MVS & Neural Radiance Fields

Video from the original ECCV'20 paper

CS180/280A: Intro to Computer Vision and Computational Photography

Angjoo Kanazawa and Alexei Efros

UC Berkeley Fall 2023

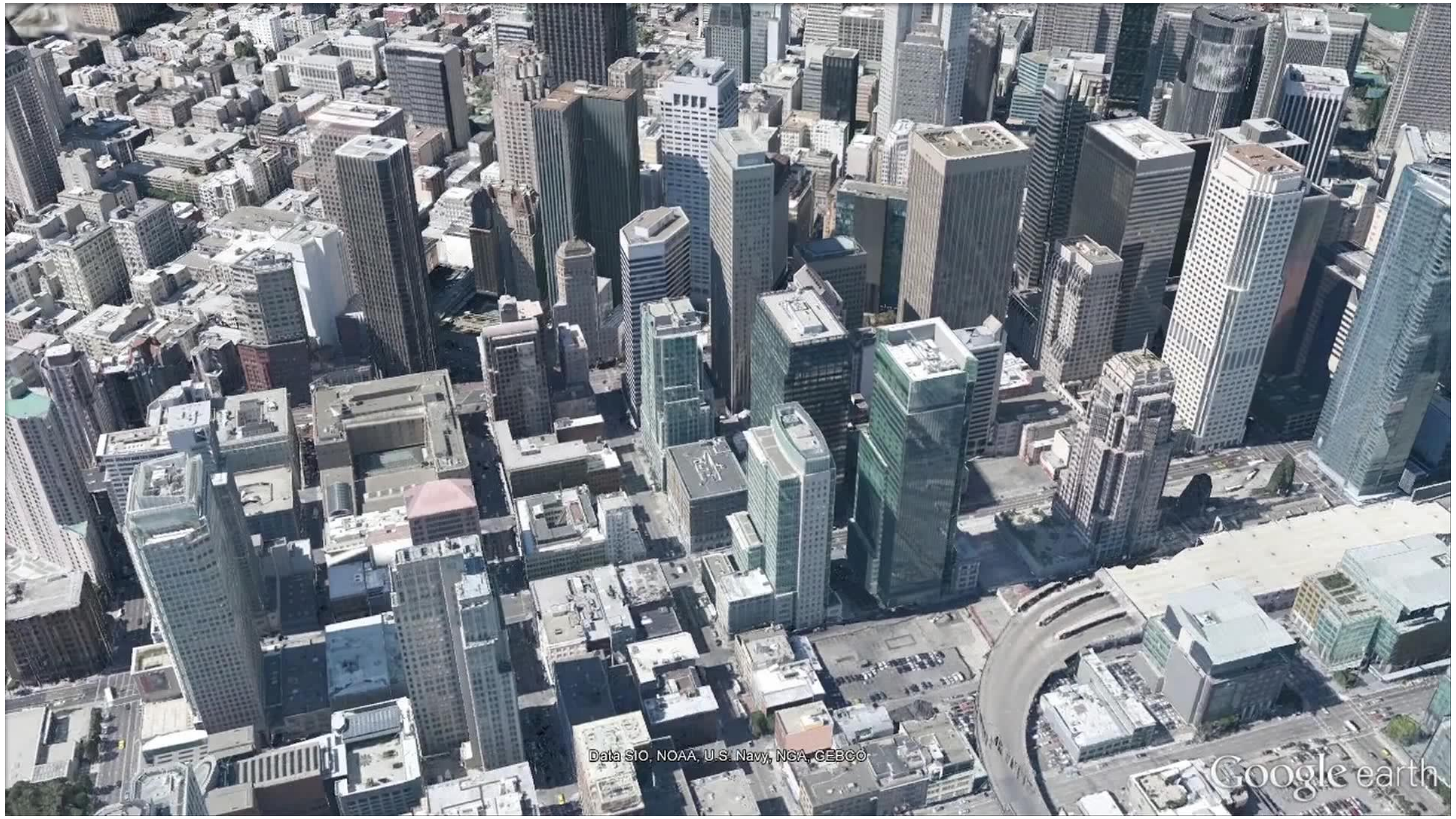
Lots of content from Noah Snavely and [ECCV 2022 Tutorial on Neural Volumetric Rendering for Computer Vision]

Logistics

- Project 4 due tonight! Good luck!

Multi-View Stereo

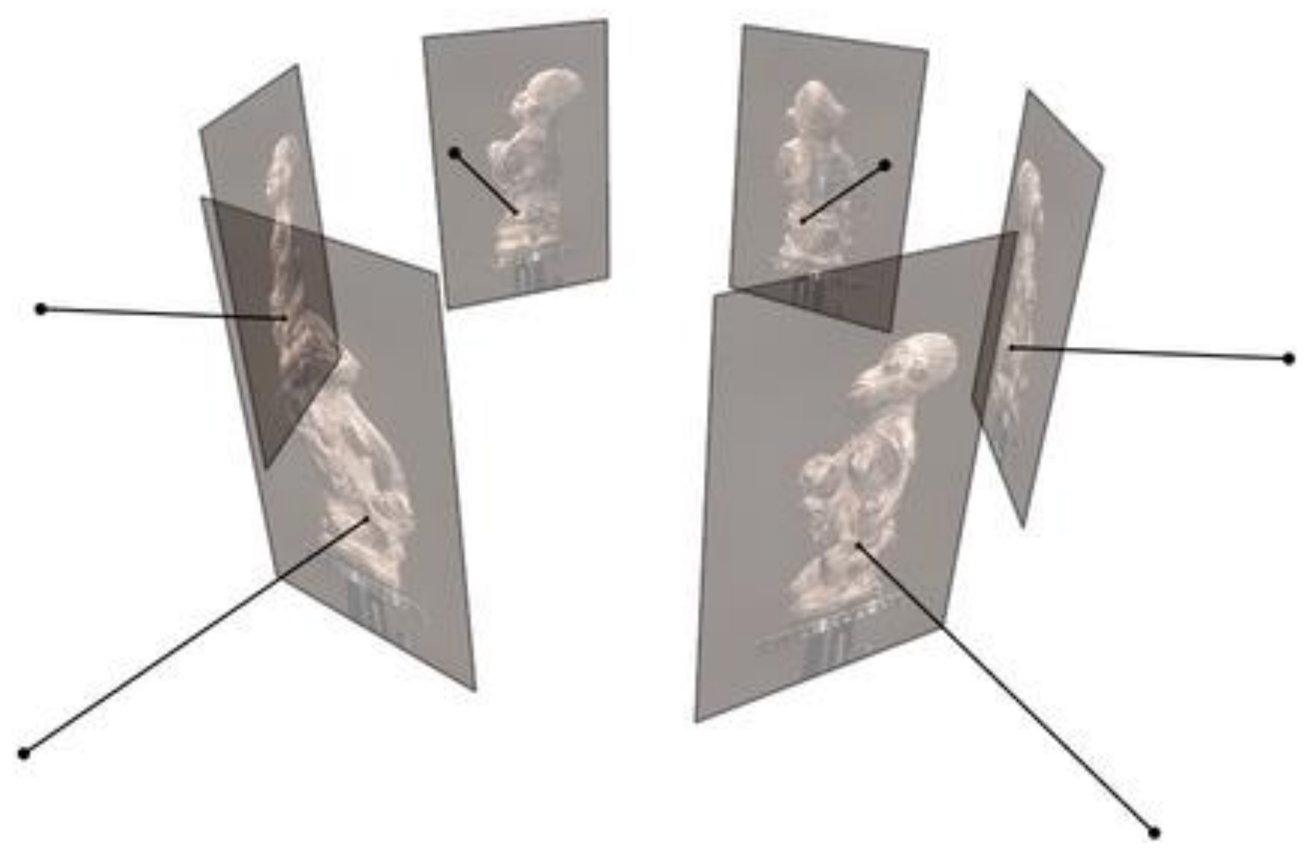
What if we want solid models?



Slide credit: Noah Snavely

Multi-view Stereo (Lots of calibrated images)

- Input: calibrated images from several viewpoints (known camera: intrinsics and extrinsics)
- Output: 3D Model

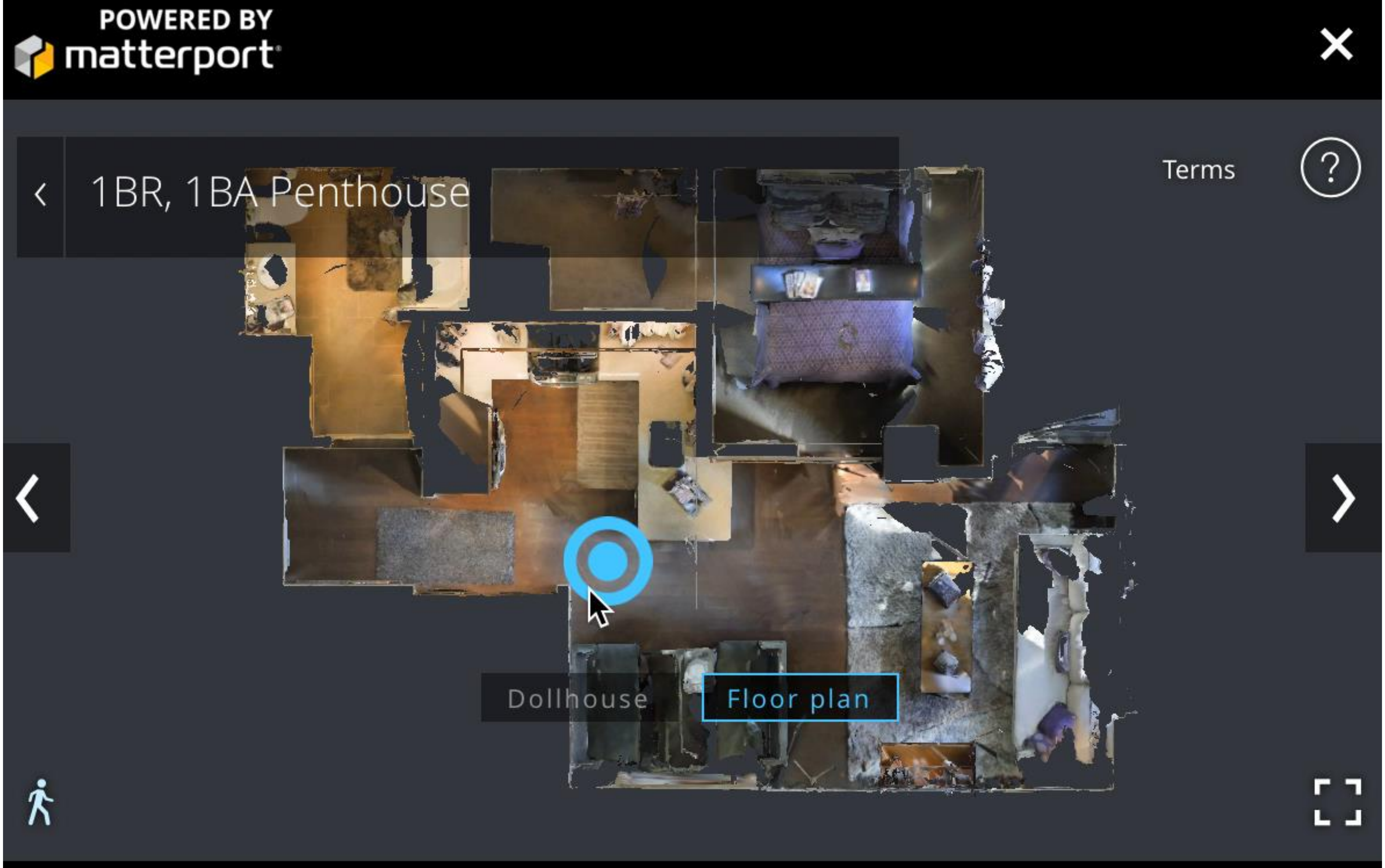


Figures by Carlos Hernandez

Slide credit: Noah Snavely

In general, conducted in a controlled environment with multi-camera setup that are all calibrated

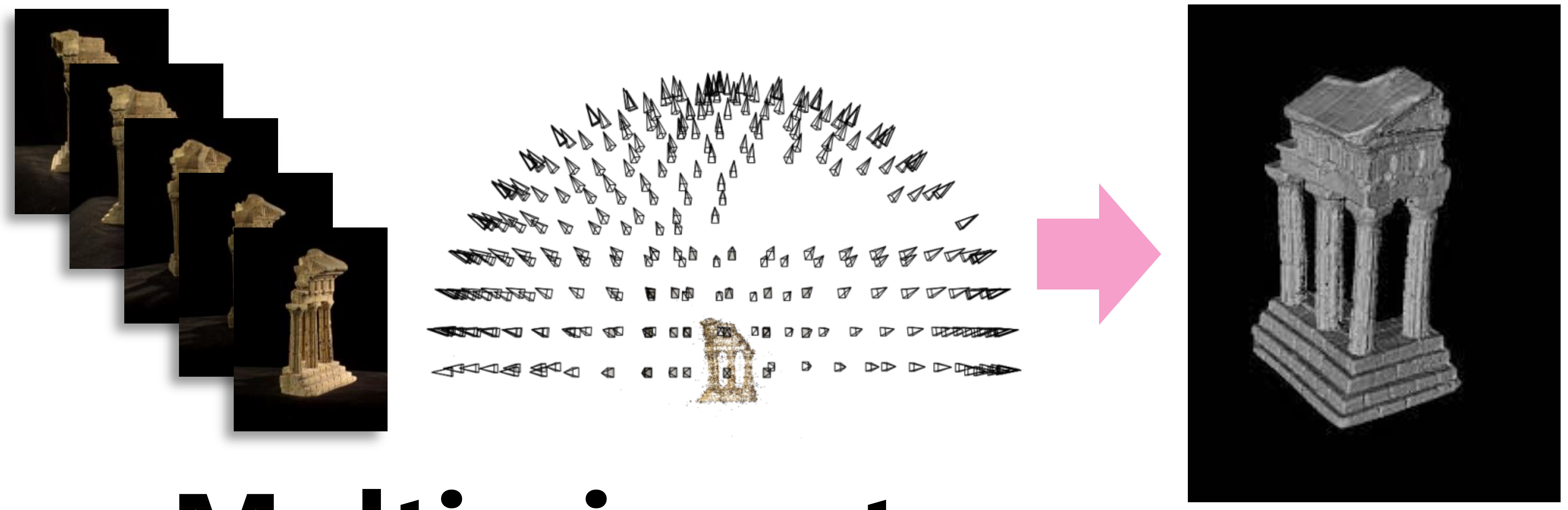
Slide credit: Noah Snavely


POWERED BY
matterport
1BR, 1BA Penthouse
Terms
Dollhouse
Floor plan

Multi-view Stereo

Problem formulation: given several images of the same object or scene, compute a representation of its 3D shape

Binocular Stereo



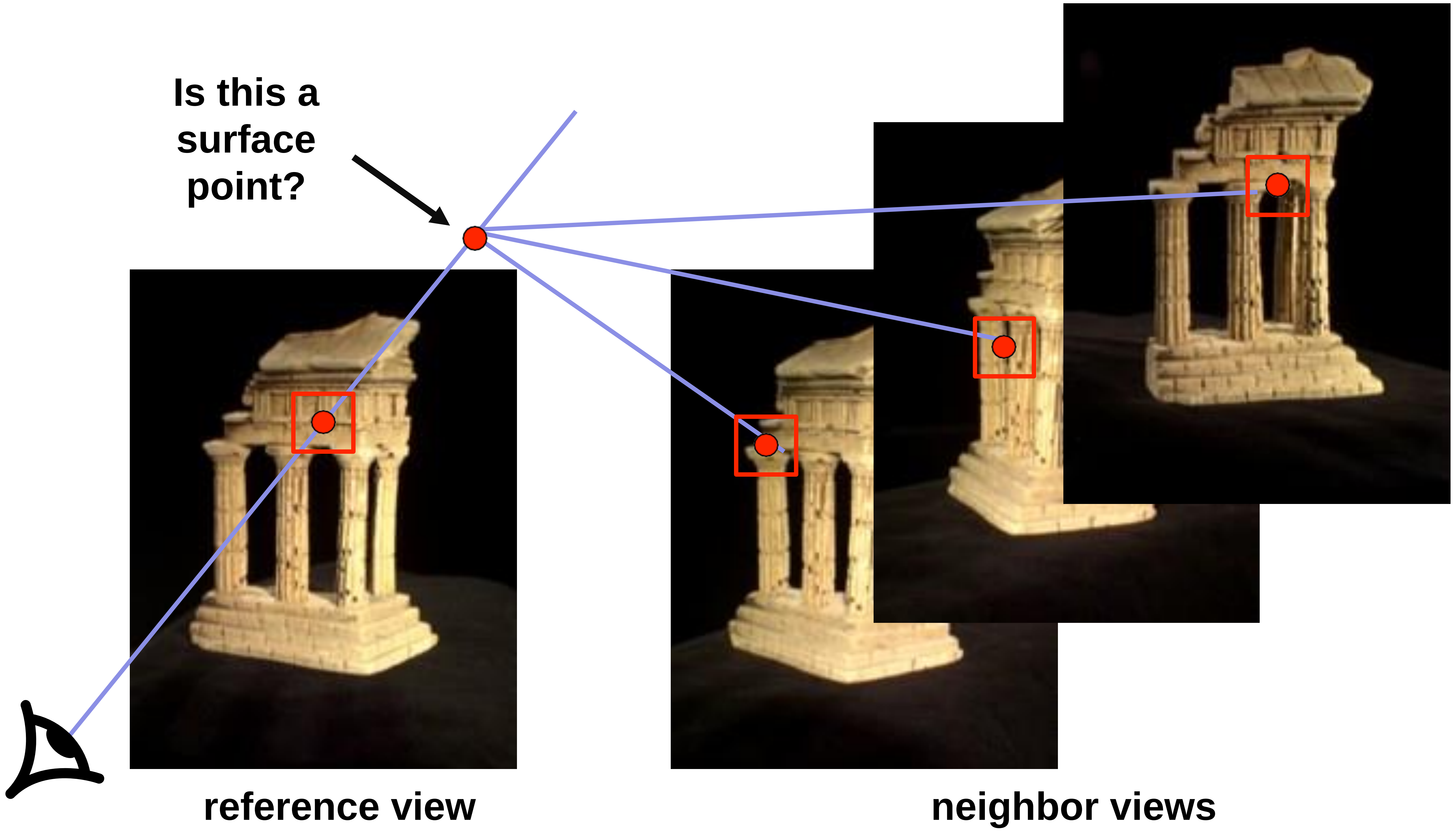
Multi-view stereo

Slide credit: Noah Snavely

Examples: Panoptic studio

http://domedb.perception.cs.cmu.edu/

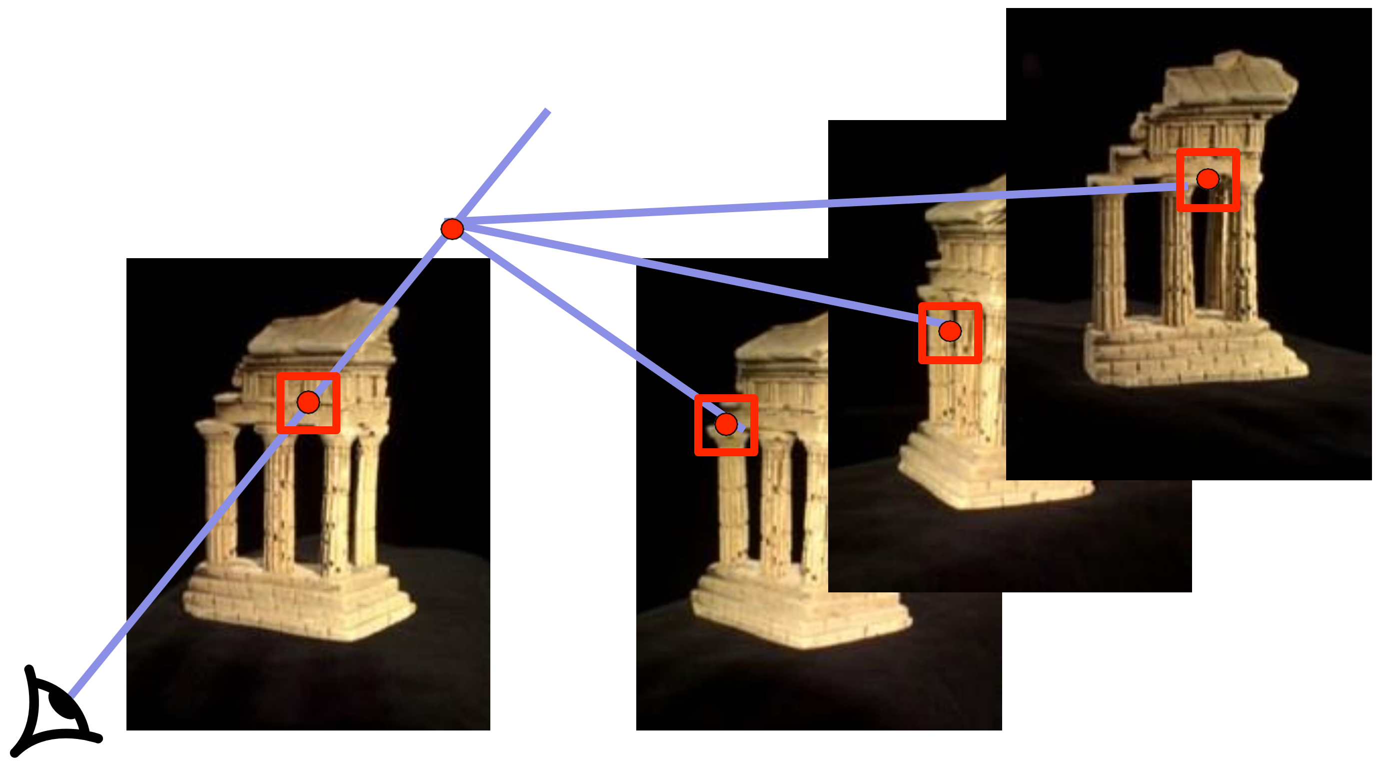
Multi-view stereo: Basic idea



Source: Y. Furukawa

Multi-view stereo: Basic idea

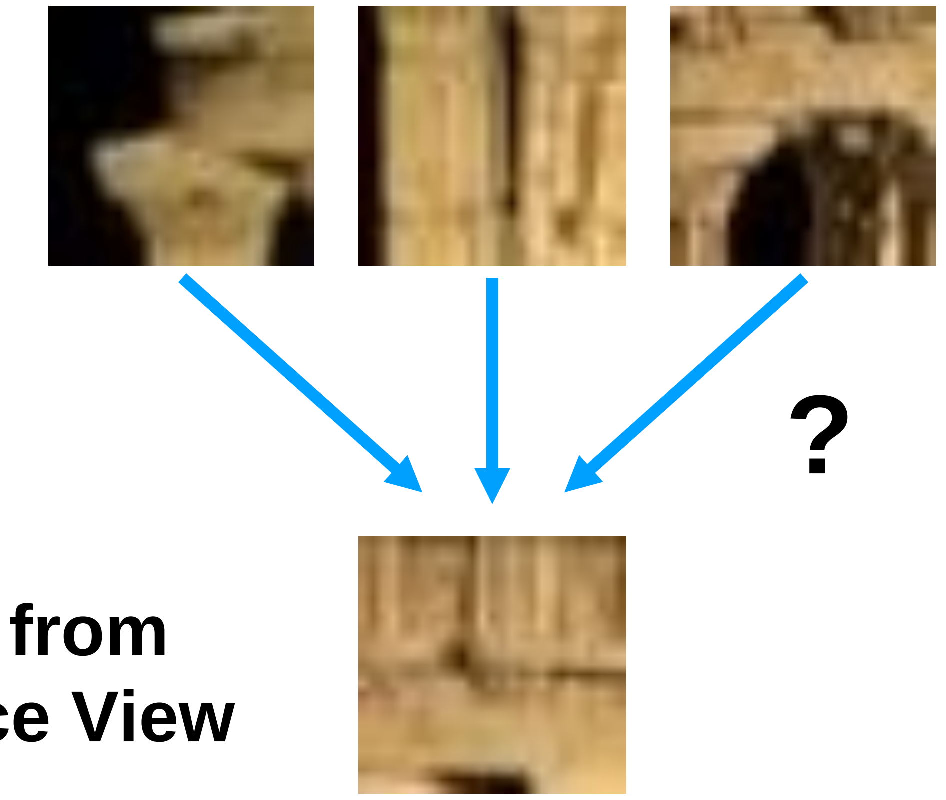
Evaluate the likelihood of geometry at a particular depth for a particular reference patch:



reference view

neighbor views

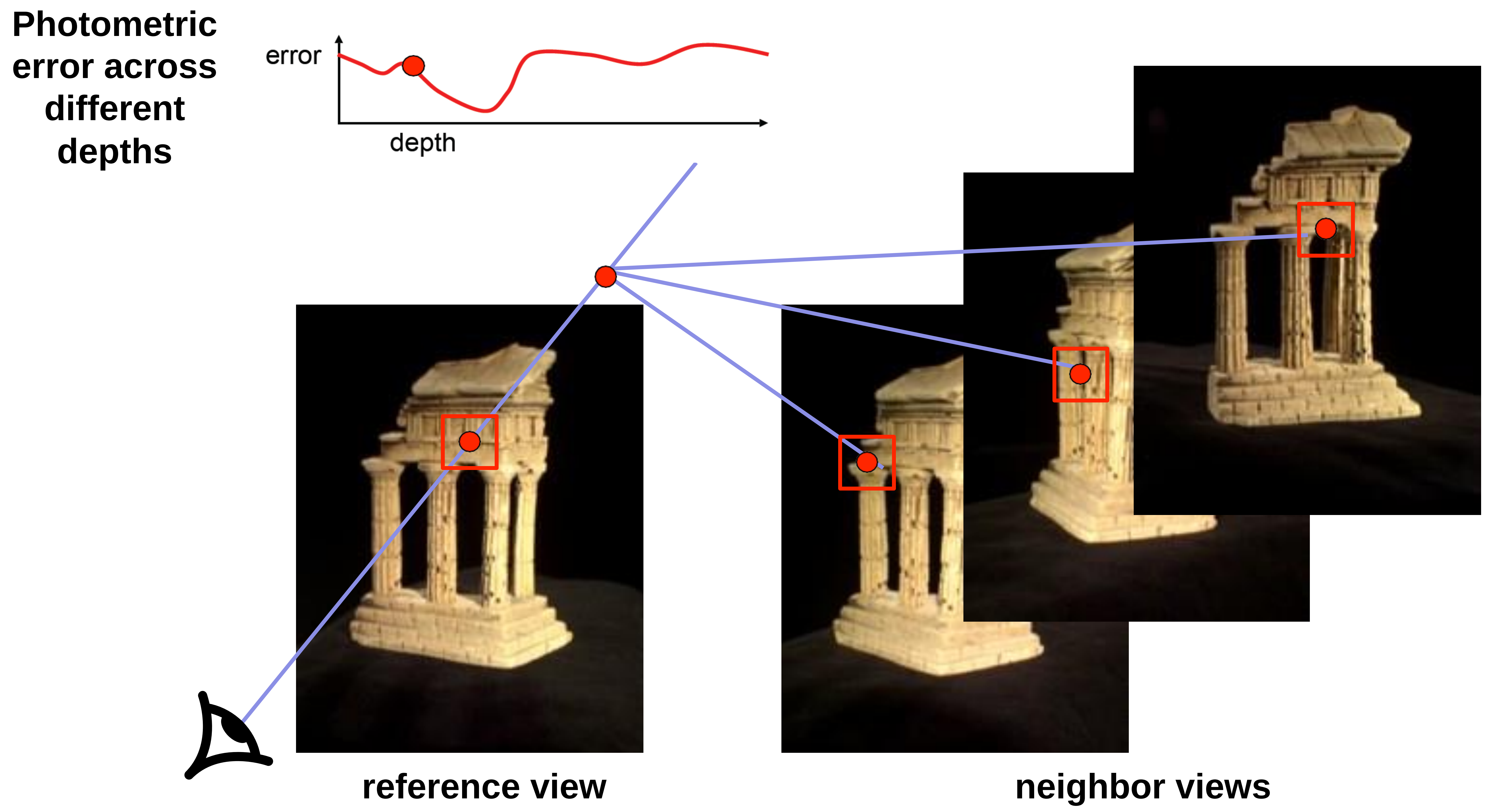
Corresponding patches at depth guess in other views



Patch from reference View

Source: Y. Furukawa

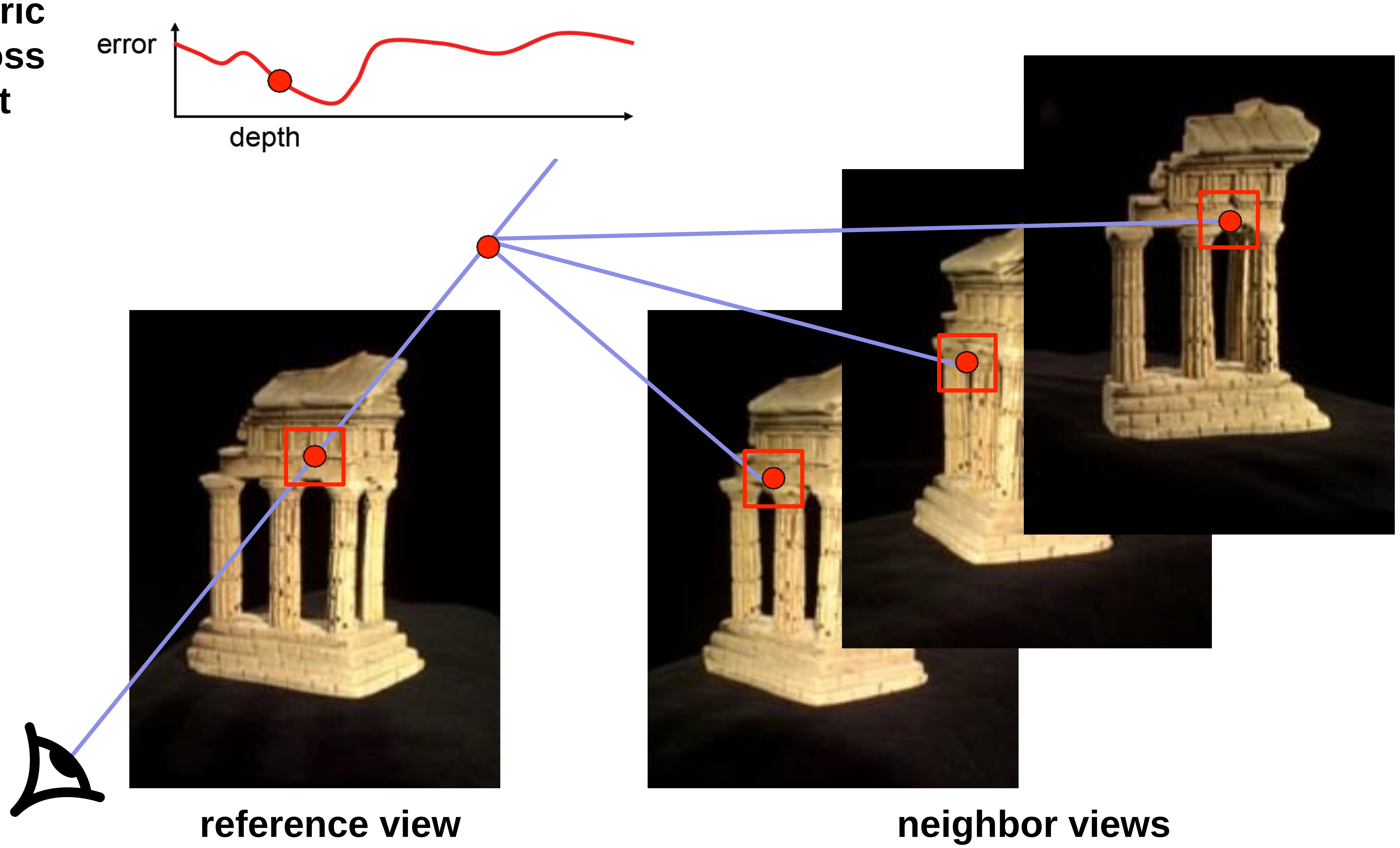
Multi-view stereo: Basic idea



Source: Y. Furukawa

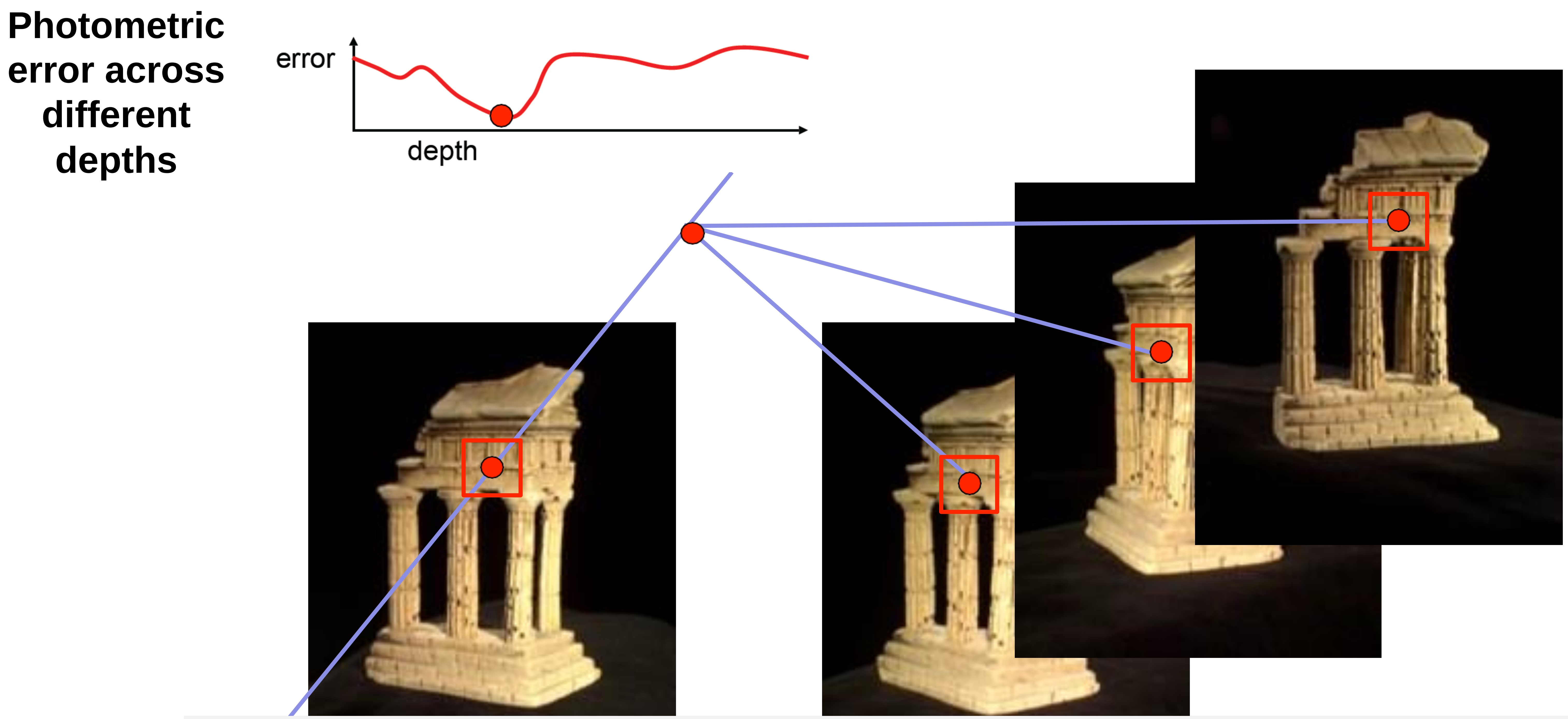
Multi-view stereo: Basic idea

Photometric error across different depths



Source: Y. Furukawa

Multi-view stereo: Basic idea



In this manner, solve for a depth map over the whole reference view

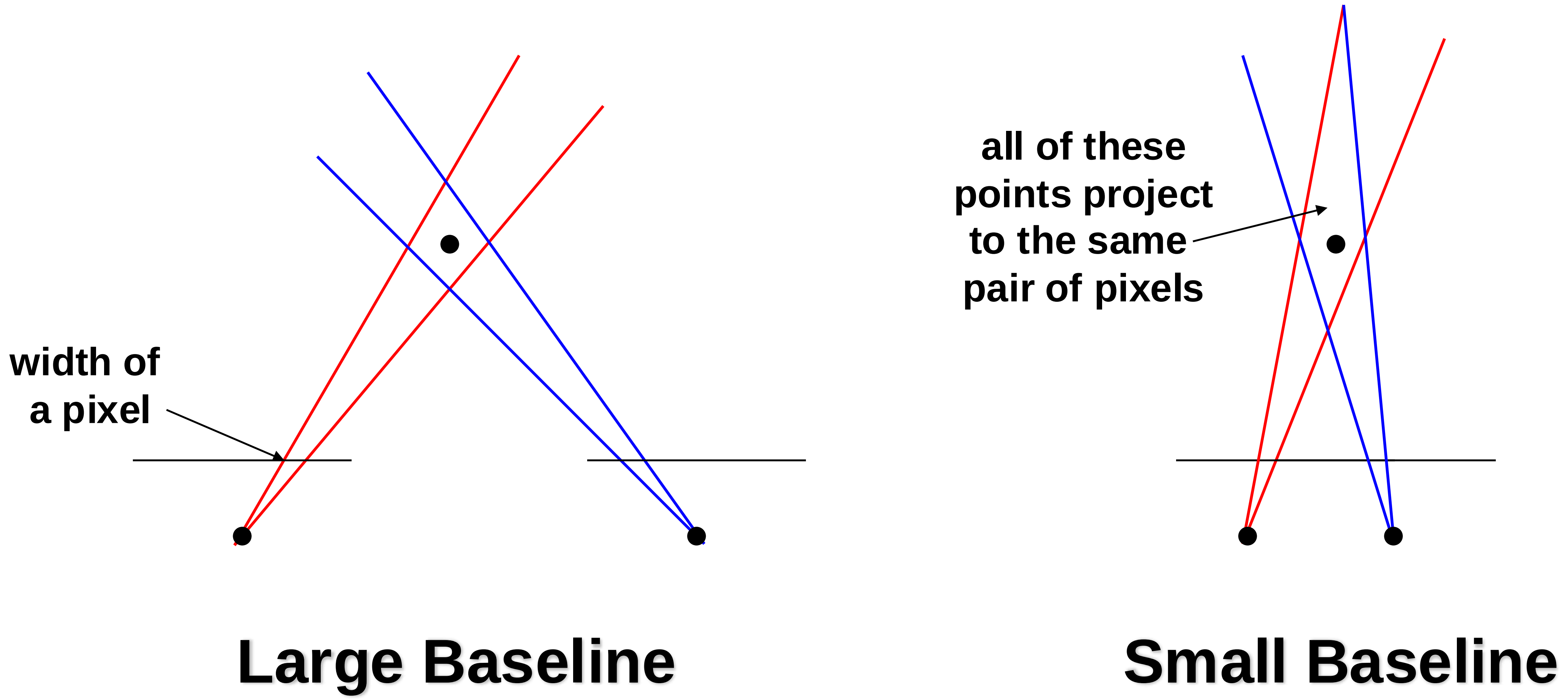
Furukawa

Multi-view stereo: advantages over 2 view

- Can match windows using more than 1 other image, giving a **stronger match signal**
- If you have lots of potential images, can **choose the best subset** of images to match per reference image
- Can reconstruct a depth map for each reference frame, and the merge into a **complete 3D model**

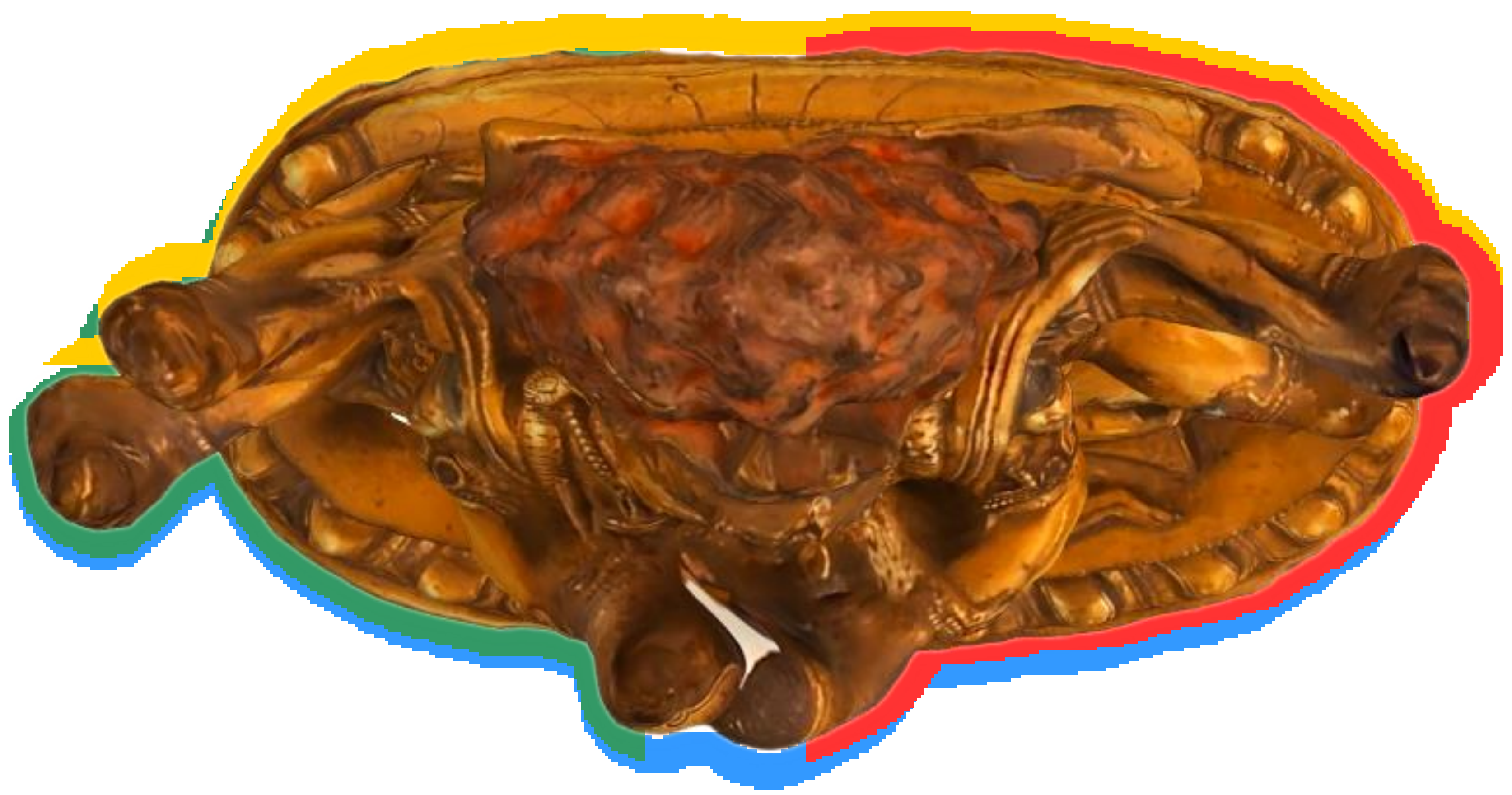
Source: Y. Furukawa

Choosing the baseline

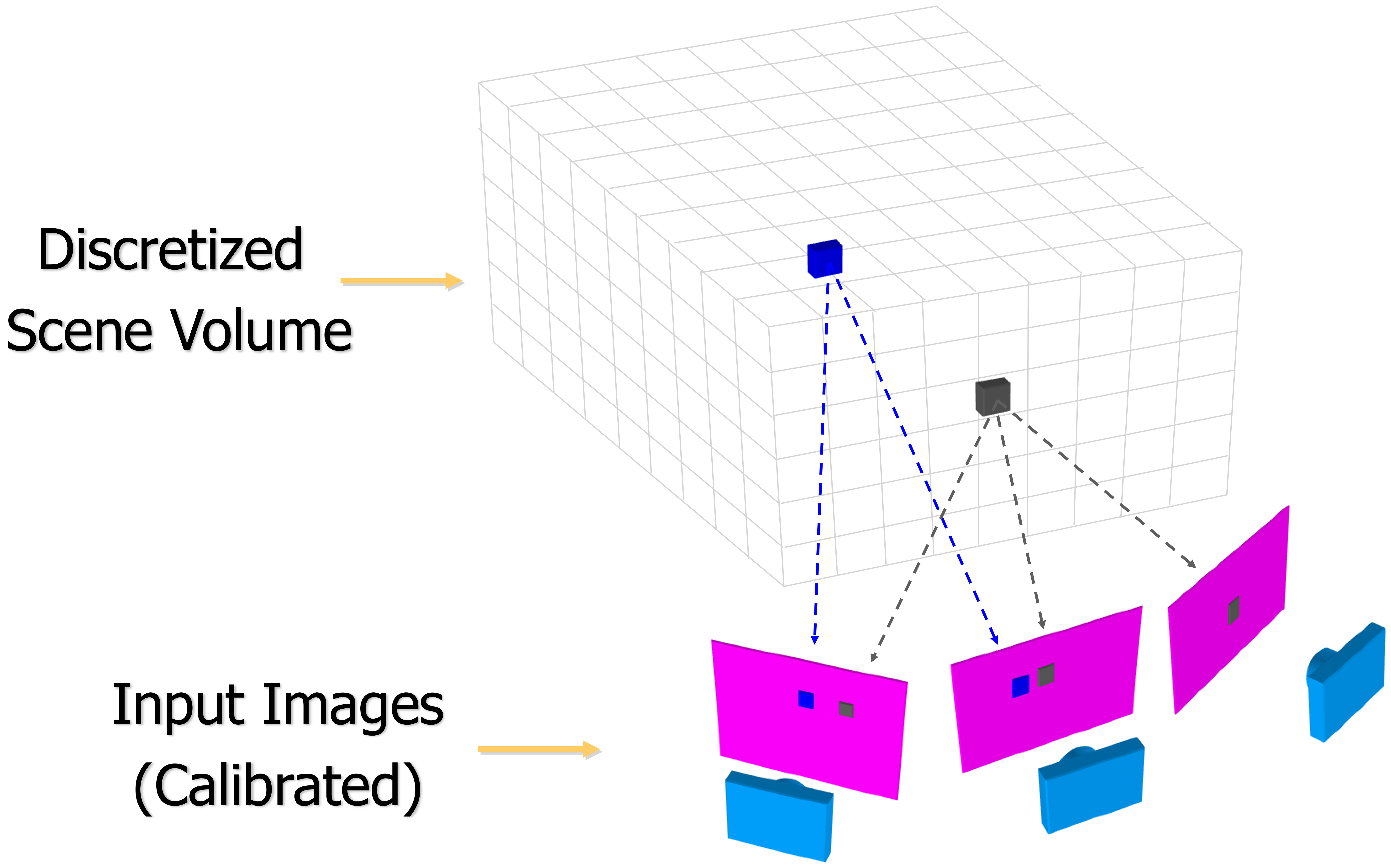


- What’s the optimal baseline?
  - Too small: large depth error
  - Too large: difficult search problem

Slide credit: Noah Snavely

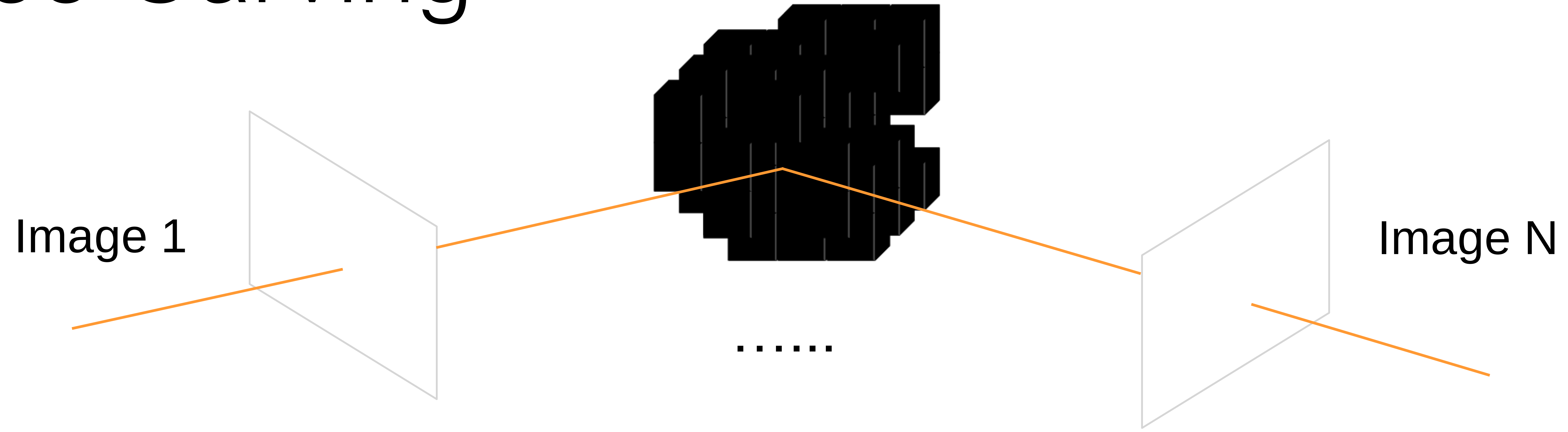


Volumetric stereo



Goal: **Assign RGB values to voxels in V**
photo-consistent **with images**

Space Carving



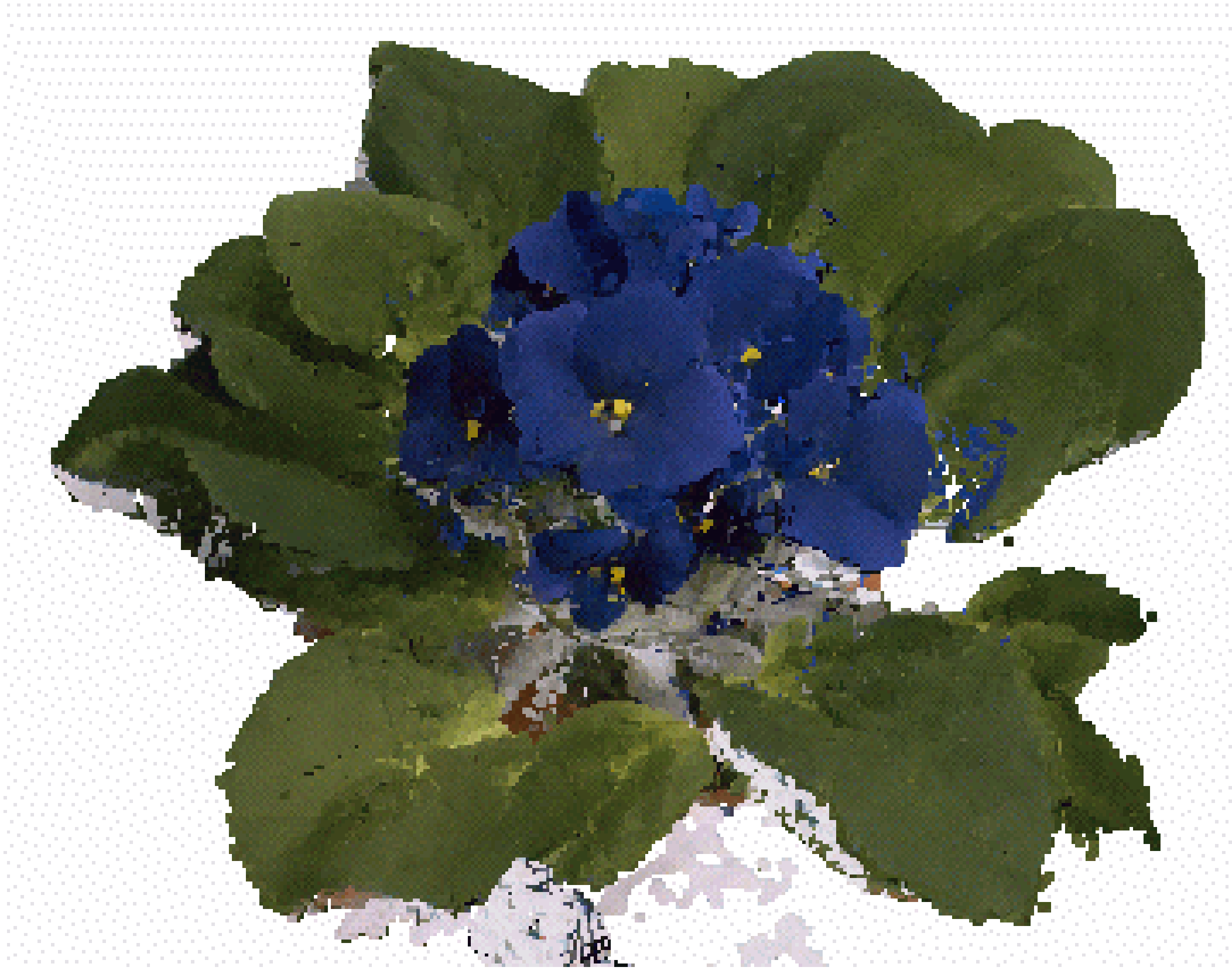
- Space Carving Algorithm
  - Initialize to a volume V containing the true scene
  - Choose a voxel on the outside of the volume
  - Project to visible input images
  - Carve if not photo-consistent
  - Repeat until convergence

K. N. Kutulakos and S. M. Seitz, **A Theory of Shape by Space Carving**, *ICCV* 1999

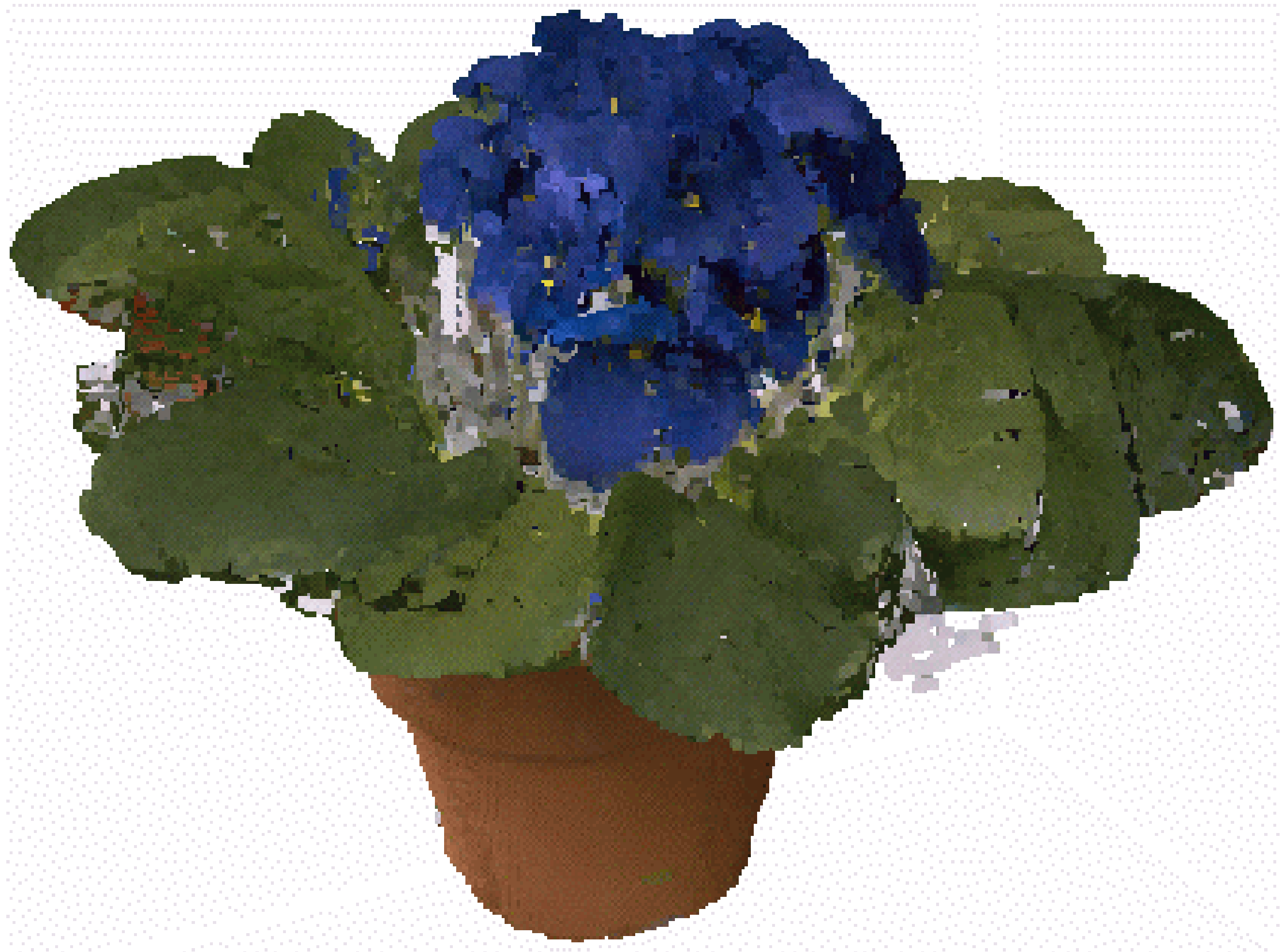
Space Carving Results

Input Image (1 of 45)

Reconstruction



Reconstruction

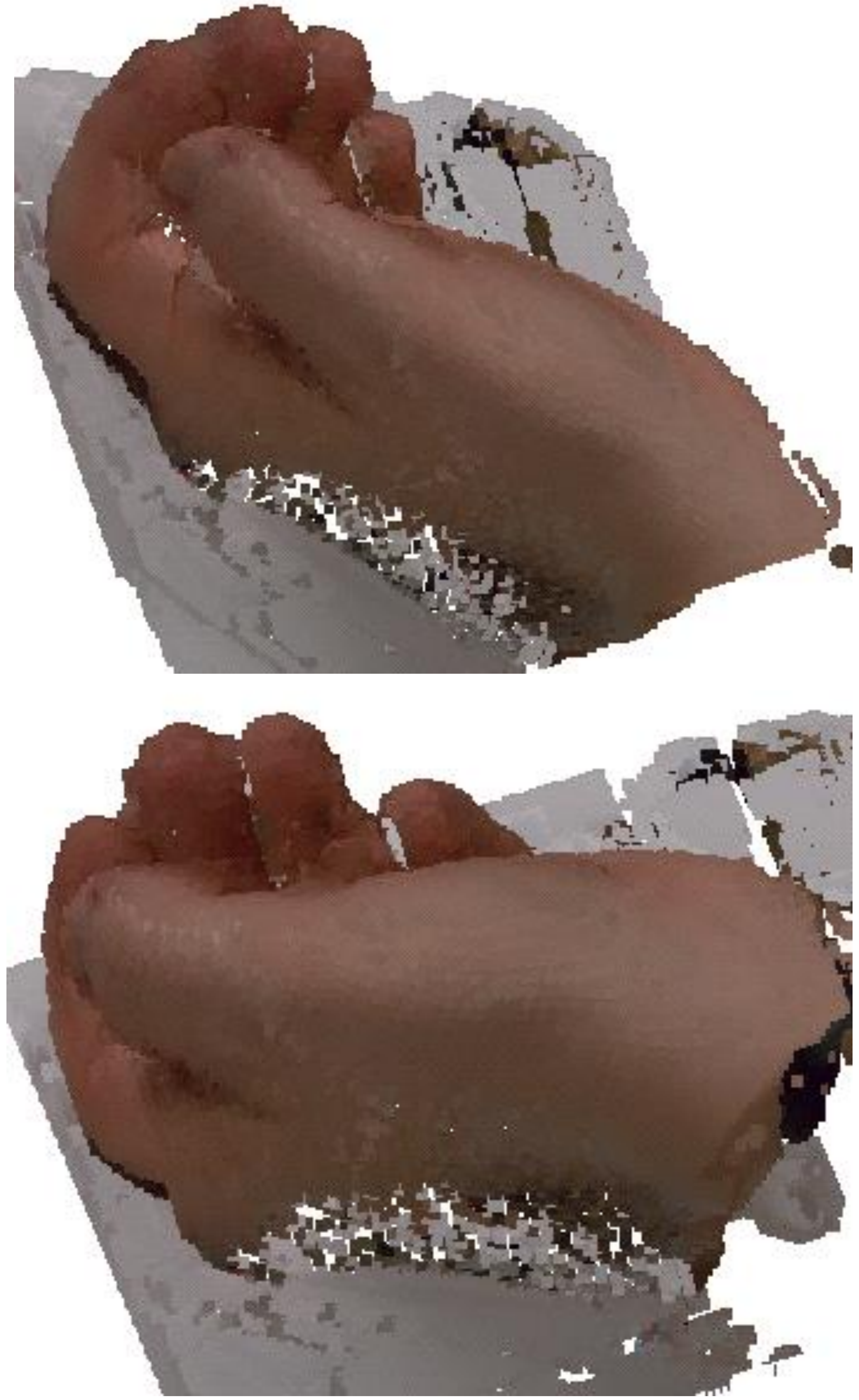


Reconstruction

Source: S. Seitz

Space Carving Results

Input Image (1 of 100)



Reconstruction

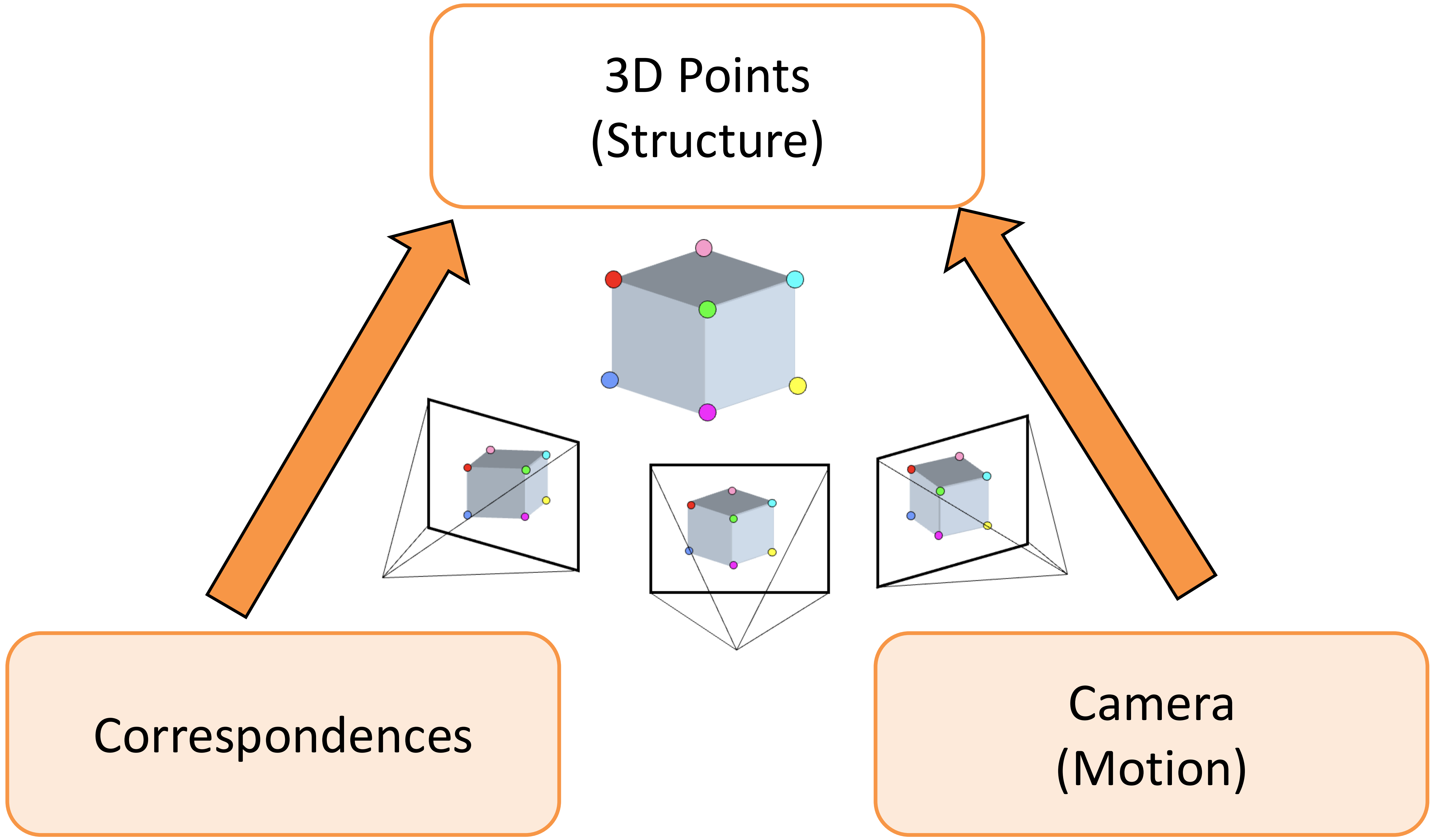
Source: S. Seitz

Tool for you: COLMAP

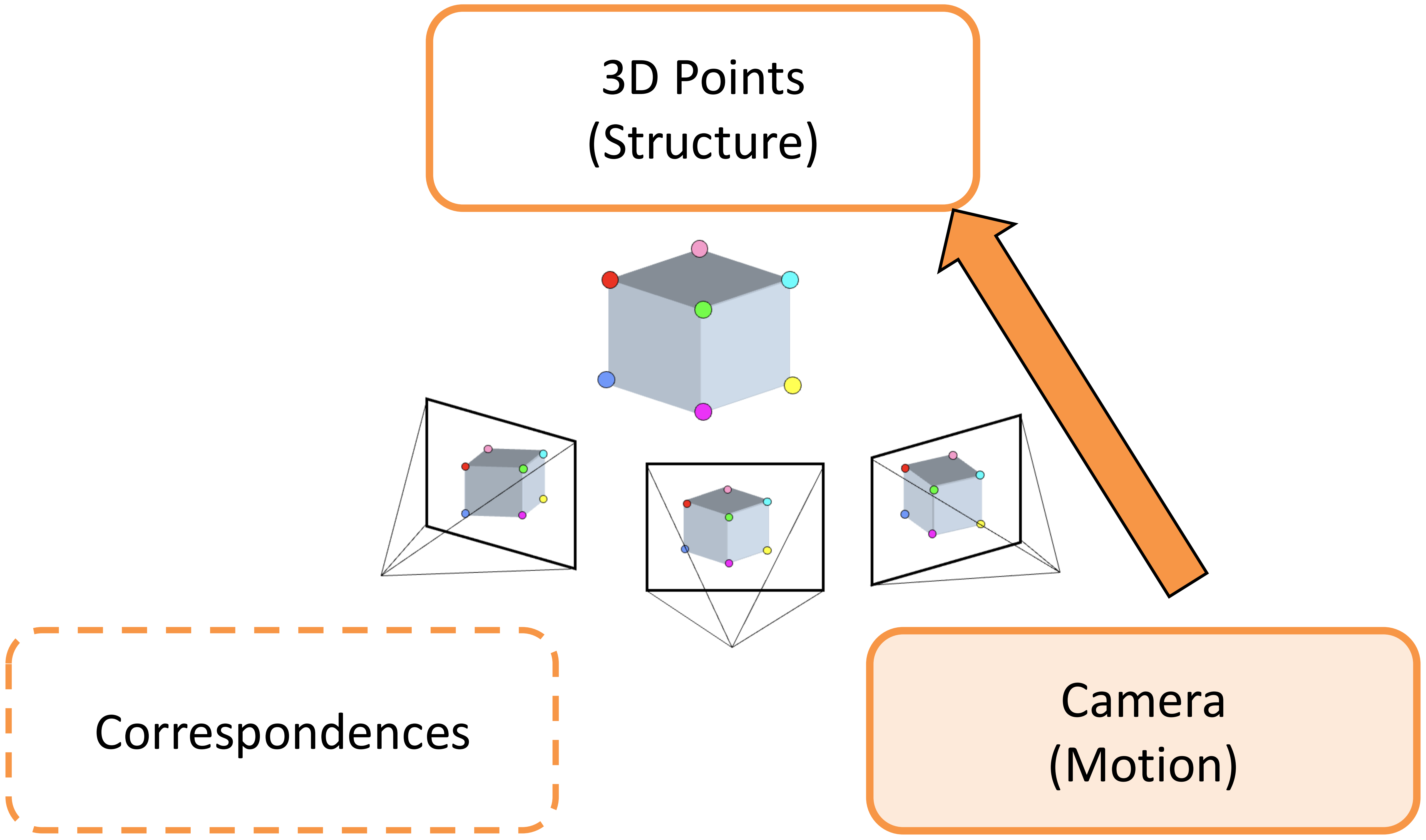
https://github.com/colmap/colmap

A general SfM + MVS pipeline

Multi-View Stereo



Volumetric “Neural” Rendering



Does not use explicit correspondences,
relies on reconstruction loss (Analysis-by-Synthesis)

Neural Radiance Fields

Video from the original ECCV'20 paper

Capturing Reality

Earliest cave painting (45,500 years old) in Sulawesi, Indonesia

Capturing Reality

Monet’s Cathedral series: study of light 1893-1894

Capturing Reality

First self-portrait Cornelius 1839

First Movie - Muybridge 1878

Capturing Reality – in 3D

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

Capturing Reality – in 3D (MVS – last lecture)

Google Earth 2016~

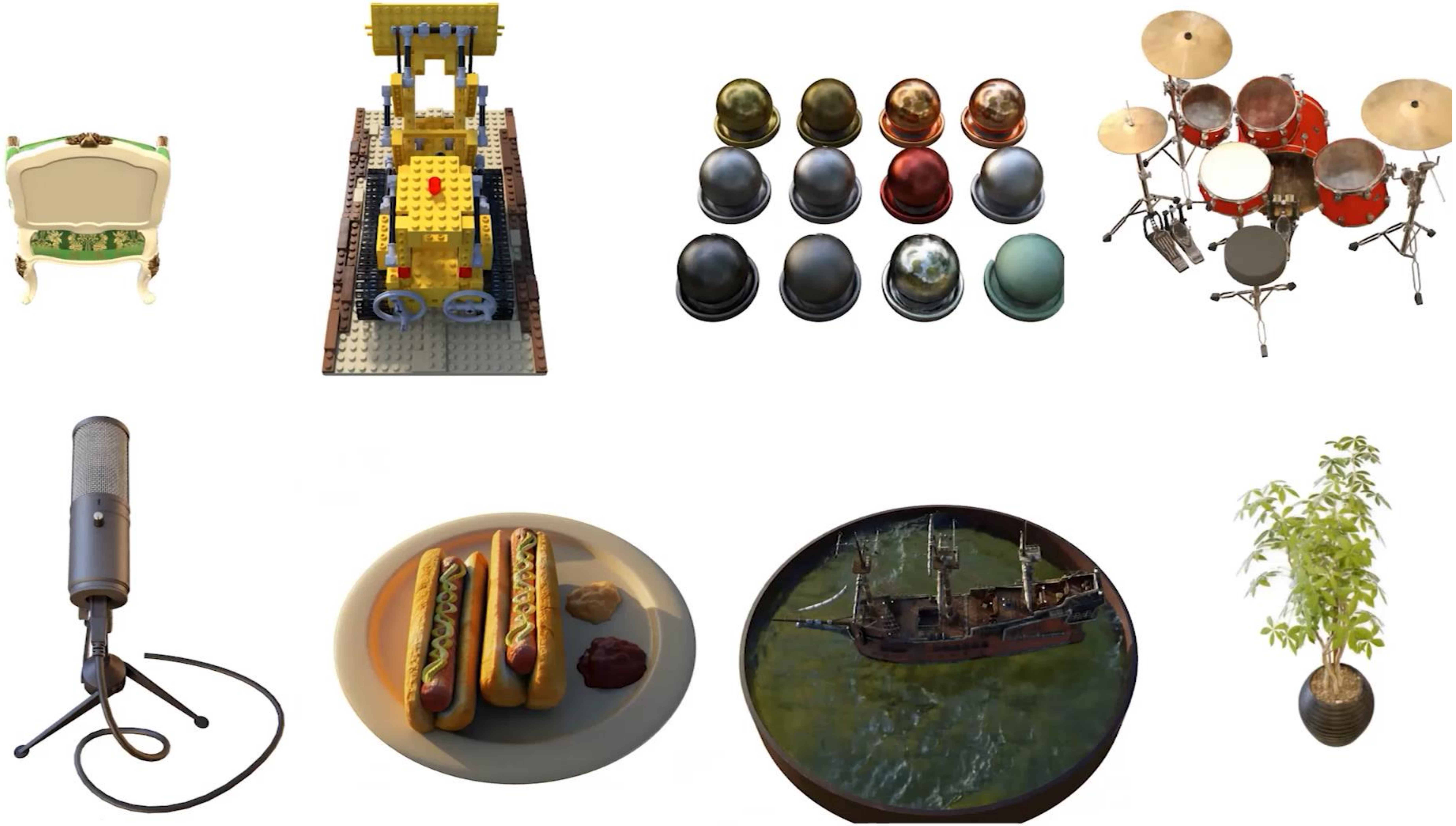
What is next?

2020: Neural Radiance Field (NeRF)

Mildenhall*, Srinivasan*, Tancik*, Barron, Ramamoorthi, Ng, ECCV 2020

It has been three years

- Original NeRF paper: 4200+ citations in 3 years



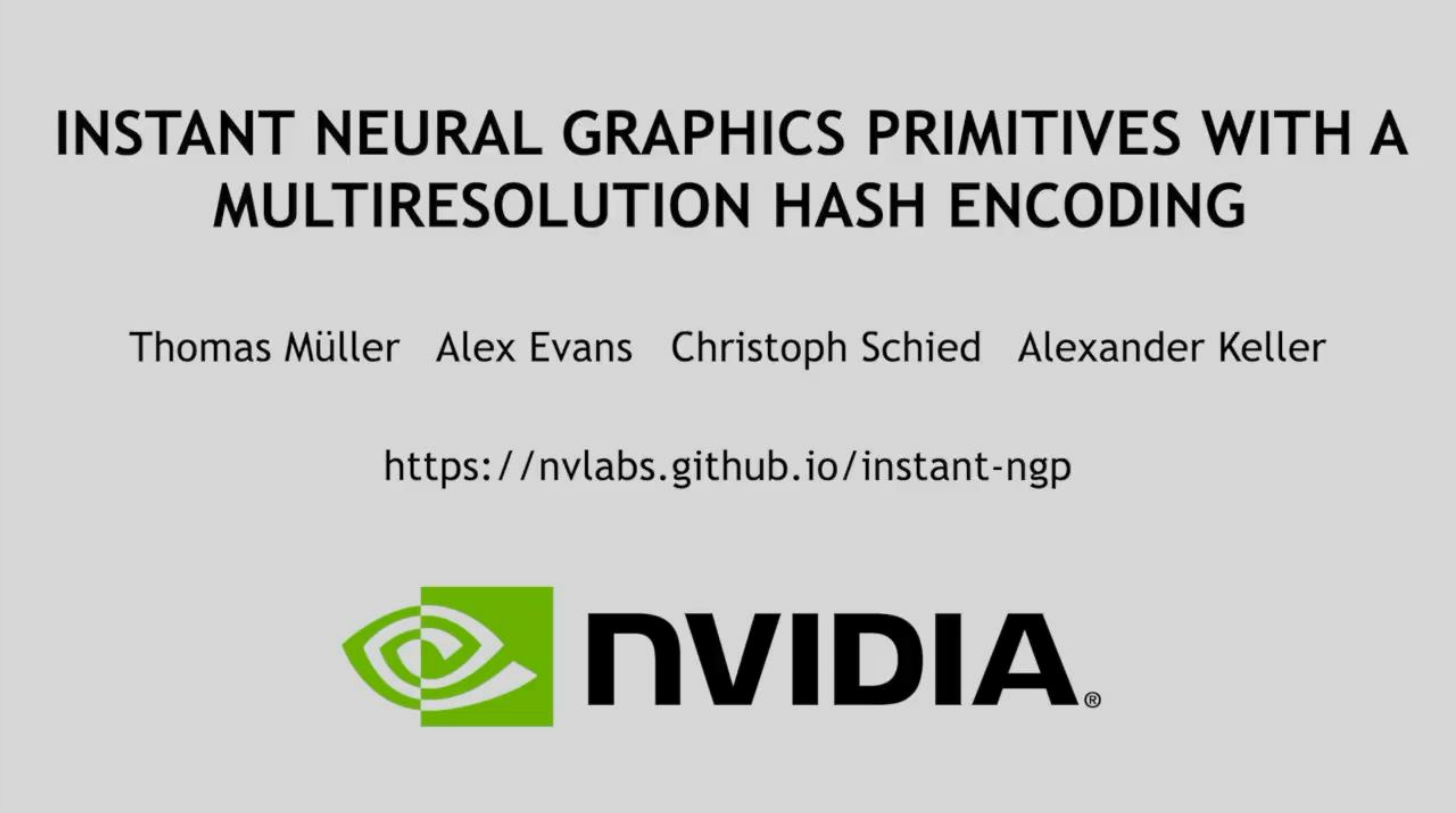
Handling Appearance Changes

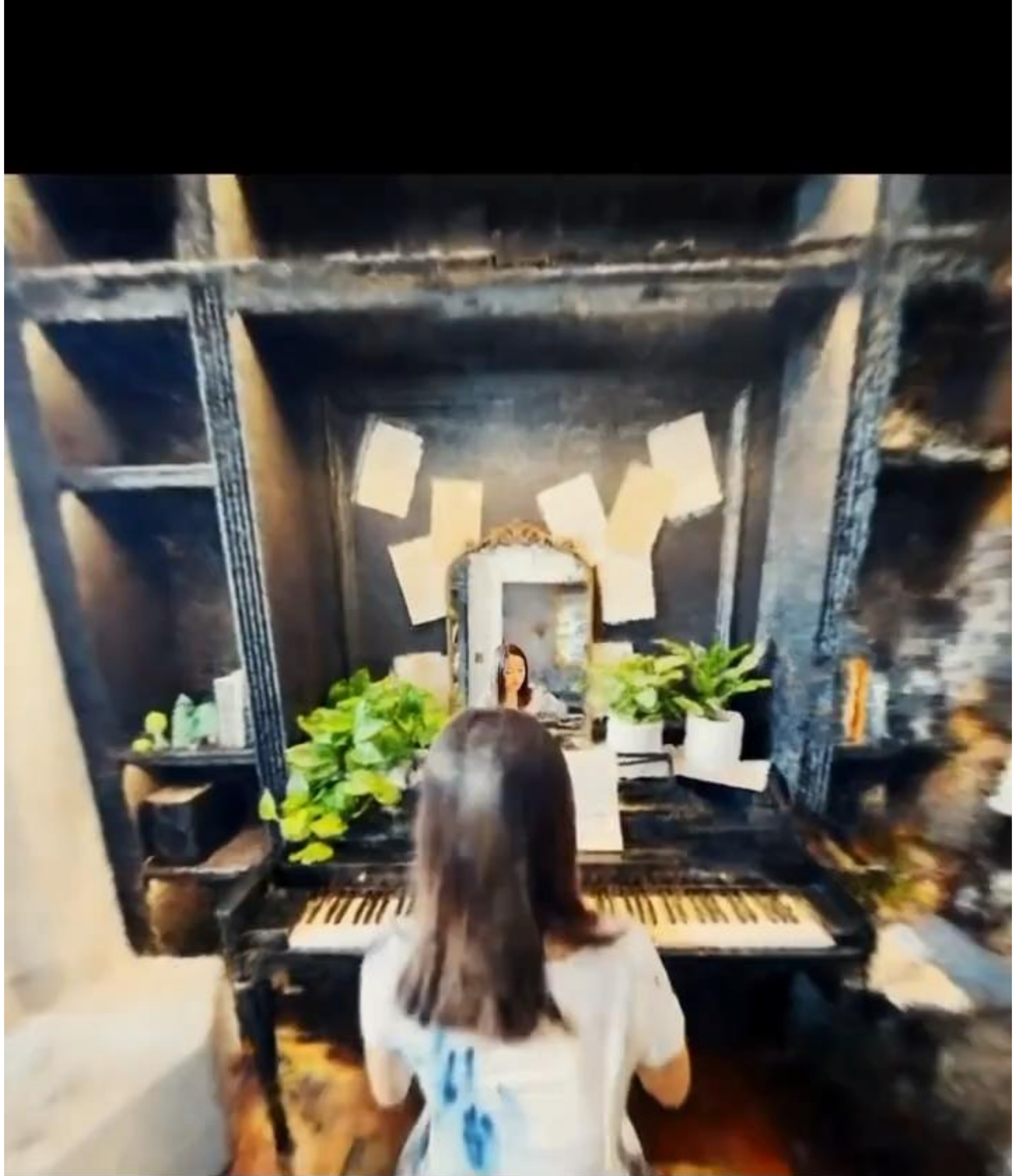
Nerf-W [Martin-Brualla et al. CVPR 2021]

Real-time Rendering

Video from PlenOctrees [Yu et al. CVPR 2021]

Real-time Inference





@karenxcheng, with InstantNGP [Müller et al., SIGGRAPH 2022]

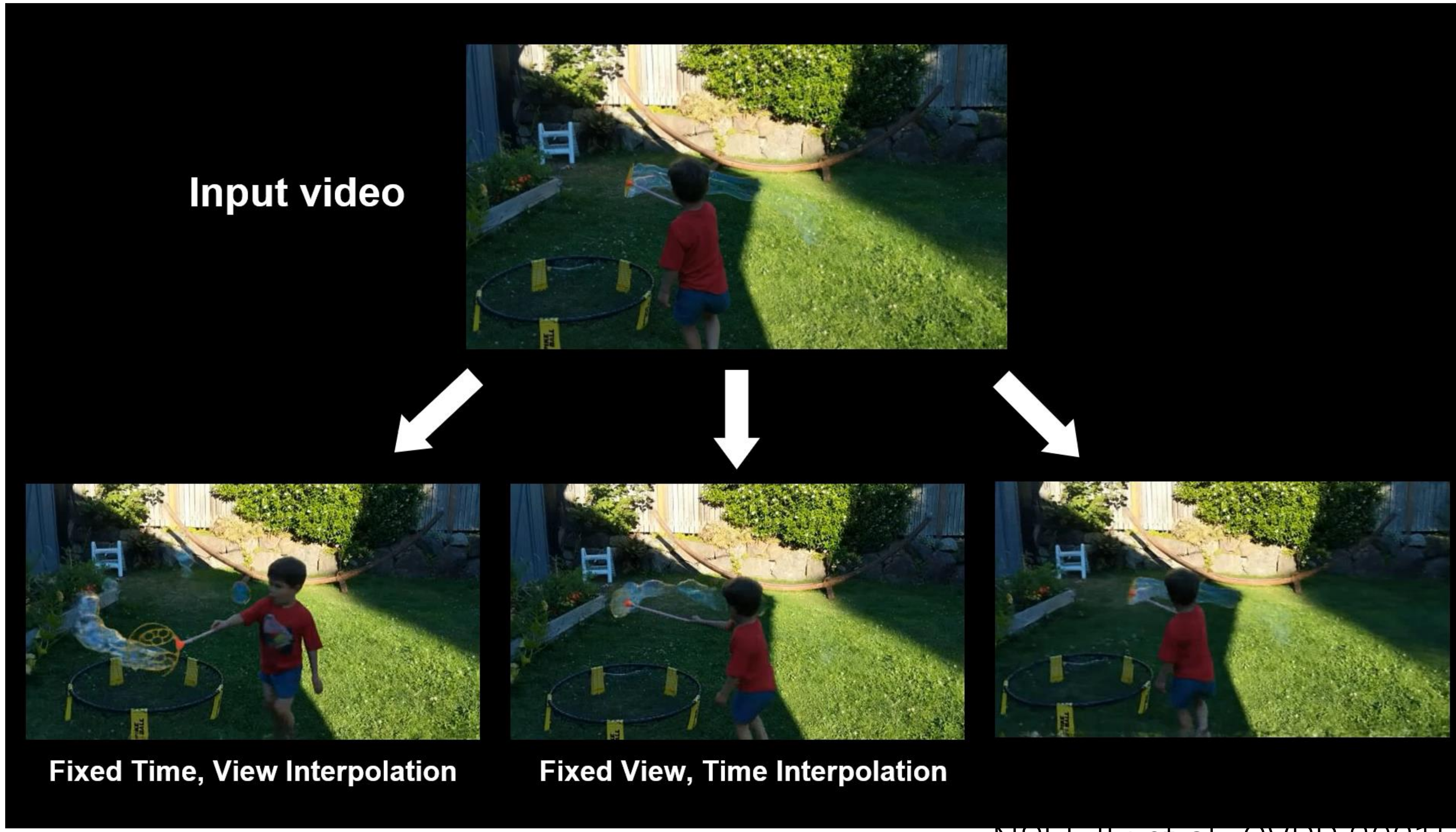

GETTING STARTED
GITHUB
DOCUMENTATION
nerfstudio
VIEWPORT
RENDER VIEW
RESUME TRAINING
Show Scene
Show Images
Refresh Page
Resolution: 640x1024px
Time Allocation: 100% spent on viewer
Server Connected
Render Connected
CONTROLS
RENDER
SCENE
LOAD PATH
EXPORT PATH
Height
1080
Width
1920
FOV
50
Seconds
4
FPS
24
ADD CAMERA
Smoothness
0.00
0
1
2
3
CAMERA 0
0
1
2
nerfstudio

Dynamic NeRFs

[Xian et al., CVPR 2021]

HyperNeRF [Park et al., SigAsia 2021]

Nerfies [Park et al., ICCV 2021]



NSFF [Li et al., CVPR 2021]

Generative 3D Faces

EG3D: Efficient Geometry-aware 3D Generative Adversarial Networks, Chan et al. CVPR 2022

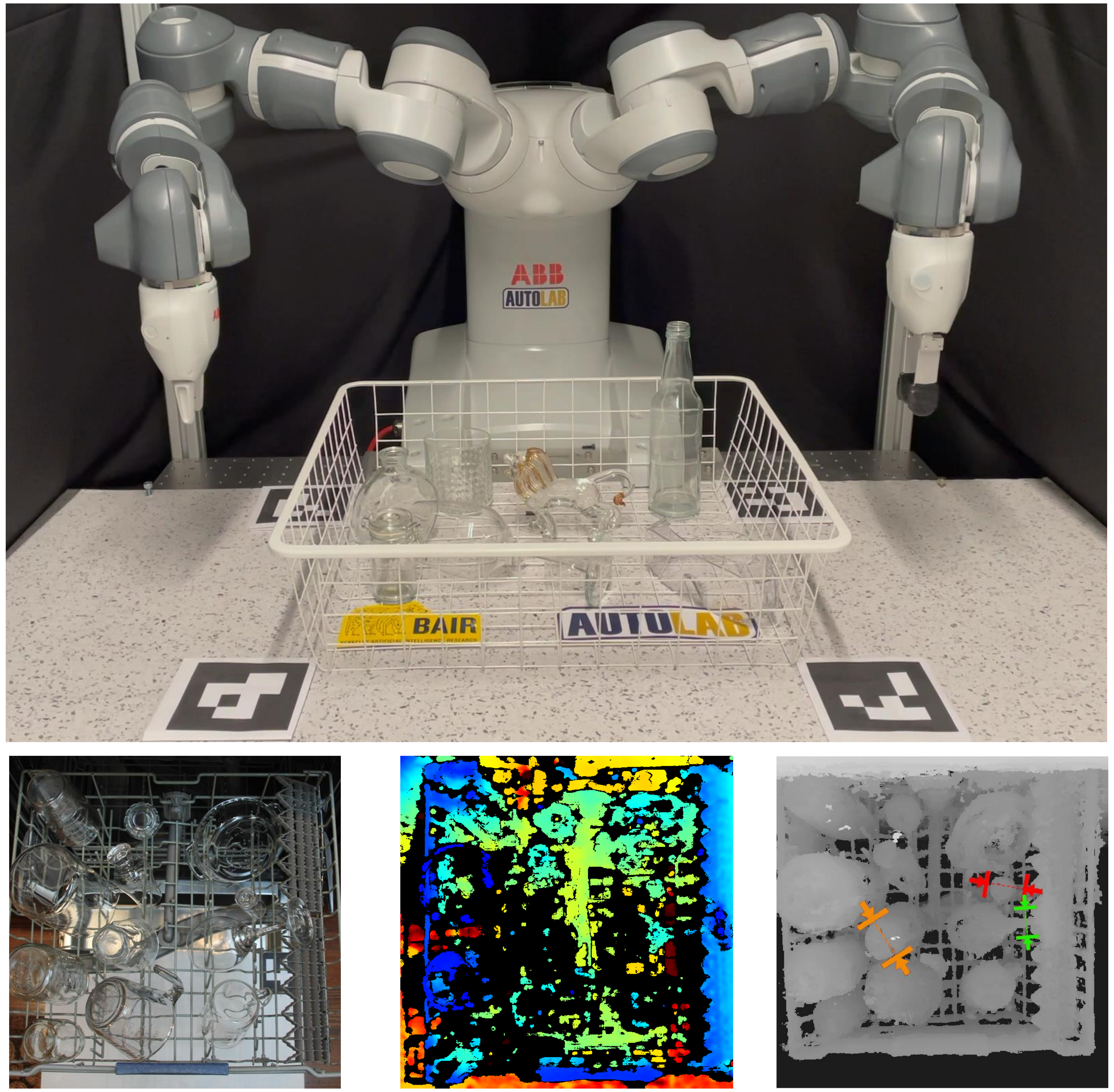
Wang et al. SIGGRAPH 2022

City-Scale NeRFs

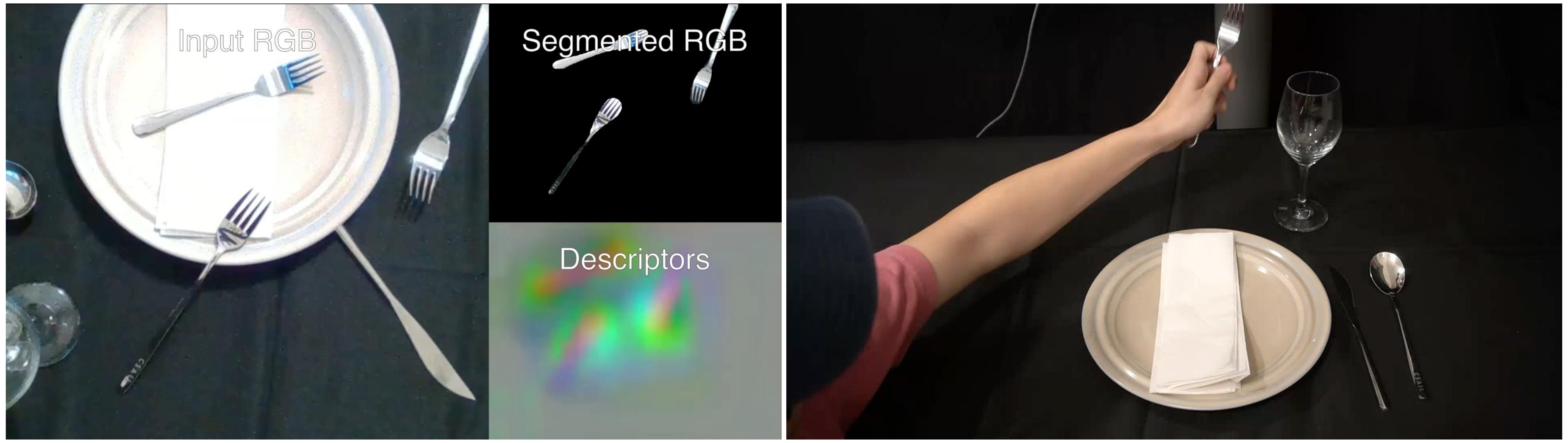
BlockNeRF
[Tancik et al.
CVPR 2022]

RawNeRF
[Mildenhall et al.
CVPR 2022]

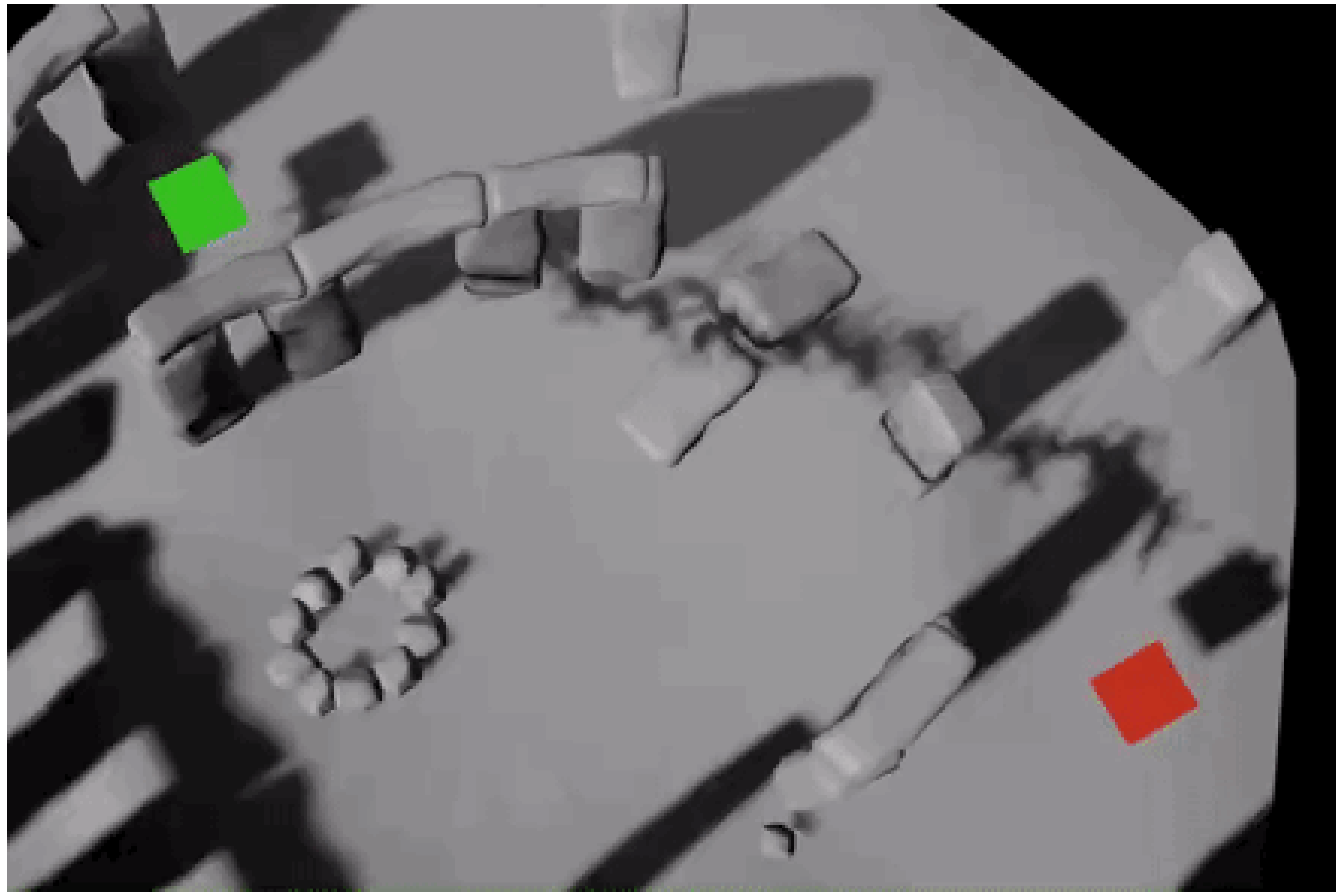
Robotics



Dex-NeRF: Using a Neural Radiance field to Grasp Transparent Objects, [Ichnowski and Avigal et al. CoRL 2021]



NeRF-Supervision: Learning Dense Object Descriptors from Neural Radiance Fields, [Yen-Chen et al. ICRA 2022]



Vision-Only Robot Navigation in a Neural Radiance World [Adamkiewicz and Chen et al. ICRA 2022]

Generating 3D scenes with diffusion models

DreamFusion [Poole et al. ICLR 2023]

Querying with Language

[Kerr* and Kim* et al. LERF, ICCV 2023]

Editing with Instructions

[Haque, Tancik, Efros, Holynski, Kanazawa, ICCV 2023]

Goals of the next few lectures

- Visit the fundamentals in Neural Volumetric Rendering by abstracting away recent developments
- Provide first principles + background for you to go and read these papers & play around with the tools
- New Project 5!! Implement these concepts yourself

nerfstudio

Capture of UC Berkeley redwoods with

Birds Eye View & Background

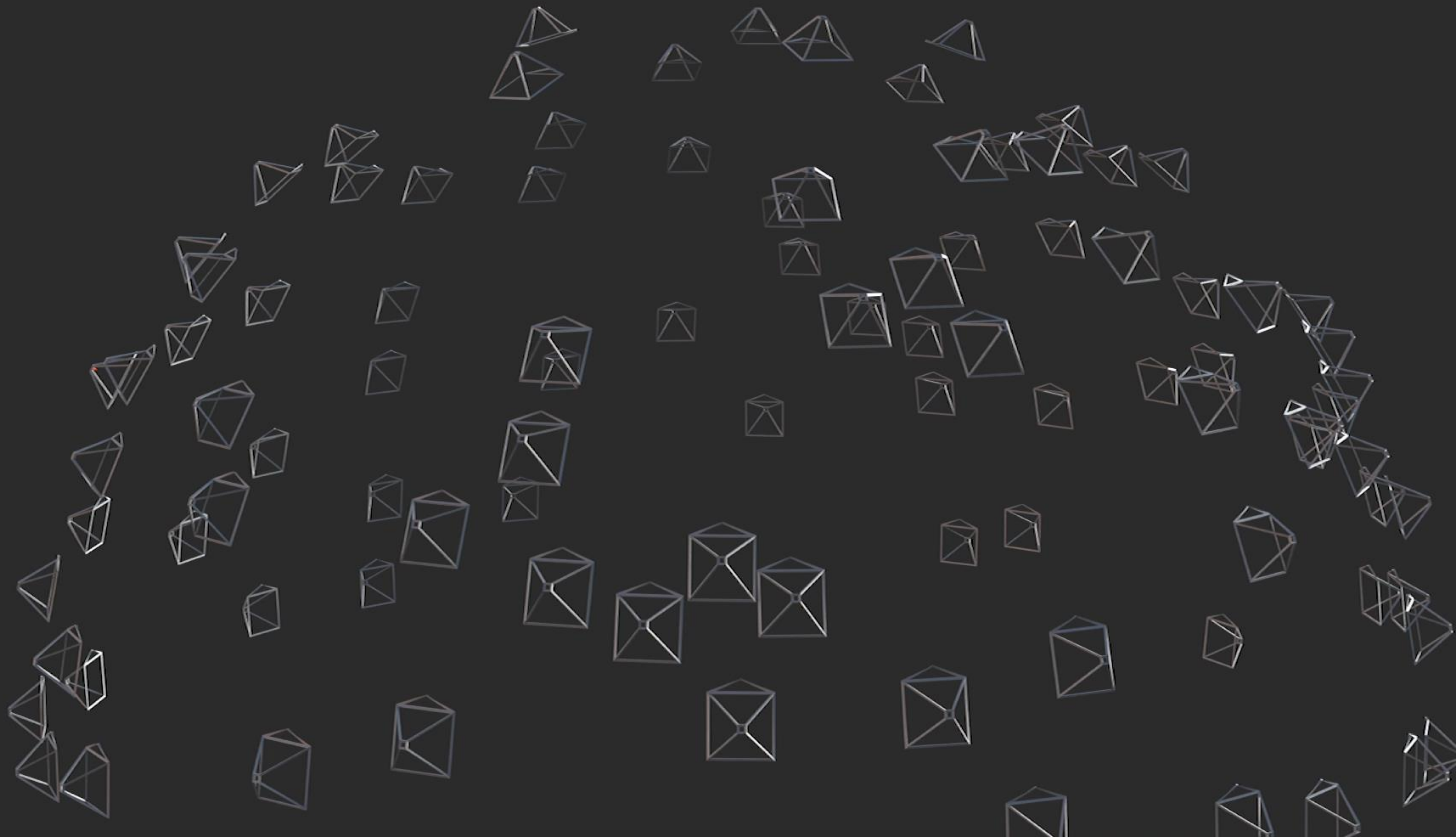
Birds Eye View

- What is NeRF?
- How is it different or similar to existing approaches?
- What is its historical context?

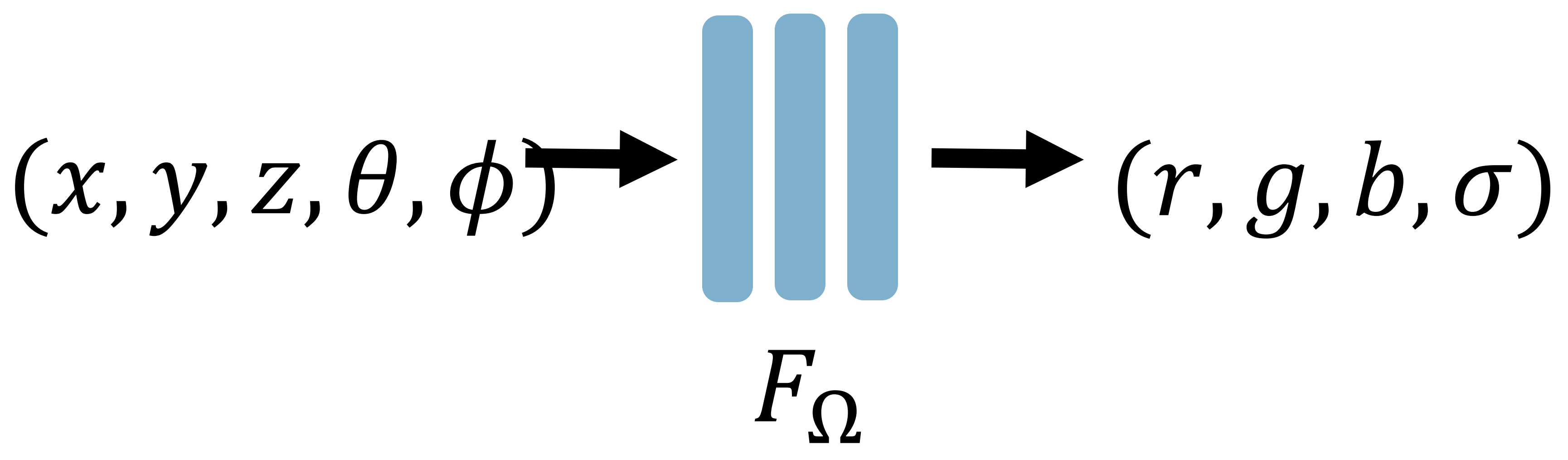
Problem Statement

Input:
A set of calibrated Images

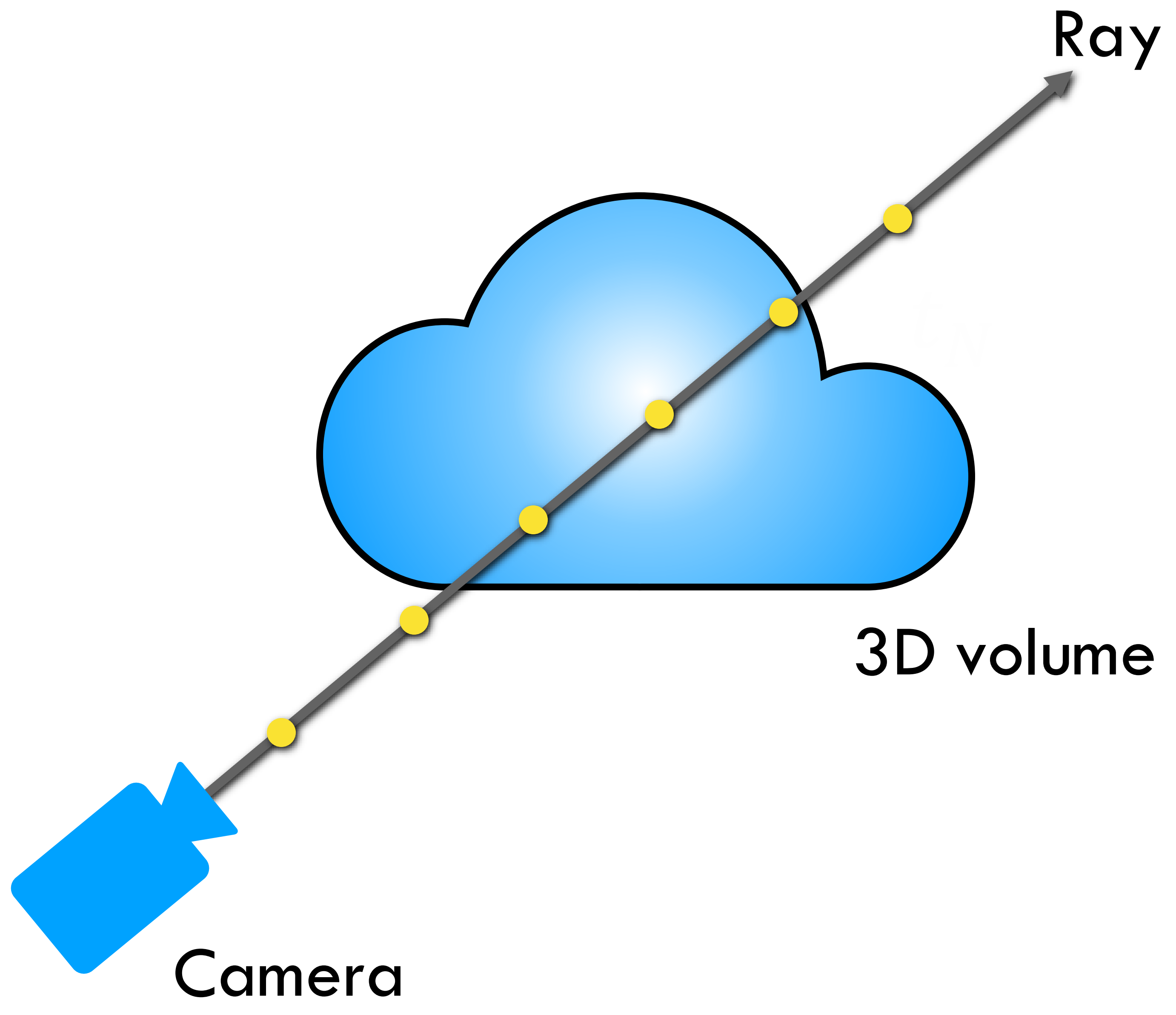
Output:
A 3D scene representation that renders novel views



Three Key Components

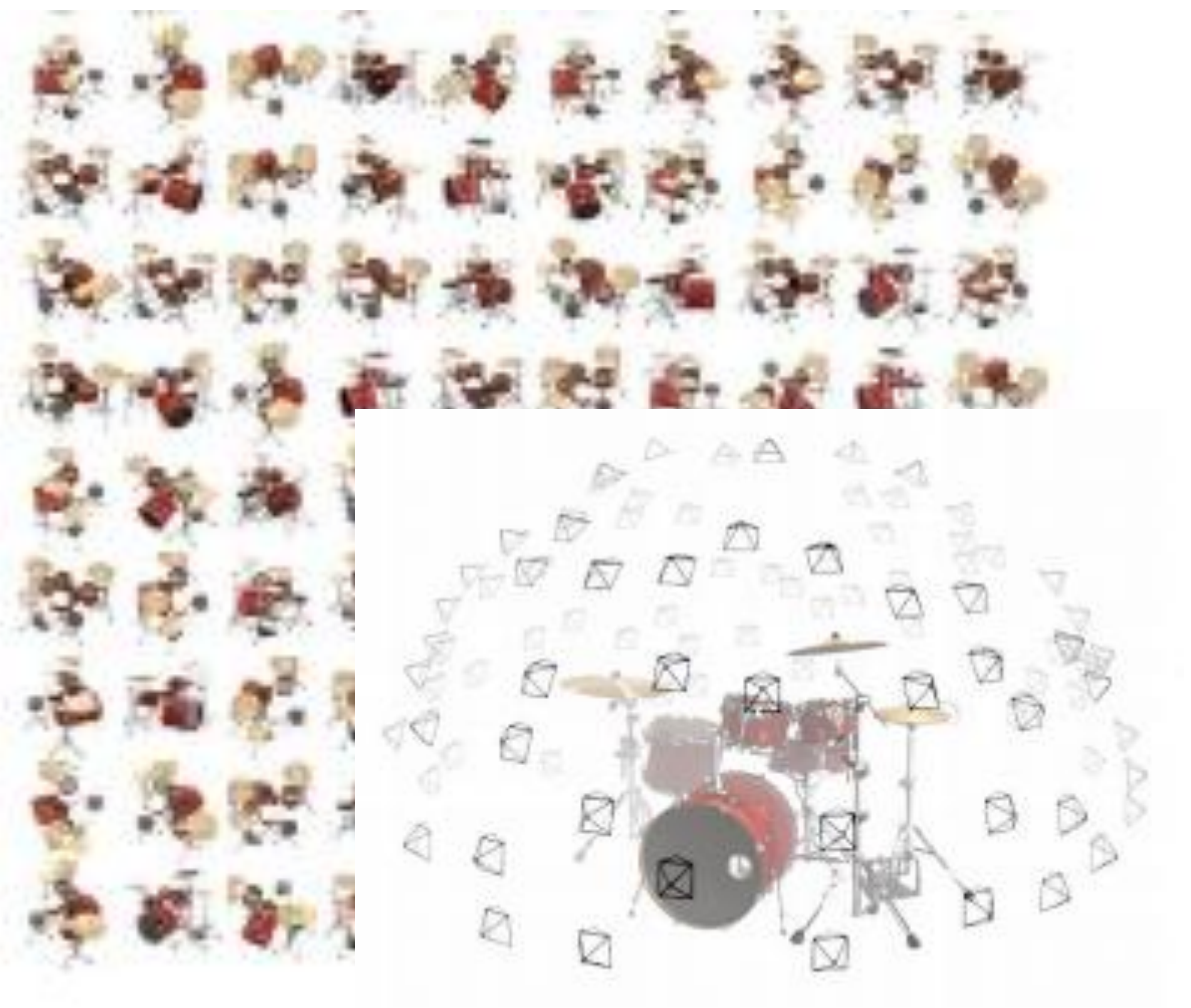


Neural Volumetric 3D Scene Representation



Differentiable Volumetric Rendering Function

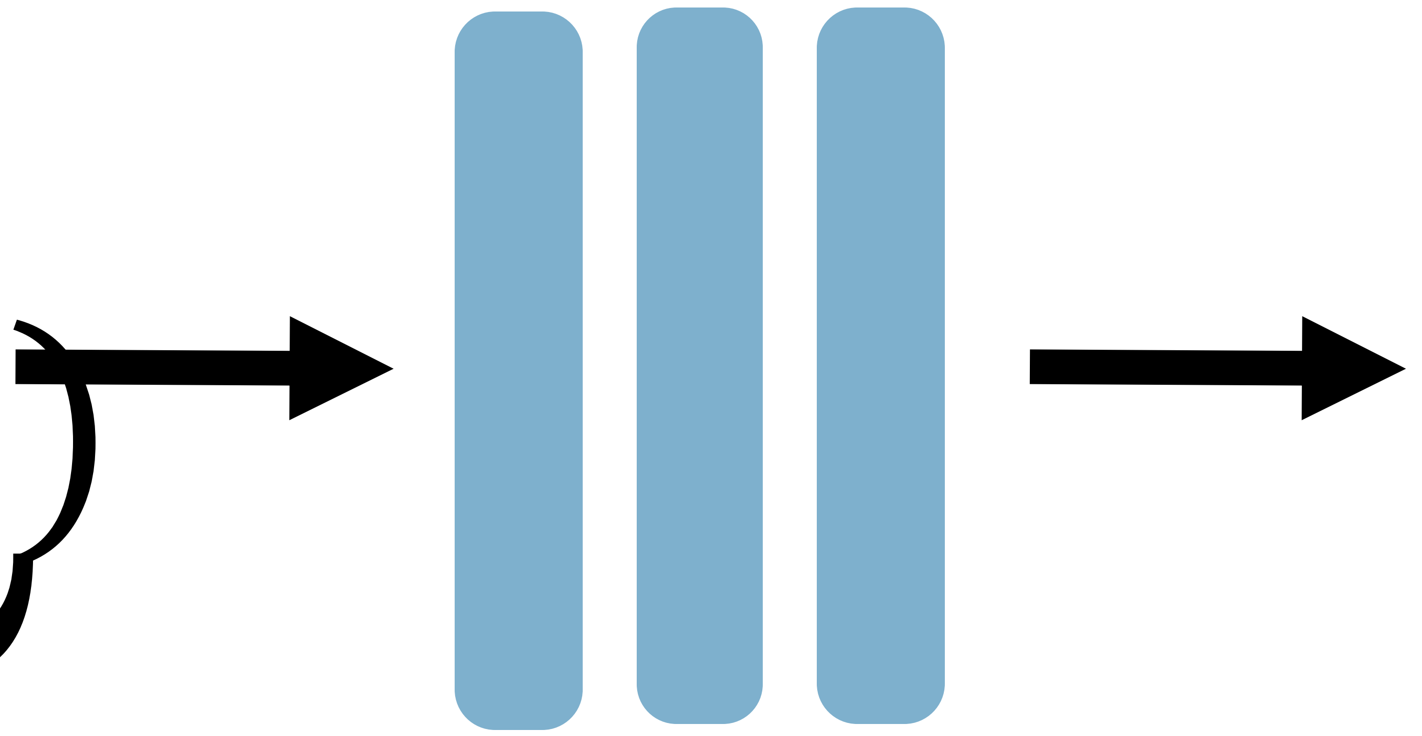
Objective: Synthesize all training views



Optimization via Analysis-by-Synthesis

Representing a 3D scene as a continuous 5D function

$$(x, y, z, \theta, \phi) \rightarrow F_\Omega \rightarrow (r, g, b, \sigma)$$

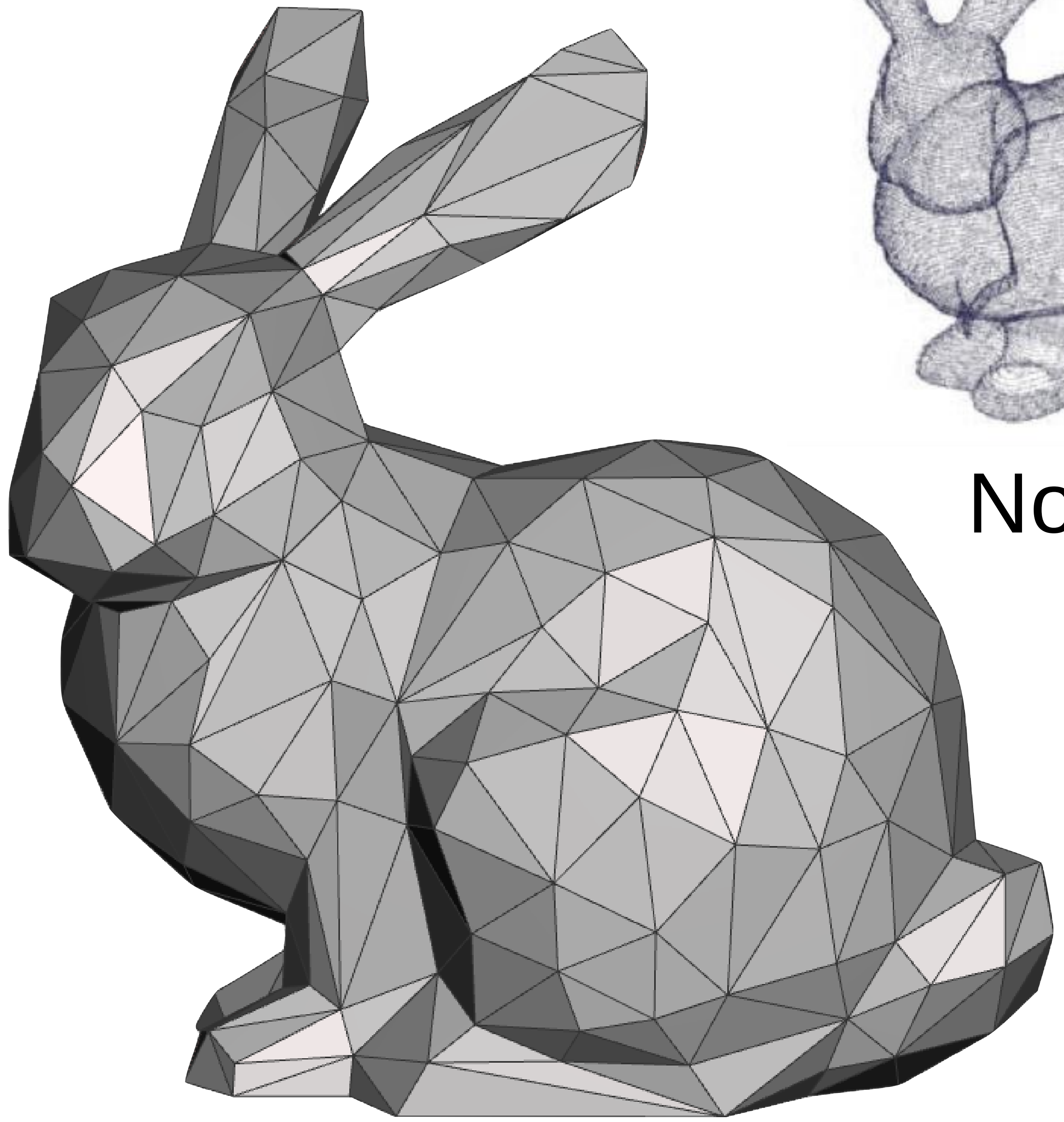


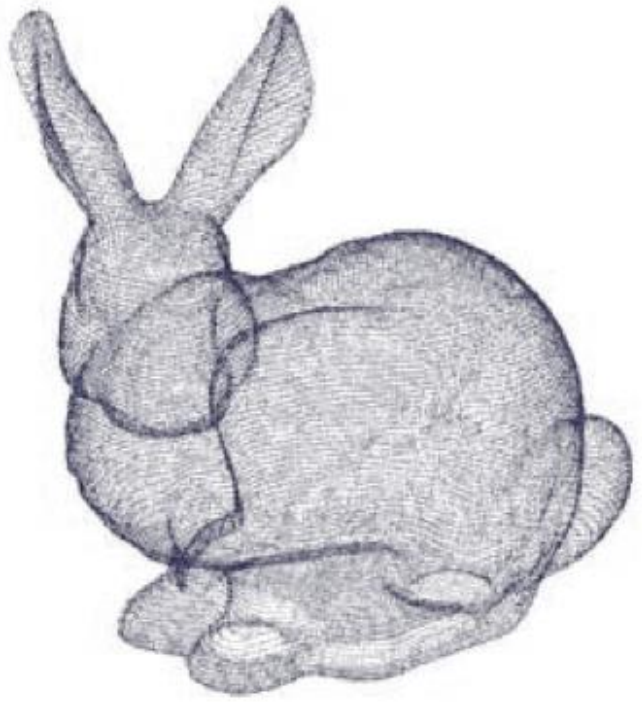
- (x, y, z): Spatial location
- (θ, ϕ): Viewing direction
- (r, g, b): Output color
- σ: Output density

F_Ω
MLP
9 layers,
256 channels

What kind of a 3D representation is this?

It is not a Mesh





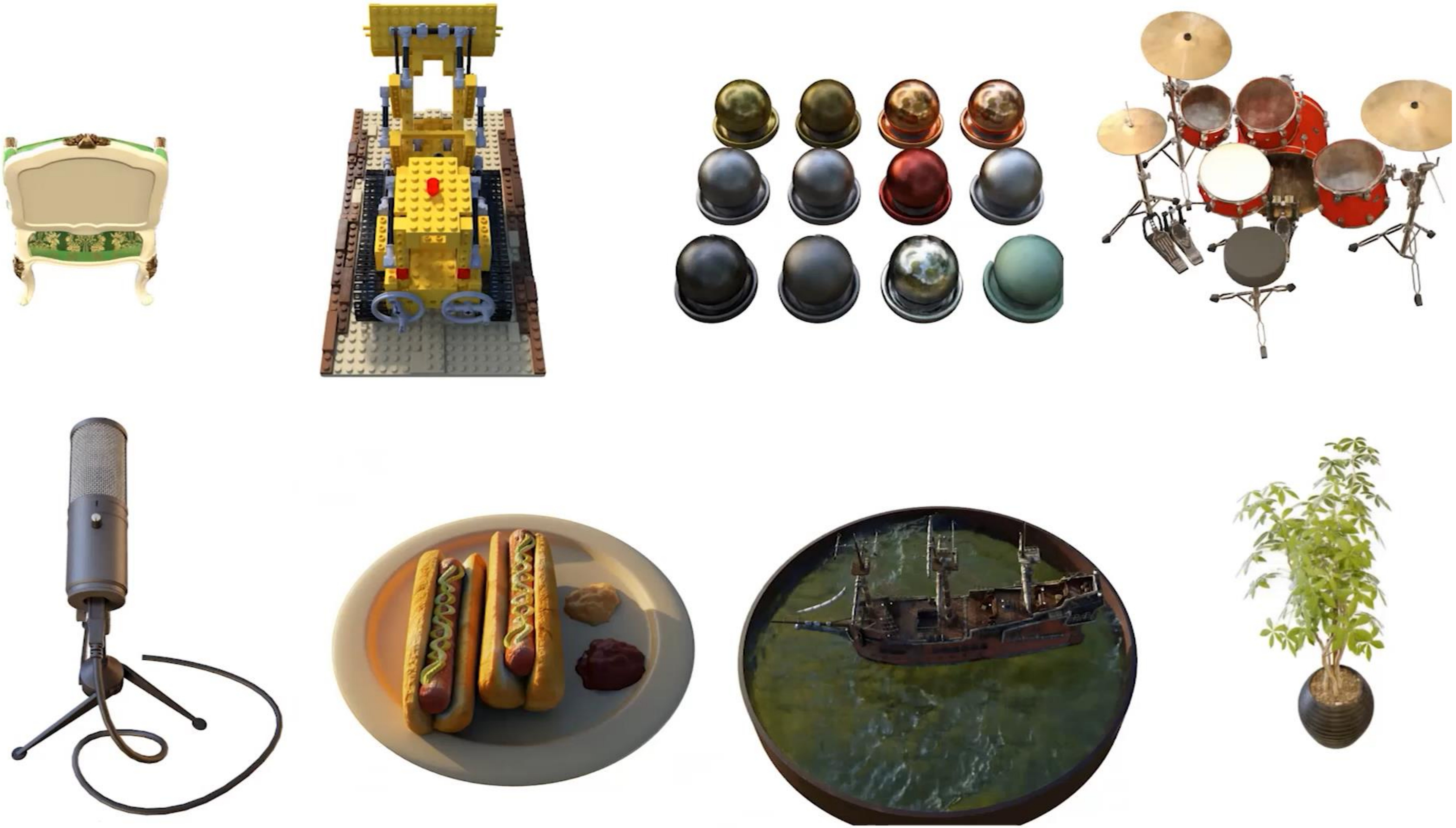
Not a point cloud either

✨🧊✨

It is volumetric

It's *continuous* voxels made of shiny transparent cubes

What is the problem that is being solved?



Plenoptic Function

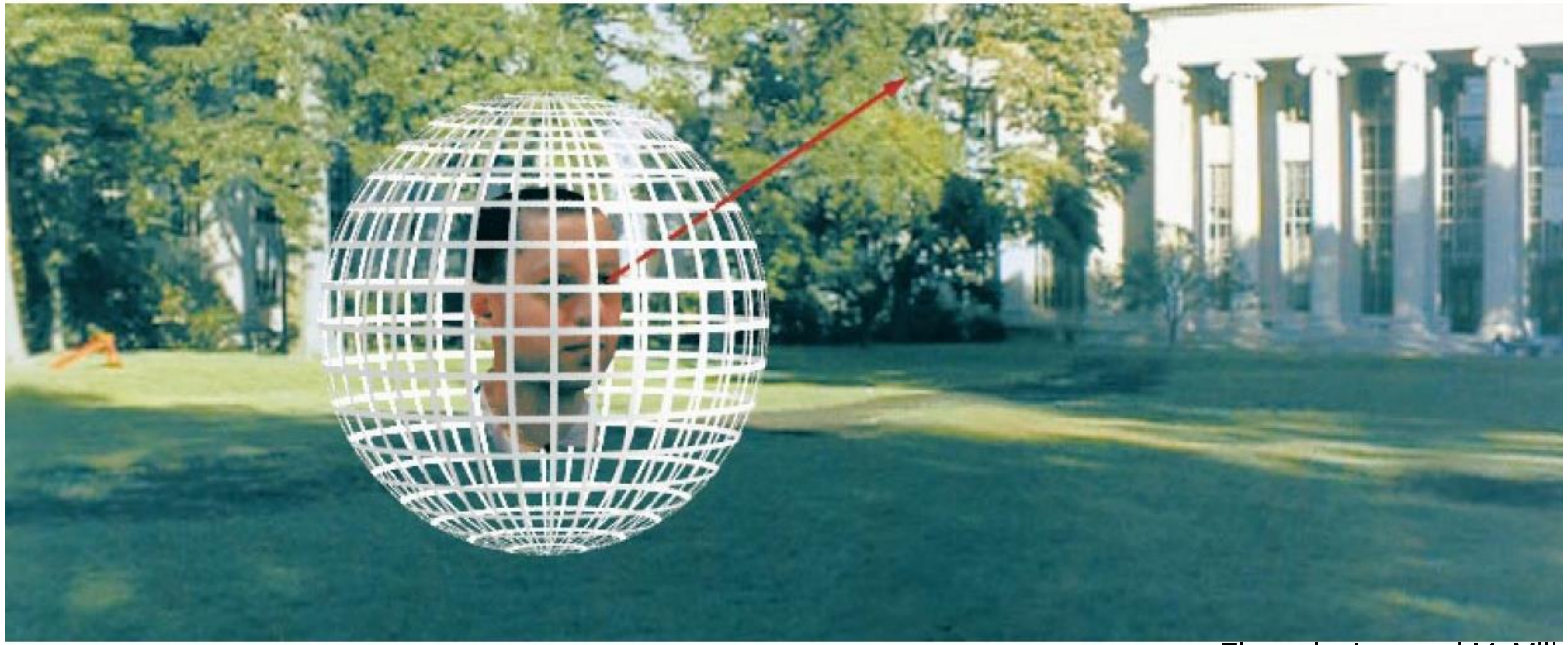


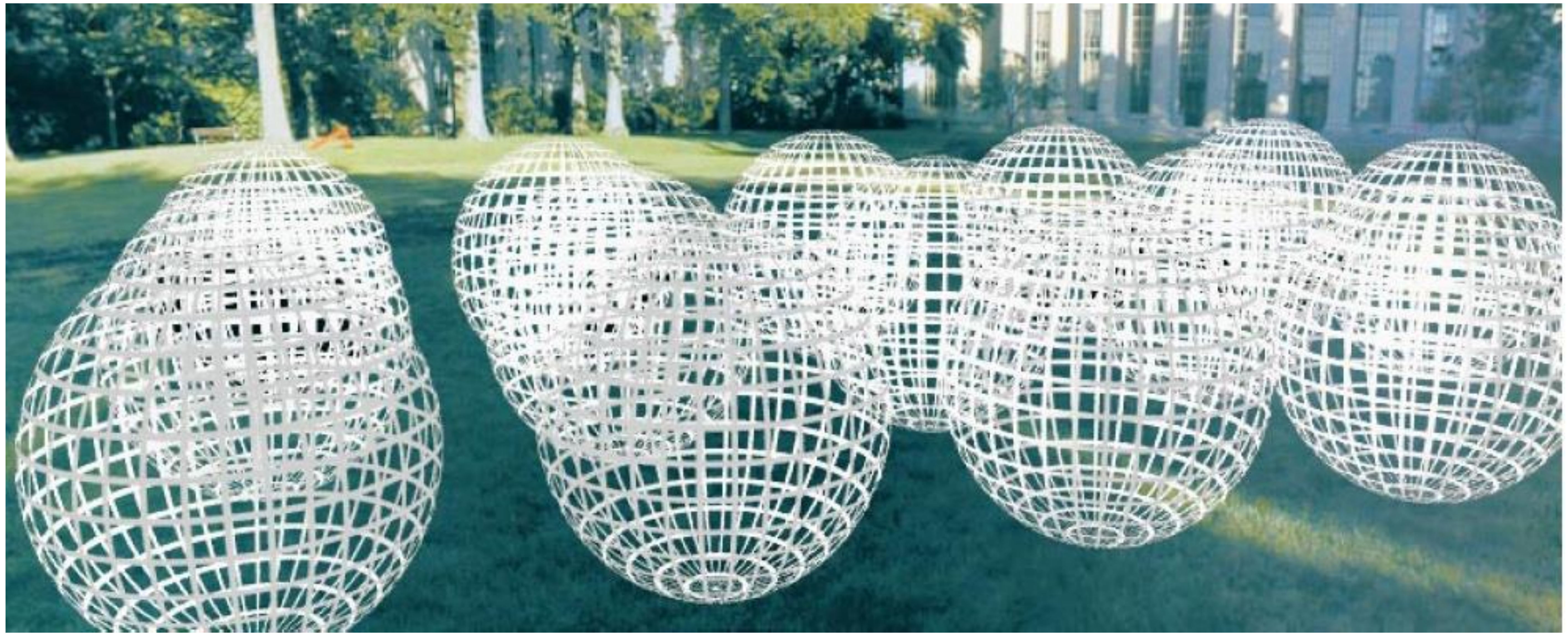
Figure by Leonard McMillan

Q: What is the set of all things that we can ever see?

A: The Plenoptic Function (Adelson & Bergen ‘91)

Slide credit:
Alyosha Efros

A holographic movie

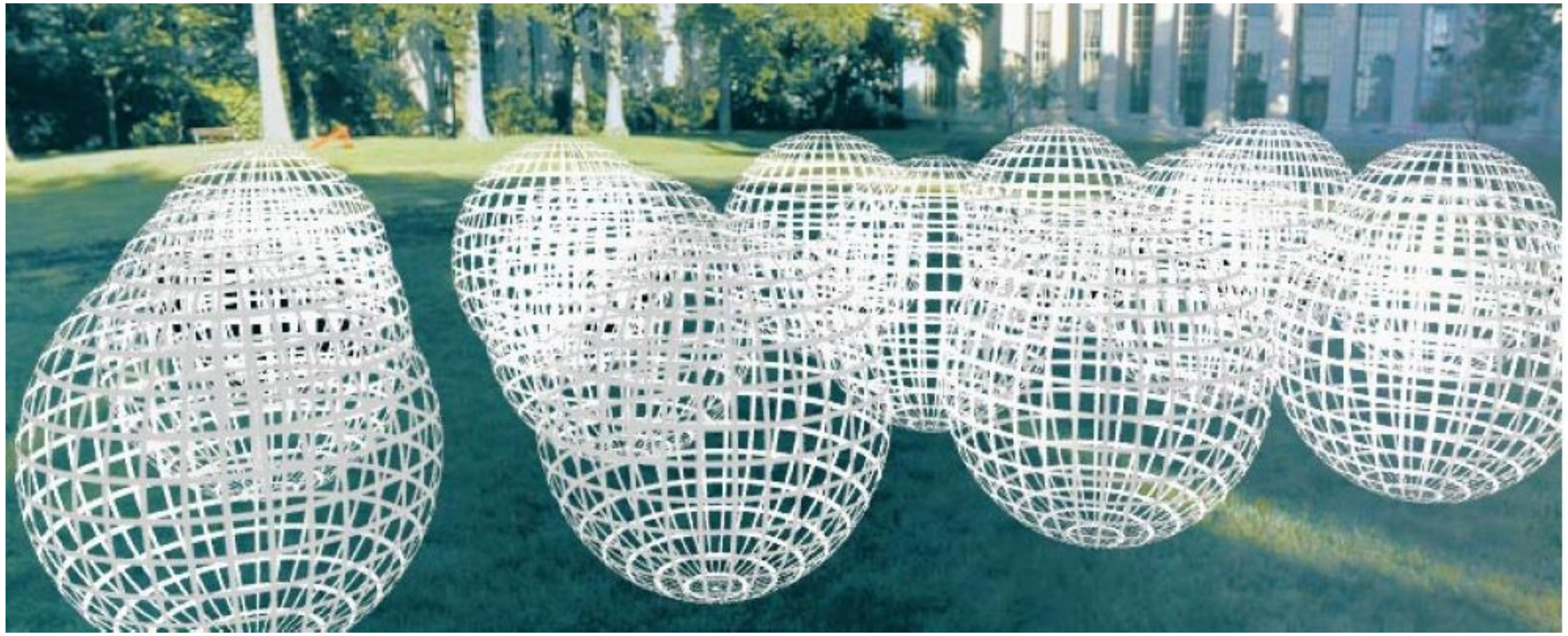


$$P(\theta,\phi,\lambda,t,V_X,V_Y,V_Z)$$

- is intensity of light
  - Seen from ANY position and direction
  - Over time
  - As a function of wavelength

Slides from Alyosha Efros

The plenoptic function

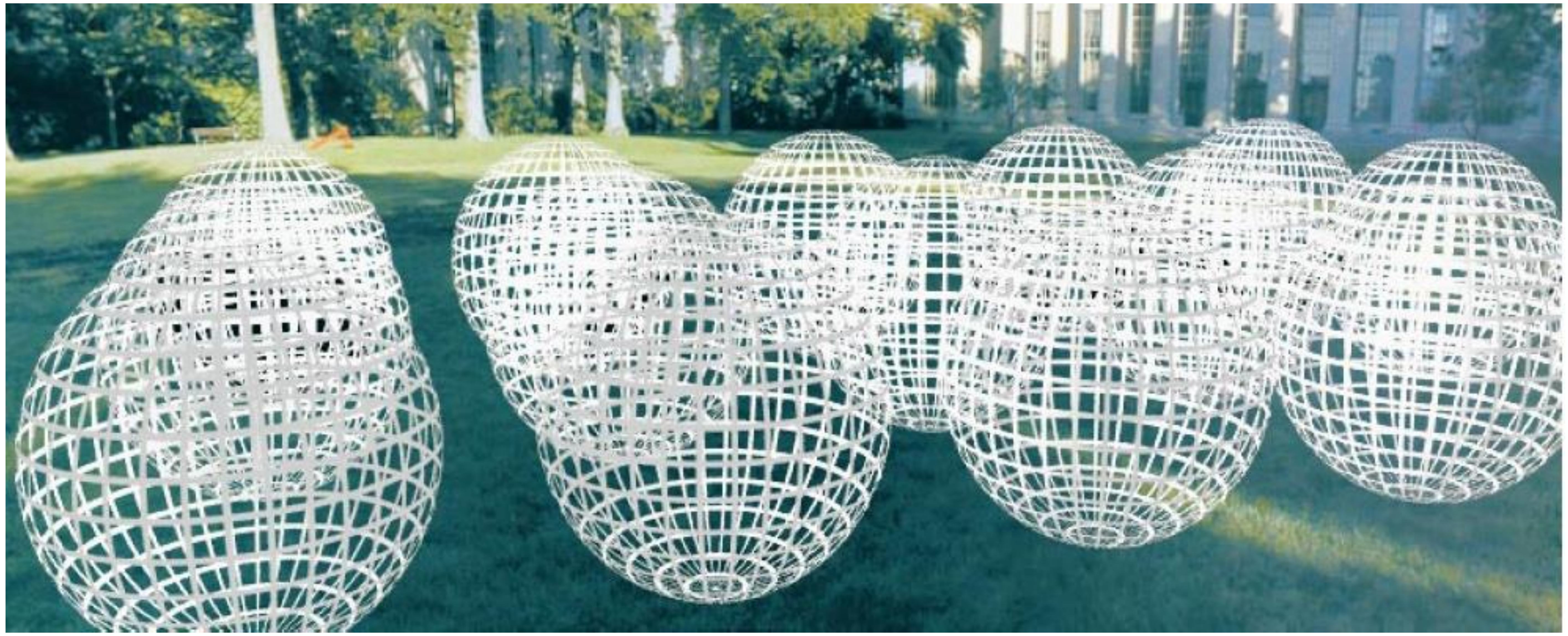


$$P(\theta,\phi,\lambda,t,V_X,V_Y,V_Z)$$

7D function, that can reconstruct every position & direction,
at every moment, at every wavelength
= it recreates the entirety of our visual reality!

Slides from Alyosha Efros

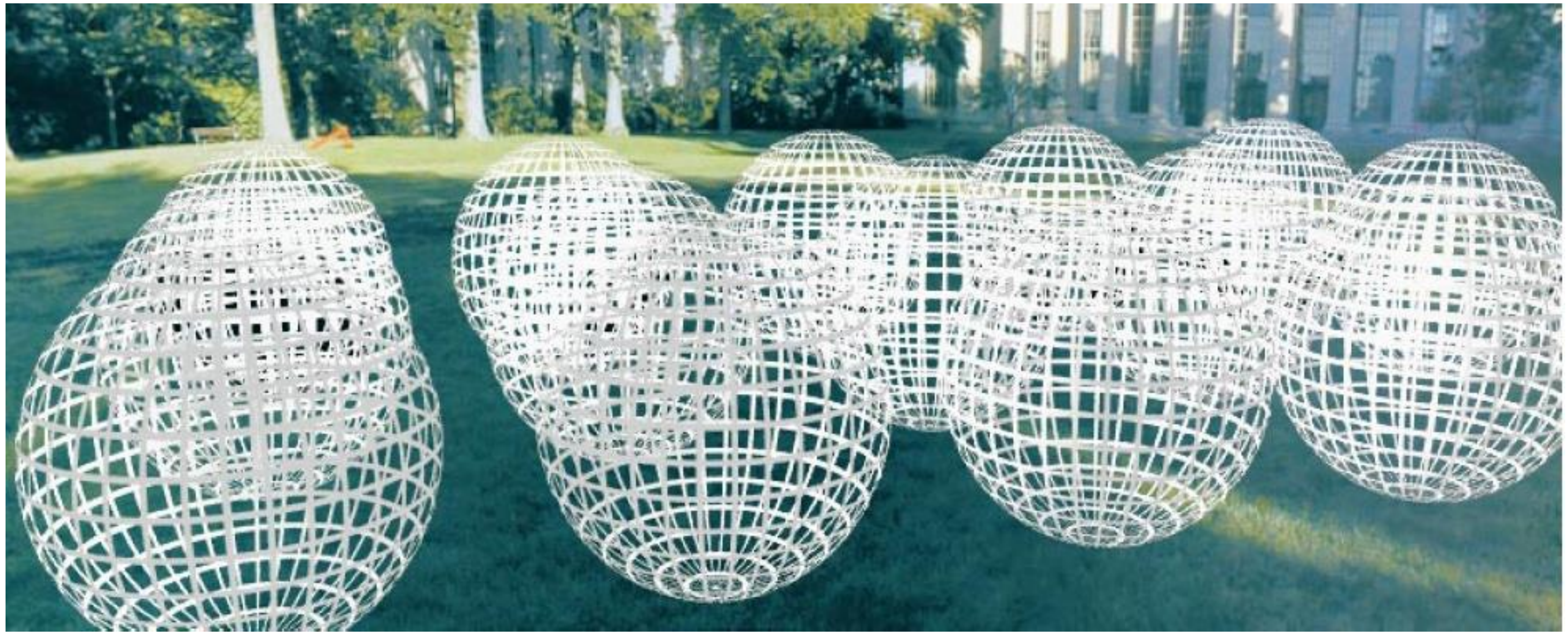
Goal: Plenoptic Function from a set of images



- Objective: Recreate the visual reality
- All about recovering photorealistic pixels, not about recording 3D point or surfaces
  —Image Based Rendering

aka **Novel View Synthesis**

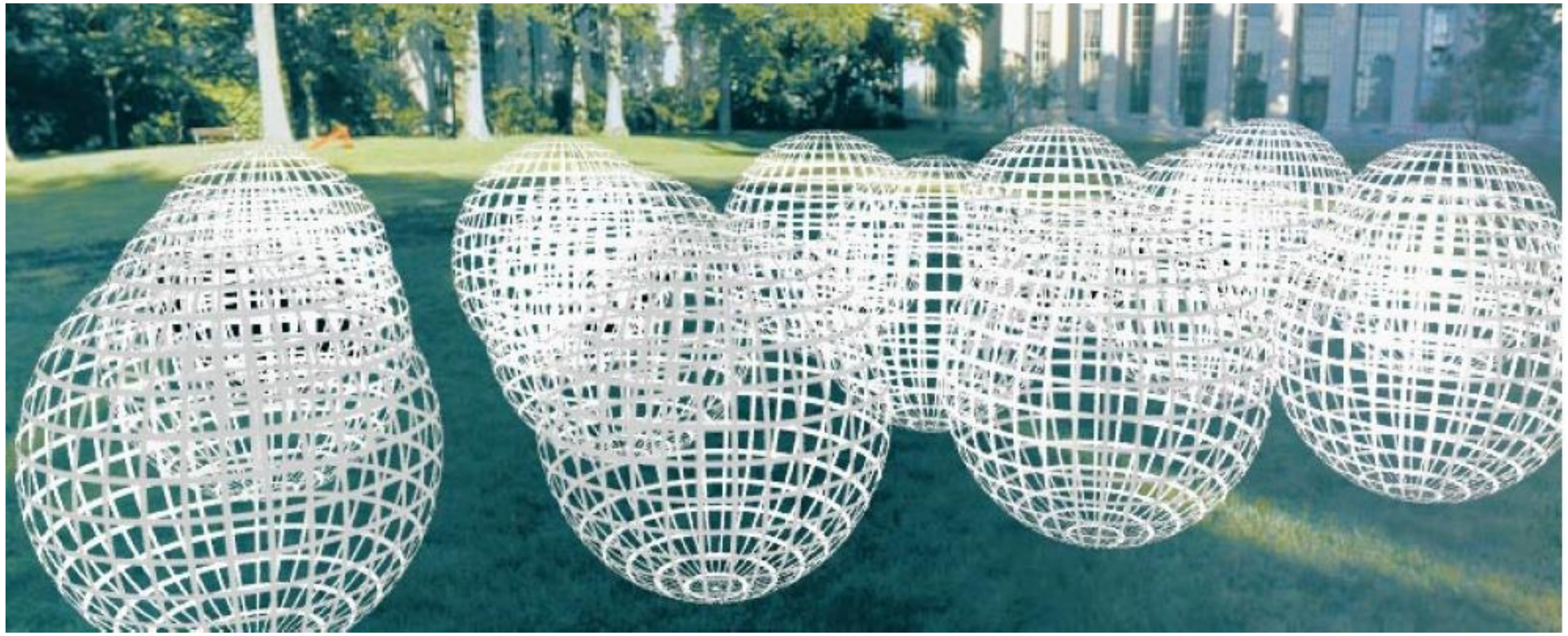
Goal: Plenoptic Function from a set of images



It is a conceptual device

Adelson & Bergen do not discuss how to solve this

Plenoptic Function



7D function:
2 – direction
1 – wavelength
1 – time
3 – location

$$P(\theta,\phi,\lambda,t,V_X,V_Y,V_Z) \longrightarrow P(\theta,\phi,V_X,V_Y,V_Z)$$

Look familiar 🙂?

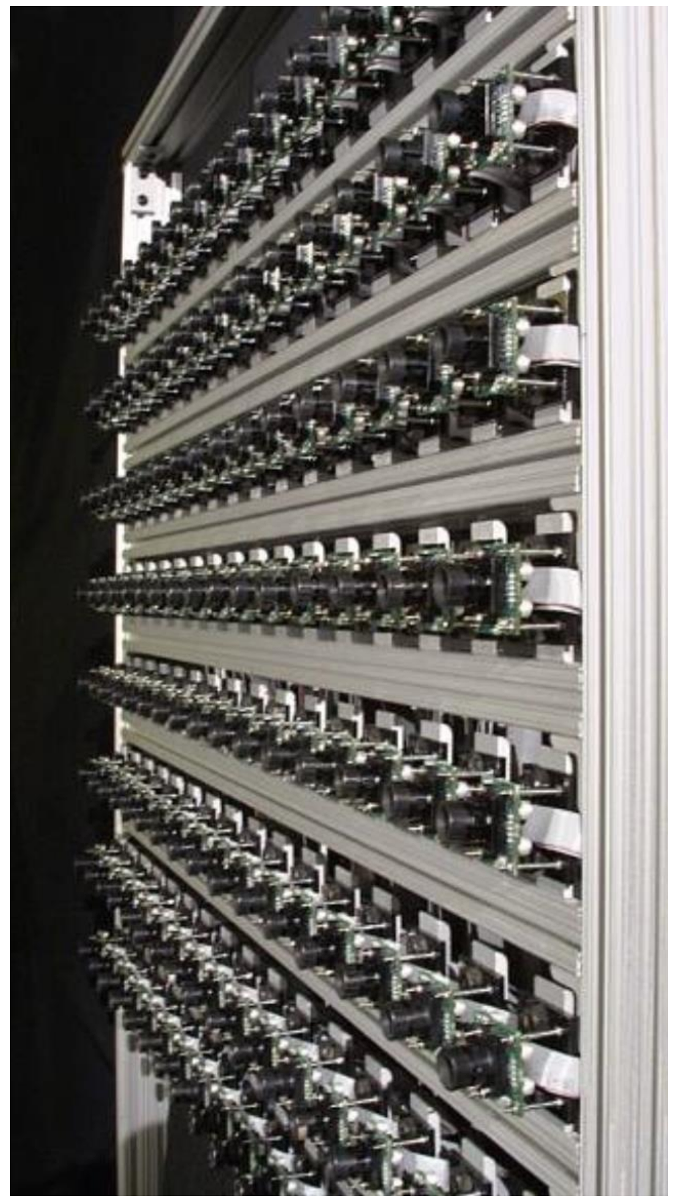
Let’s simplify:

1. Remove the time
2. Remove the wavelength & let the function output RGB colors

Levoy and Hanrahan, SIGGRAPH 1996
Gortler et al. SIGGRAPH 1996

Lightfield / Lumigraph

- Previous approaches for modeling the Plenoptic Function
- Take a lot of pictures from many views
- Interpolate the rays to render a novel view



Stanford Gantry
128 cameras

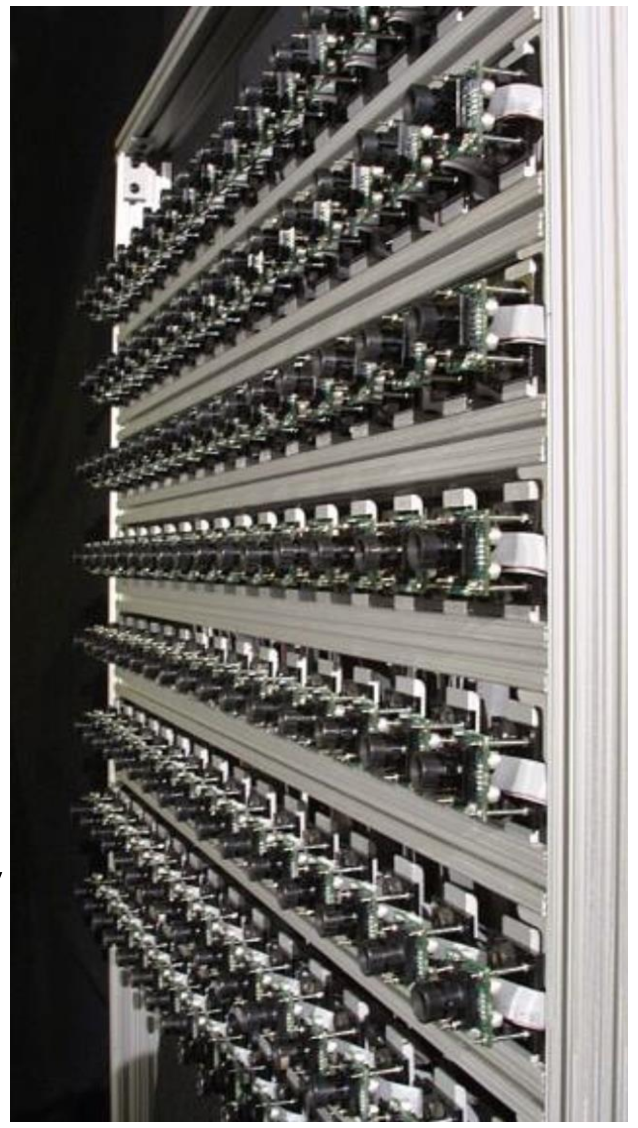
Lytro camera

Levoy and Hanrahan, SIGGRAPH 1996
Gortler et al. SIGGRAPH 1996

Lightfield / Lumigraph

- Previous approaches for modeling the Plenoptic Function
- Take a lot of pictures from many views
- Interpolate the rays to render a novel view

Stanford Gantry
128 cameras



Lytro camera

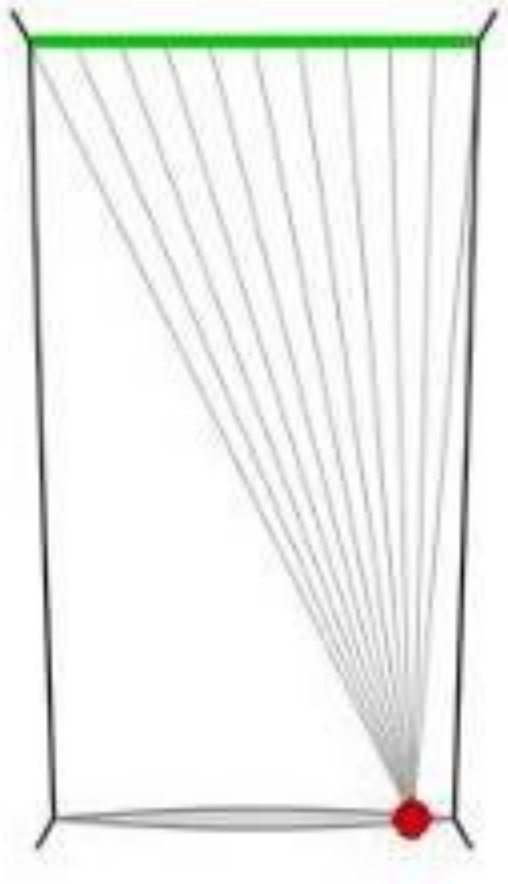
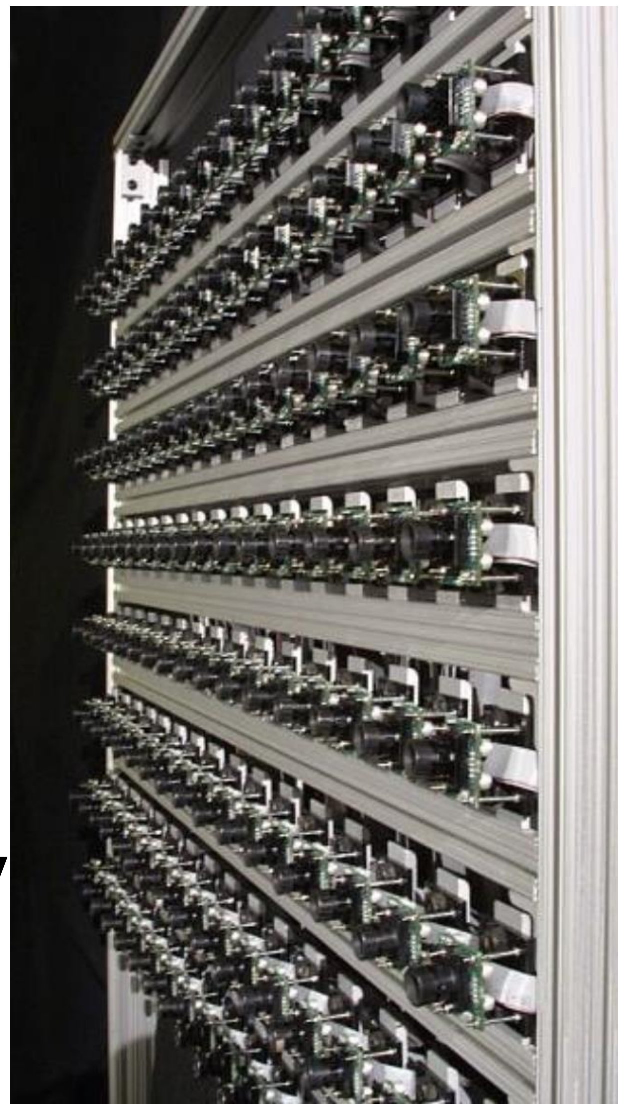


Figure from Marc Levoy

Levoy and Hanrahan, SIGGRAPH 1996
Gortler et al. SIGGRAPH 1996

Lightfield / Lumigraph

- Previous approaches for modeling the Plenoptic Function
- Take a lot of pictures from many views
- Interpolate the rays to render a novel view



Stanford Gantry
128 cameras

Lytro camera

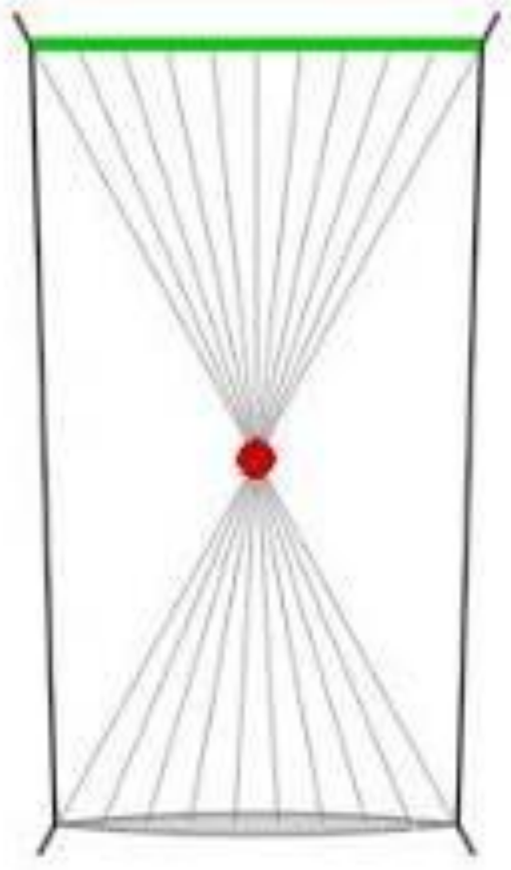
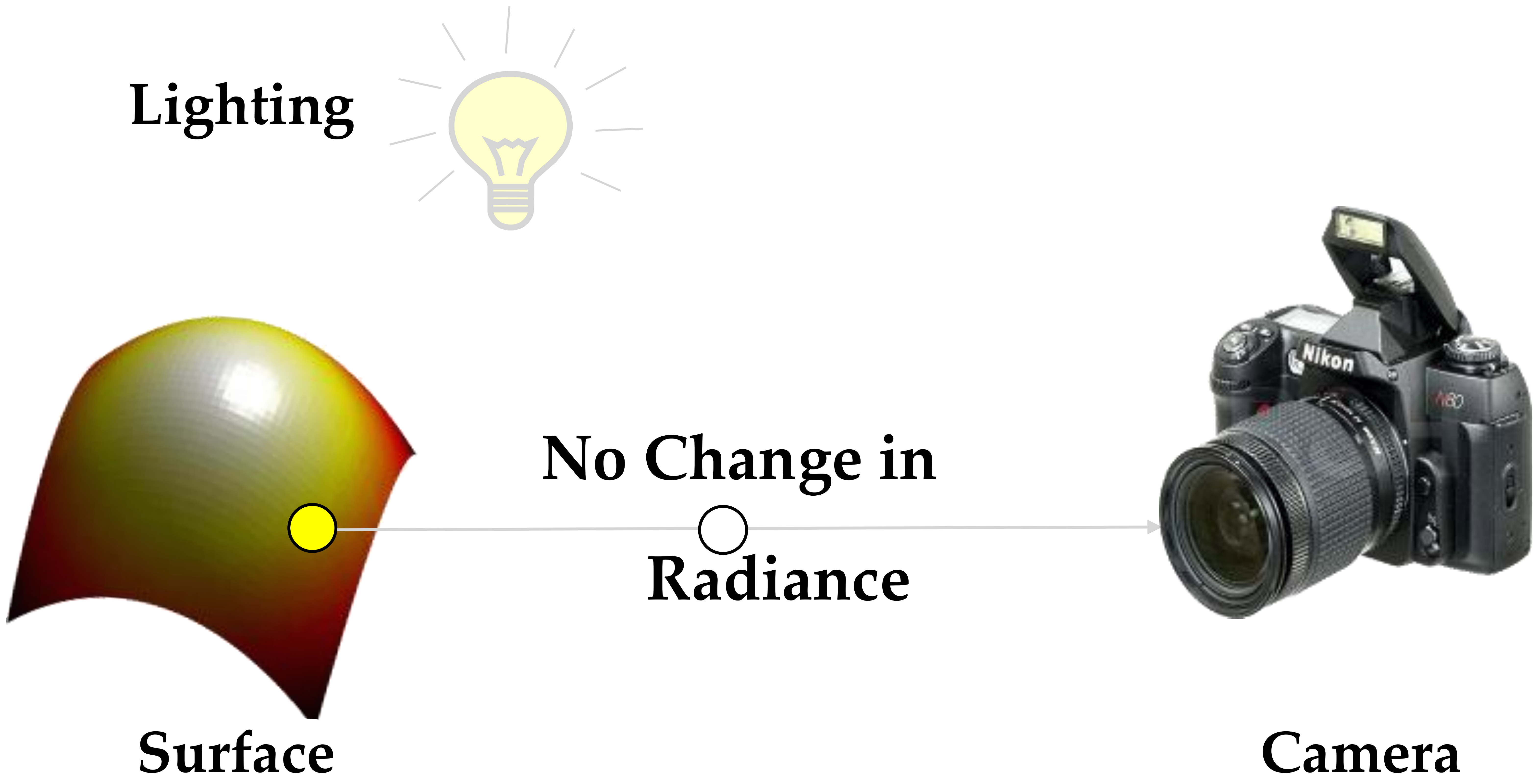


Figure from Marc Levoy

Big Assumption: a ray does not change color

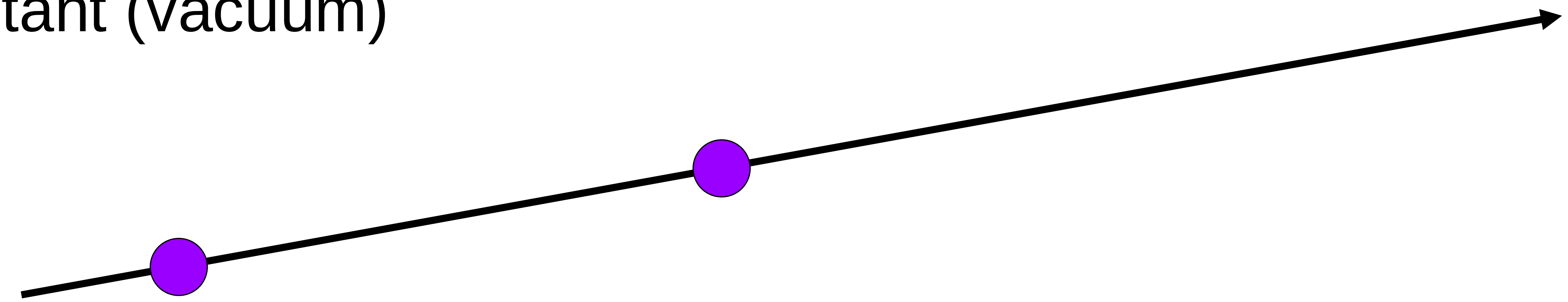


True if there is no occlusion or fog

With this assumption: Ray Reuse

Infinite line

- Assume light is constant (vacuum)



The 5D function

- 3D position
- 2D direction

is now 4D

- 2D direction
- 2D position
- non-dispersive medium

Slide by Rick Szeliski and Michael Cohen

Levoy and Hanrahan, SIGGRAPH 1996
Gortler et al. SIGGRAPH 1996

Ray Reuse Assumption

Because of this it only models the plenoptic surface:

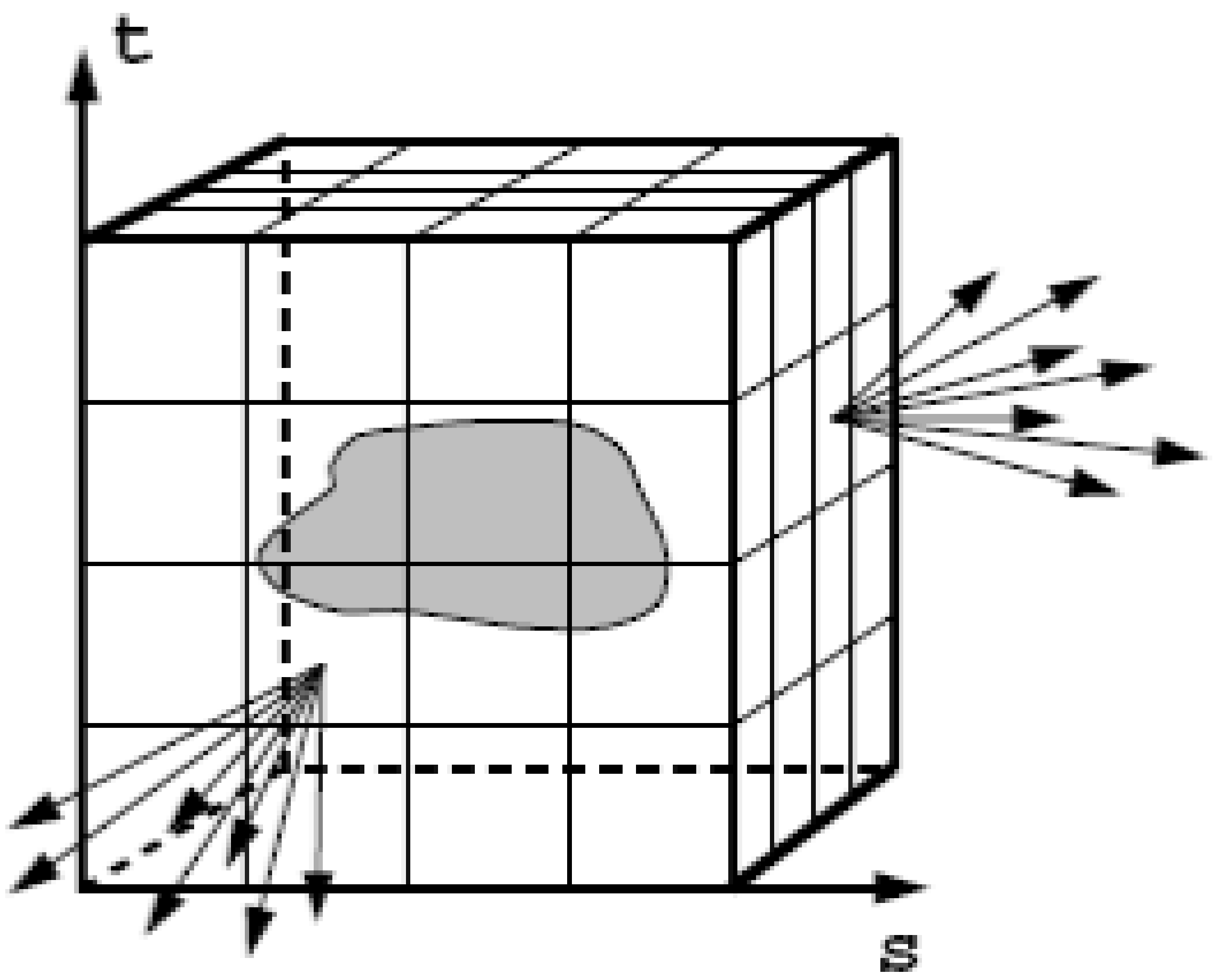
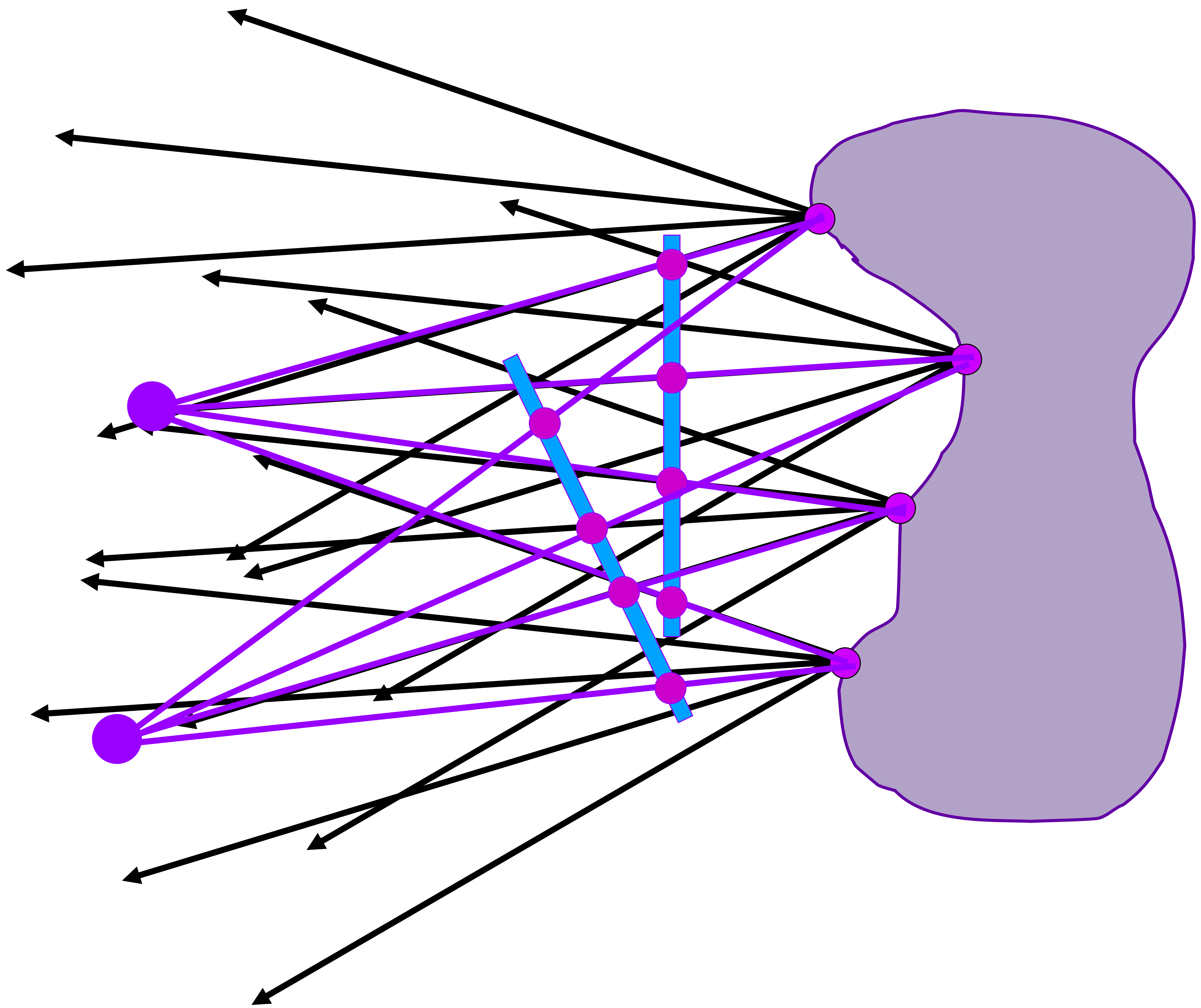


Figure 1: The surface of a cube holds all the radiance information due to the enclosed object.

It's like

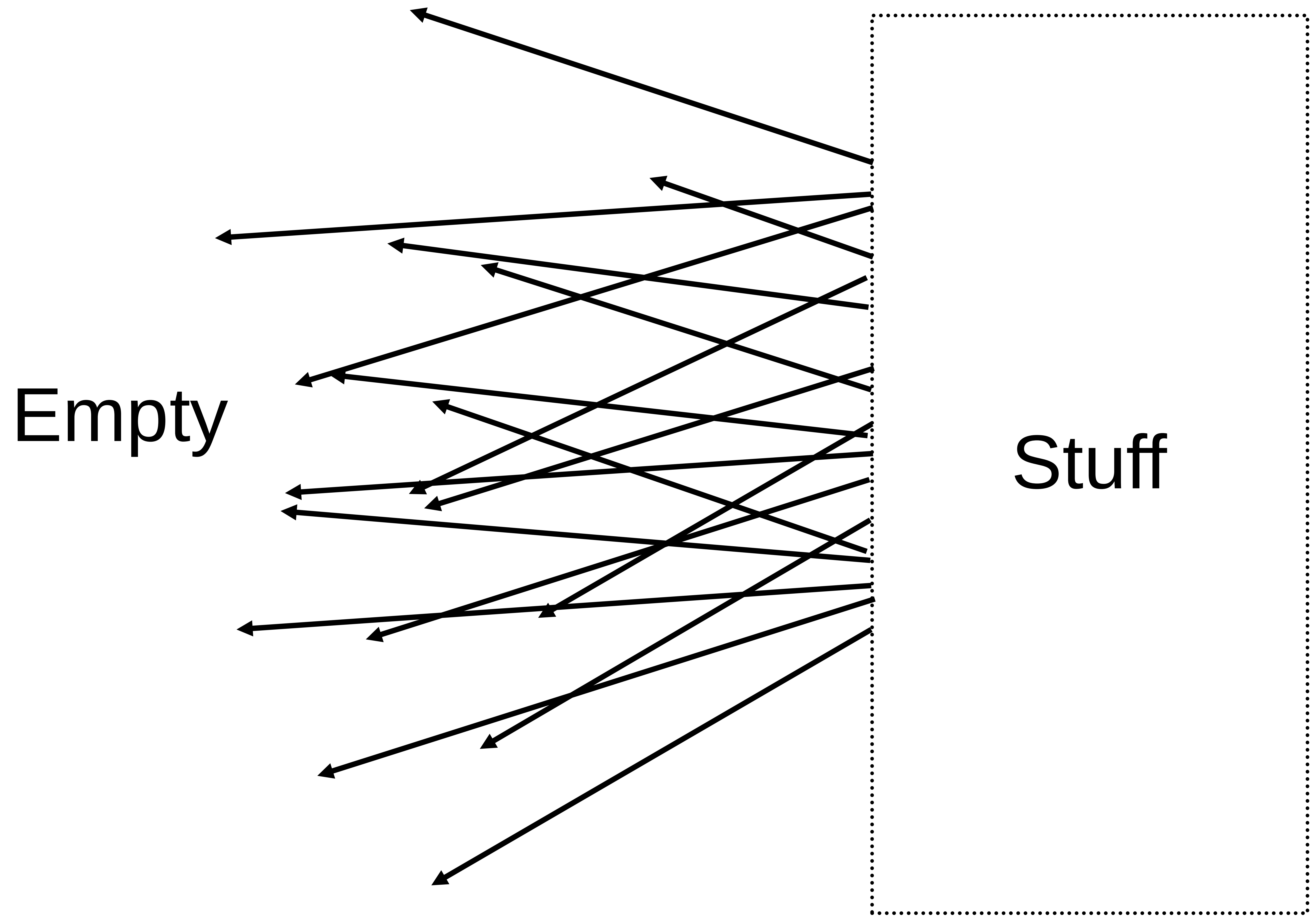
Synthesizing novel views



Slide by Rick Szeliski and Michael Cohen

Lumigraph / Lightfield

Outside convex space

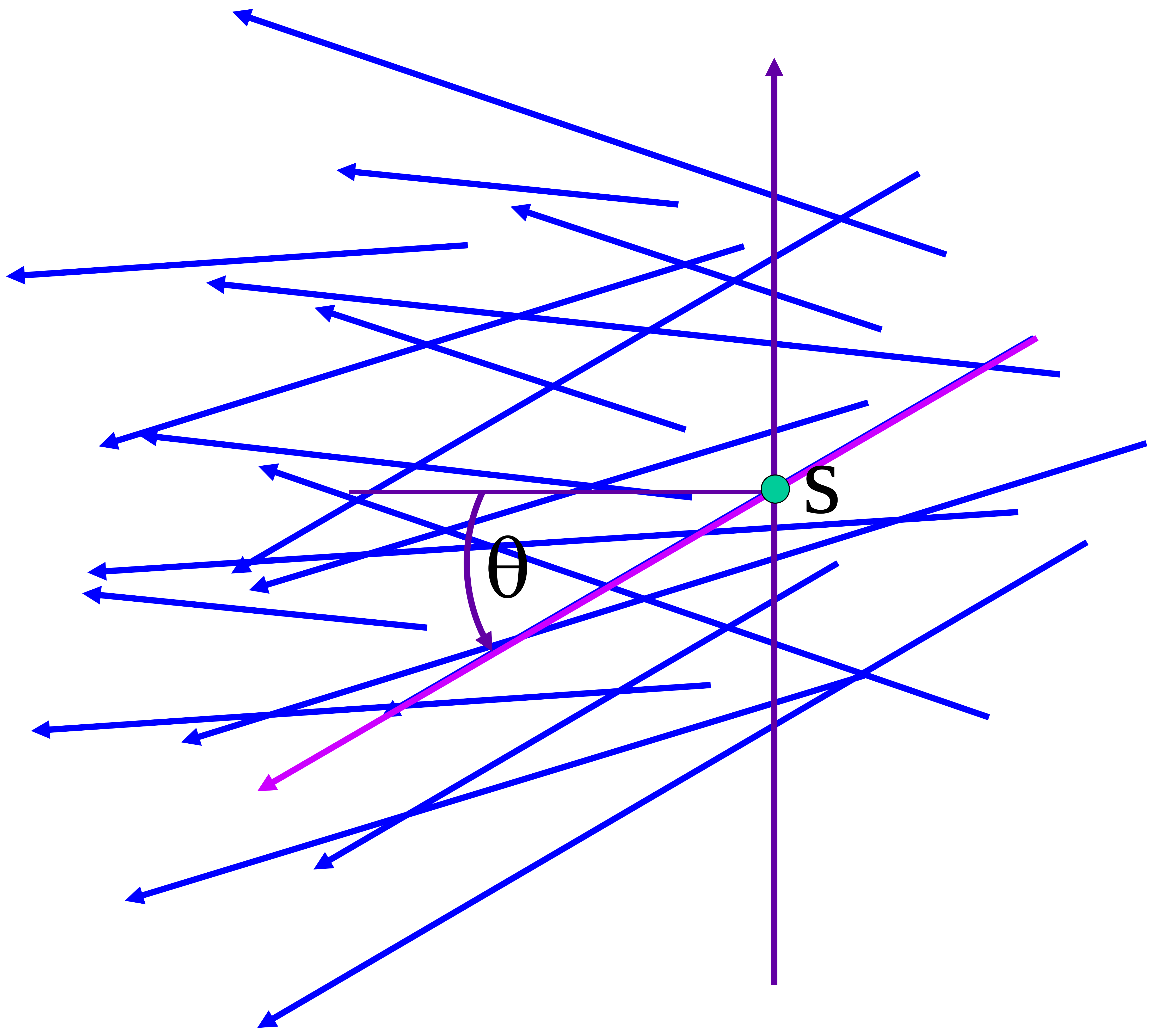


4D

Lumigraph - Organization

2D position

2D direction

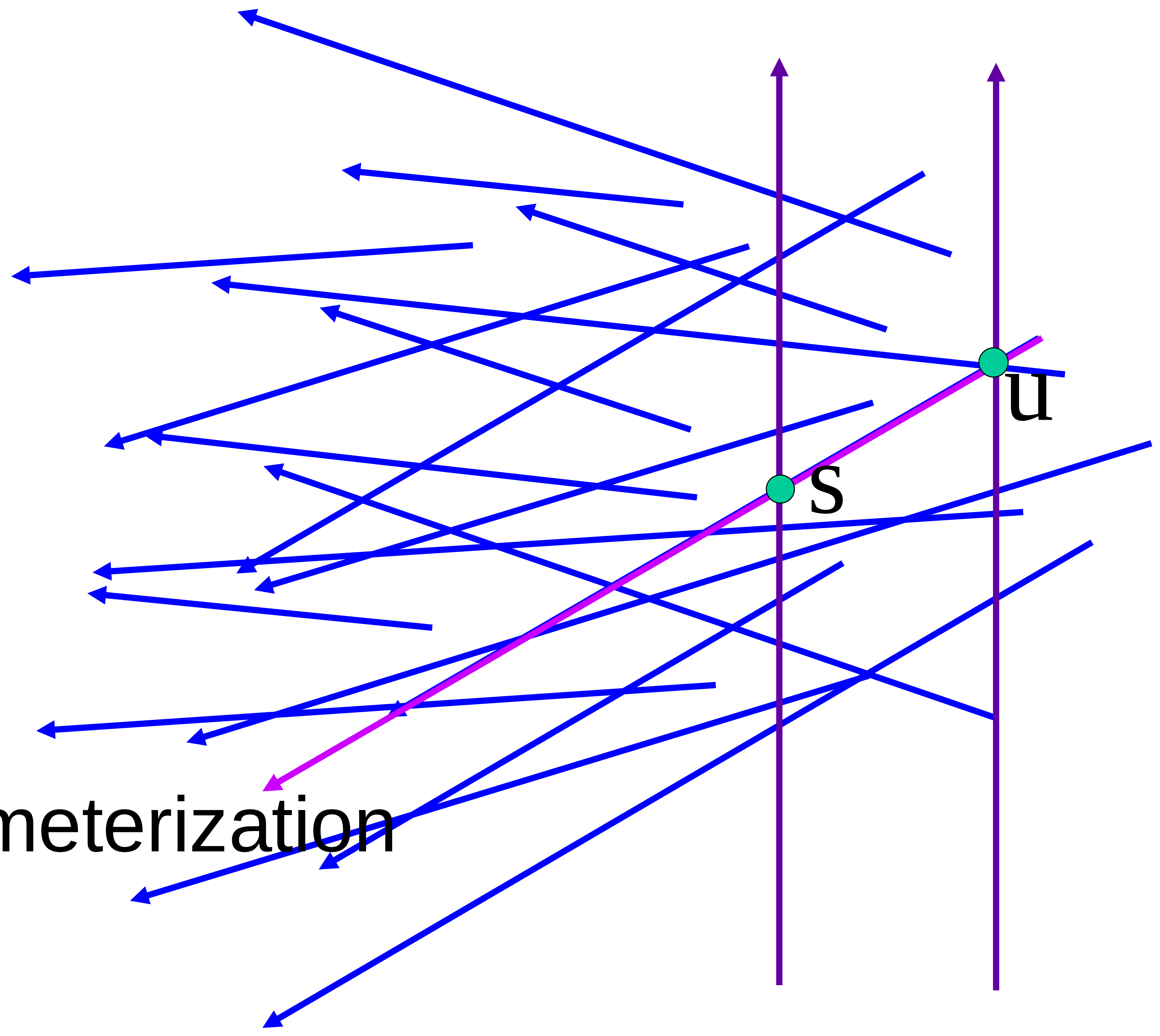


Slide by Rick Szeliski and Michael Cohen

Lumigraph - Organization

2D position

2D position



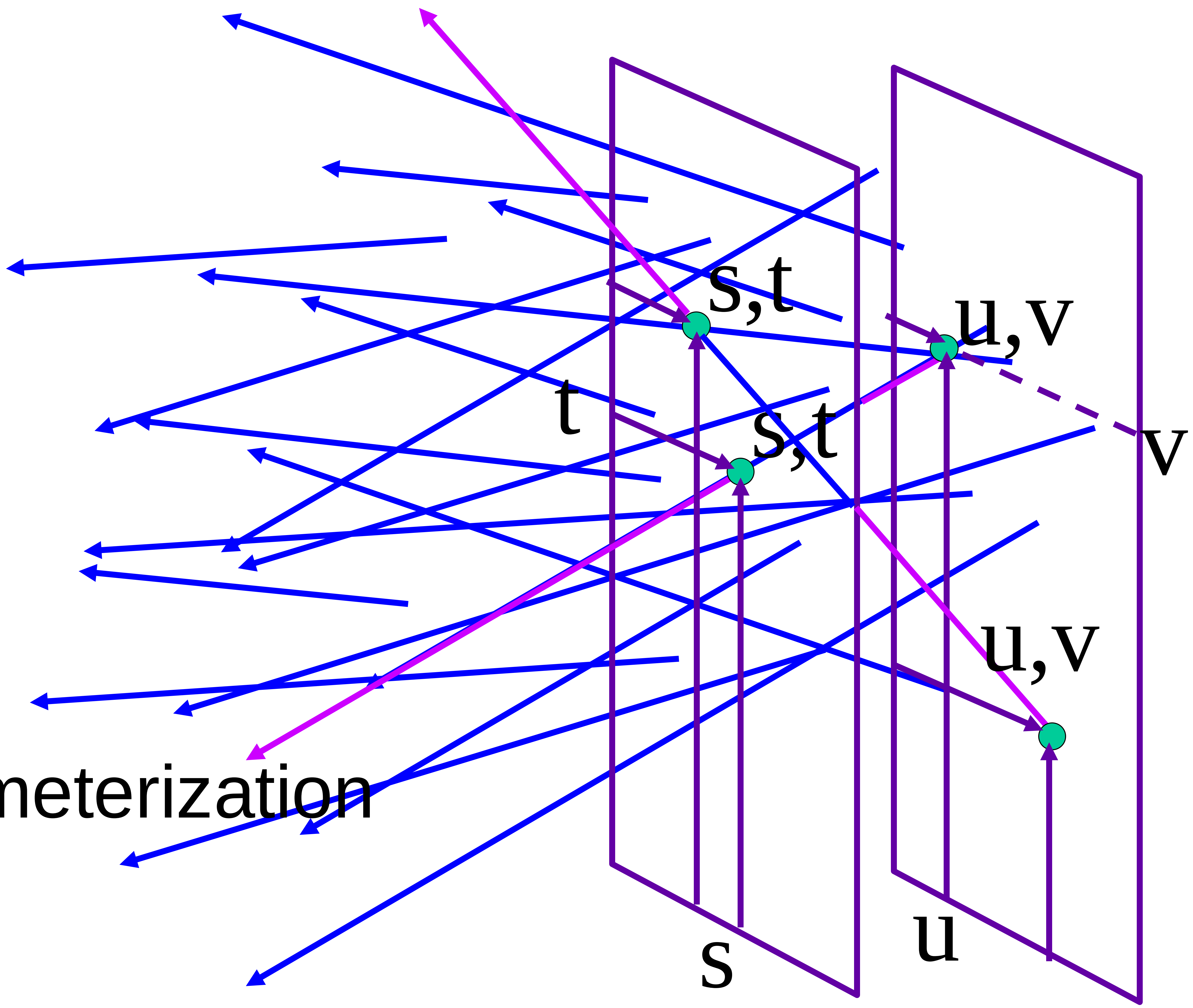
2 plane parameterization

Lumigraph - Organization

2D position

2D position

2 plane parameterization



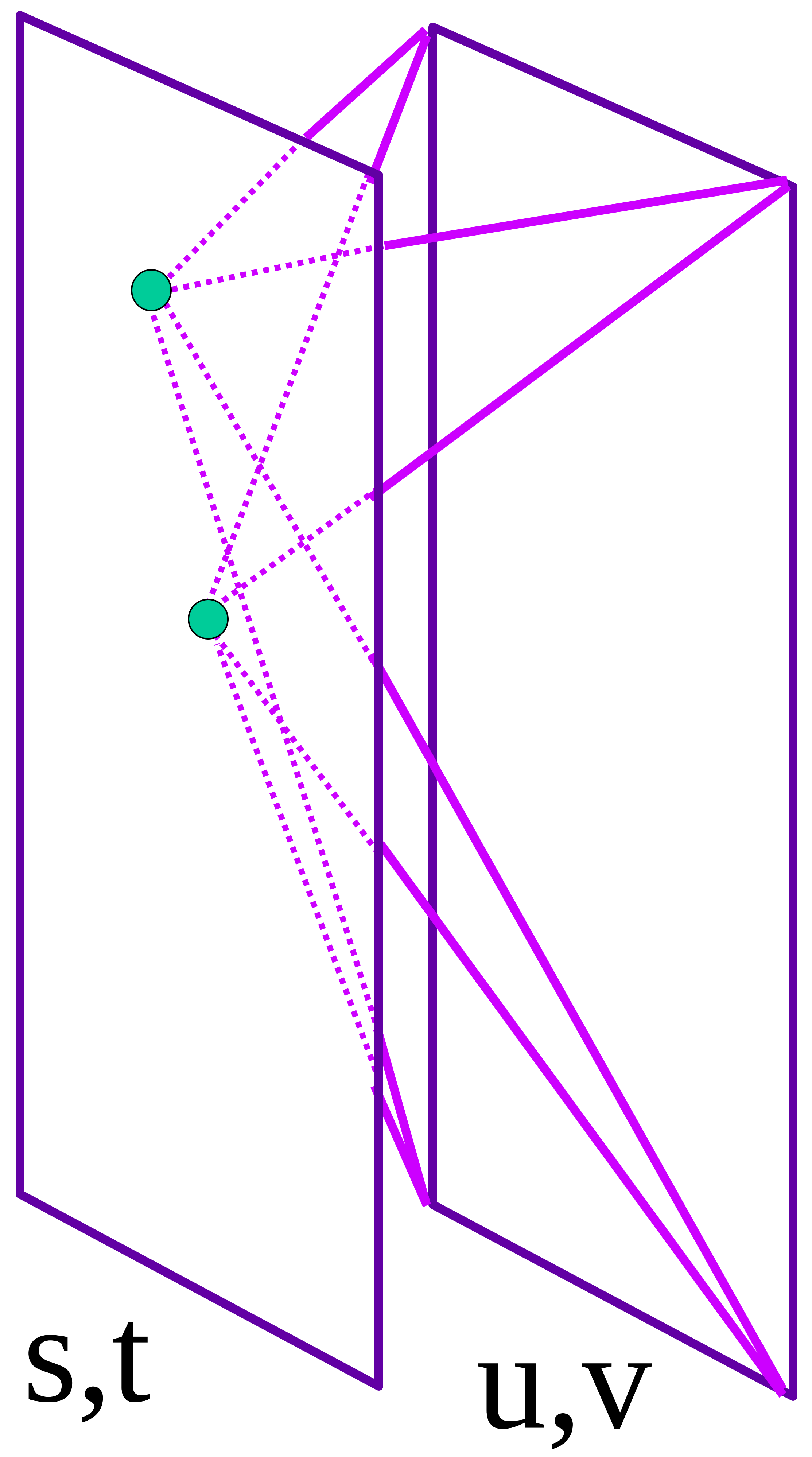
Slide by Rick Szeliski and Michael Cohen

Lumigraph - Organization

Hold s,t constant

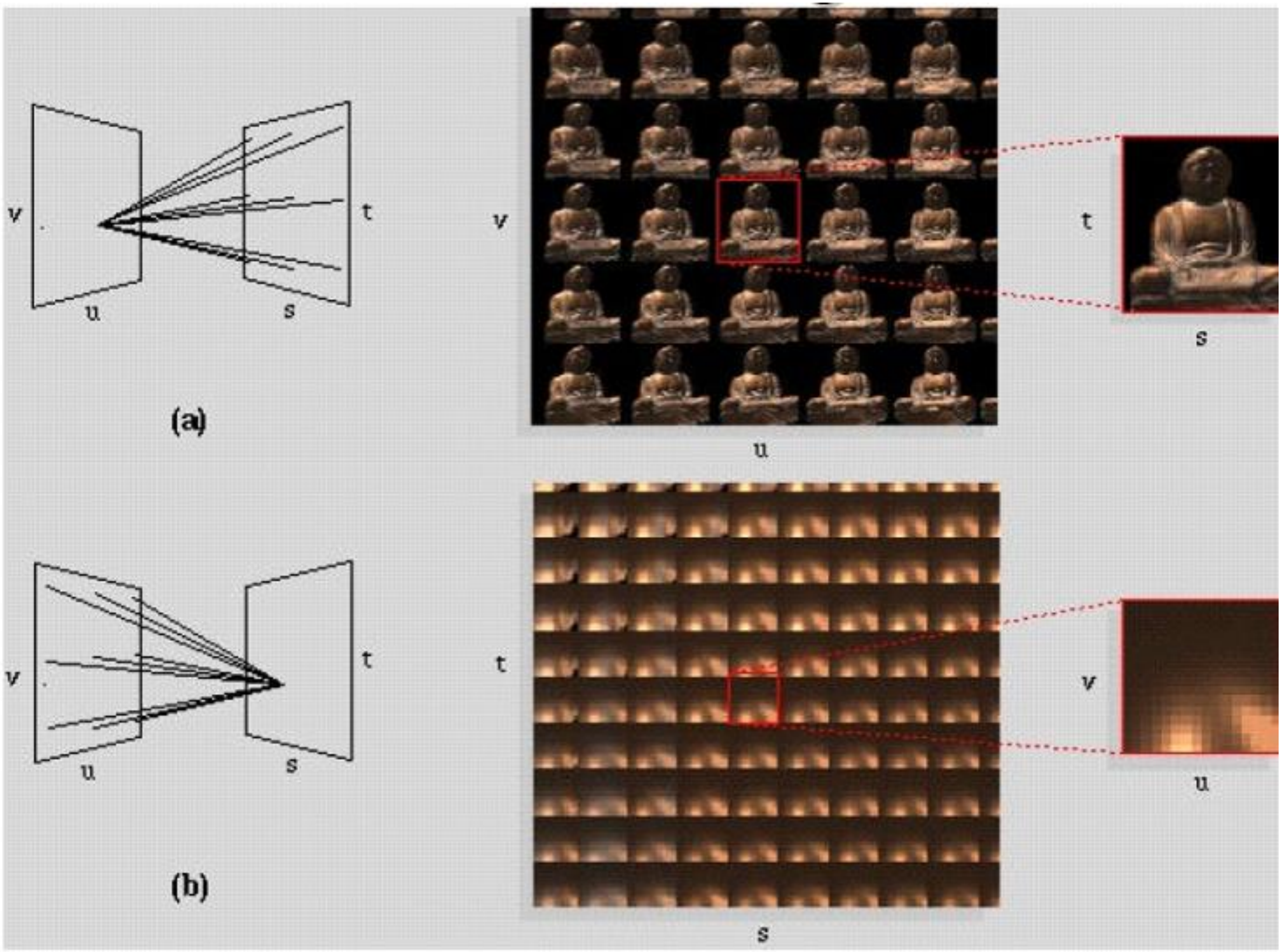
Let u,v vary

An image



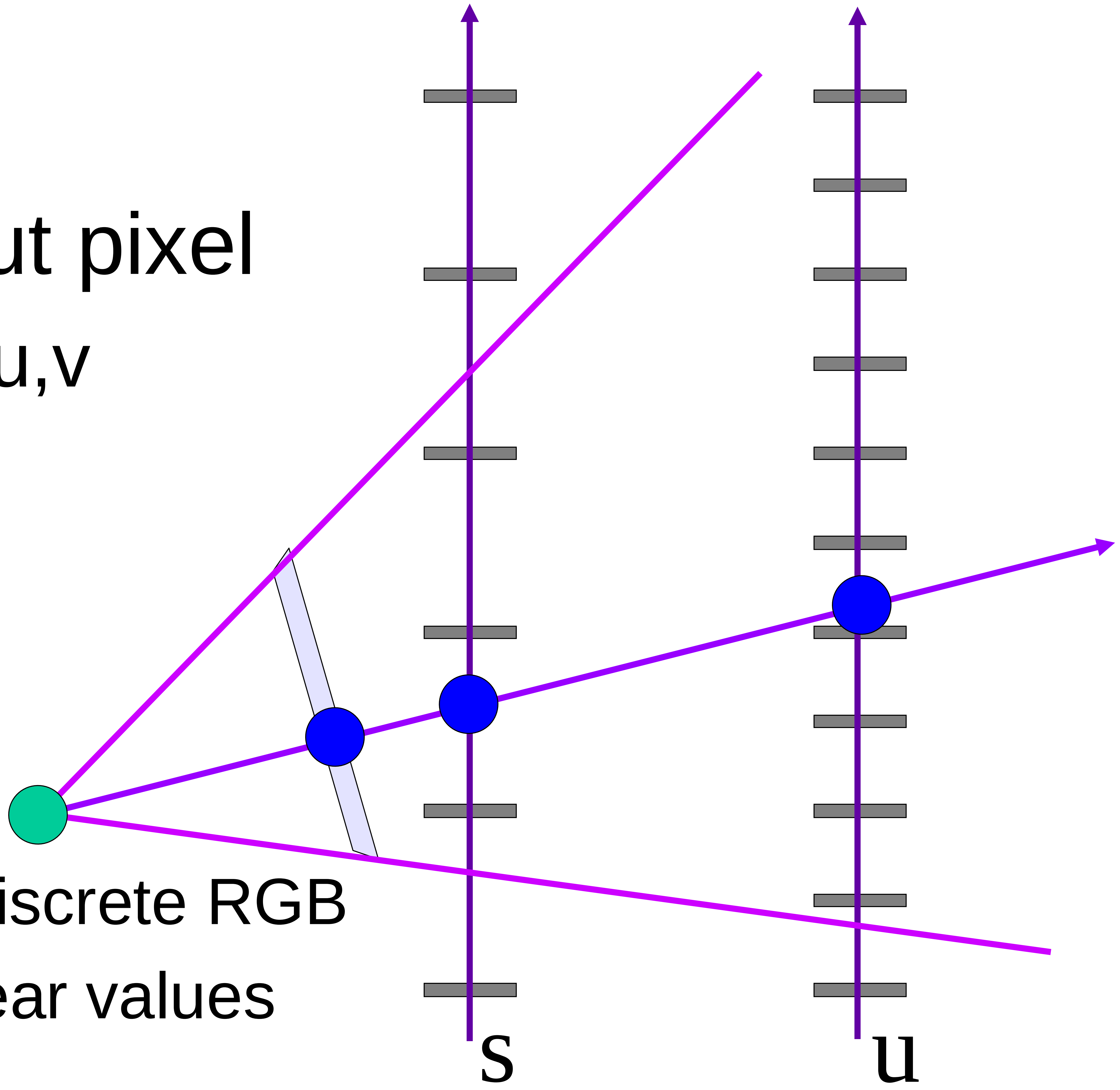
Slide by Rick Szeliski and Michael Cohen

Lumigraph / Lightfield



Novel View Synthesis

- For each output pixel
  - determine s,t,u,v
  - either
    - use closest discrete RGB
    - interpolate near values



Slide by Rick Szeliski and Michael Cohen

How NeRF models the Plenoptic Function

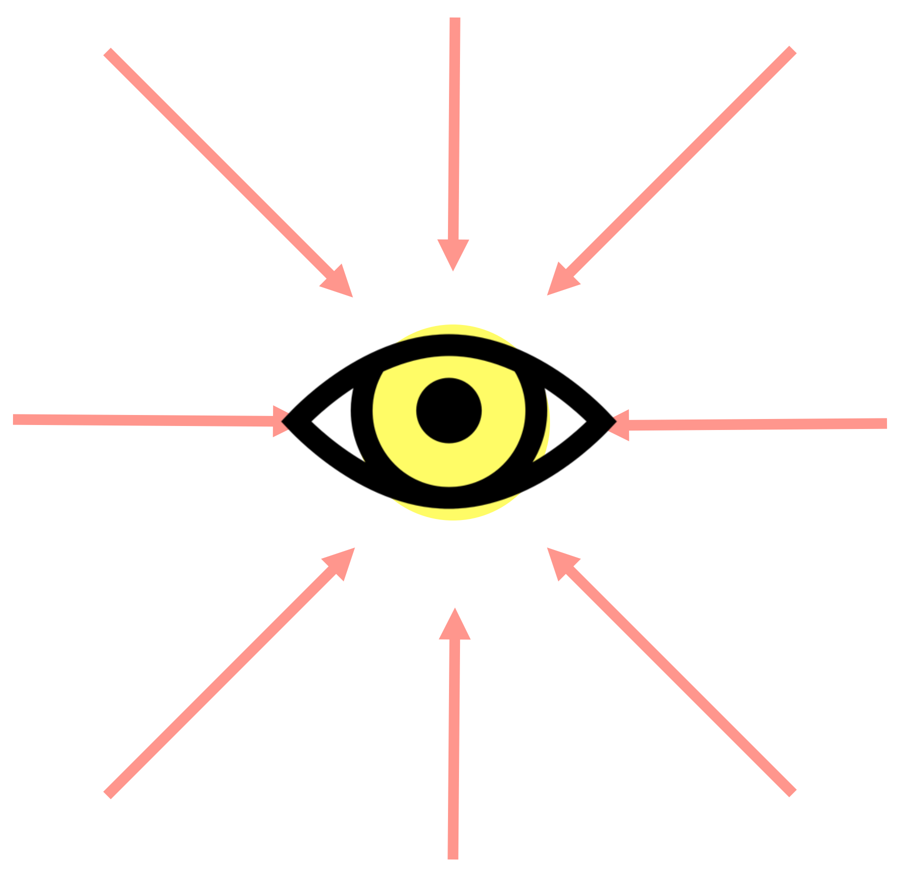
$$P(\theta,\phi,V_X,V_Y,V_Z)$$

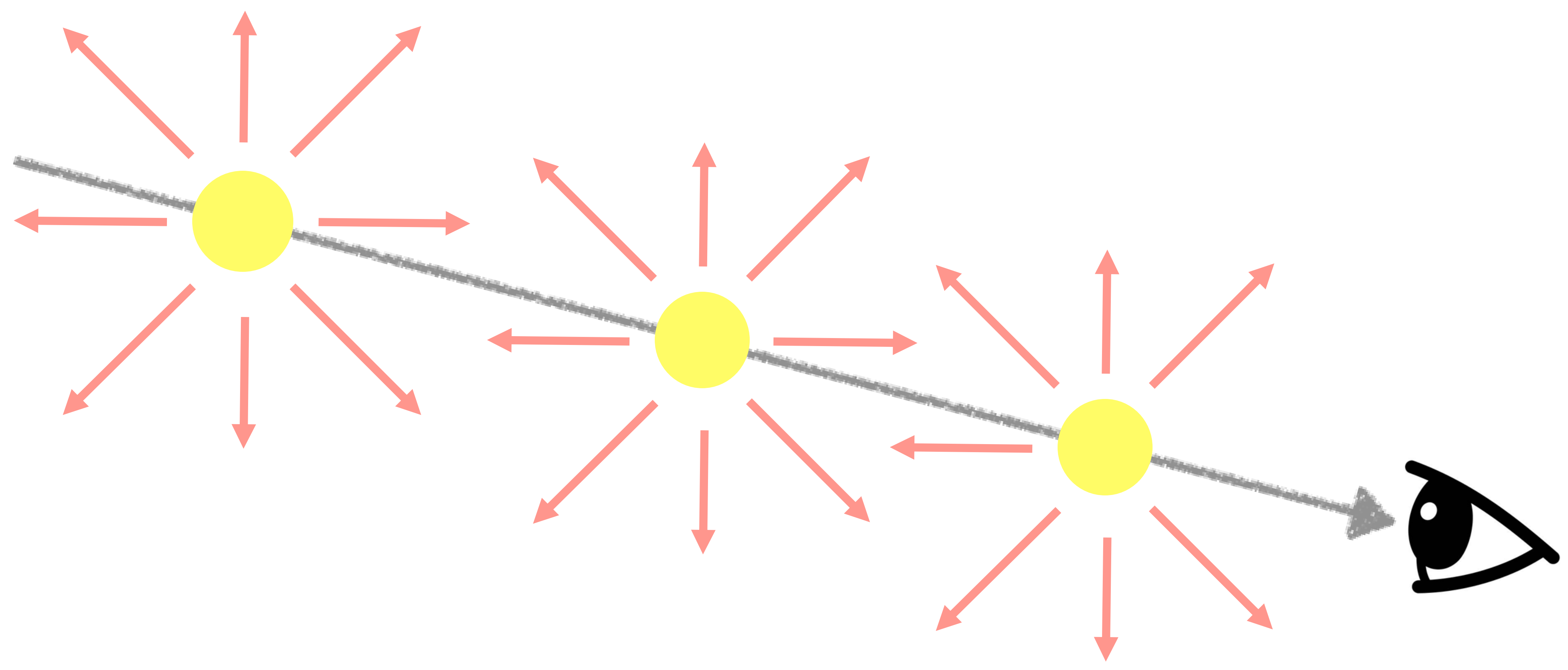
Look familiar 🙂?

NeRF takes the same input as the Plenoptic Function!

It allows rays to change color. Hence we can fly into the glass bowl (if we had enough observation)

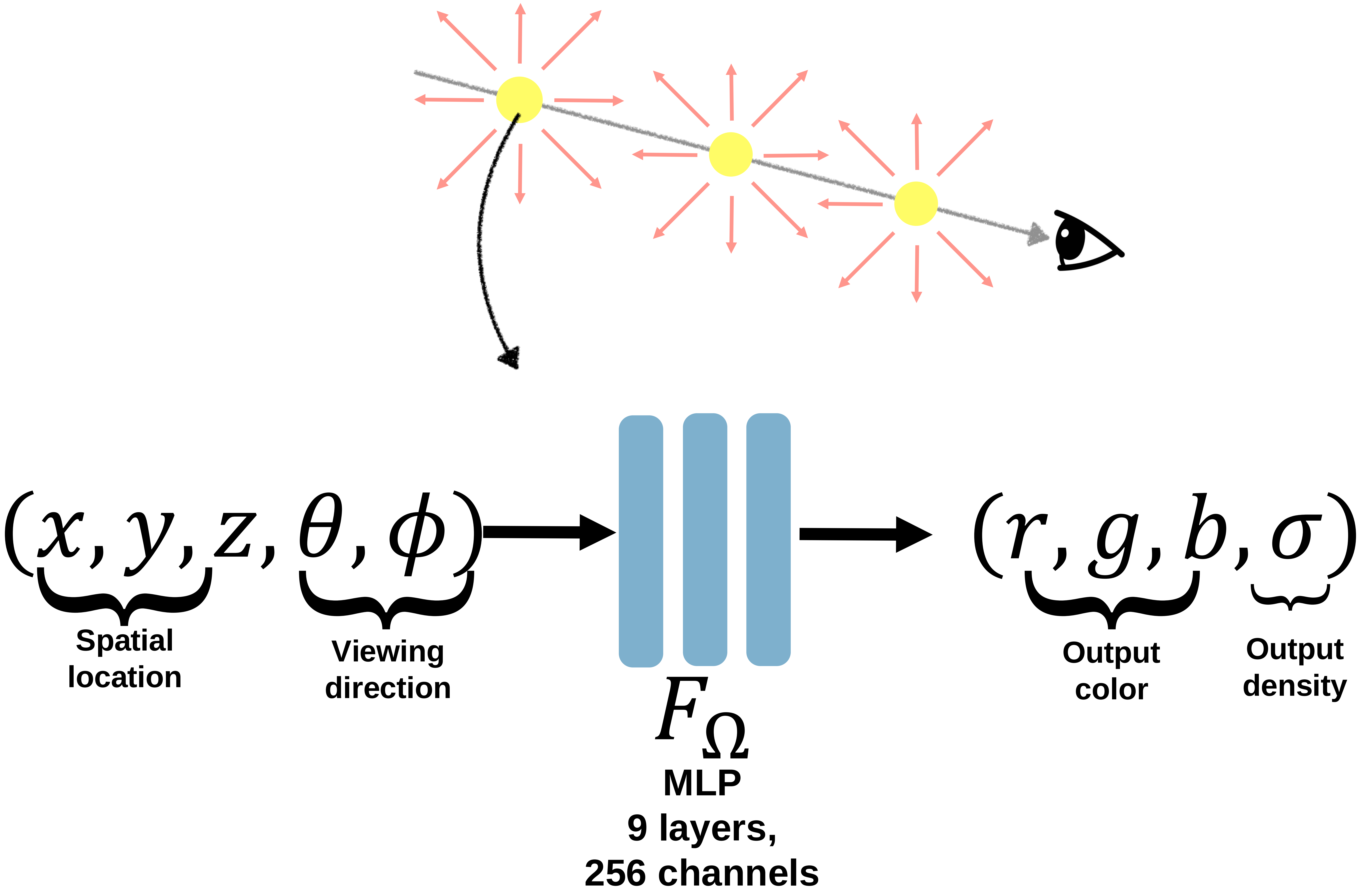
A subtle difference





So NeRF requires the integration along the viewing ray to compute the Plenoptic Function

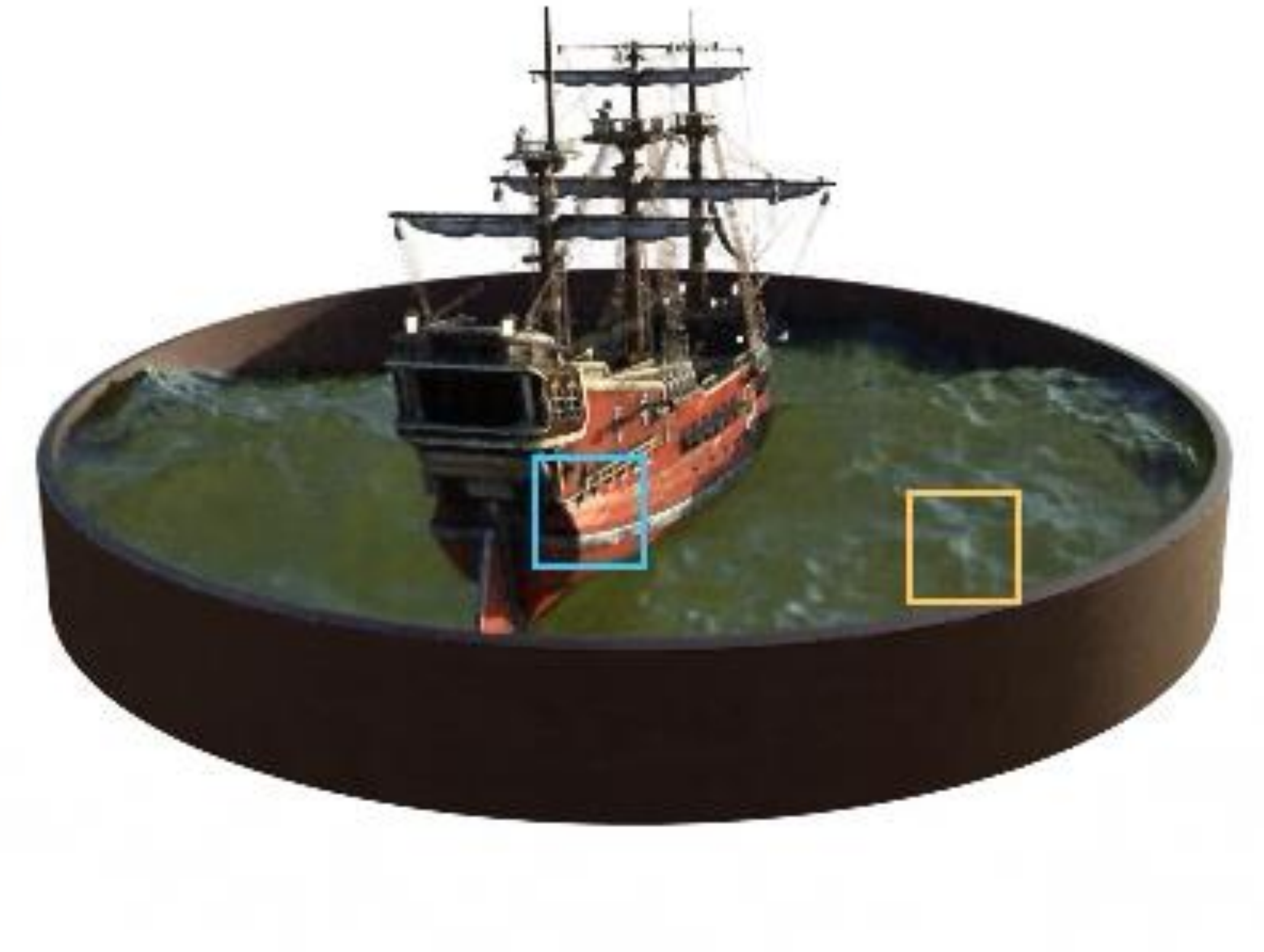
Bottom line: it models the full (5D) plenoptic function!

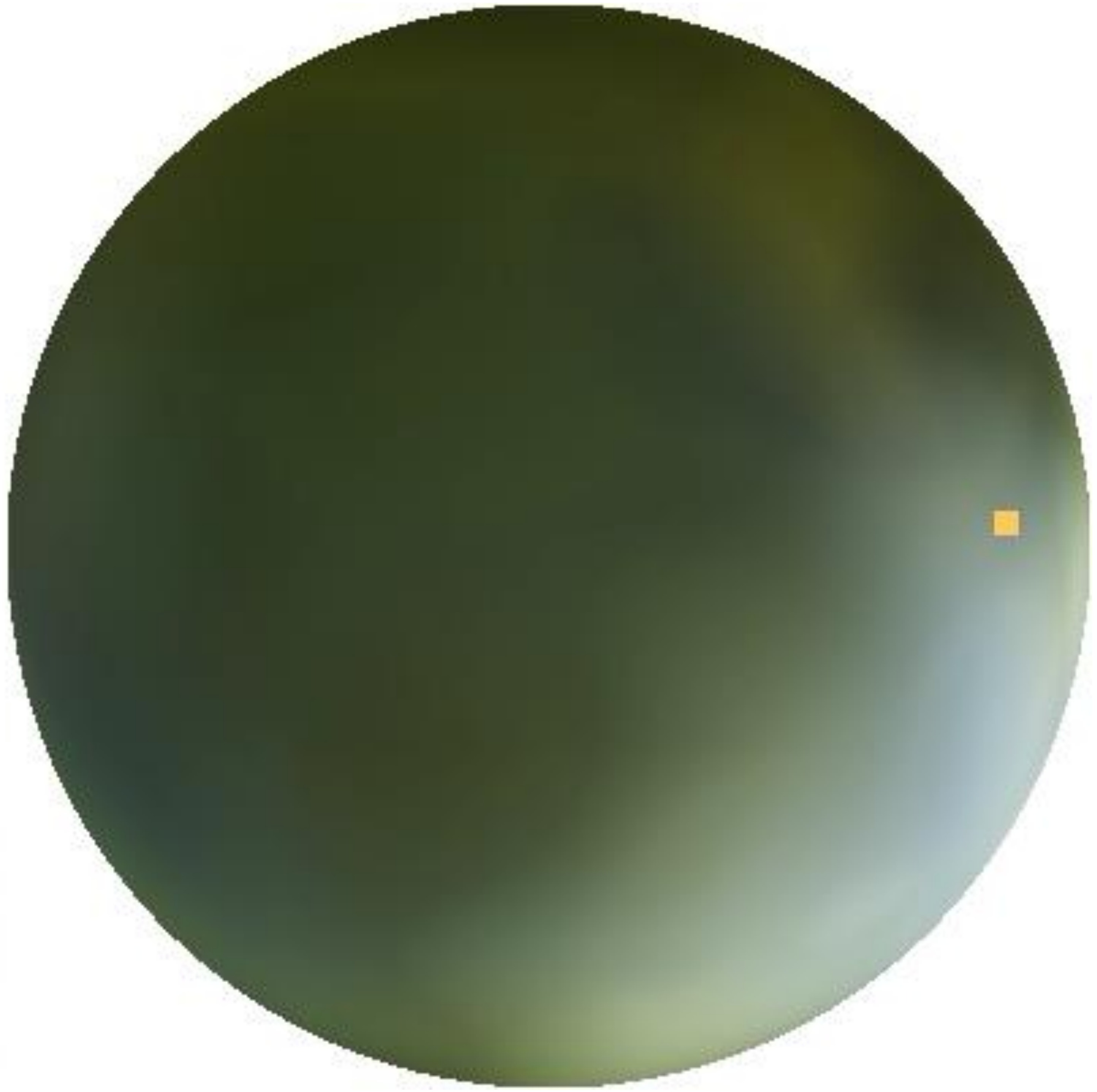


(x, y, z, θ, ϕ)
Spatial location
Viewing direction
F_Ω
MLP
9 layers,
256 channels
(r, g, b, σ)
Output color
Output density

Visualizing the 2D function on the sphere

Outgoing radiance distribution for point on side of ship





Outgoing radiance distribution for point on water's surface

Baking in Light

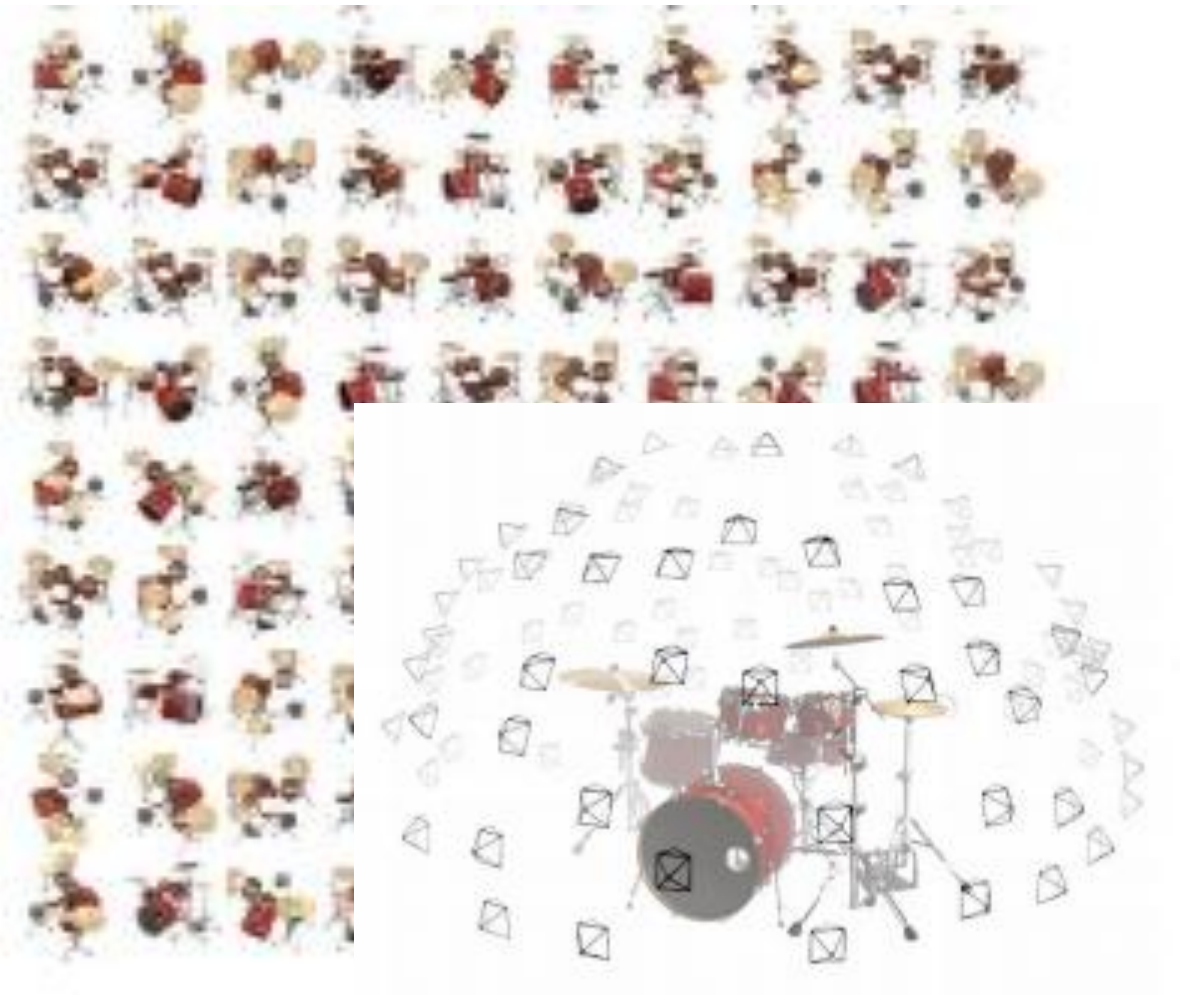
- NeRF can capture non-Lambertian (specular, shiny surfaces) because it models the color in a view-dependent manner
- This is hard to do with meshes unless you model the physical materials & lighting interactions
- But, with Image Based Rendering — All lighting effects are baked in

NeRF in a Slide

Volumetric 3D Scene Representation

Differentiable Volumetric Rendering Function

Objective: Reconstruct all training views



Optimization via Analysis-by-Synthesis

Unmentioned caveat so far

- Training a NeRF requires a **calibrated** camera!!!!
- Need to know the camera parameters: extrinsic (viewpoint) & intrinsics (focal length, distortion, etc)

How do we get this from images?

Structure from Motion

Or Photogrammetry (1850~)
Long history in Computer Vision

Proc. R. Soc. Lond. B. **203**, 405–426 (1979)
Printed in Great Britain

The interpretation of structure from motion

BY S. ULLMAN

Artificial Intelligence Laboratory, Massachusetts Institute of Technology,
545 *Technology Square* (*Room* 808), *Cambridge, Massachusetts* 02139 *U.S.A.*

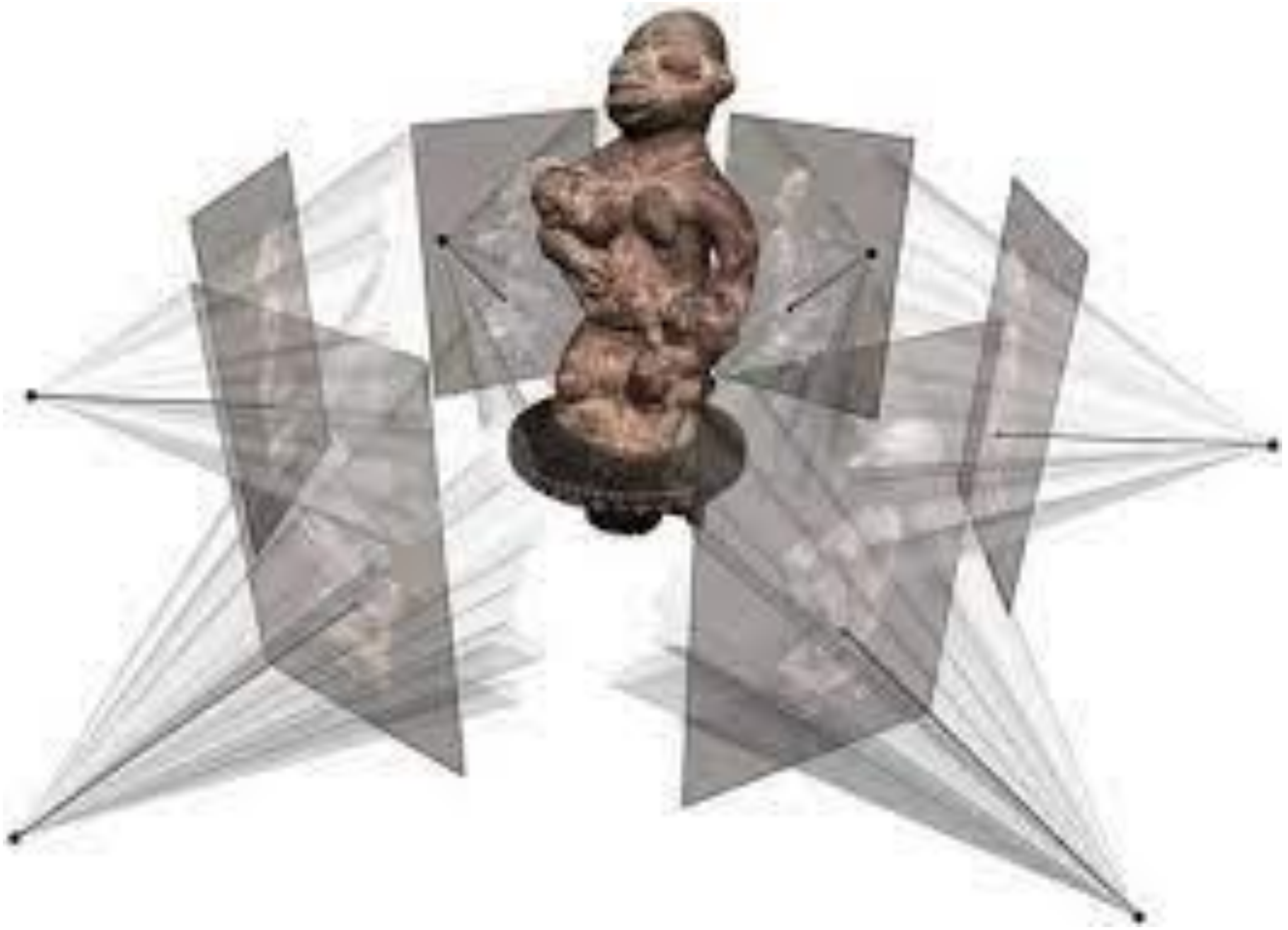
NeRF is AFTER Structure from Motion

- In order to train NeRF you need to run SfM/SLAM on the images to estimate the camera parameters
- In this sense, the problem category is same as that of **Multi-view Stereo**

Colmap: Schönberger et al. 2016

Multi-view Stereo

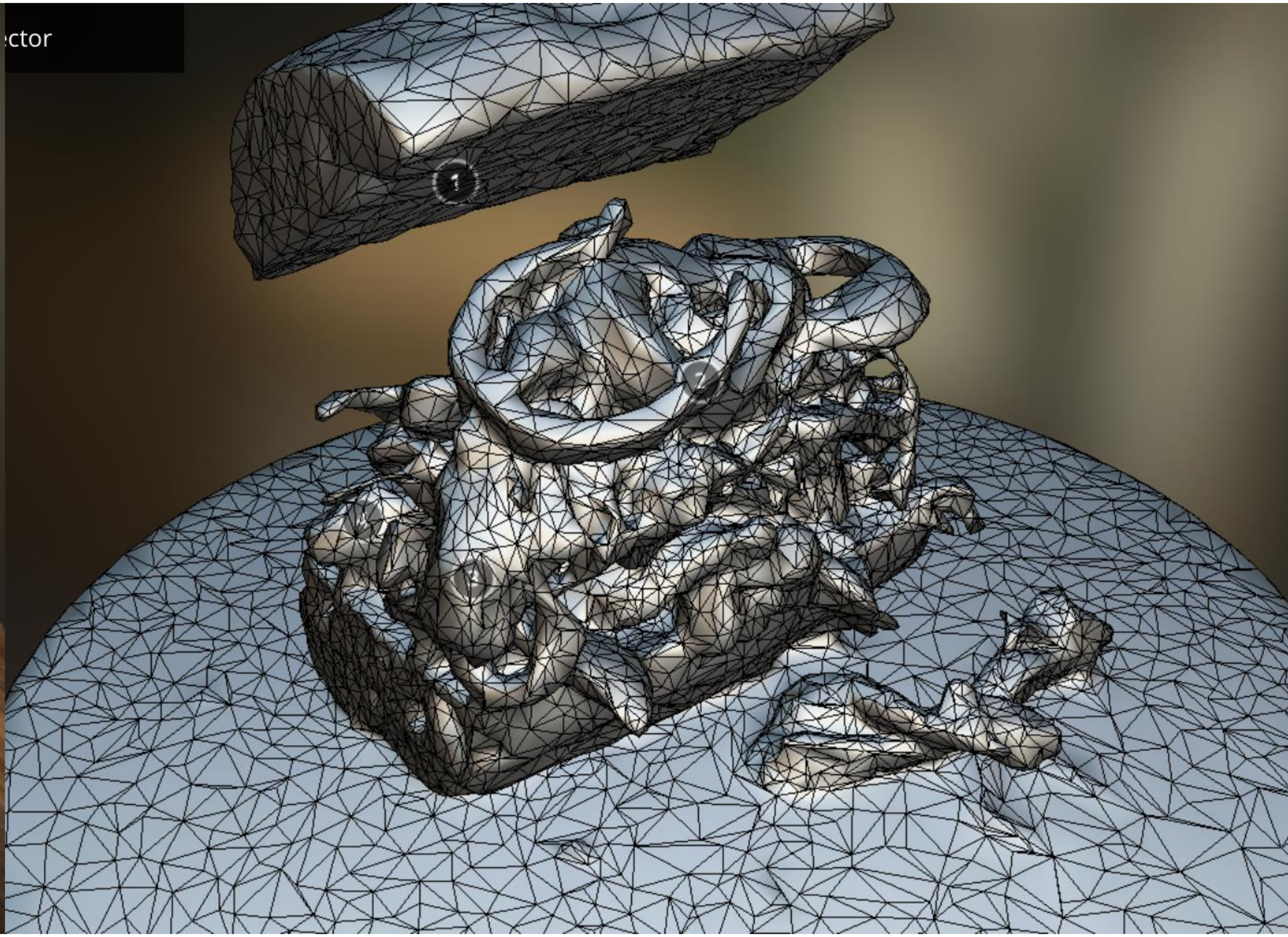
- Problem: Given calibrated cameras, recover highly detailed 3D **surface** model
- Dense photogrammetry, often the output is textured meshes



Figures by Carlos Hernandez, Yasutaka Furukawa

Multi-View Stereo

Solutions to MVS is what you see for any existing 3D scanning system, ie sketchfab, or what’s in your video game



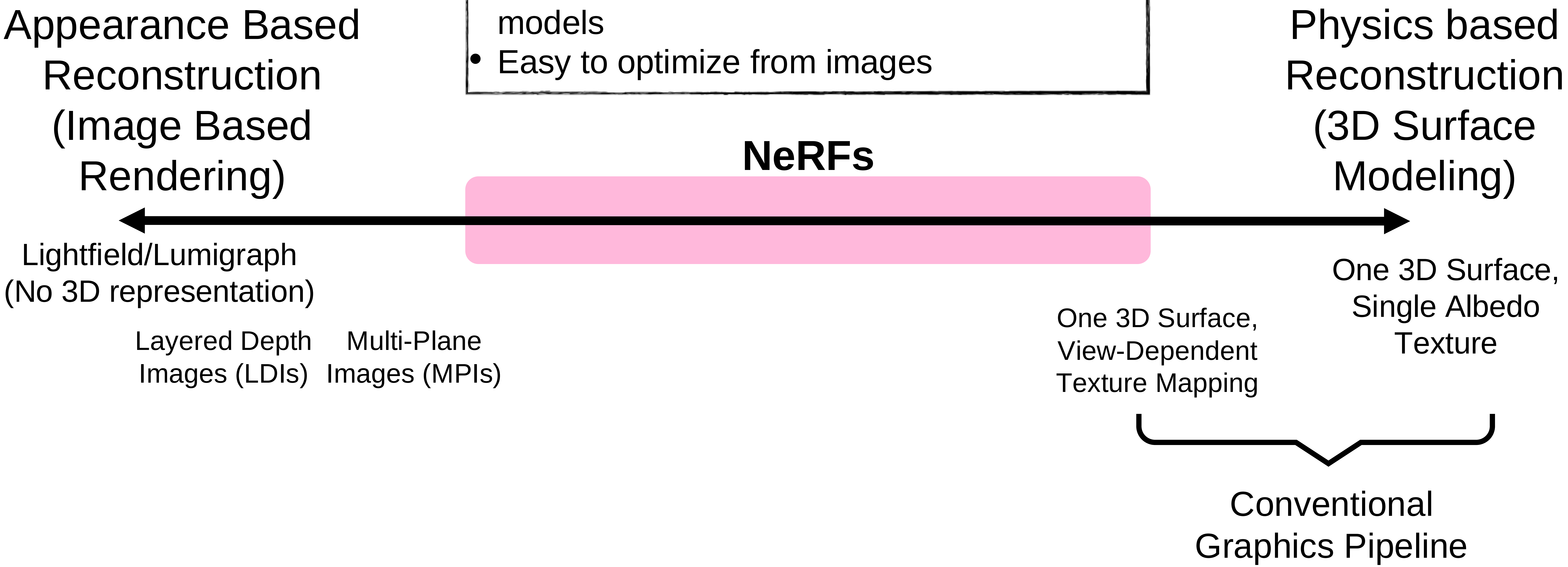
Multi-View Stereo

Because they often model surfaces, struggles on Thin / Amorphus / Shiny objects


nerfstudio

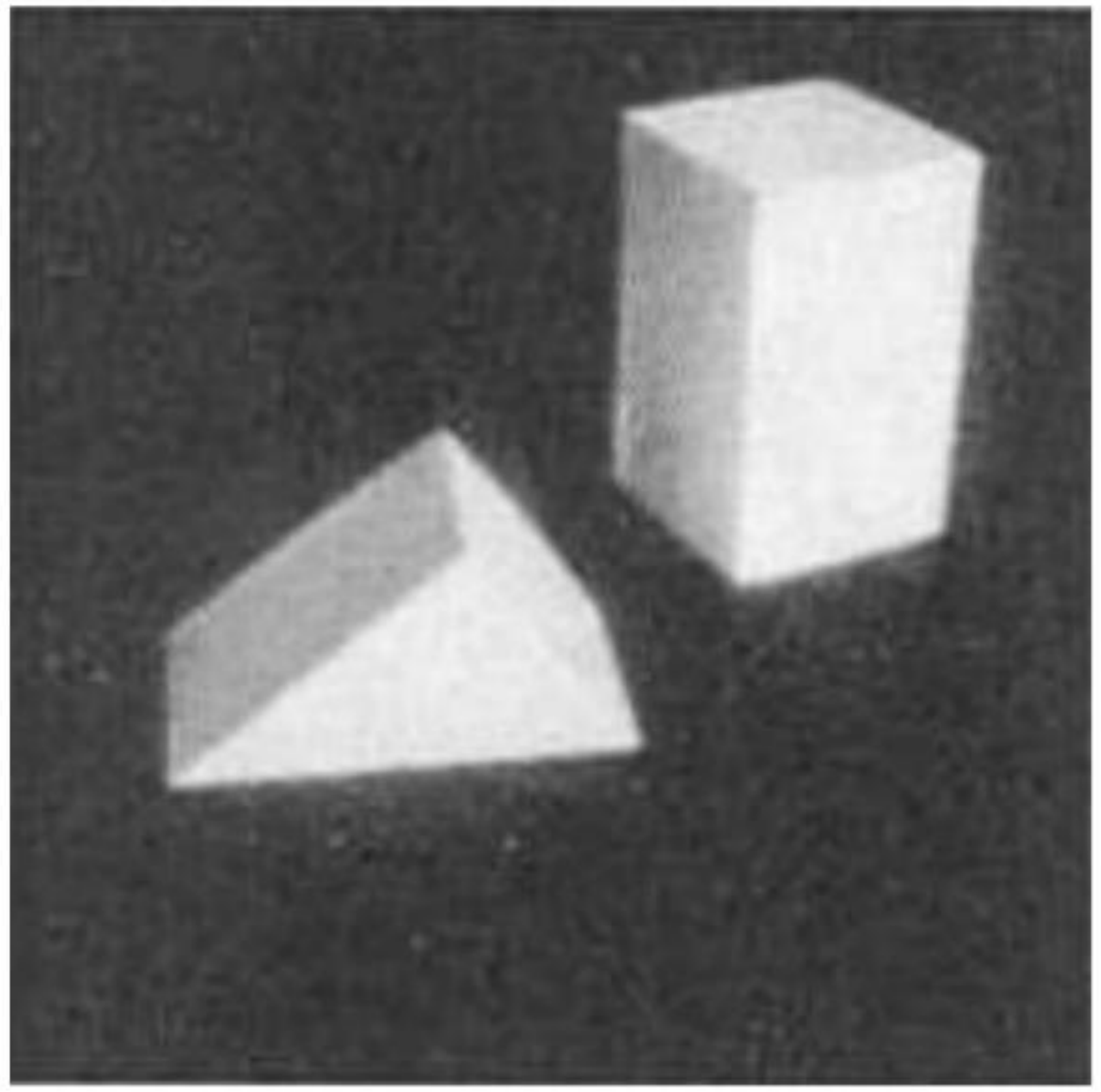
Where NeRF stands

- can do Image Based Rendering well, while also being a 3D representation
- Does not suffer from limitations of surface models
- Easy to optimize from images

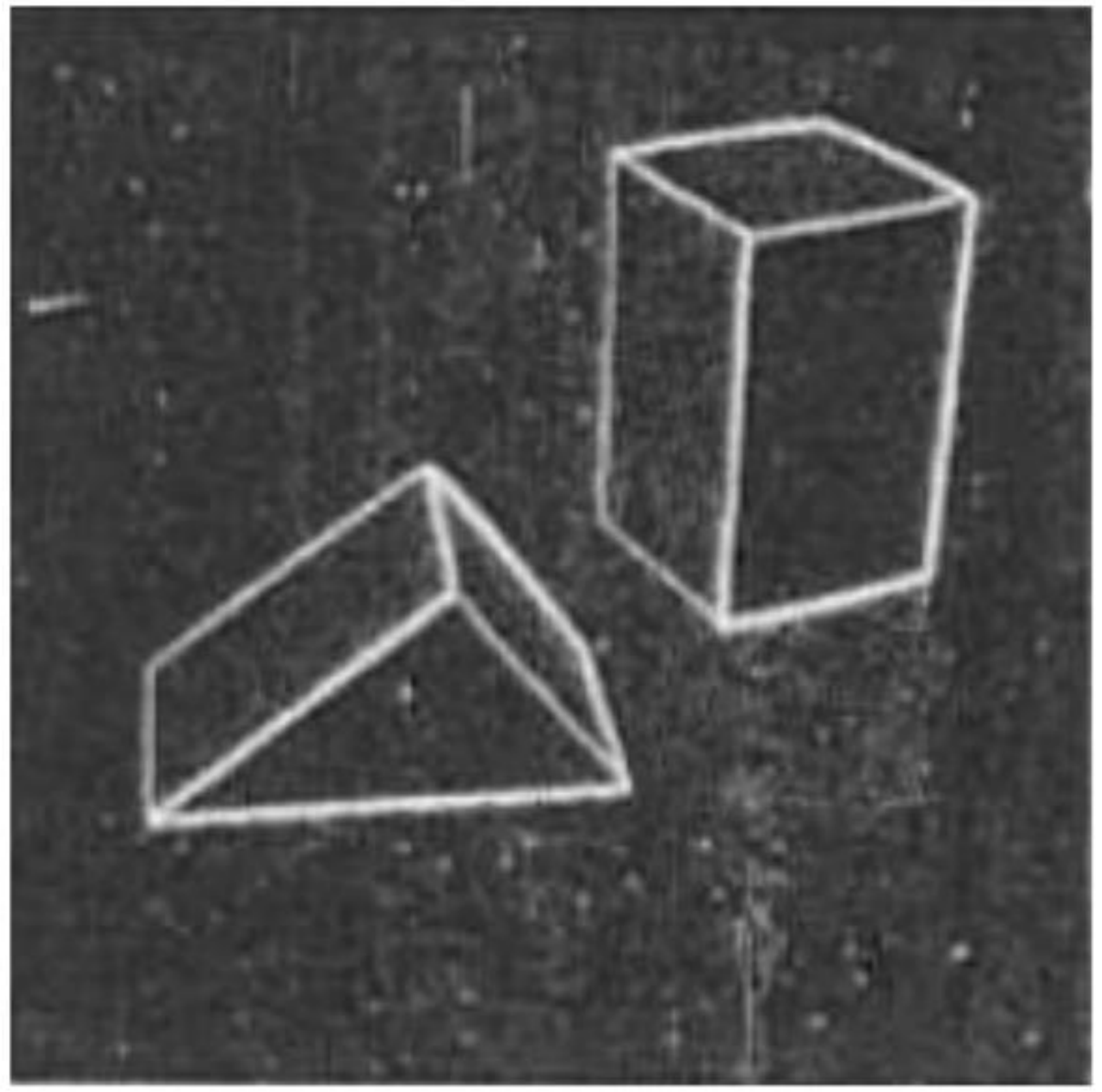


Analysis-by-Synthesis

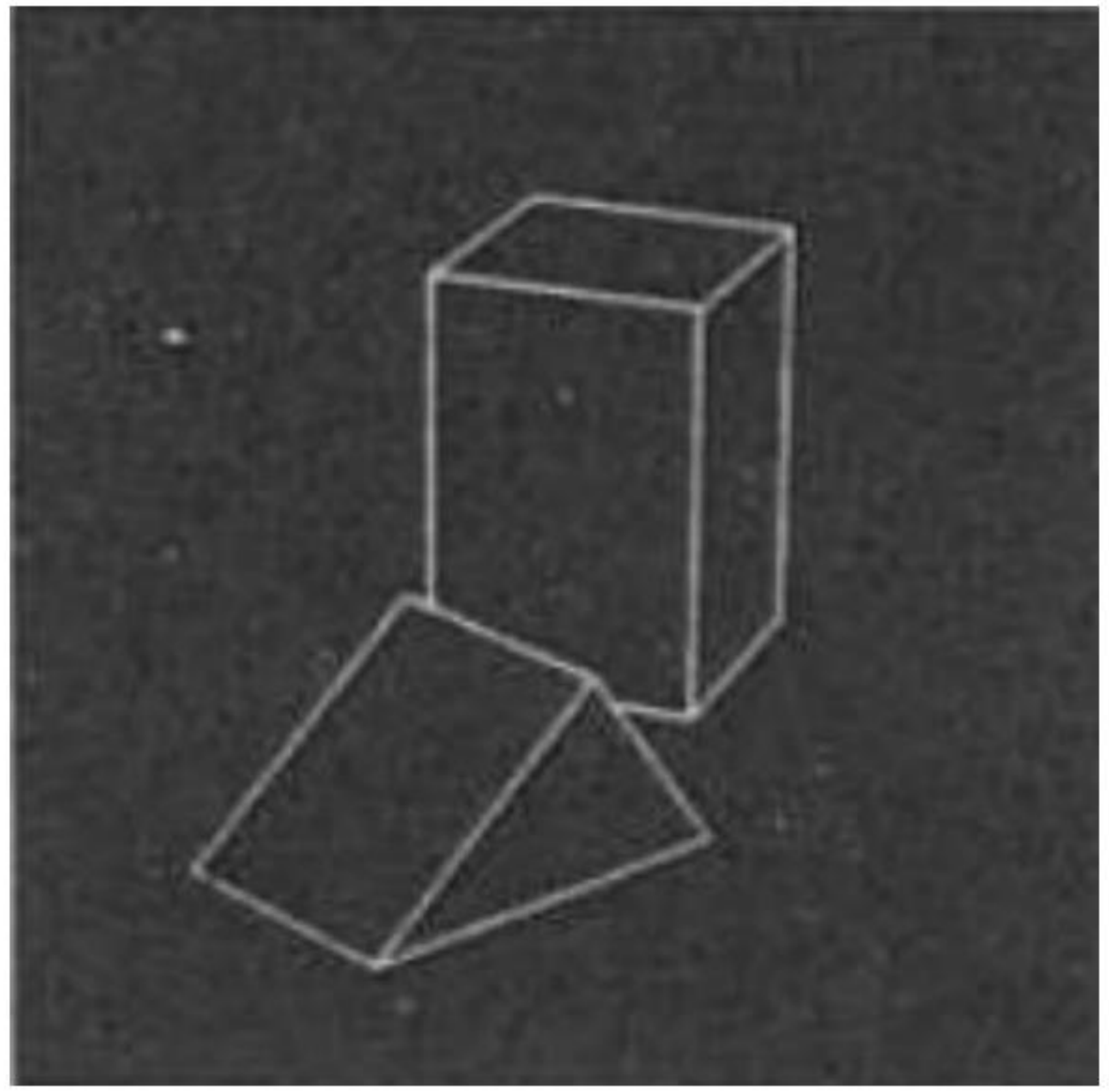
Larry Roberts
"Father of Computer Vision"



Input image



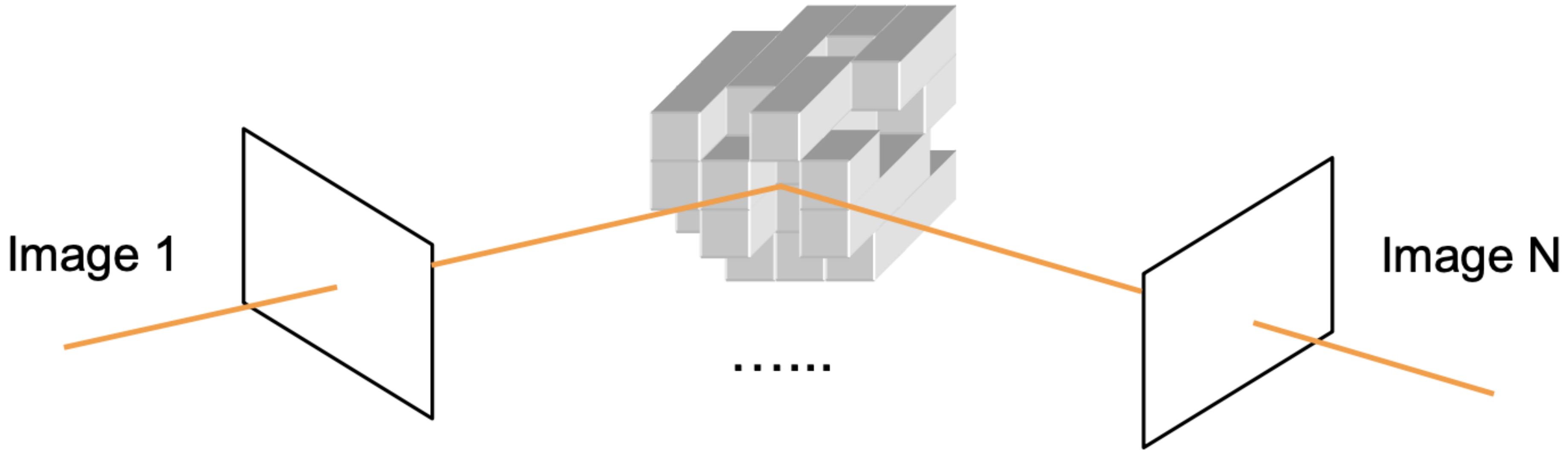
2x2 gradient operator



computed 3D model
rendered from new viewpoint

- History goes way back to the **first** Computer Vision paper!
  Roberts: Machine Perception of Three-Dimensional Solids, MIT, 1963

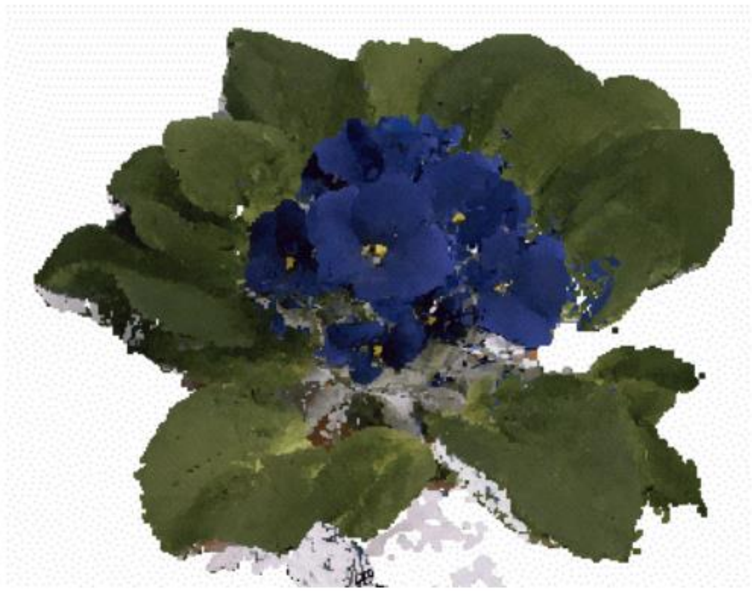
Power of Analysis-by-Synthesis



- Space Carving: A MVS method that used Colored voxels
- But the optimization method was bottom up then.
- Key is optimization via Analysis-by-Synthesis [Plenoxels, Yu et al. 2022]


Input Image (1 of 45)


Reconstruction


Reconstruction

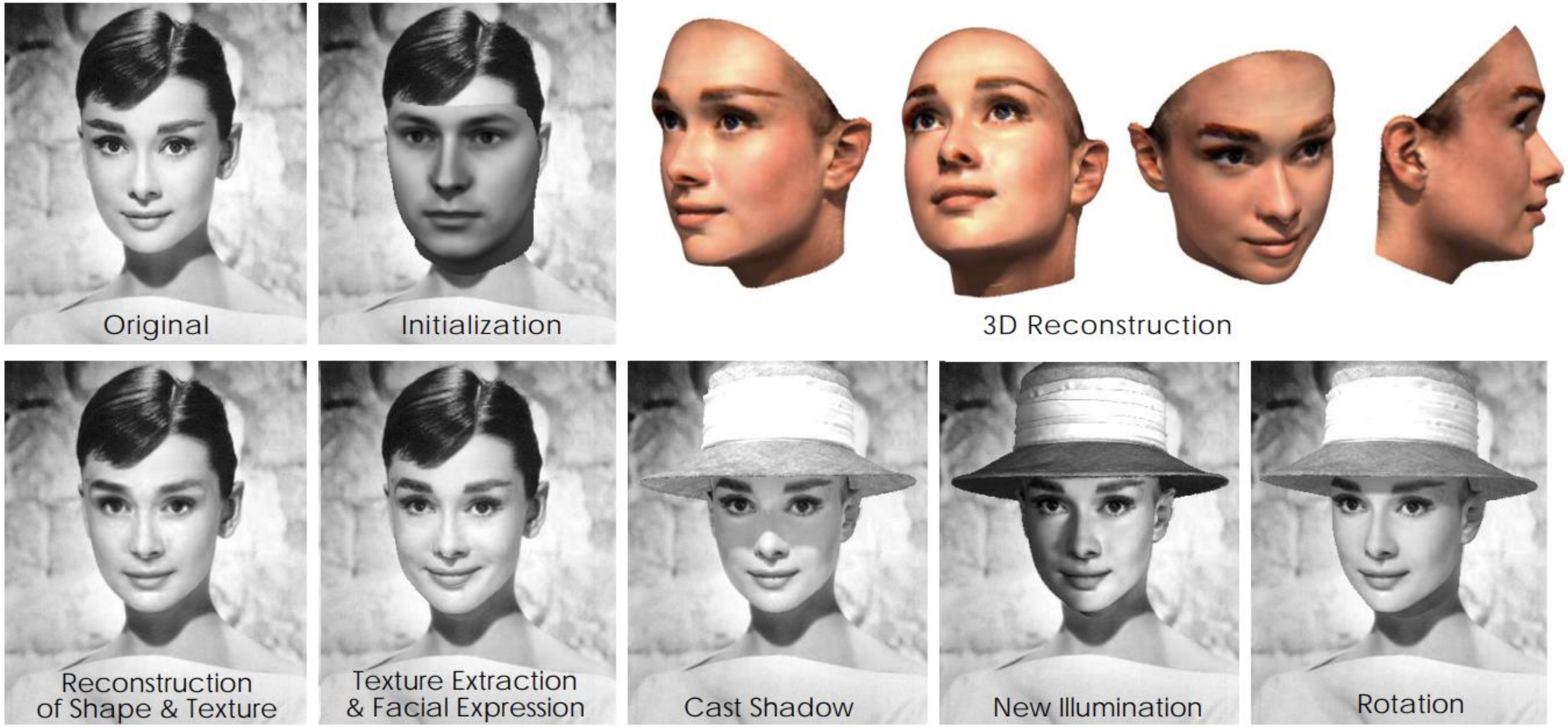

Reconstruction


Input Image
(1 of 100)


Views of Reconstruction

Kultulakos and Seitz, A Theory of Shape by Space Carving IJCV 2000

Analysis-by-Synthesis



Blanz & Vetter 1999

- With custom differentiable renders

Analysis by Synthesis Requires Differentiable Renderers

Next: Deep dive into Volumetric Rendering Function

Where we are

1. Birds Eye View & Background
2. **Volumetric Rendering Function**
3. Encoding and Representing 3D Volumes
4. Signal Processing Considerations
5. Challenges & Pointers

Volume Rendering

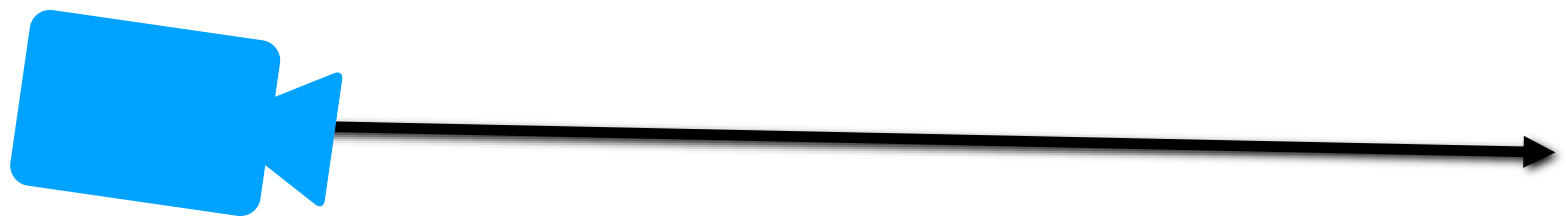
"... in 10 years, all rendering will be volume rendering."

Jim Kajiya at SIGGRAPH '91

Neural Volumetric Rendering

Neural Volumetric Rendering

computing color along rays
through 3D space



What color is this pixel?