Understand the environment



Meta Quest

Applications of 3D Vision - **Virtual Reality**

Track user's hands and body

[Predicting 4D Hand Trajectory from Monocular Videos](), Ye et al. 2025

Applications of 3D Vision - **Human Pose Estimation**

Track humans from monocular videos

Neural Localizer Fields for Continuous 3D Human Pose and Shape Estimation Sarandi et al. NeurIPS '24

Applications of 3D Vision - **Human Object Interaction**

Interaction Replica: Tracking human–object interaction and scene changes from human motion







Vladimir Guzov[1,2] Julian Chibane[1,2] Riccardo Marin[1] Yannan He[1] Yunus Saracoglu[1] Torsten Sattler[3] Gerard Pons-Moll[1,2]

[1] University of Tübingen, Tübingen AI Center, Germany

[2] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

[3] CIIRC, Czech Technical University in Prague, Czech Republic

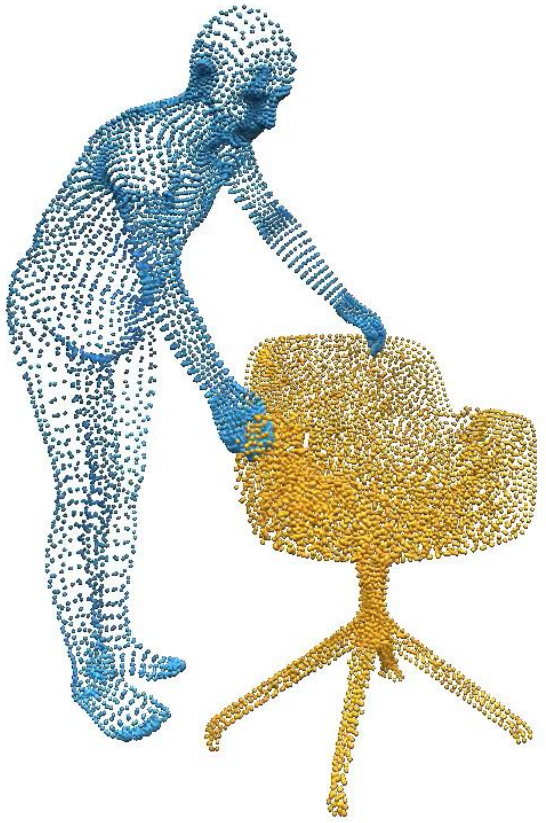
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

max planck institut informatik

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

Interaction Replica, Guzov et al. 3DV 2024

Applications of 3D Vision - **Human Object Interaction**



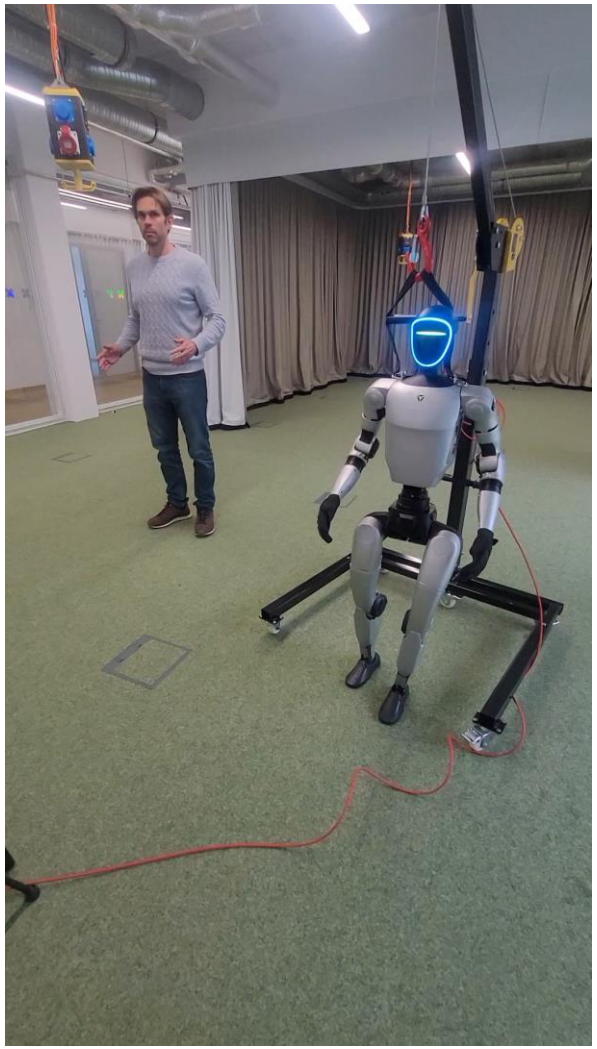
InterTrack: Tracking Human Object Interaction without Object Templates, Xie et al. 3DV 2025

Applications of 3D Vision - **Robotics**

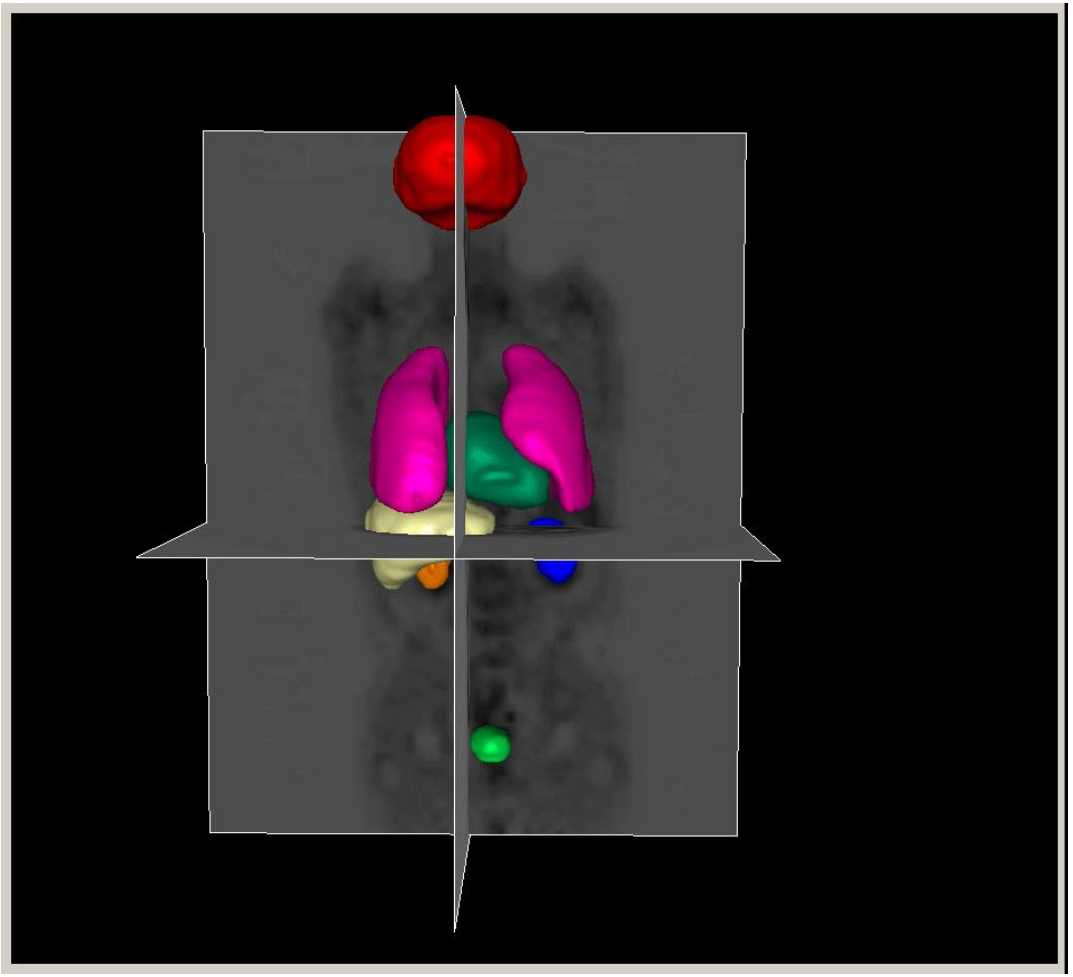
Robots interacting with environment need 3D scene understanding

OmniH2O: Universal and Dexterous Human-to-Humanoid Whole-Body Teleoperation and Learning, He et al. CORL '24

Applications of 3D Vision - **Robotics**



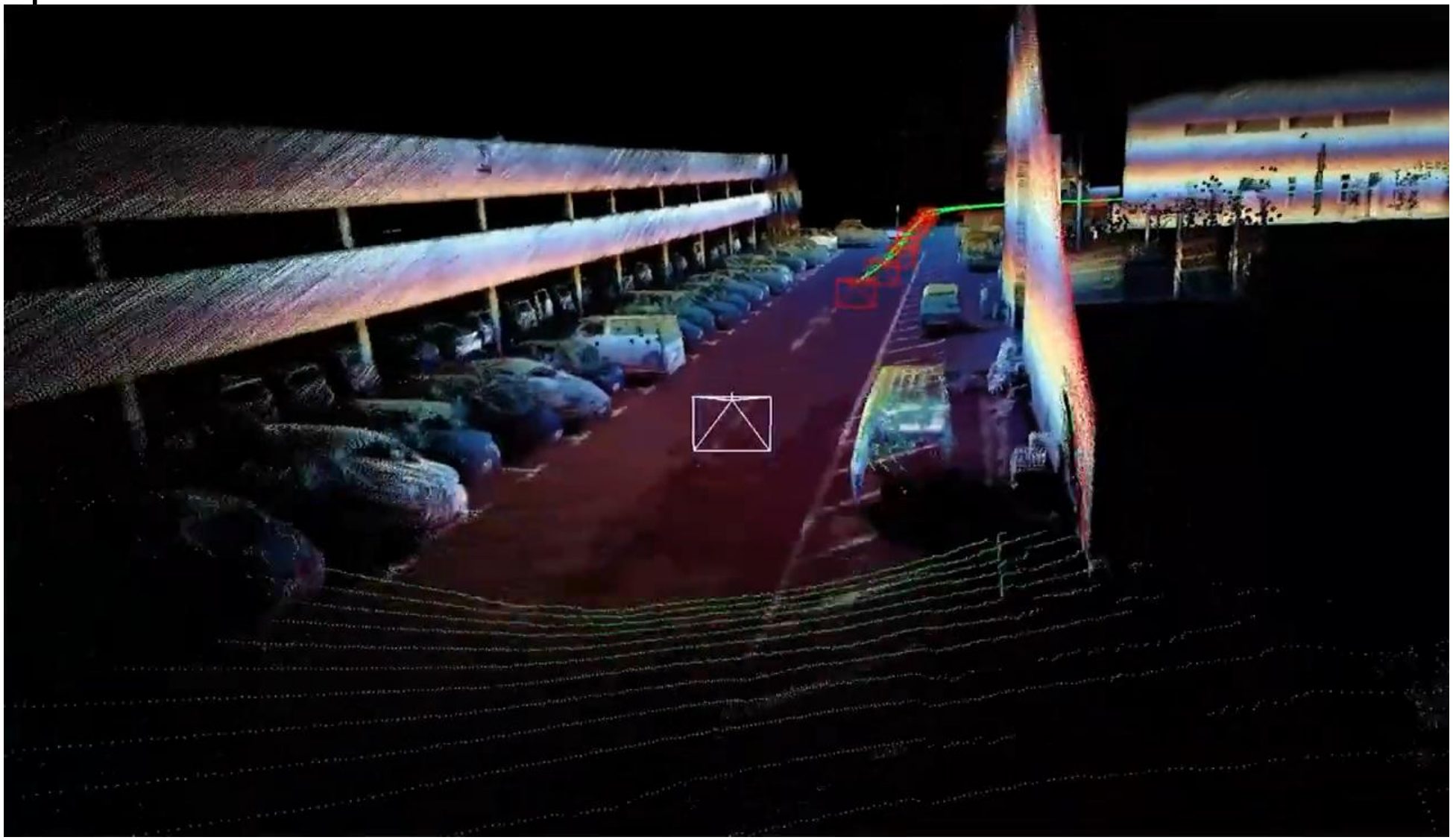
Applications of 3D Vision - **Healthcare**



[3D Modeling, Simulation and Segmentation in Medical Images]

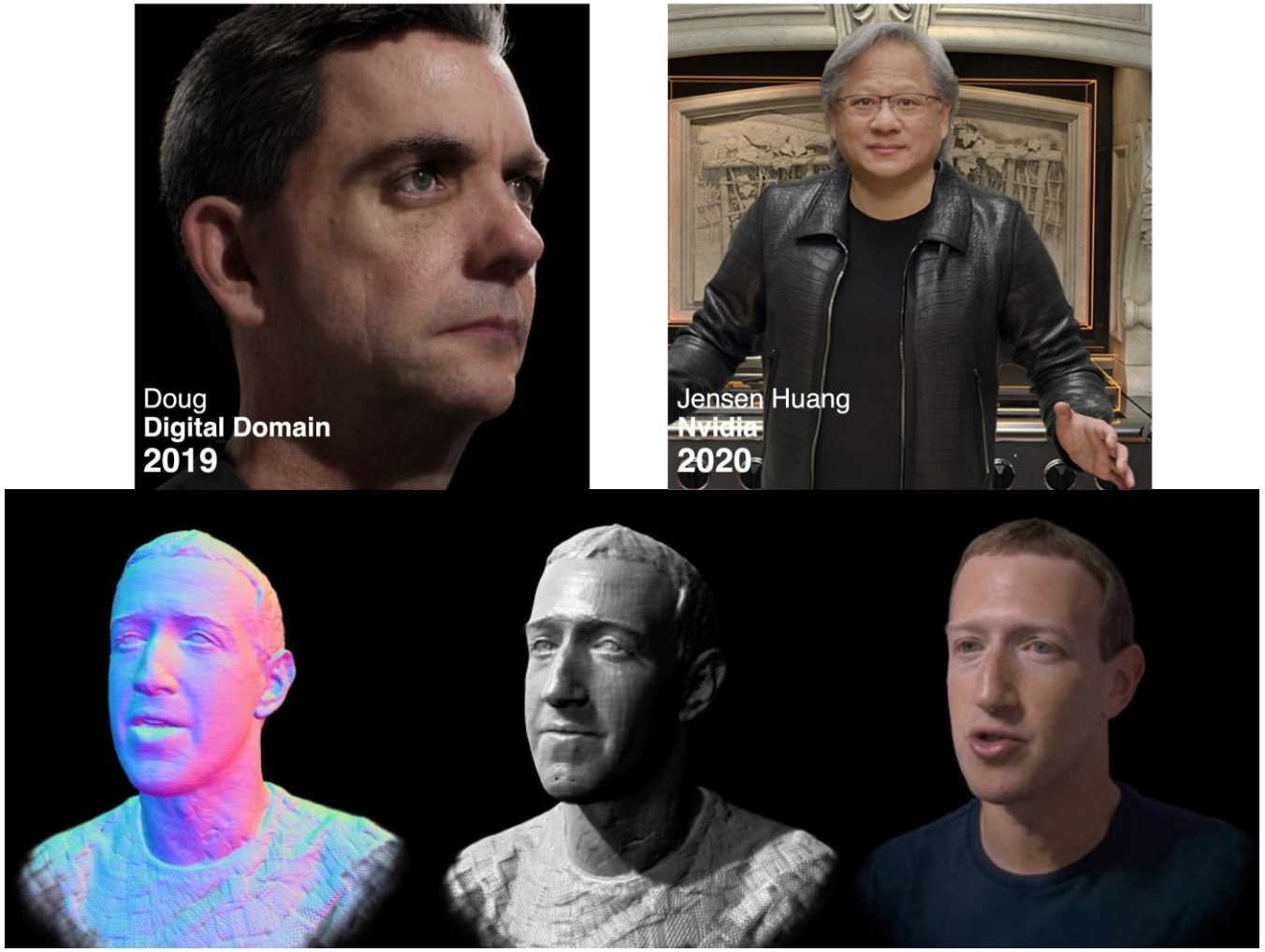
[3D Cone Beam CT Scan and Animation]

Applications of 3D Vision - **Scene Reconstruction**



Real-time Odometry-less 3D LiDAR SLAM with Generalized ICP and Pose-Graph Optimization

Applications of 3D Vision - **Human Avatar Creation**



[Relightable Gaussian Codec Avatars](), Saito et al. CVPR 2024

Applications of 3D Vision - **Human Avatar Creation**

Relightable Gaussian Codec Avatars, Saito et al. CVPR 2024

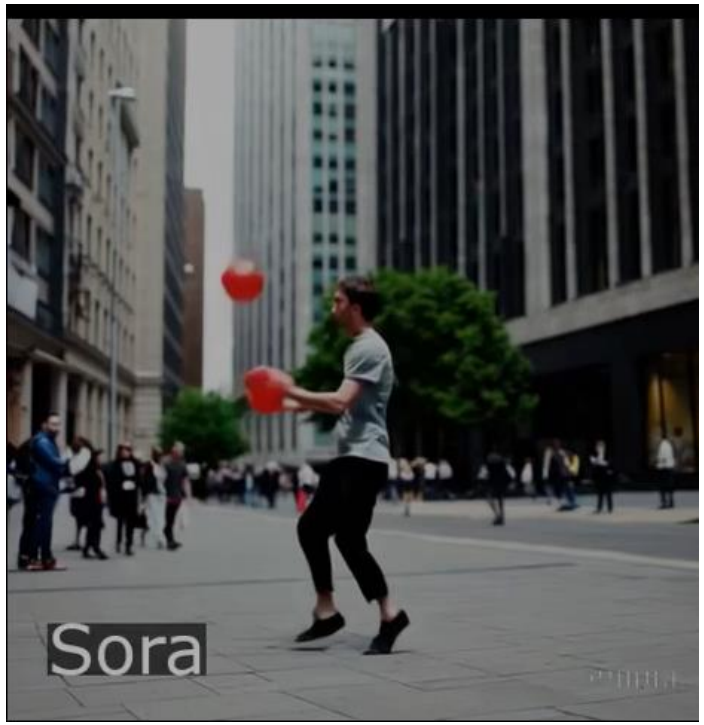
3D critics

Common argument you will hear:

3D is a man made representation, just like segmentation, detection, etc. Ultimately, it will be unnecessary for end goals like robot learning. All you need is pixels in, action out. No 3D

Video source: videojam

Scale is all we need? No need for 3D. Really?



Video source: videojam

How to make AI our partners

- Make AI understand **physical interaction**
- Give AI the **appearance** and behavior of humans

A Brief History of 3D Vision

A Brief History of 3D Vision

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

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

THE SUMMER VISION PROJECT

Seymour Papert

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

Goals - General

The primary goal of the project is to construct a system of programs which will divide a vidisector picture into regions such as

likely objects

likely background areas

chaos.

We shall call this part of its operation FIGURE-GROUND analysis.

It will be impossible to do this without considerable analysis of shape and surface properties, so FIGURE-GROUND analysis is really inseparable in practice from the second goal which is REGION DESCRIPTION.

The final goal is OBJECT IDENTIFICATION which will actually name objects by matching them with a vocabulary of known objects.

60s – 80s: Basic Image Processing & Recognition

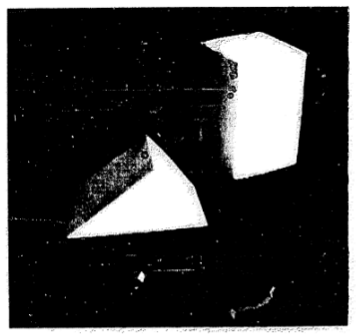
MACHINE PERCEPTION OF THREE-DIMENSIONAL SOLIDS

by

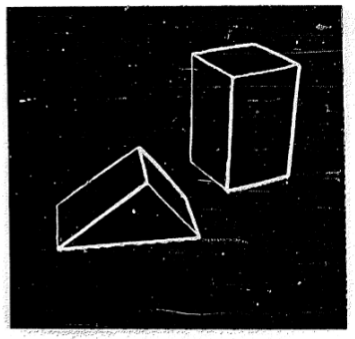
LAWRENCE GILMAN ROBERTS

S.B., Massachusetts Institute of Technology
(1961)

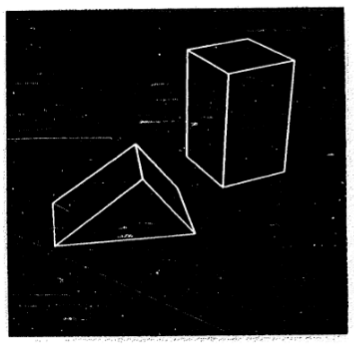
M.S., Massachusetts Institute of Technology
(1961)



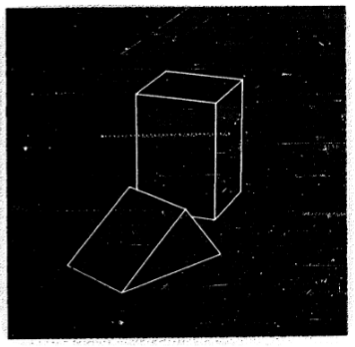
A. Original Picture



B. Differentiated Picture



C. Line Drawing



D. Rotated View

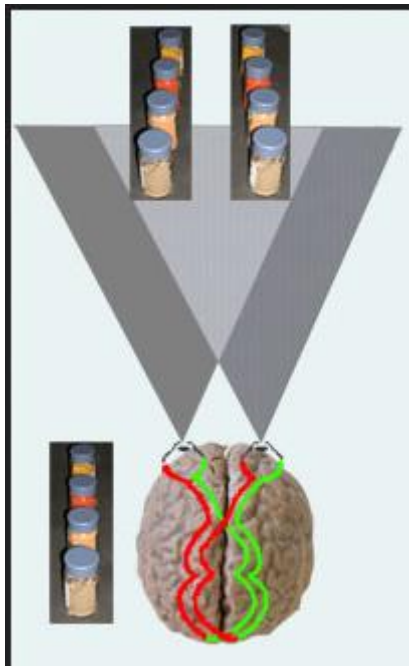
70s – 80s: Stereo Vision

A computer algorithm for reconstructing a scene from two projections

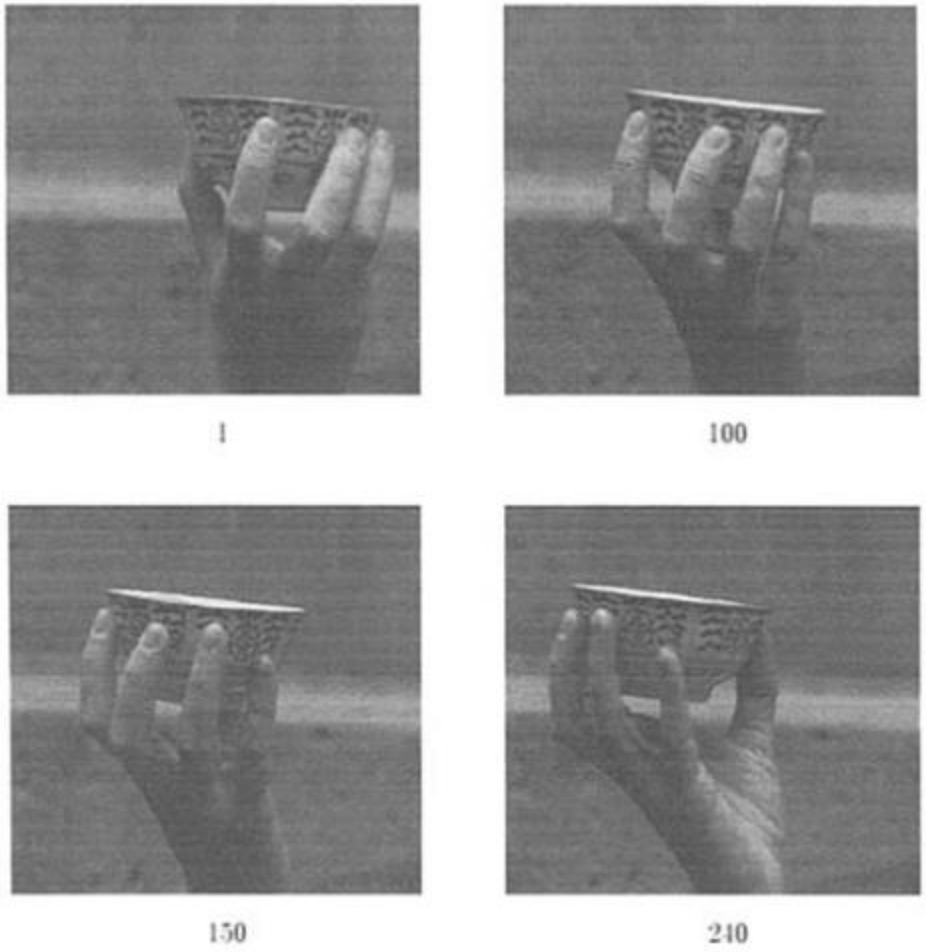
H. C. Longuet-Higgins

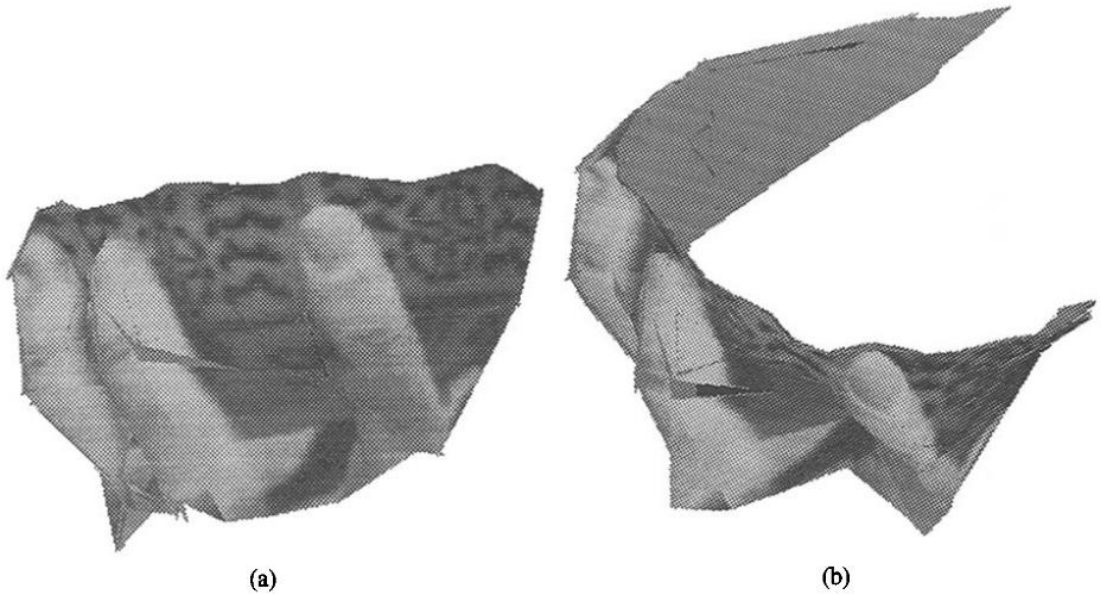
Laboratory of Experimental Psychology, University of Sussex, Brighton BN1 9QG, UK

A simple algorithm for computing the three-dimensional structure of a scene from a correlated pair of perspective projections is described here, when the spatial relationship between the two projections is unknown. This problem is relevant not only to photographic surveying[1] but also to binocular vision[2], where the non-visual information available to the observer about the orientation and focal length of each eye is much less accurate than the optical information supplied by the retinal images themselves. The problem also arises in monocular perception of motion[3], where the two projections represent views which are separated in time as well as space. As Marr and Poggio[4] have noted, the fusing of two images to produce a three-dimensional percept involves two distinct processes: the establishment of a 1:1 correspondence between image points in the two views—the 'correspondence problem'—and the use of the associated disparities for determining the distances of visible elements in the scene. I shall assume that the correspondence problem has been solved; the problem of reconstructing the scene then reduces to that of finding the relative orientation of the two viewpoints.



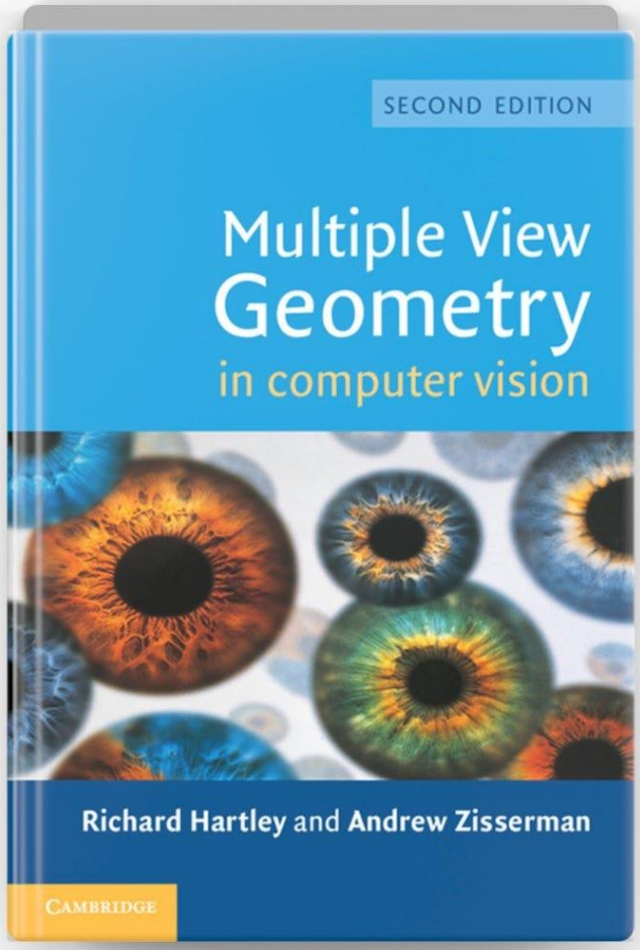
90s – 00s: Structure from Motion





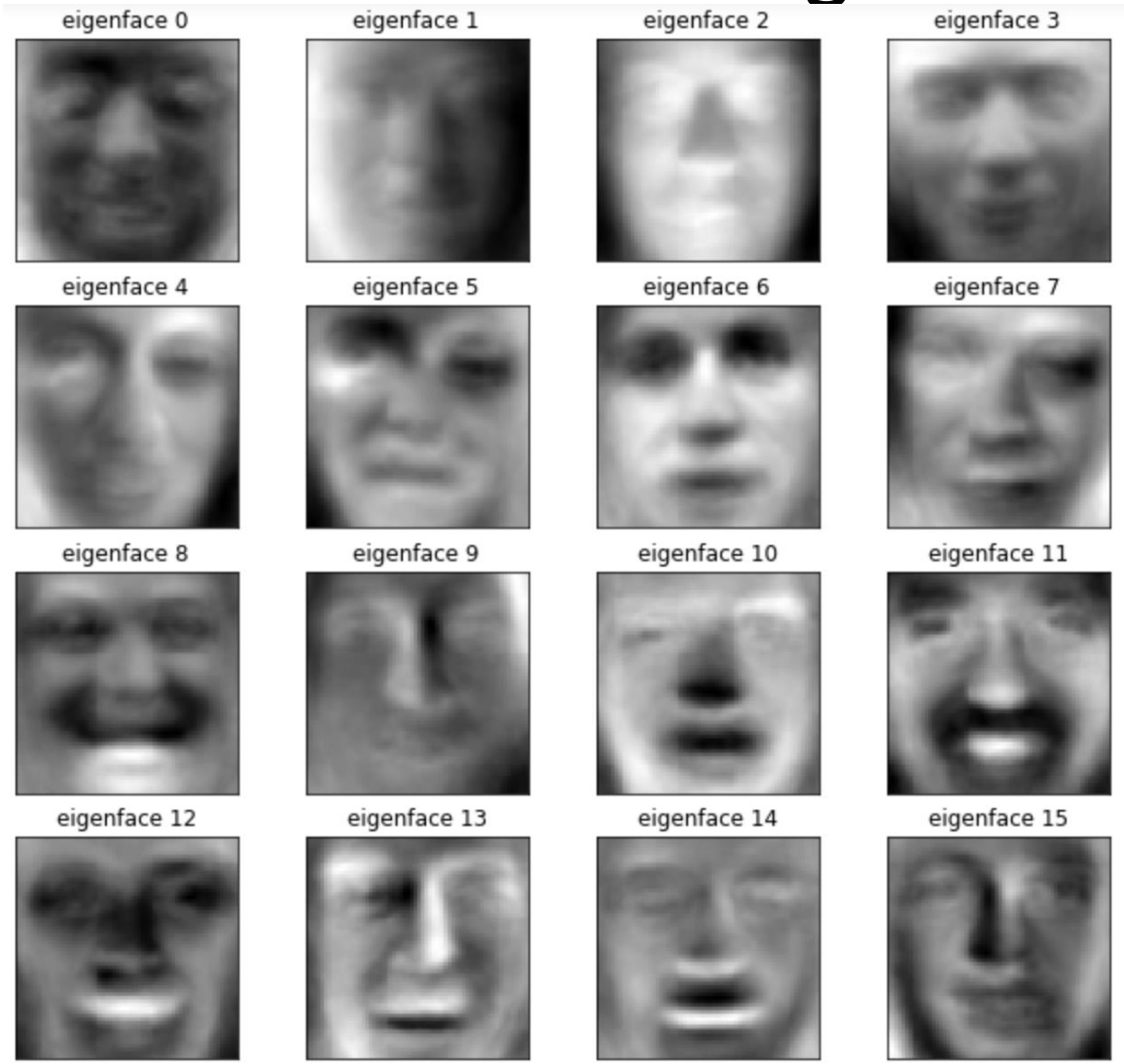
Shape and Motion from Image Streams under Orthography: a Factorization Method. Tomasi IJCV 1992

Multiple View Geometry as Established CV Discipline



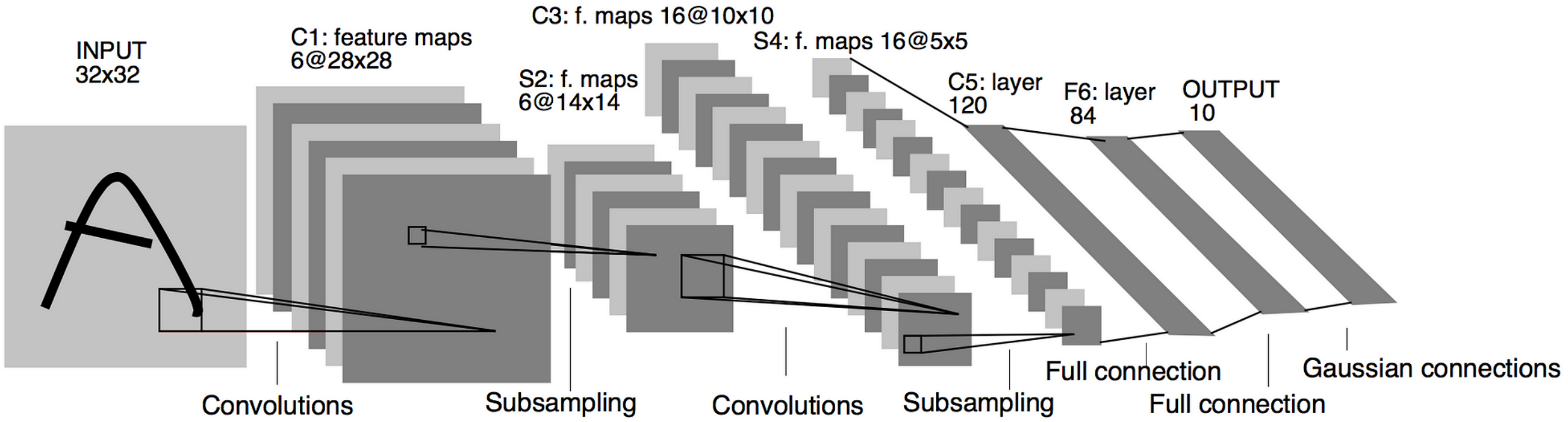
Multiple View Geometry

90s – 00s: Rise of Statistical Learning Methods



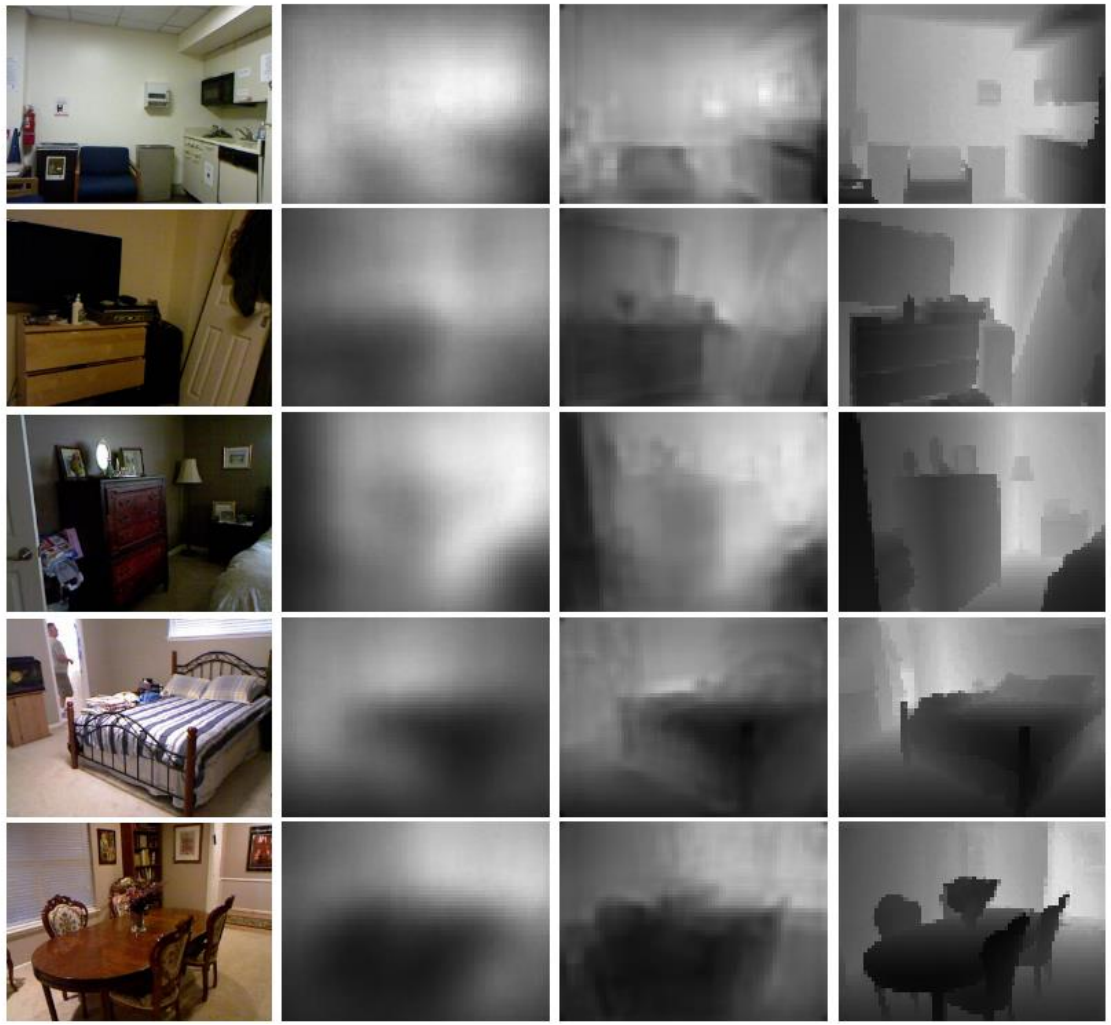
https://medium.com/data-science/eigenfaces-recovering-humans-from-ghosts-17606c328184

90s – 00s: Early Deep Learning for CV



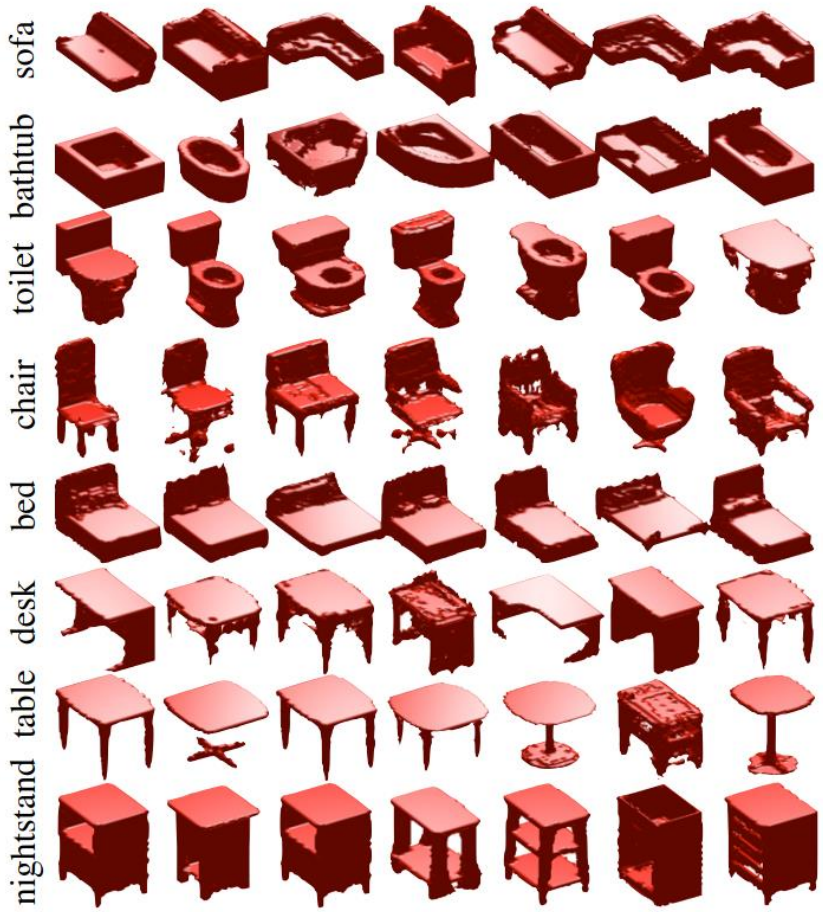
Gradient-Based Learning Applied to Document Recognition, LeCun et al. Proceedings of the IEEE 1998

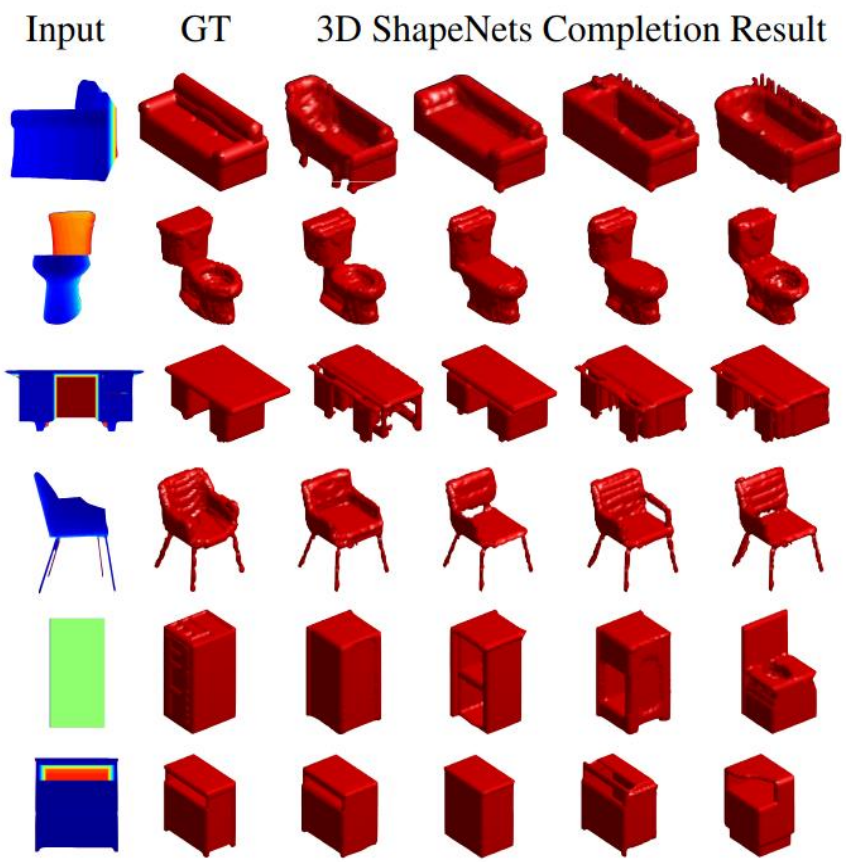
2010s: Deep Learning Meets 3D Vision



Depth Map Prediction from a Single Image
using a Multi-Scale Deep Network. Eigen et al. NeurIPS '14

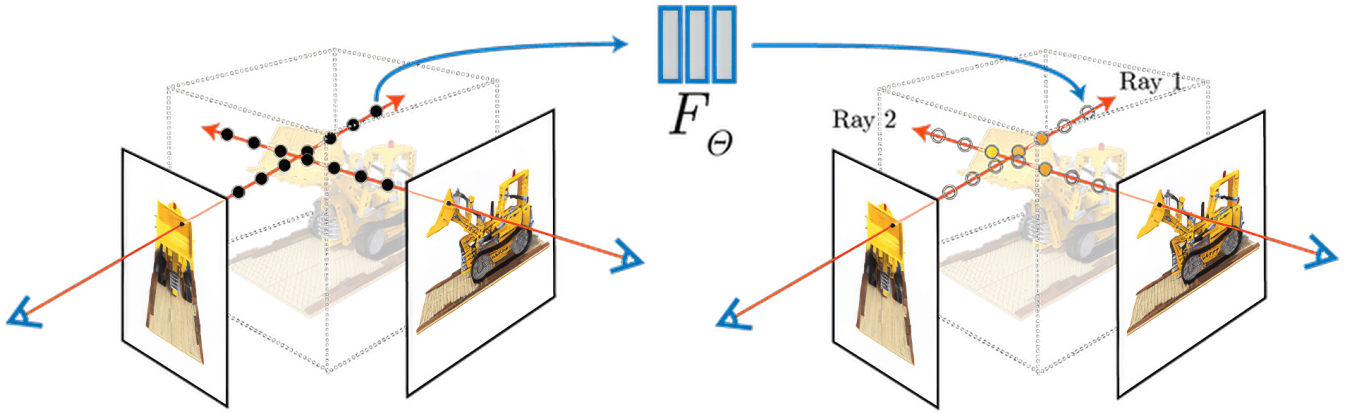
2010s: Deep Learning Meets 3D Vision





3D ShapeNets: A Deep Representation for Volumetric Shapes. Wu et al. CVPR 2015

20s – Now: Neural Representations



NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Mildenhall et al. ECCV 2020

20s-Now: Gaussian Splatting

3D Gaussians

https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

20s – Now: Generative Models

Video from Jon Barron's talk at 3DV 2025

20s – Now: Generative Models

Video from Jon Barron's talk at 3DV 2025. Veo 2: "An advertisement for a shoe made of broccoli romanesco, rotating on a turntable"

20s – Now: Generative Models

Veo 2: "A banana posing in a bodybuilding competition, striking different poses with its floppy little peel-arms"

20s – Now: Generative Models

Veo 2: "A colorized 1950s TV commercial. Left: a dad struggling to pick up a heavy box. Right: that dad wearing a full-body mechanical exoskeleton, lifting the box with ease"

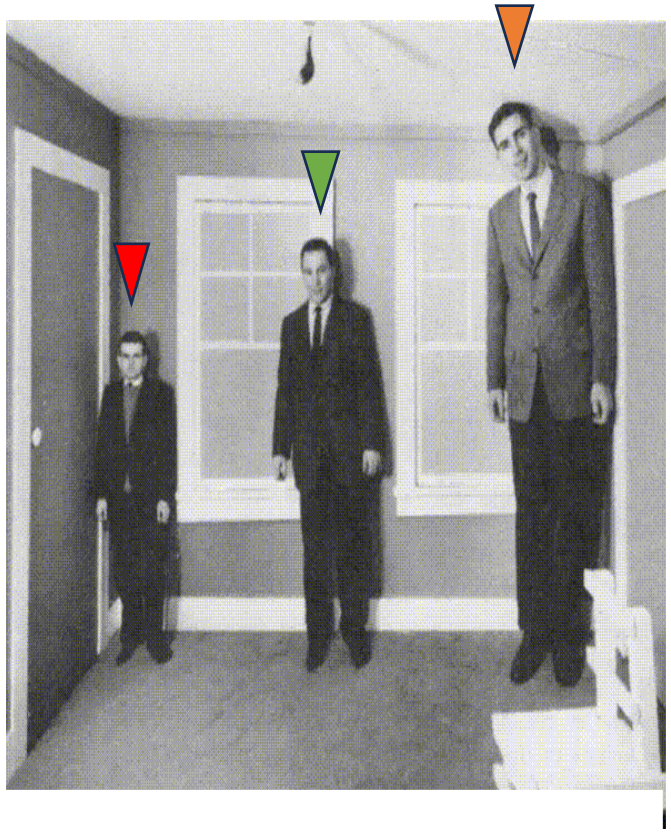
20s – Now: Generative Models

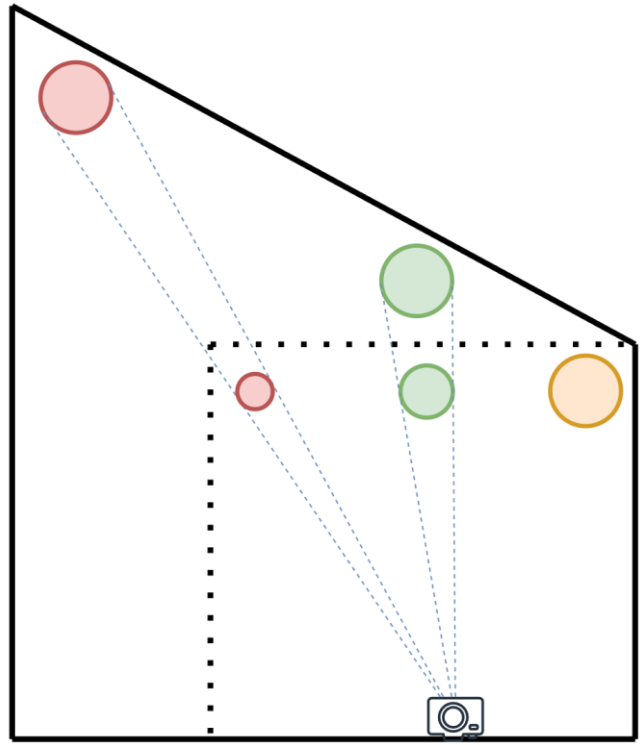
Asking sota Video Generation models to generate a video of a man writing "hi" on a blackboard

Challenges of 3D Computer Vision

Challenges of 3D Computer Vision

Depth ambiguity





Testing sota method “Depth Anything V2” (Yang et al. NeurIPS 2024) on Ames room illusion.
Demo: https://huggingface.co/spaces/depth-anything/Depth-Anything-V2

Challenges of 3D Computer Vision

Image Artifacts

Chromatic aberration

Motion Blur

near-camera reflection (backscatter)

Challenges of 3D Computer Vision

Context

Describe the scene in this picture

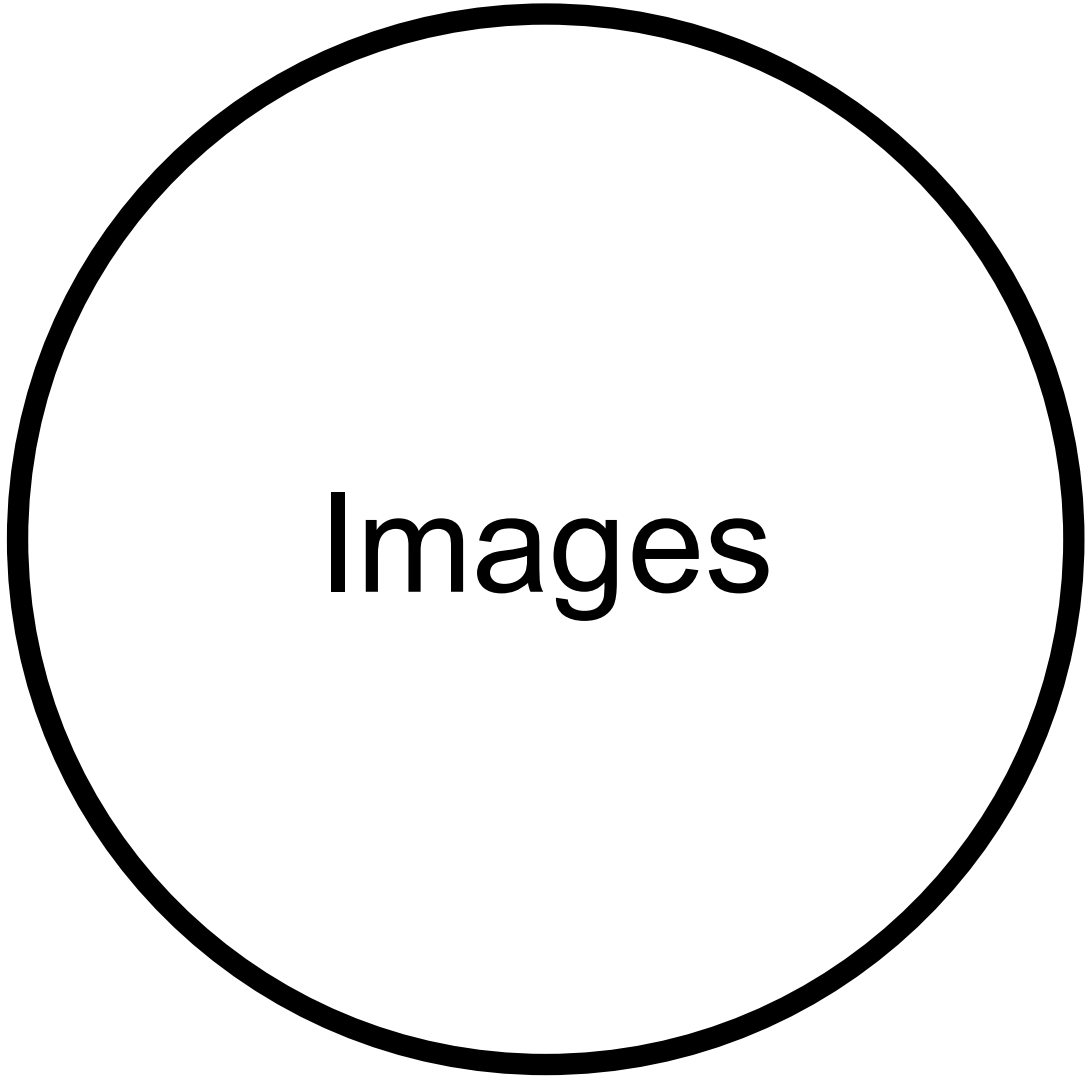
A group of formally dressed individuals, including Barack Obama, are in a hallway with checkered flooring and mirrors, engaging in what appears to be a lighthearted activity involving a scale. The atmosphere is casual and friendly, with smiles and laughter visisble, suggesting an informal or wellness-related event taking place in an institutional setting.

Why are the people laughing? What is so funny?

The people are laughing likely due to a lighthearted moment involving the scale activity, possibly sparked by humor or camaraderie in an informal setting.

Challenges of 3D Computer Vision

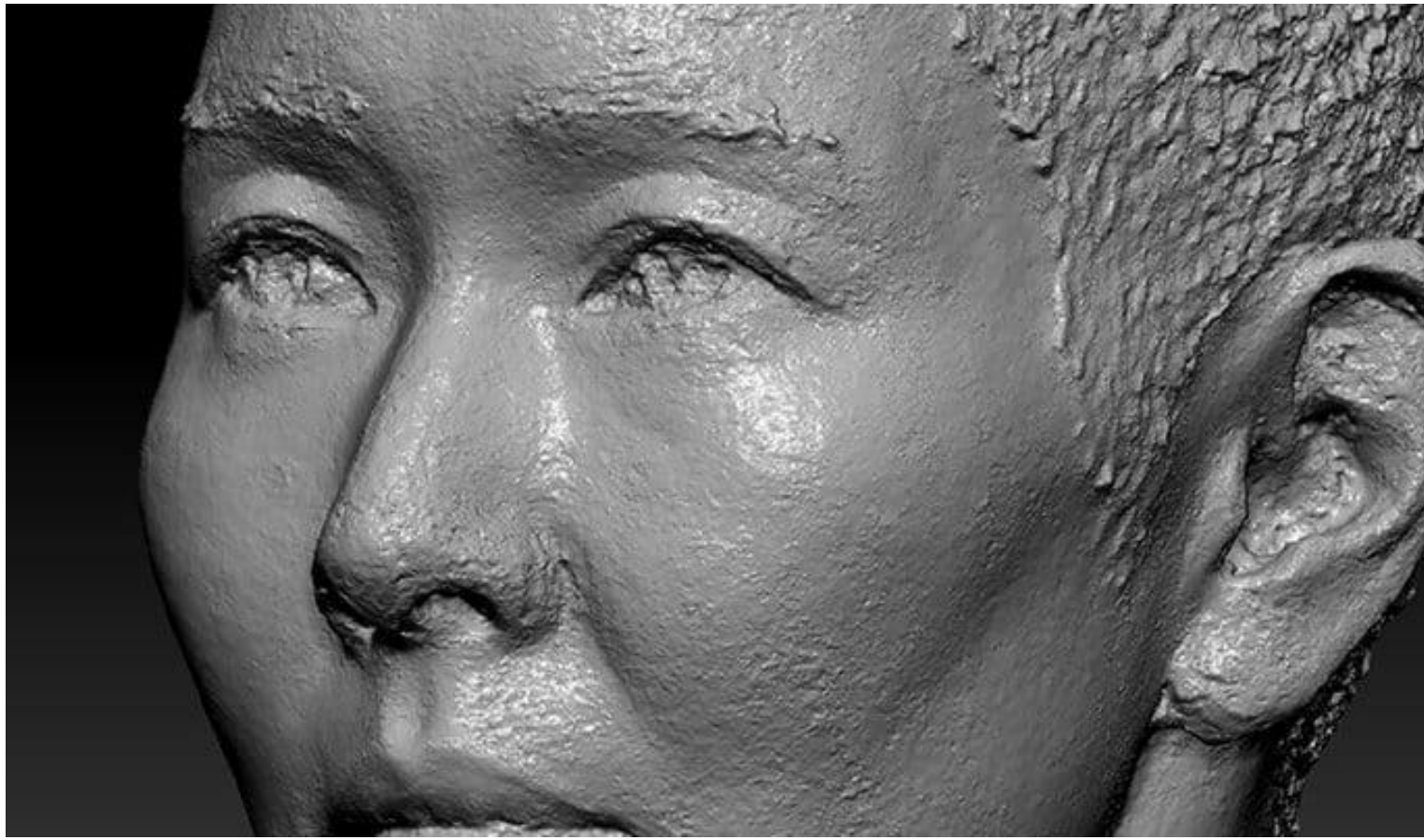
Data



3D Models →

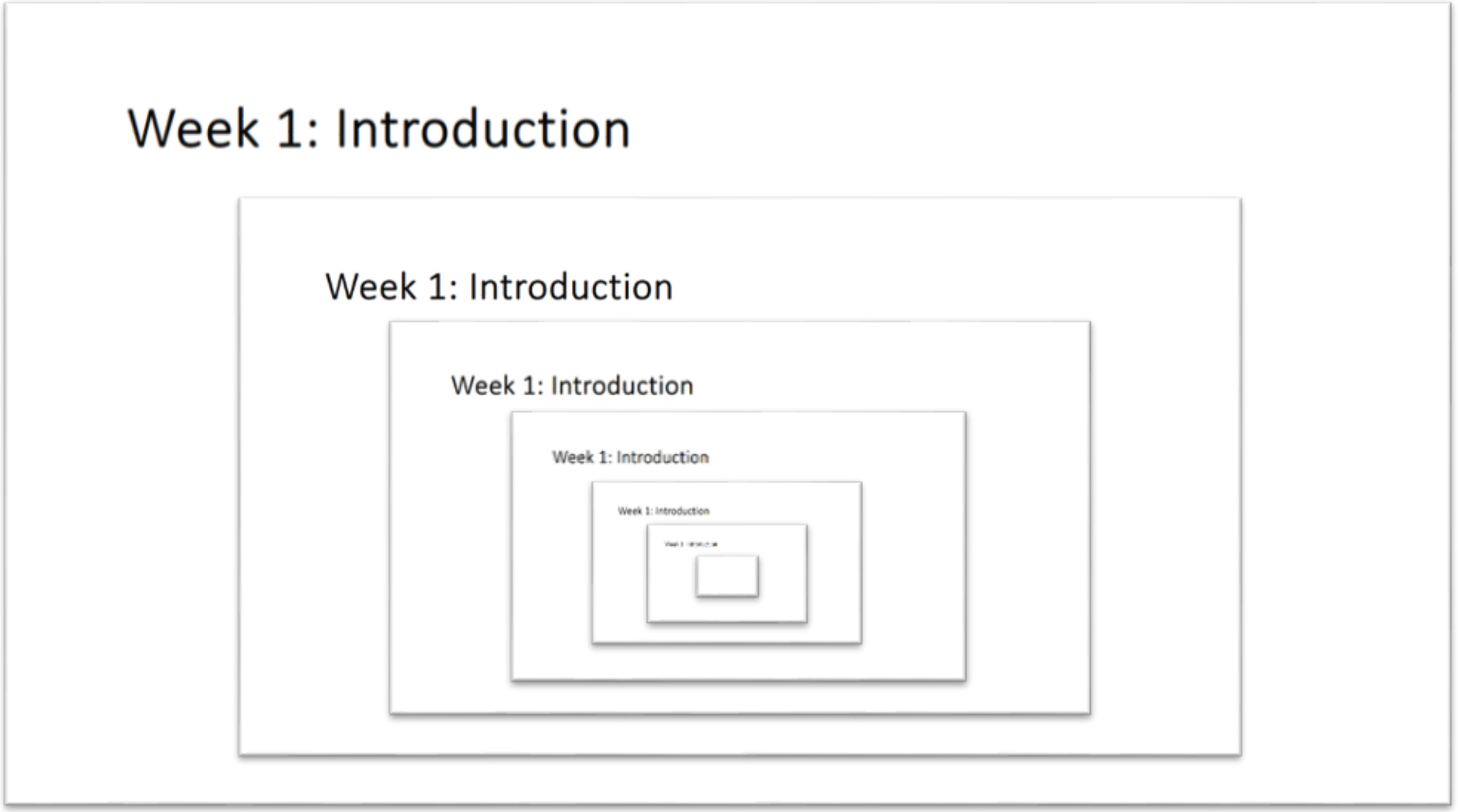
Challenges of 3D Computer Vision

3D Artifacts



Overview of the course

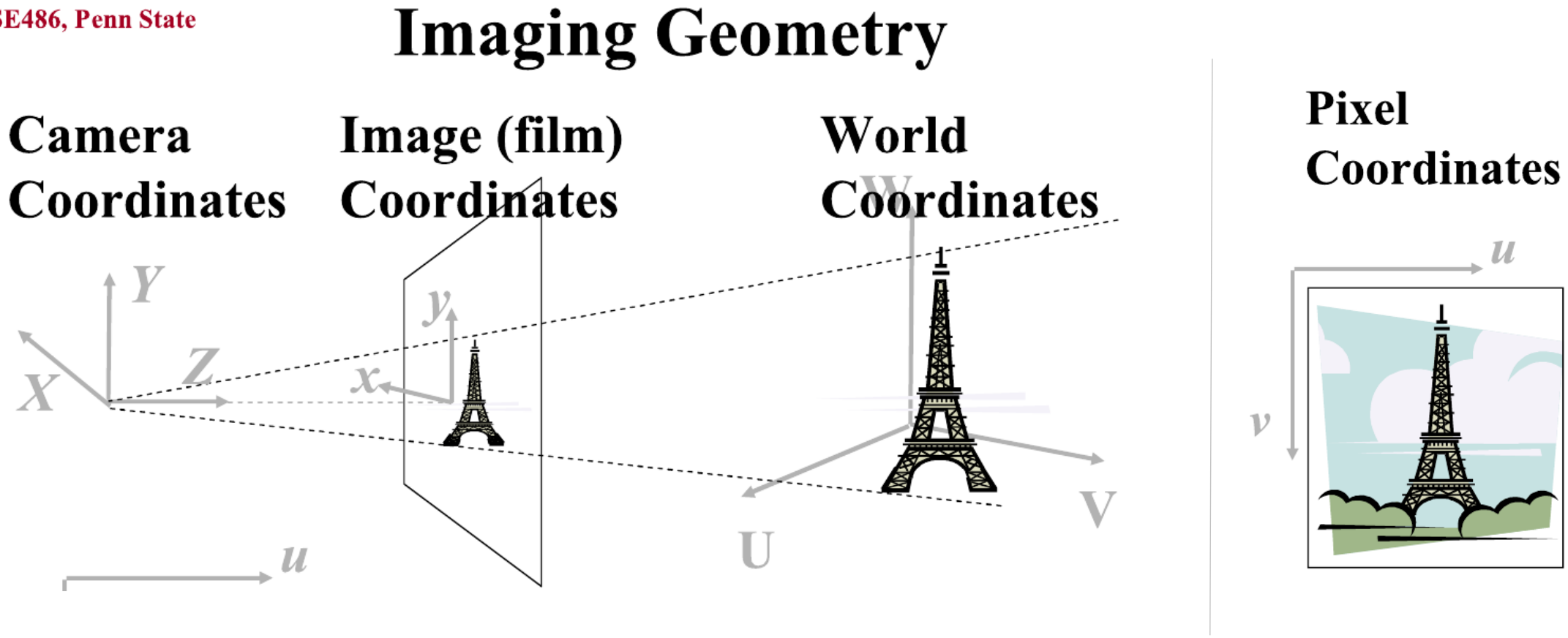
Week 1: Introduction



Week 2: Camera Models and Coordinate Systems

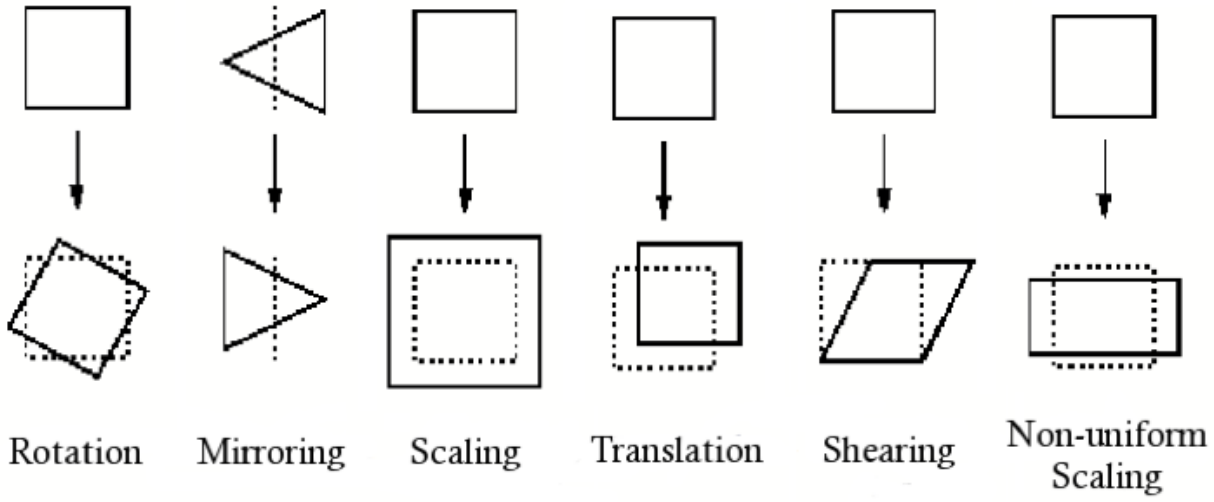
World to Image Transformations

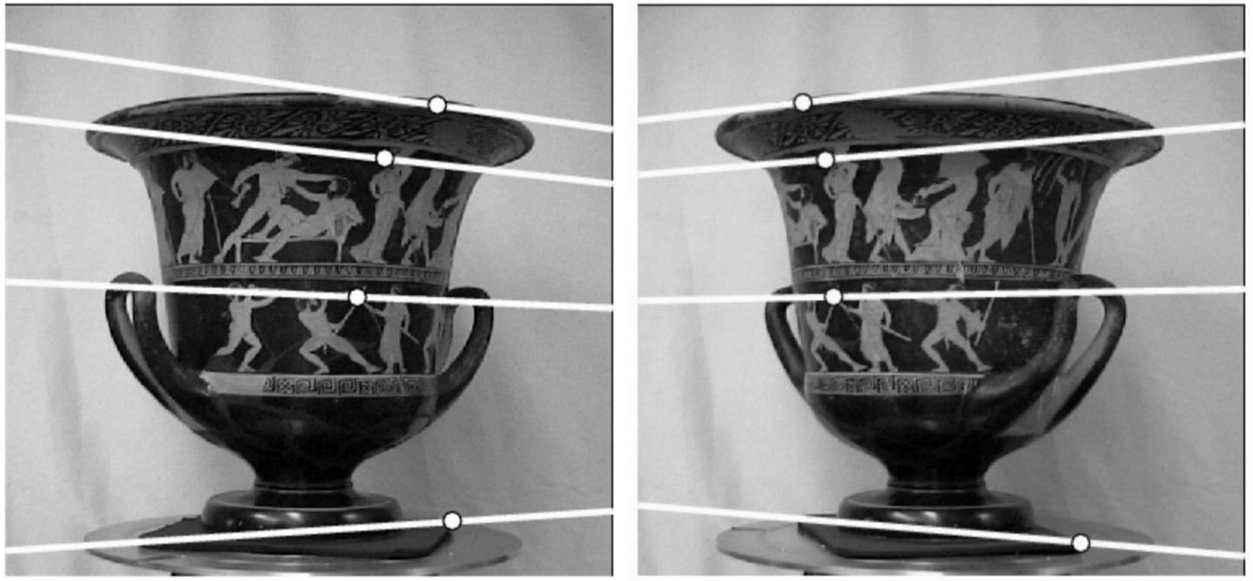
Robert Collins
CSE486, Penn State



Week 2: Camera Models and Coordinate Systems

Transformations and epipolar geometry



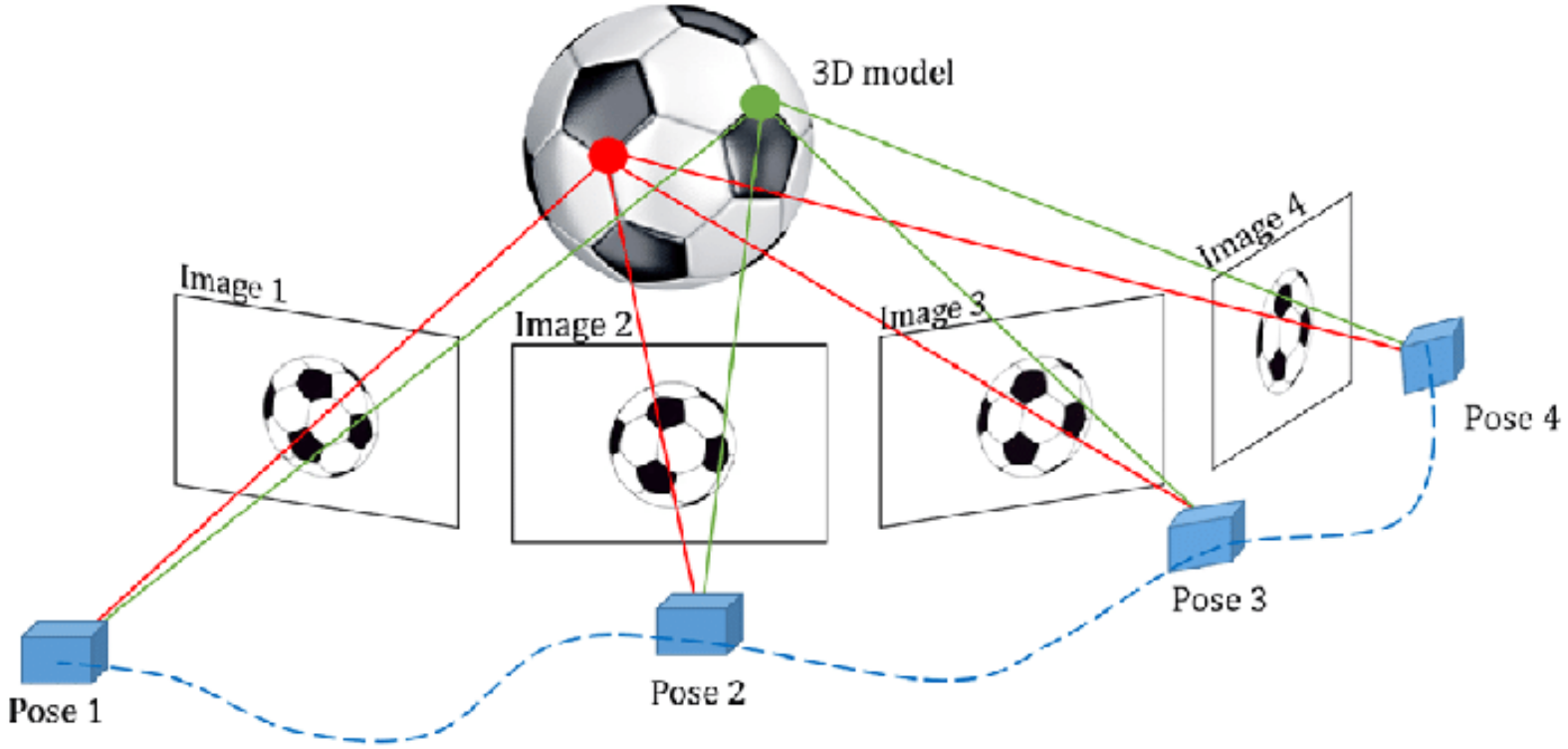


https://cs.brown.edu/courses/cs129/labs/lab_stereolab/

Week 3: Classical 3D Reconstruction

Traditional Multiview Reconstruction

https://live.staticflickr.com/3620/3384374475_725bf46a23_z.jpg



Week 3: Classical 3D Reconstruction

Structure From Motion

Video extracted from [The structure from motion pipeline]

Week 4: From Classical to Modern Stereo Vision and Depth Estimation

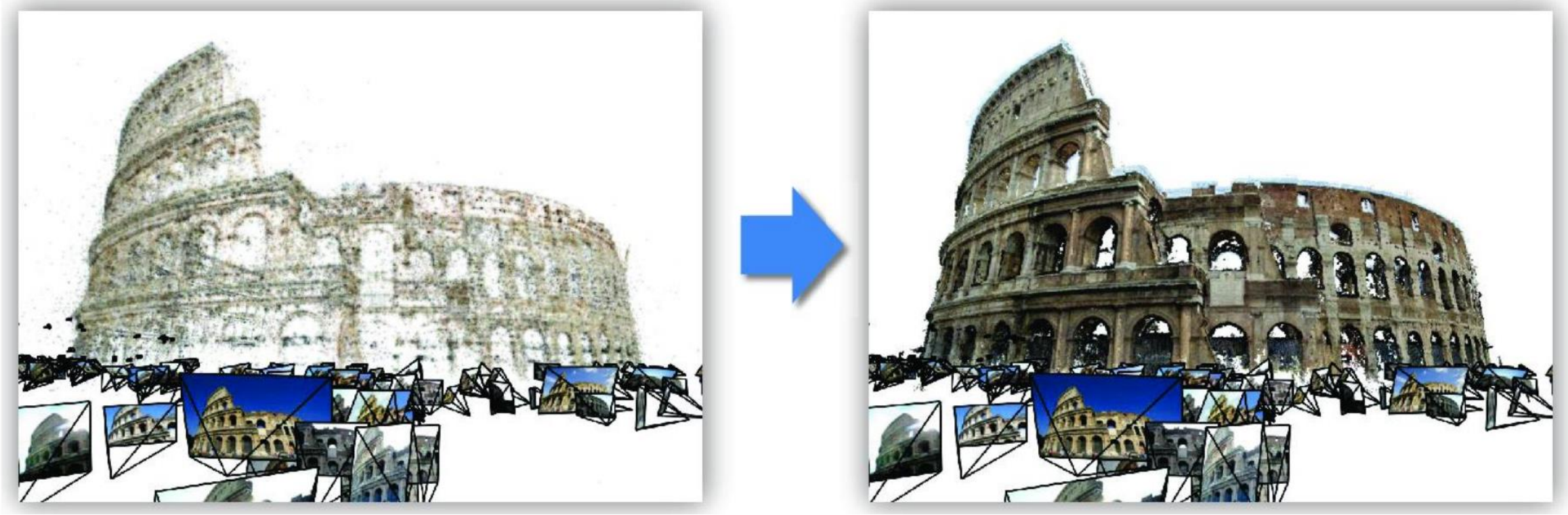
Left Image

Right Image

Depth Map

Week 4: From Classical to Modern Stereo Vision and Depth Estimation

Multi-View Stereo Reconstruction



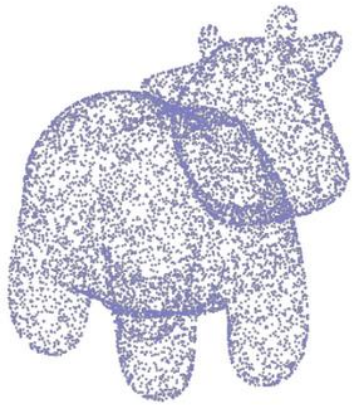
Building Rome in a Day, Agarwal et al. ICCV 2009

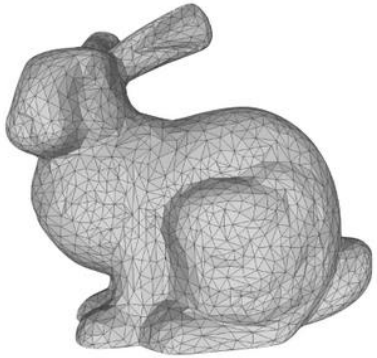
Week 4: From Classical to Modern Stereo Vision and Depth Estimation

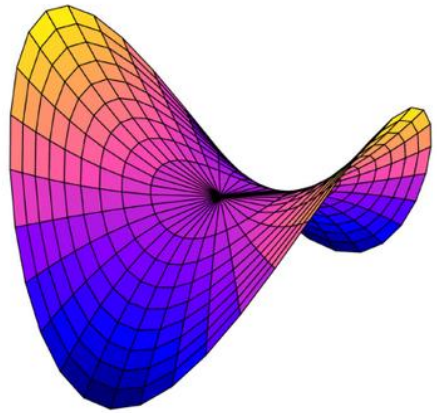
Monocular Depth Estimation

Depth Anything V2, Yang et al. Demo

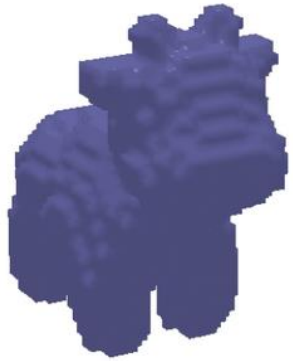
Week 5: Surface Reconstruction and Procrustes Alignment

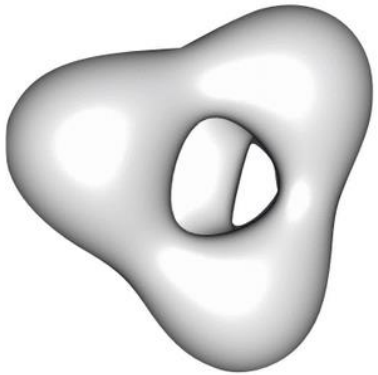






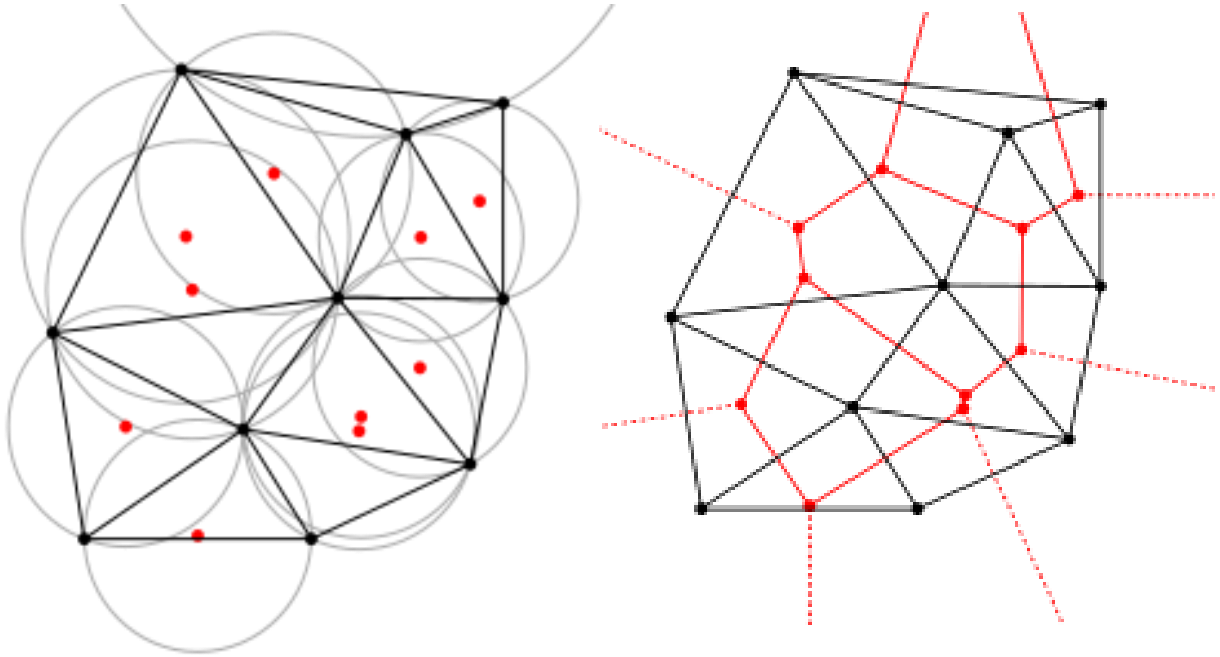
Boundary Representations for Shapes



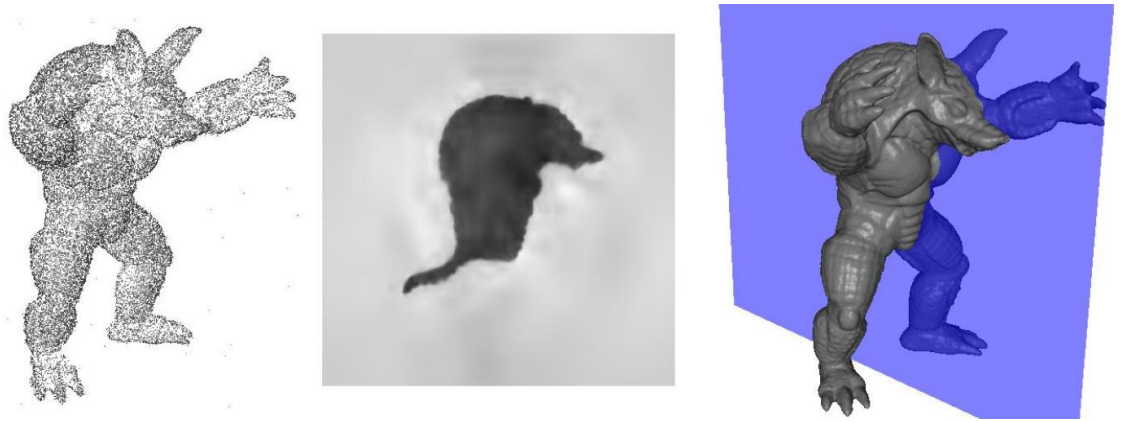


Volumetric Representations for Shapes

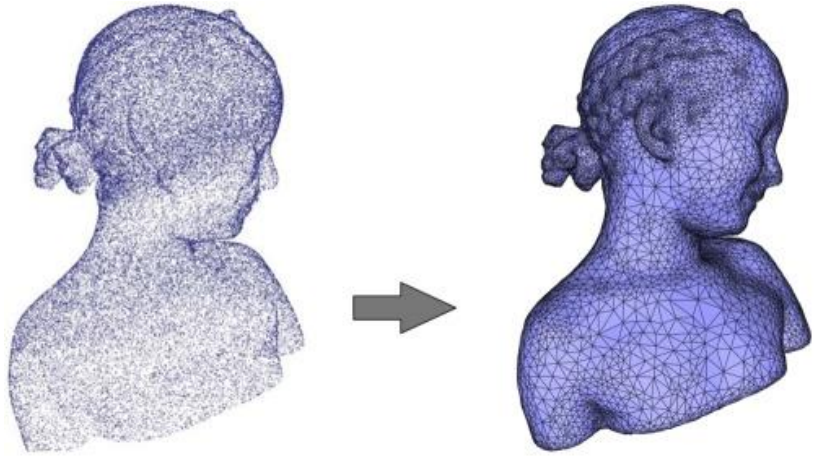
Week 5: Surface Reconstruction and Procrustes Alignment



https://en.wikipedia.org/wiki/Delaunay_triangulation

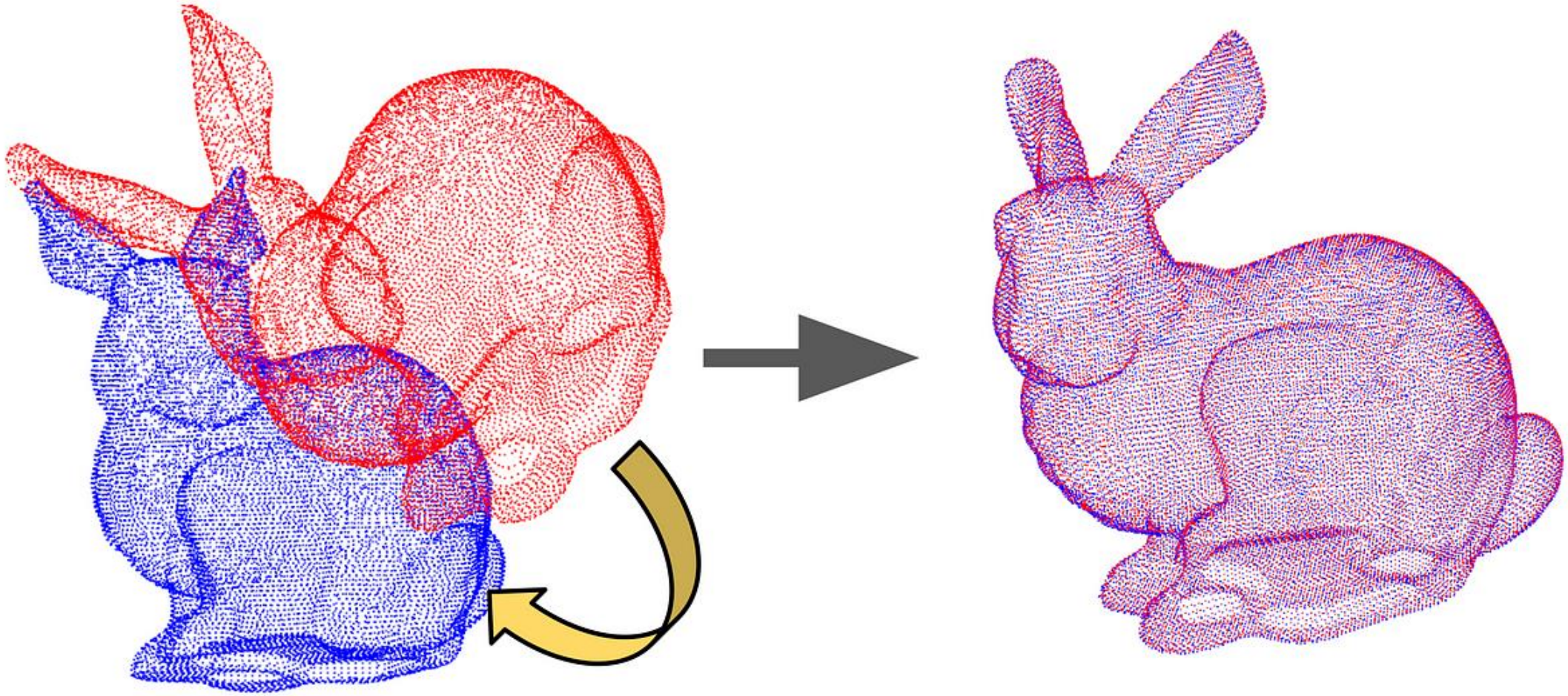


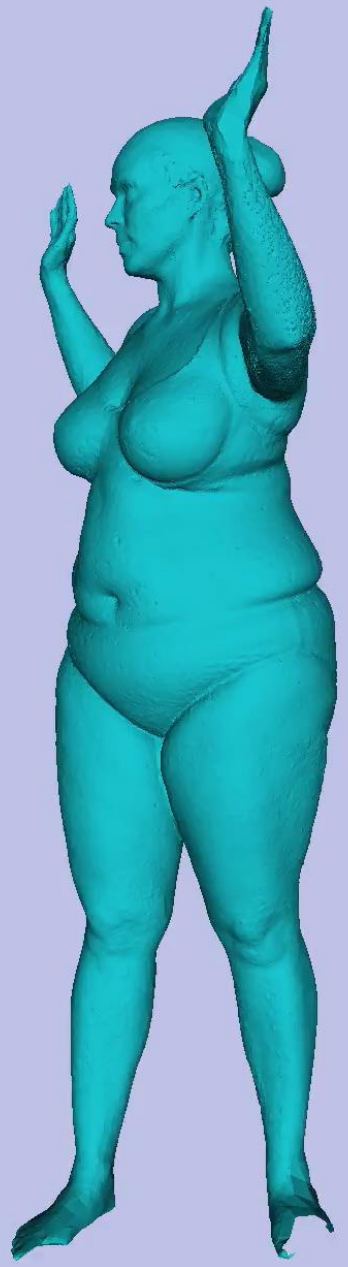
Poisson Surface Reconstruction, Kazhdan et al. SGP '06



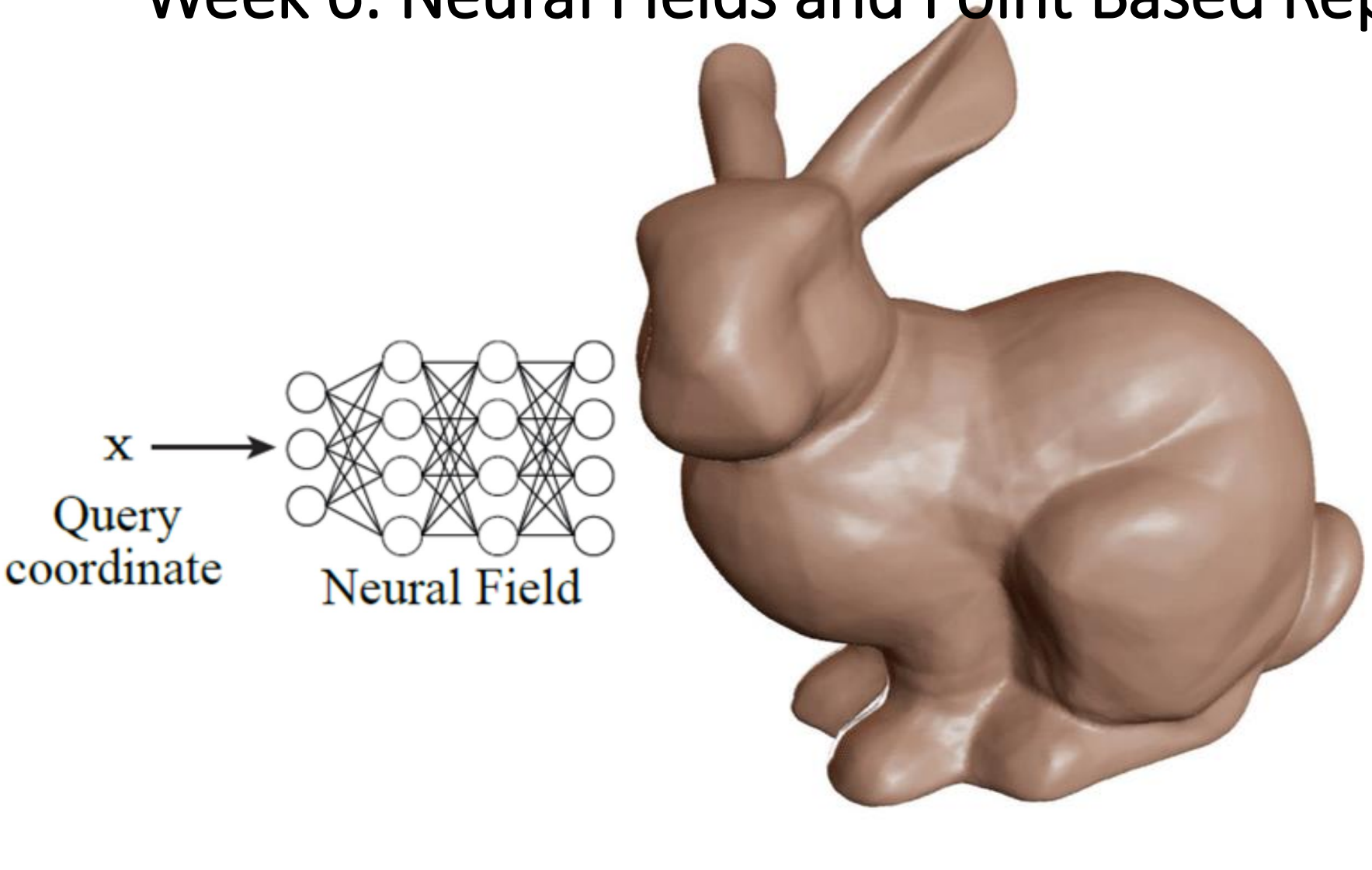
https://doc.cgal.org/latest/Poisson_surface_reconstruction_3/bimba.jpg

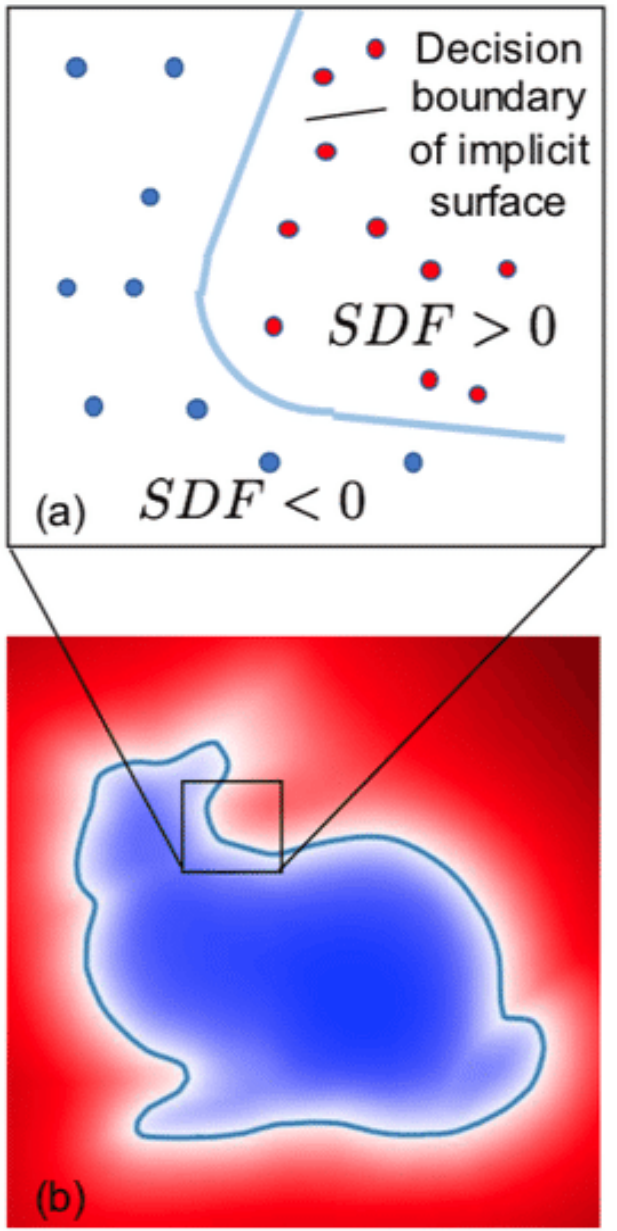
Week 5: Surface Reconstruction and Procrustes Alignment





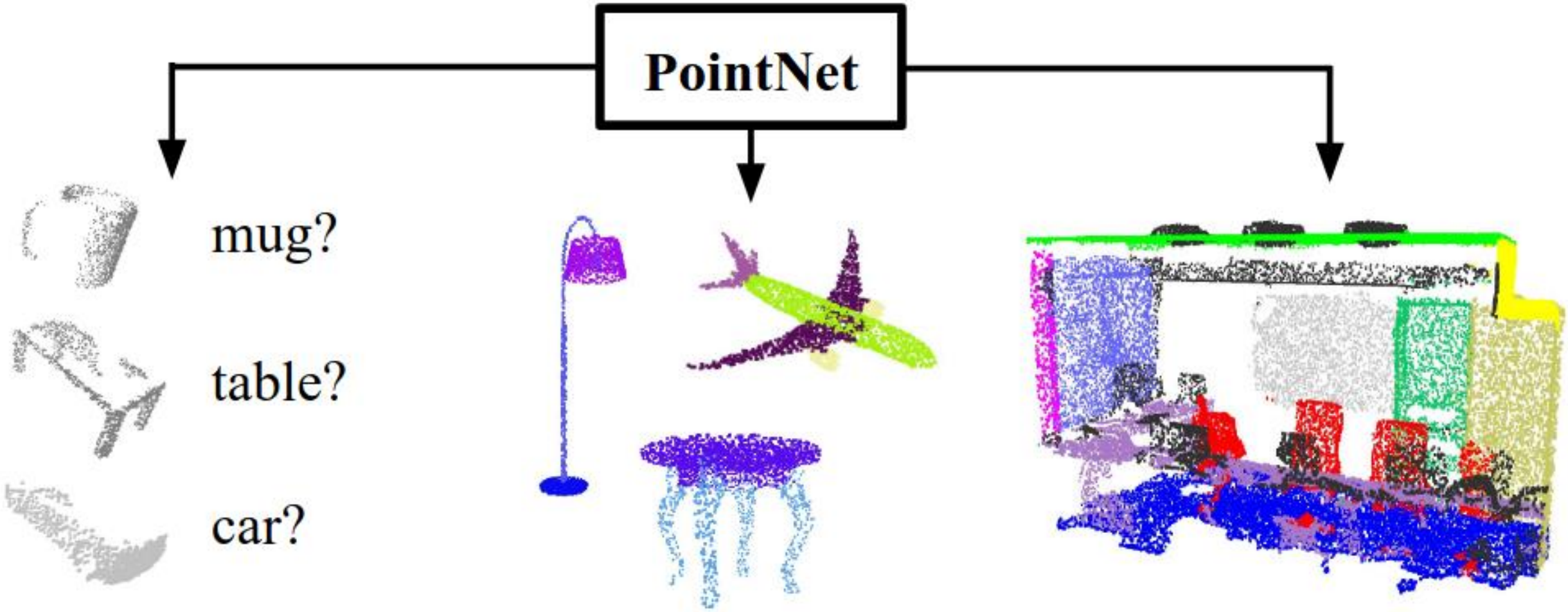
Week 6: Neural Fields and Point Based Repr





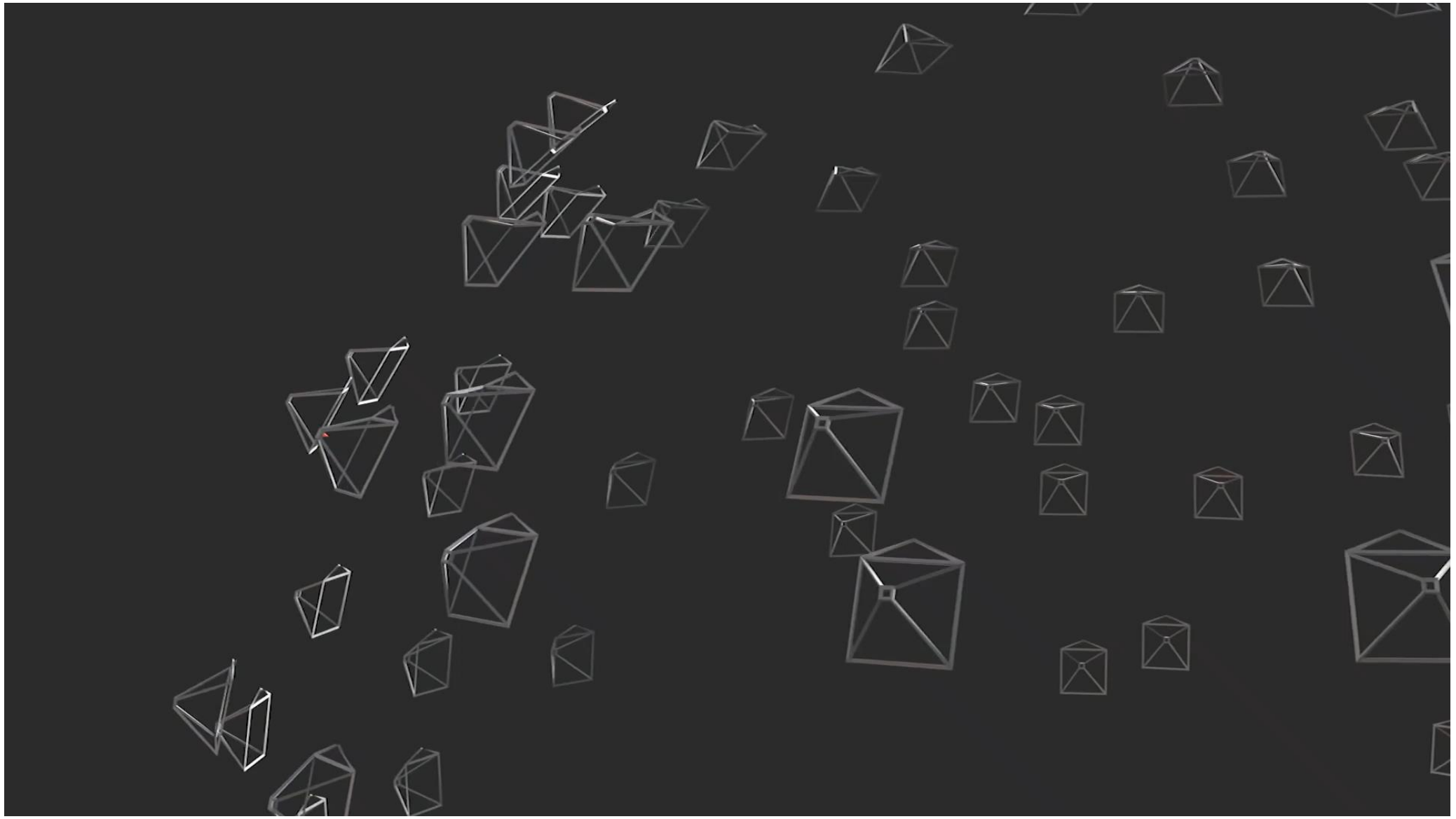
DeepSDF: Learning Continuous Signed Distance Functionscfor Shape Representation. Park et al. CVPR 2019

Week 6: Neural Fields and Point Based Representations



PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. Qi et al. CVPR 2017

Week 7: Neural Radiance Fields

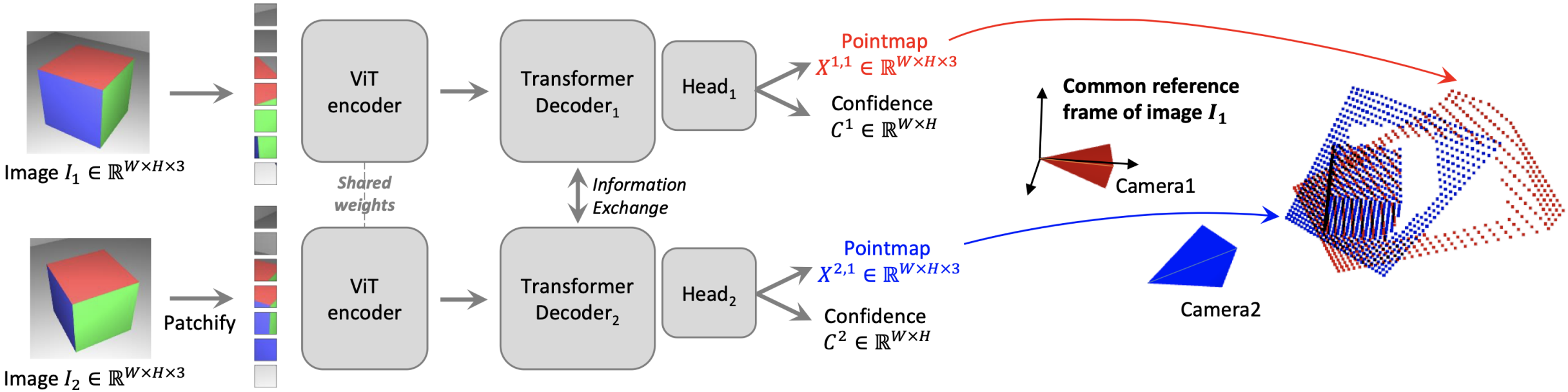


NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. Mildenhall et al. ECV 2020

Week 8: Gaussian Splatting and Point Clouds

3D Gaussian Splatting for Real-Time Radiance Field Rendering, Kerbl et al. SIGGRAPH 2023

Week 9: Advanced Methods in Learning-Based Reconstruction



Dust3R CVPR2024

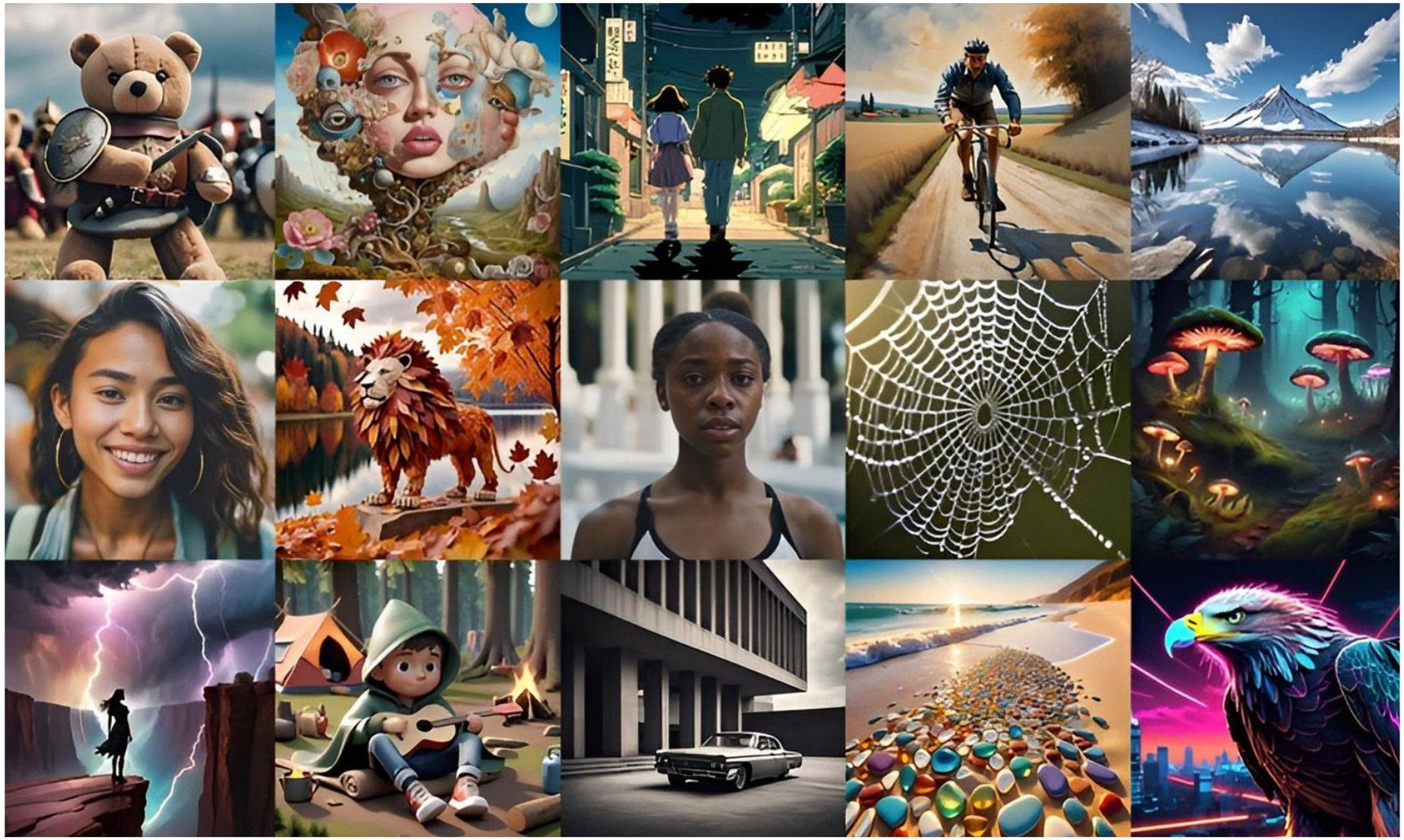
DUSt3R:
Geometric 3D Vision Made Easy

S. Wang [1], V. Leroy [2], Y. Cabon [2], B. Chidlovskii [2] and J. Revaud [2]

[1] Aalto University

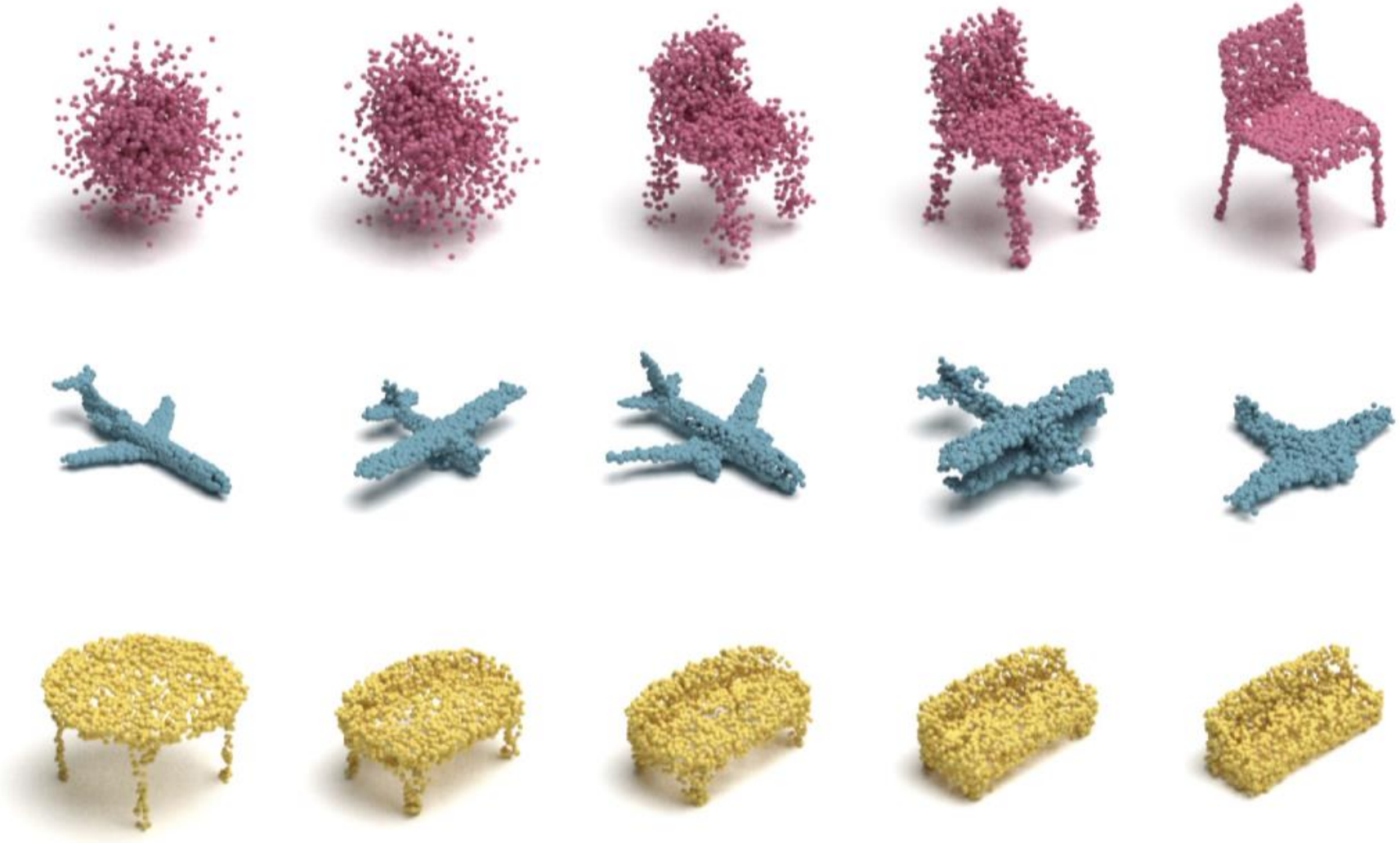
2 Naver Labs Europe

Week 10: Generative Models



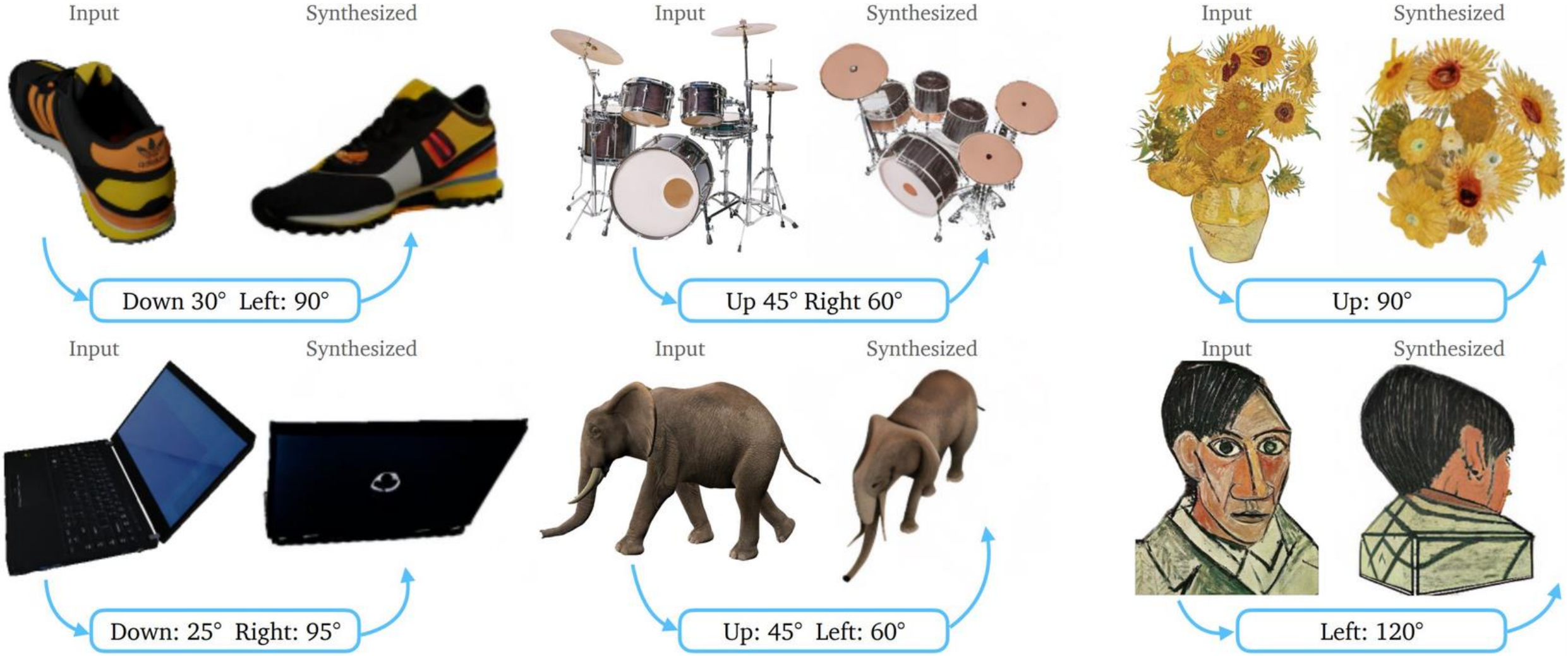
https://www.awwwards.com/inspiration/stability-ai-1

Week 10: Generative Models



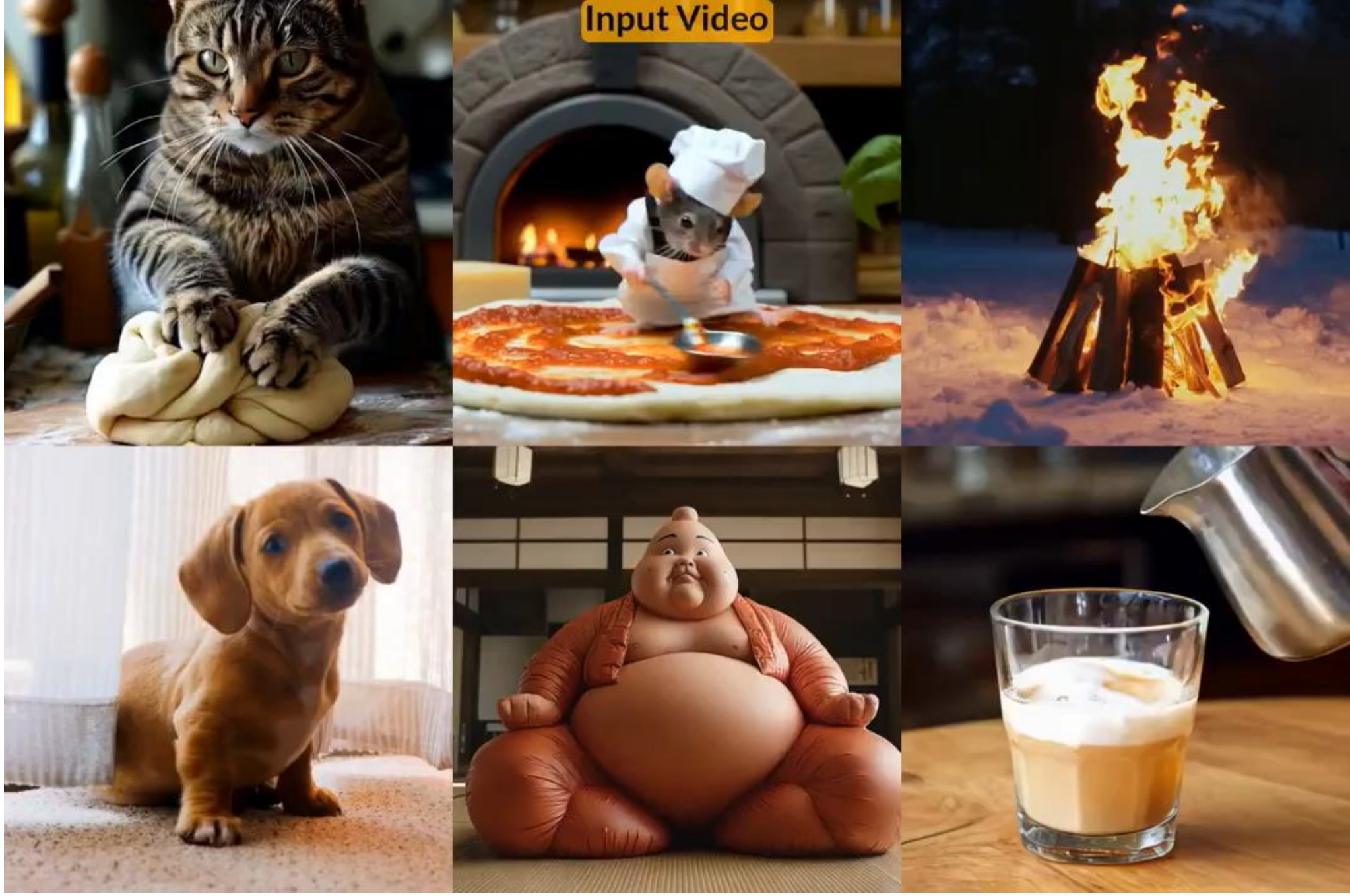
Diffusion Probabilistic Models for 3D Point Cloud Generation. Luo et al. CVPR '21

Week 10: Generative Models



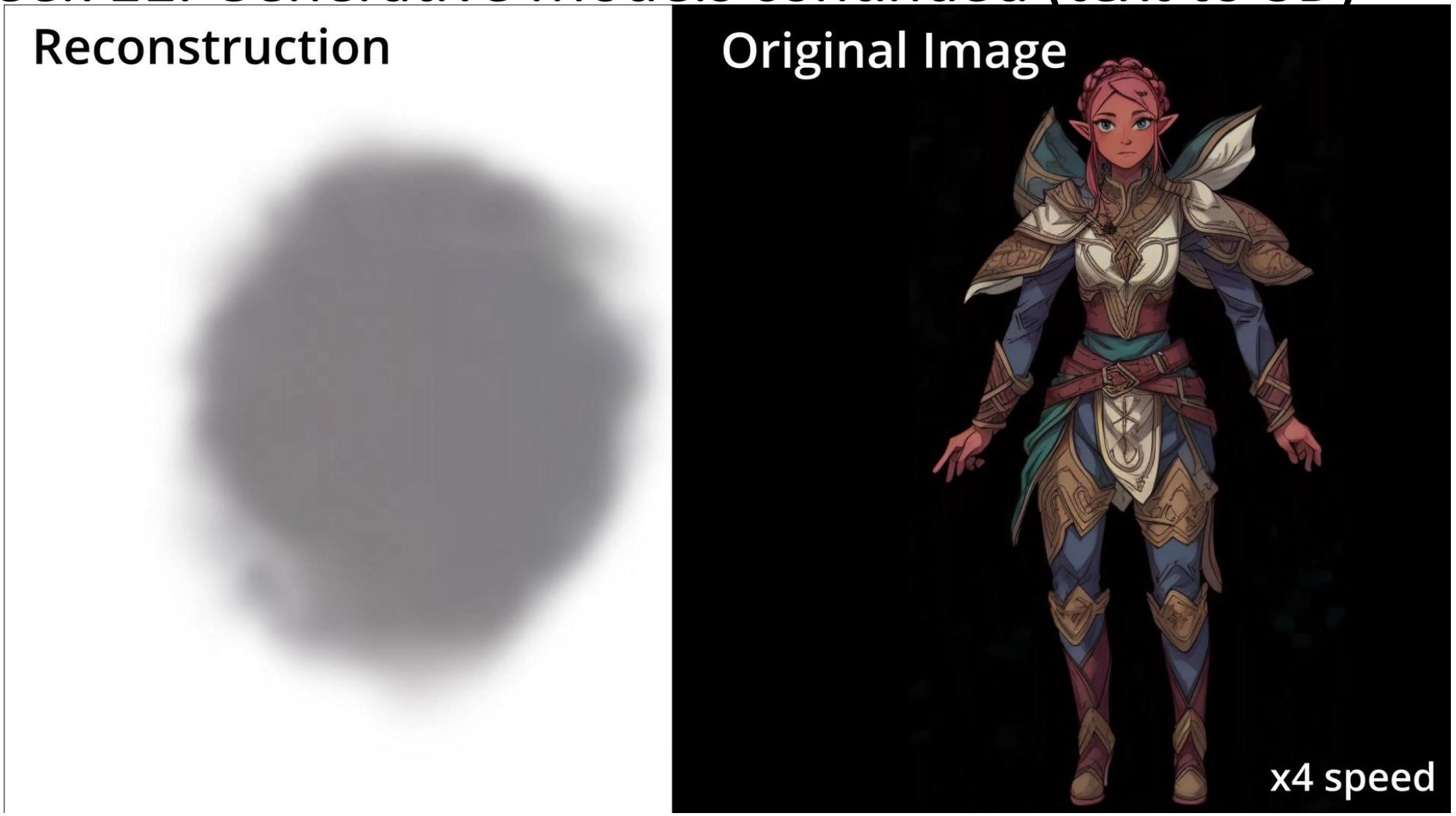
Zero-1-to-3: Zero-shot One Image to 3D Object, Liu et al. ICCV 2023

Week 10: Generative Models



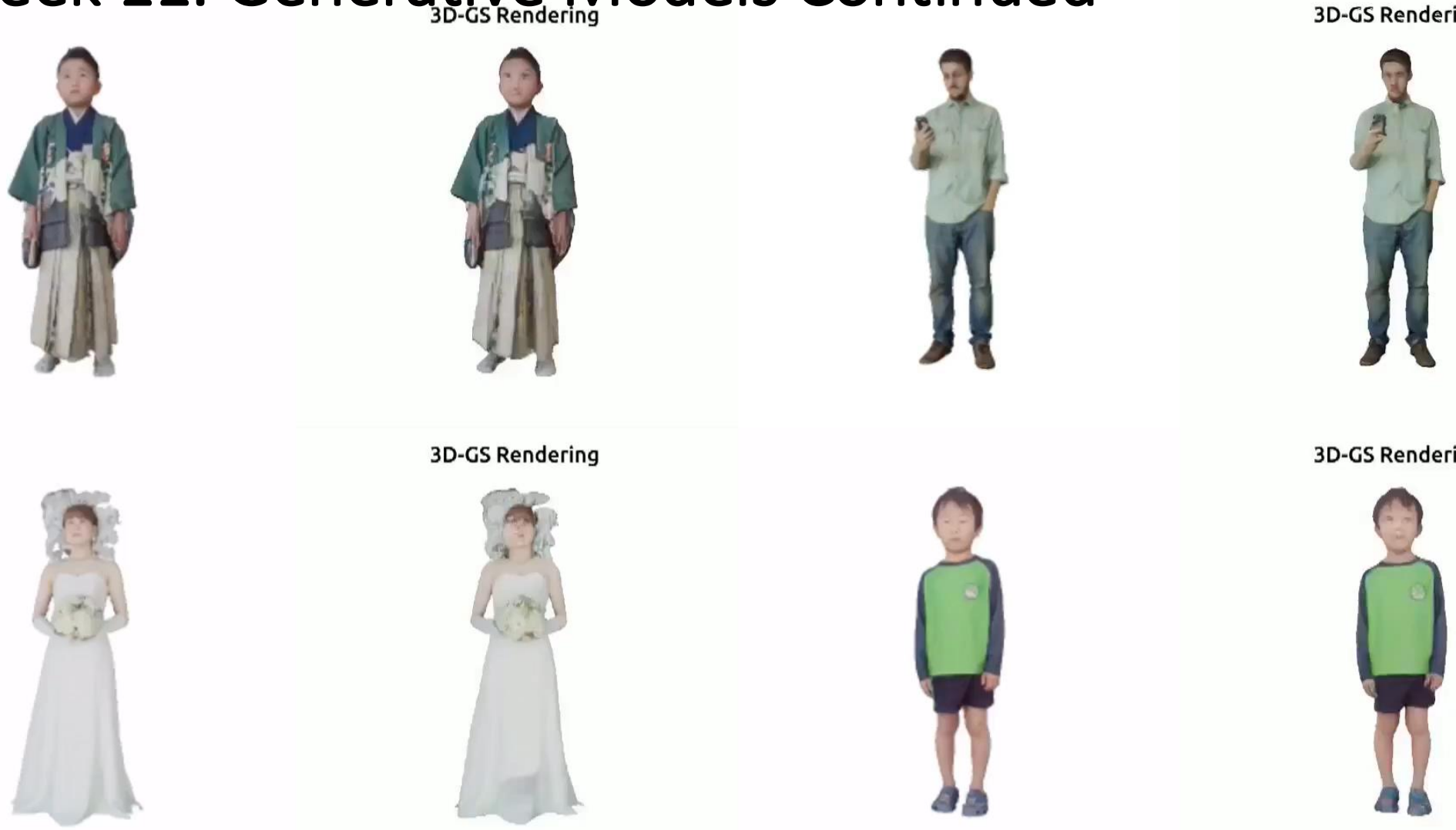
CAT4D: Create Anything in 4D with Multi-View Video Diffusion Models. Wu et al.

Week 11: Generative Models continued (text to 3D)



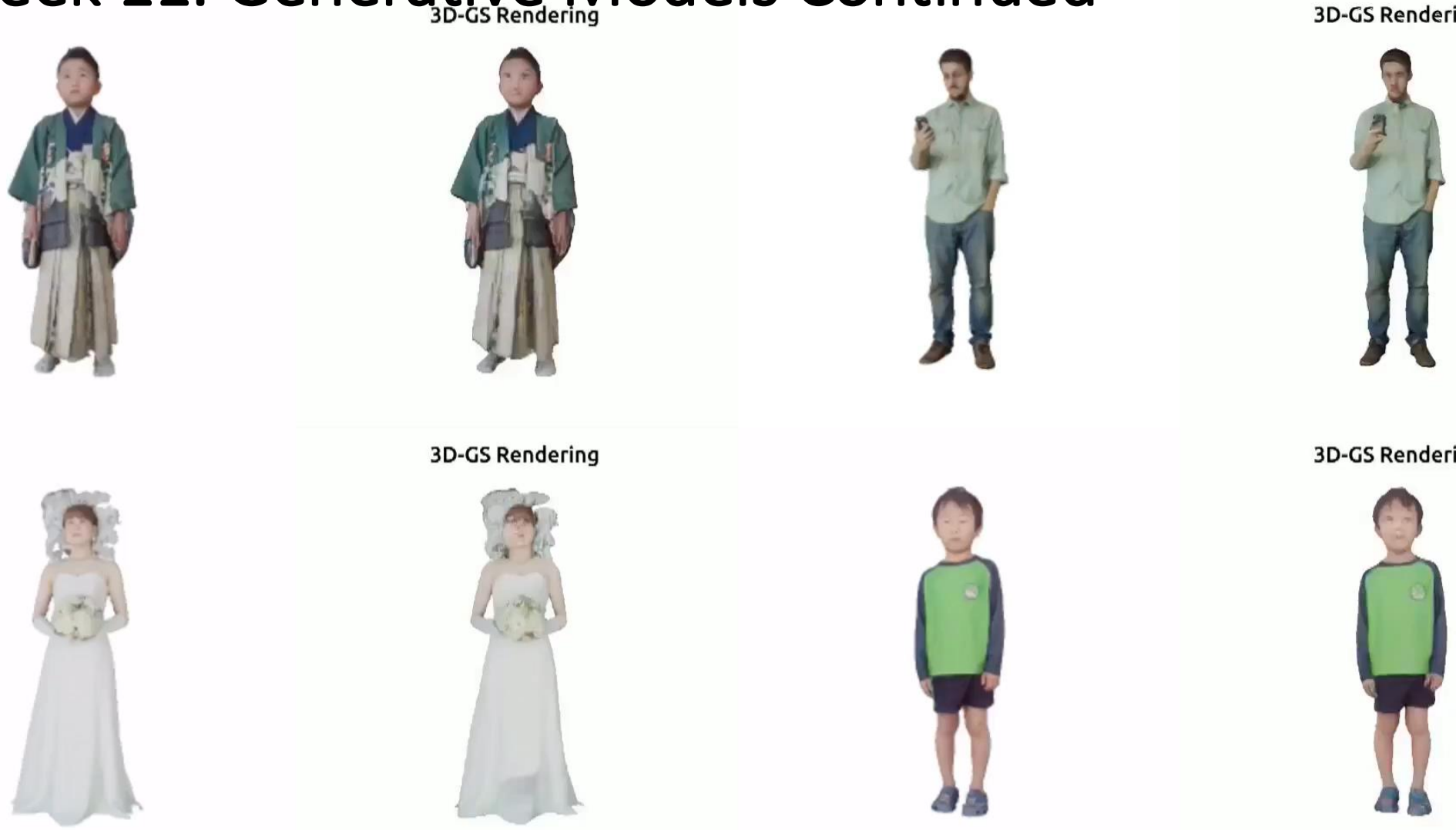
DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. Tang et al. ICLR 2024

Week 11: Generative Models Continued



Human 3Diffusion: Realistic Avatar Creation via Explicit 3D Consistent Diffusion Models. Xue at al. NeurIPS 2024

Week 11: Generative Models Continued



Human 3Diffusion: Realistic Avatar Creation via Explicit 3D Consistent Diffusion Models. Xue at al. NeurIPS 2024

Xue et al, InfiniHuman. SiggraphAsia’25

Xue et al, InfiniHuman. SiggraphAsia’25

Xue et al, InfiniHuman. SiggraphAsia’25

Learning for 3D Vision - Course Objectives

Learning for 3D Vision - Course Objectives

Understand Classical and Modern papers

Implement papers

Develop new ideas

Textbooks

- **Multiple View Geometry in Computer Vision, 2nd ed.**, Hartley & Zisserman
- **Deep Learning**, Godfellow, Bengio, Courville
- **Computer Vision: Algorithms and Applications, 2nd ed.,** Szeliski
- **Pattern Recognition and Machine Learning,** Bishop
- **Mathematics for machine learning,** Faisal & Delsenroth

Feedback

Feedback is always appreciated

Questions?

HiWi / Thesis / Research Project at Continuous Learning on Multimodal Data Streams Chair

- We offer thesis and research projects in:
  - 3D scene representation
  - Human motion modelling
  - Humans and clothing
  - 3D human pose estimation and tracking using deep learning
- Feel free to directly contact TAs, PhD students, PostDocs or at gpmintern@listserv.uni-tuebingen.de