清華大學
Tsinghua University
交叉信息研究院
Institute for Interdisciplinary Information Sciences

Lecture 7: Single Image to 3D

Li Yi

Apr 3, 2025

Recap

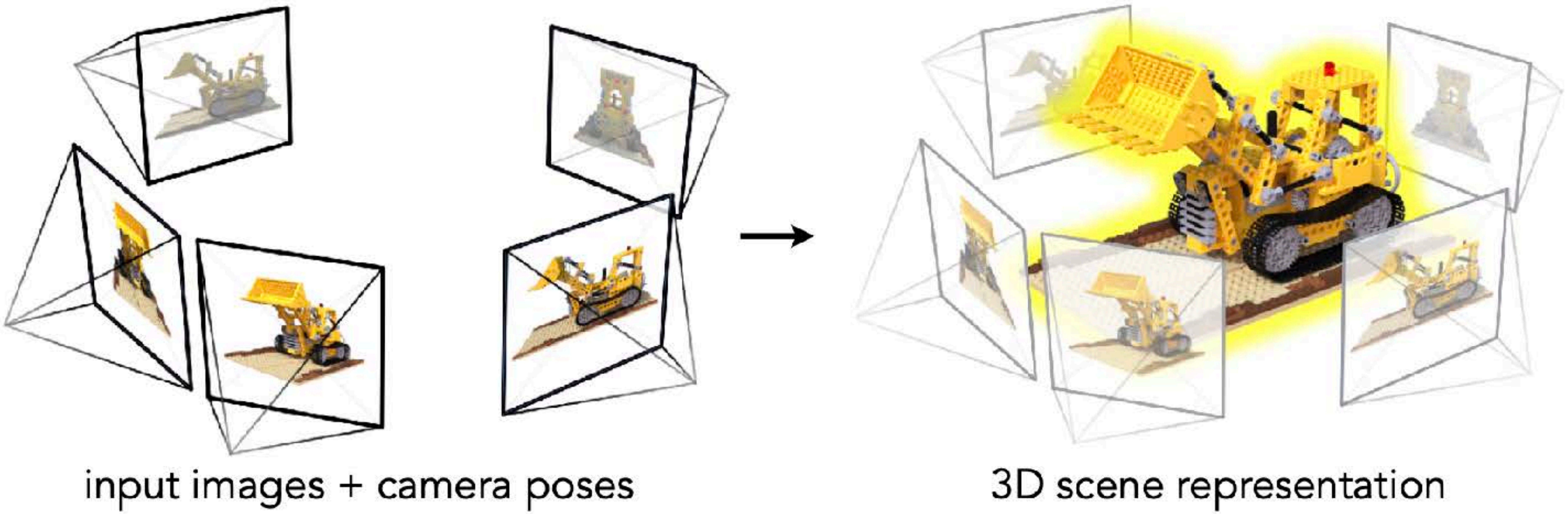
Input images

Find matched points across images

Structure from Motion (SFM)

dense reconstruction

Recap

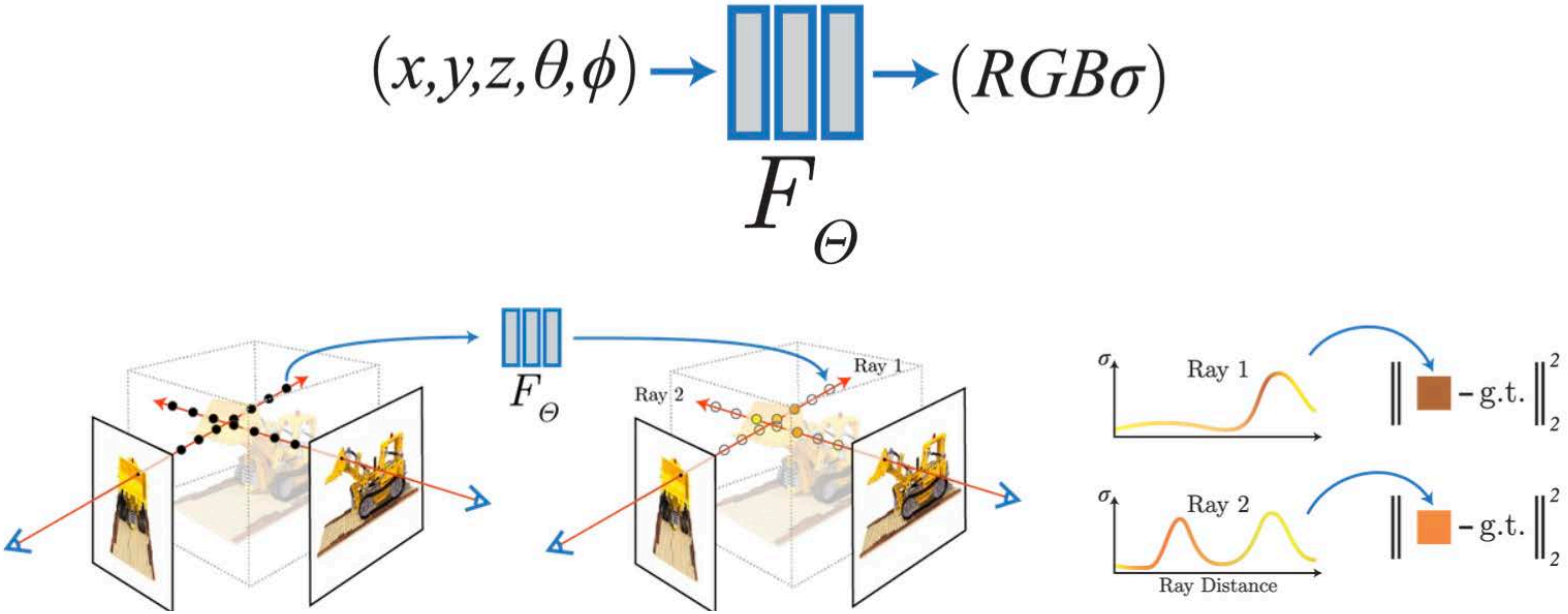


NeRF

Recap

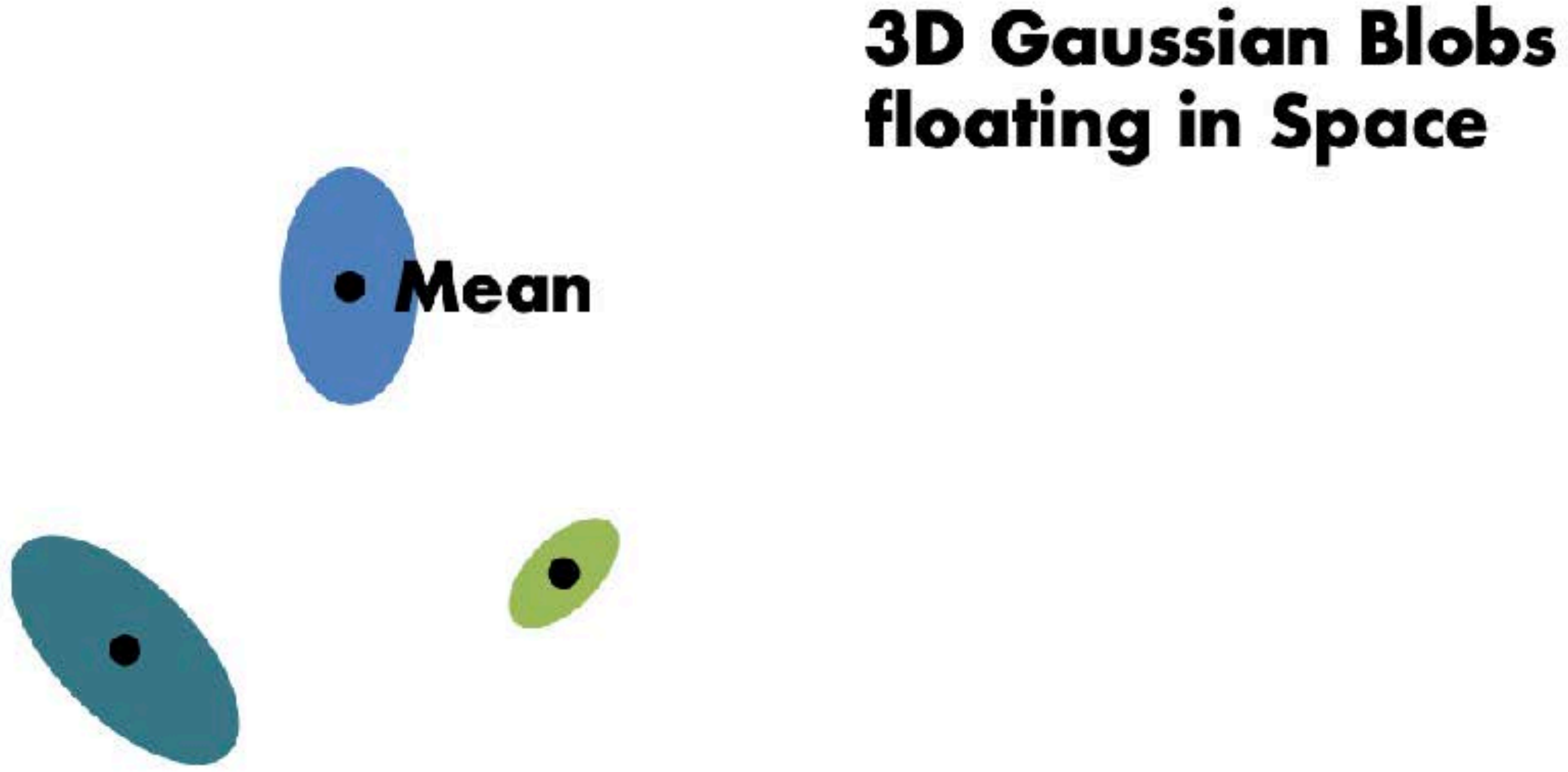
- Neural Radiance Field (NeRF)
  - Implicit Functions: an Illustration with 3D Surface Representation
  - Volume Rendering with Ray Marching
  - Learning NeRF
- NeRF Extensions
  - Handling dynamic scenes when acquiring calibrated views
  - One network trained per scene - no generalization

NeRF: Parameterize Radiance Field Densely



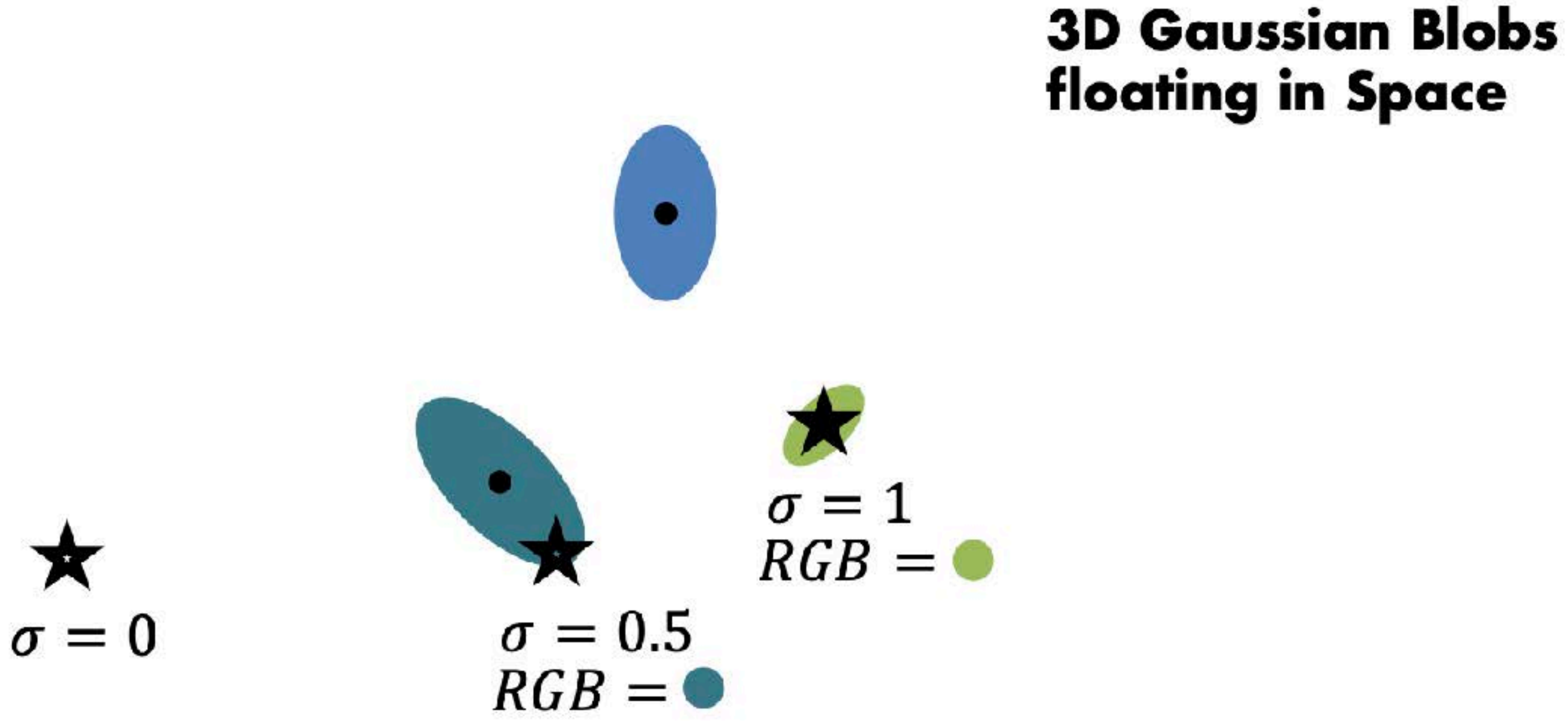
3D Gaussian Splatting (3DGS)

Key Idea: Parameterize Radiance Field *sparsely, only where density is nonzero*

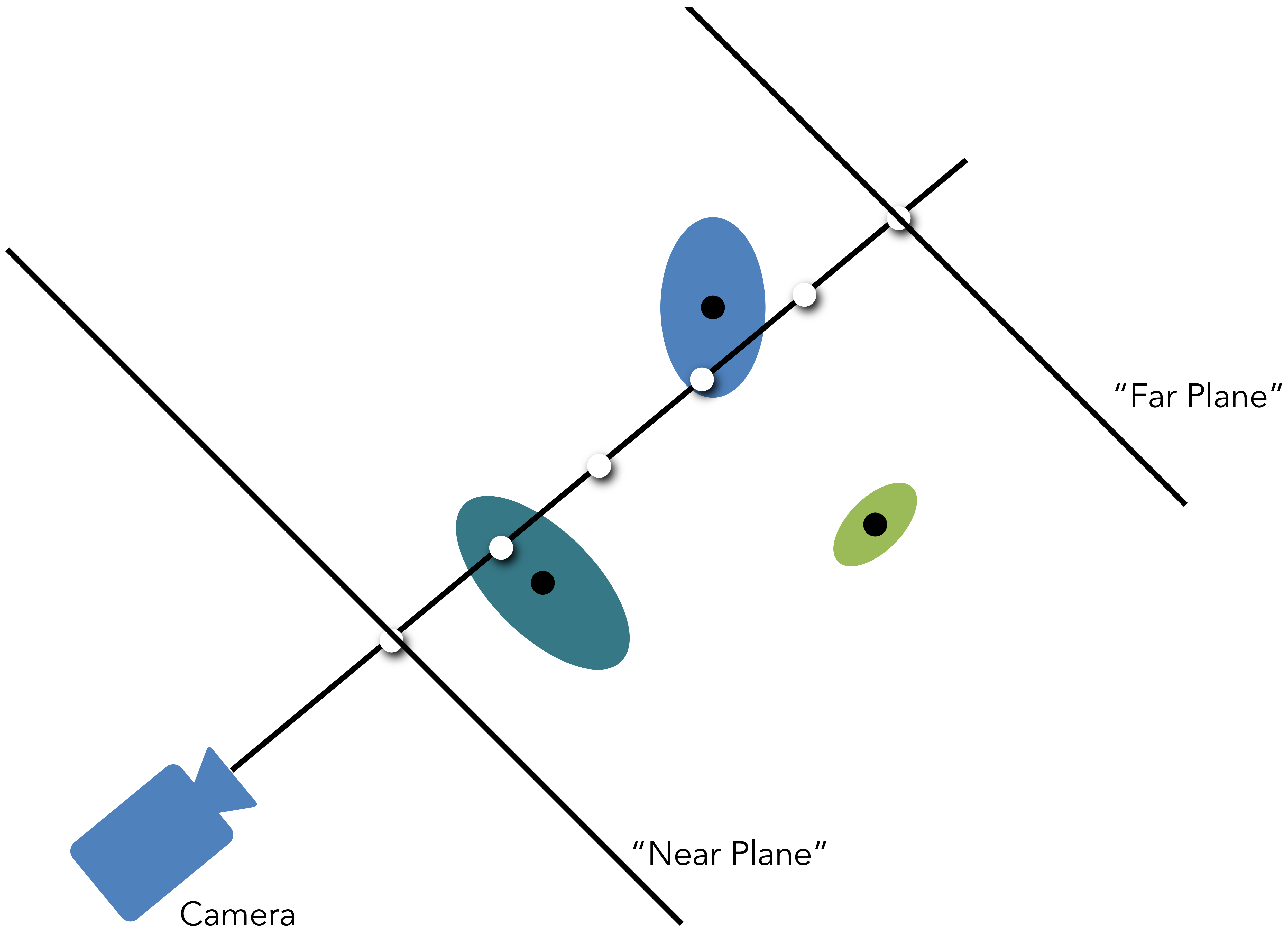


3D Gaussian Splatting (3DGS)

Key Idea: Parameterize Radiance Field *sparsely*,
only where density is nonzero

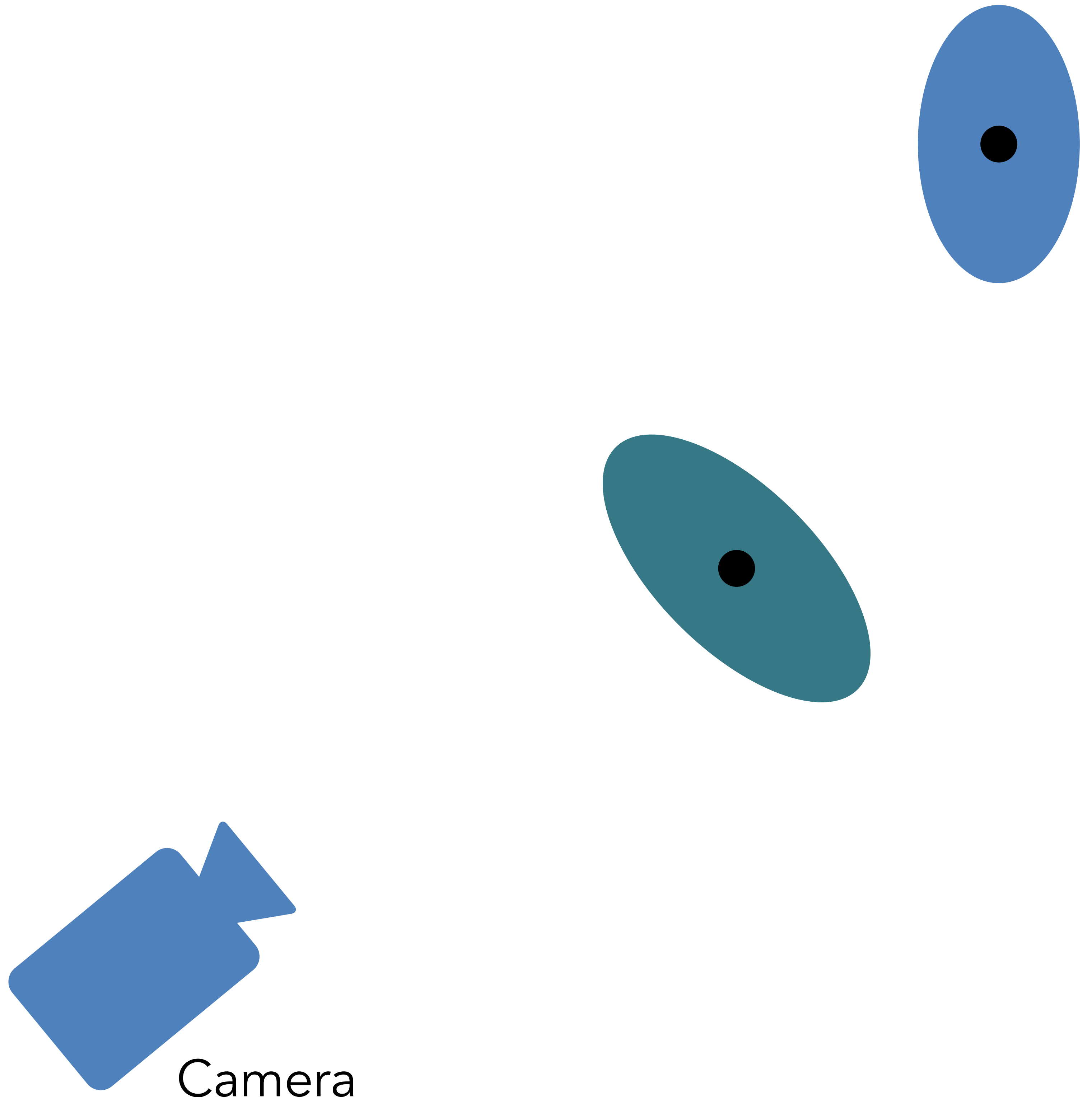


Still Volume Rendering?



Computation Properties of Gaussians

Gaussians are closed under affine transforms, integration



$$\mathcal{G}_{\mathbf{V}}(\mathbf{x}-\mathbf{p}) = \frac{1}{2\pi|\mathbf{V}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{p})^T \mathbf{V}^{-1}(\mathbf{x}-\mathbf{p})}$$

3D Covariance!

Affine mapping $\Phi = \mathbf{M}\mathbf{x} + \mathbf{p}$ of coordinates (such as cam2world matrix!):

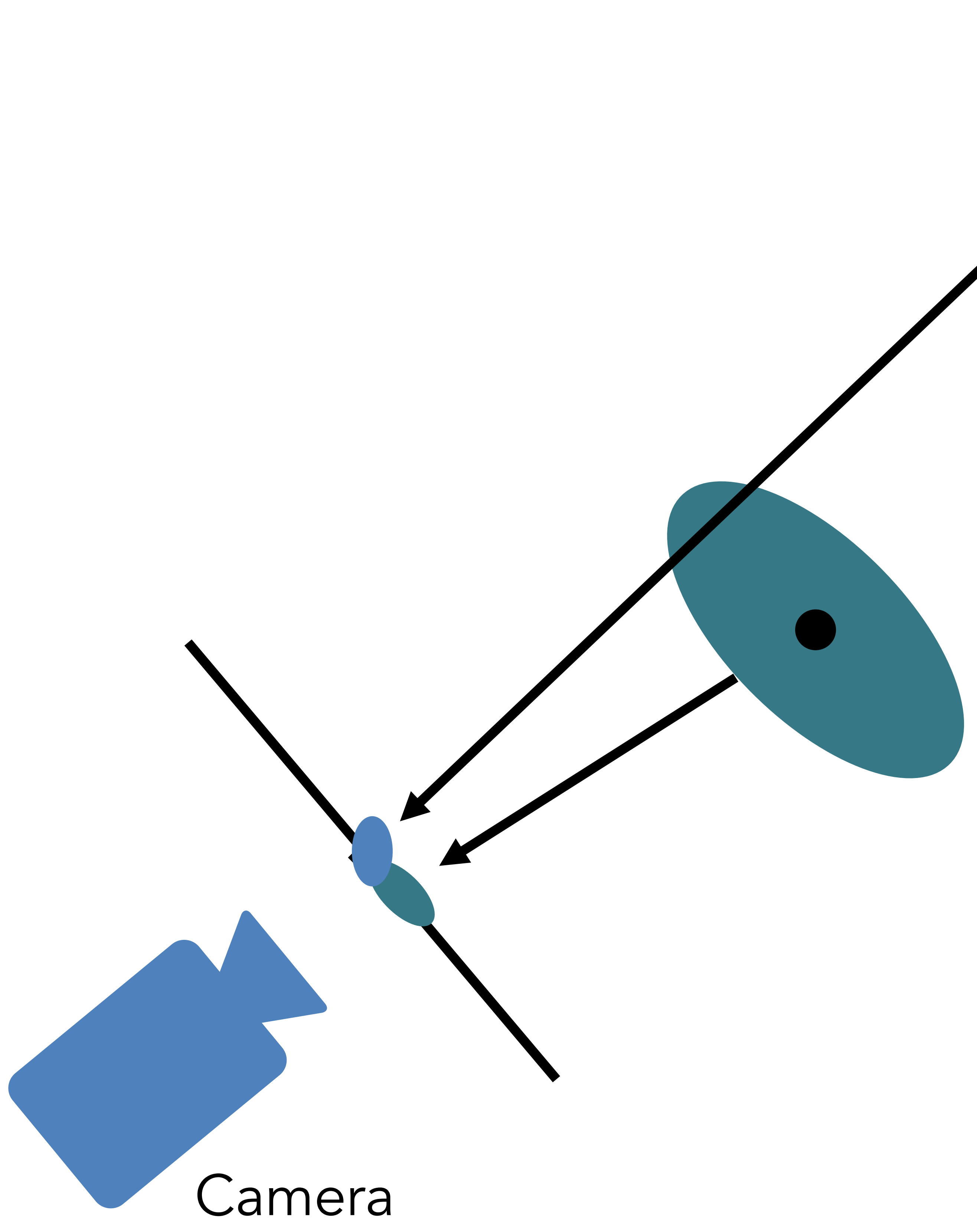
$$\mathcal{G}_{\mathbf{V}}(\Phi^{-1}(\mathbf{u}) - \mathbf{p}) = \frac{1}{|\mathbf{M}^{-1}|} \mathcal{G}_{\mathbf{M}\mathbf{V}\mathbf{M}^T}(\mathbf{u} - \Phi(\mathbf{p}))$$

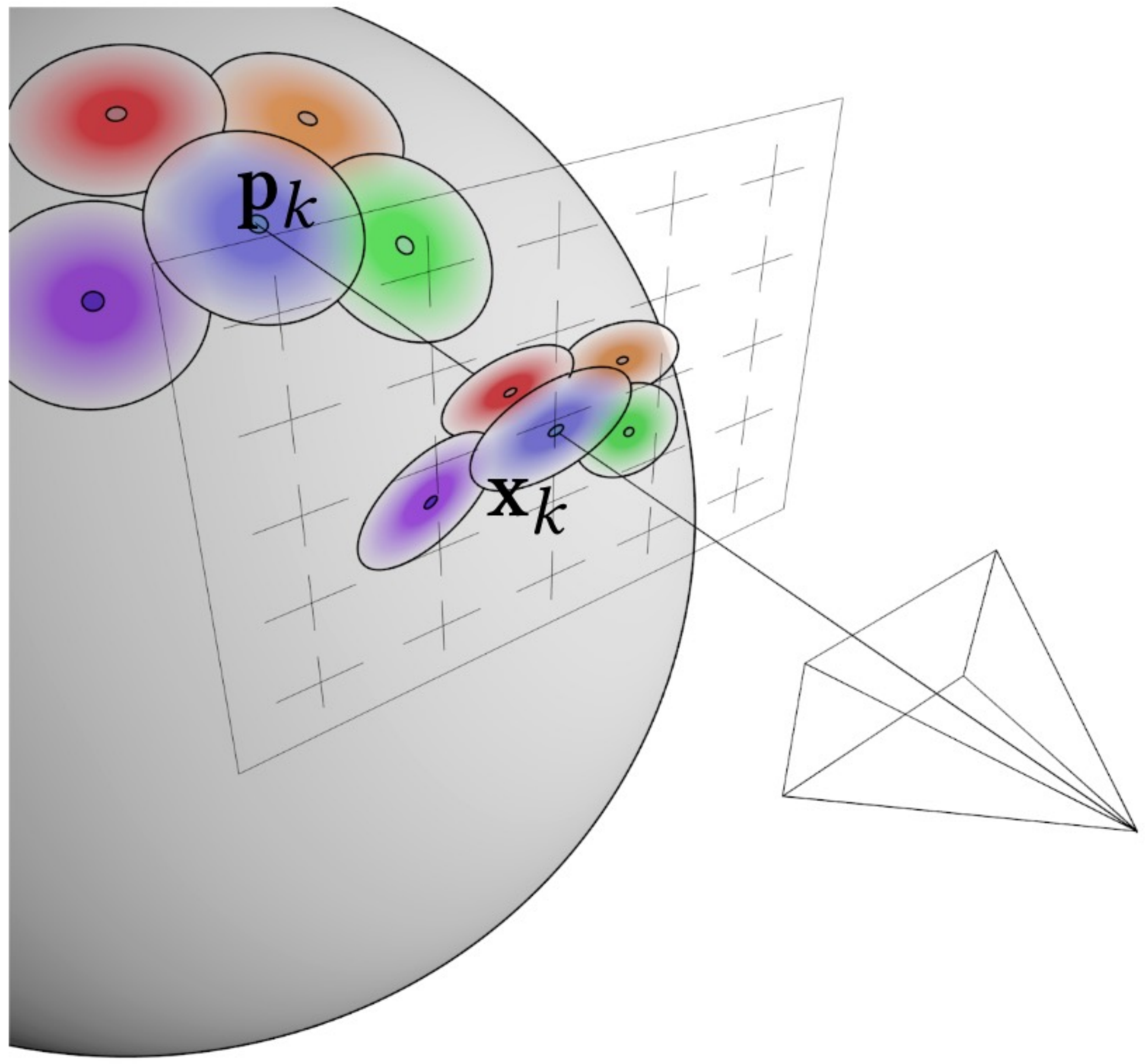
Integrate along axis:

$$\int_{\mathbb{R}} \mathcal{G}^3_{\mathbf{V}}(\mathbf{x}-\mathbf{p})\, dx_2 = \mathcal{G}^2_{\hat{\mathbf{V}}}(\hat{\mathbf{x}} - \hat{\mathbf{p}})$$

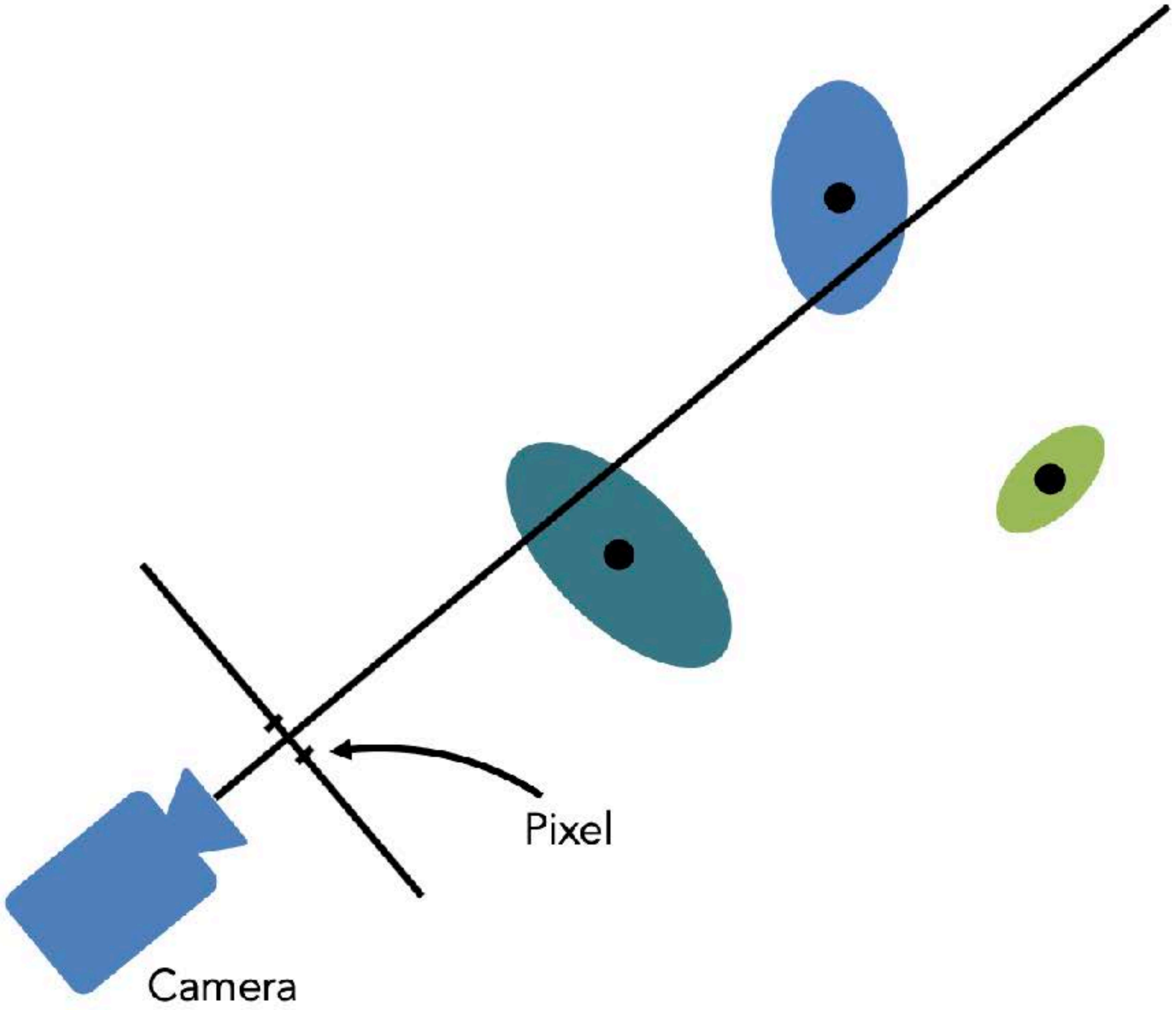
$$\mathbf{V} = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix} \Leftrightarrow \begin{pmatrix} a & b \\ b & d \end{pmatrix} = \hat{\mathbf{V}}$$

Projected 3D Gaussian makes 2D Gaussian

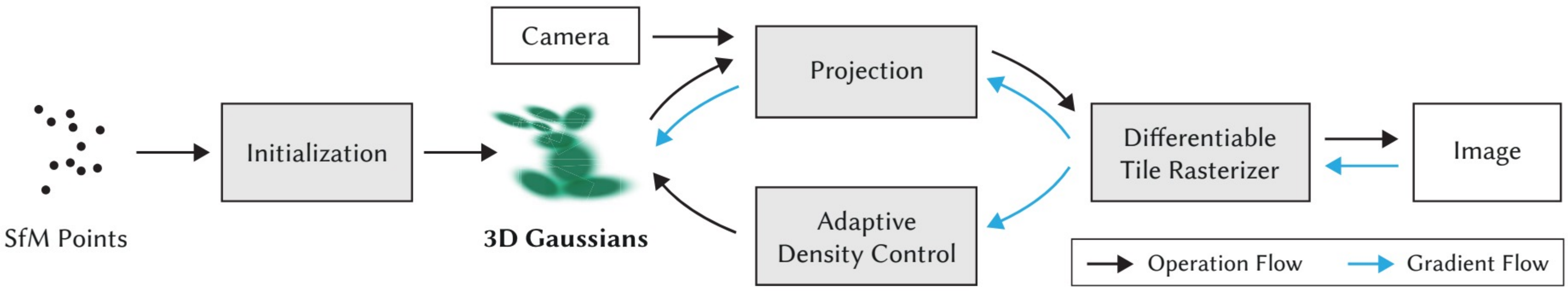




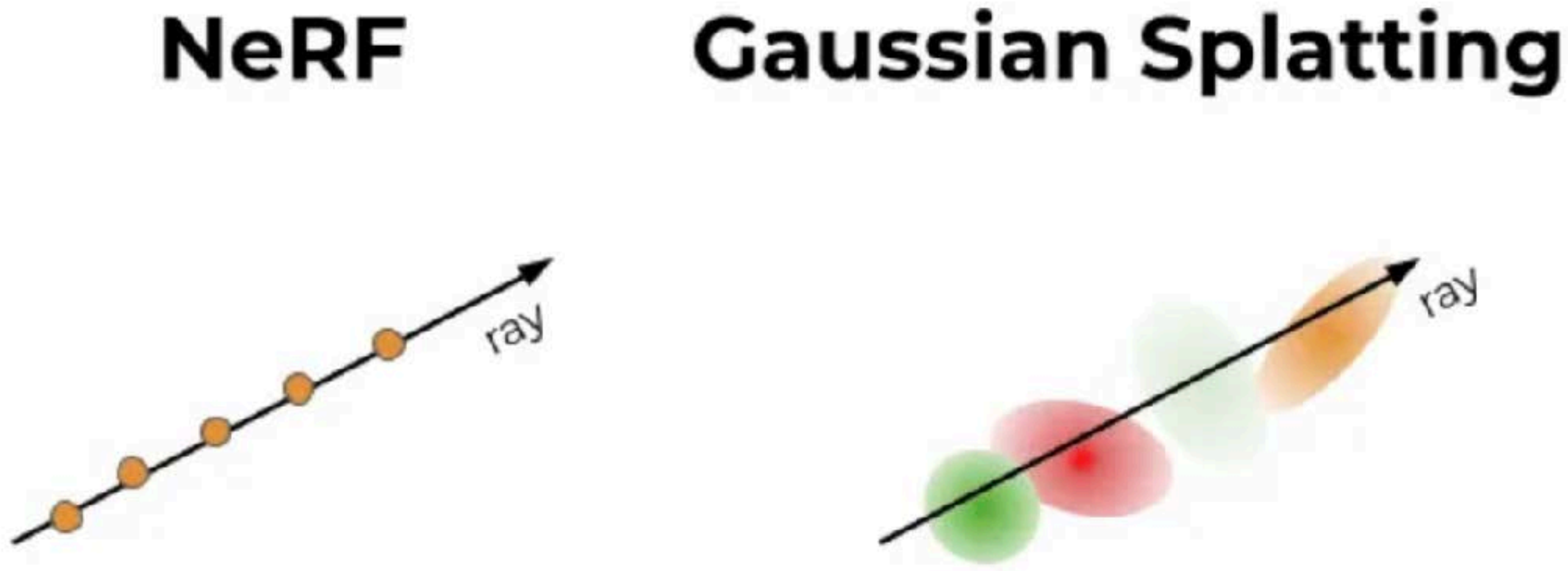
Using Rasterization Instead of Volume Rendering



3DGS Framework



3D Gaussian Splatting



Today’s Focus

Single Image to 3D

Outline

- ***Task***
- Synthesis-for-Learning Pipeline
- Single-image to Point Cloud
- Single-image to Mesh

Review: Multi-View Stereo

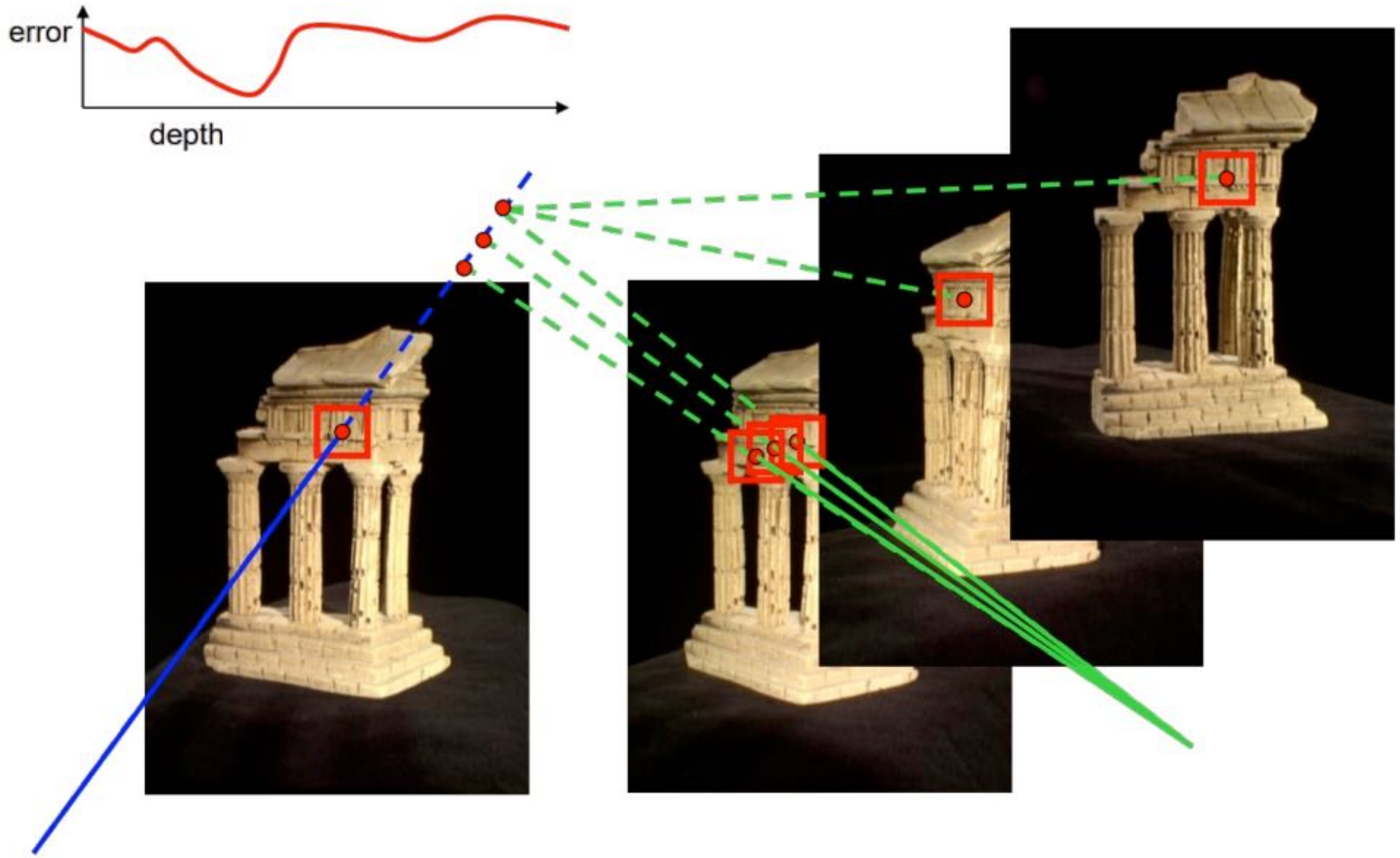


Image source: UW CSE455

Task

Can We Infer 3D from just a **Single** Image?

Many Cues that Allow 3D Estimation

contrast

color

texture

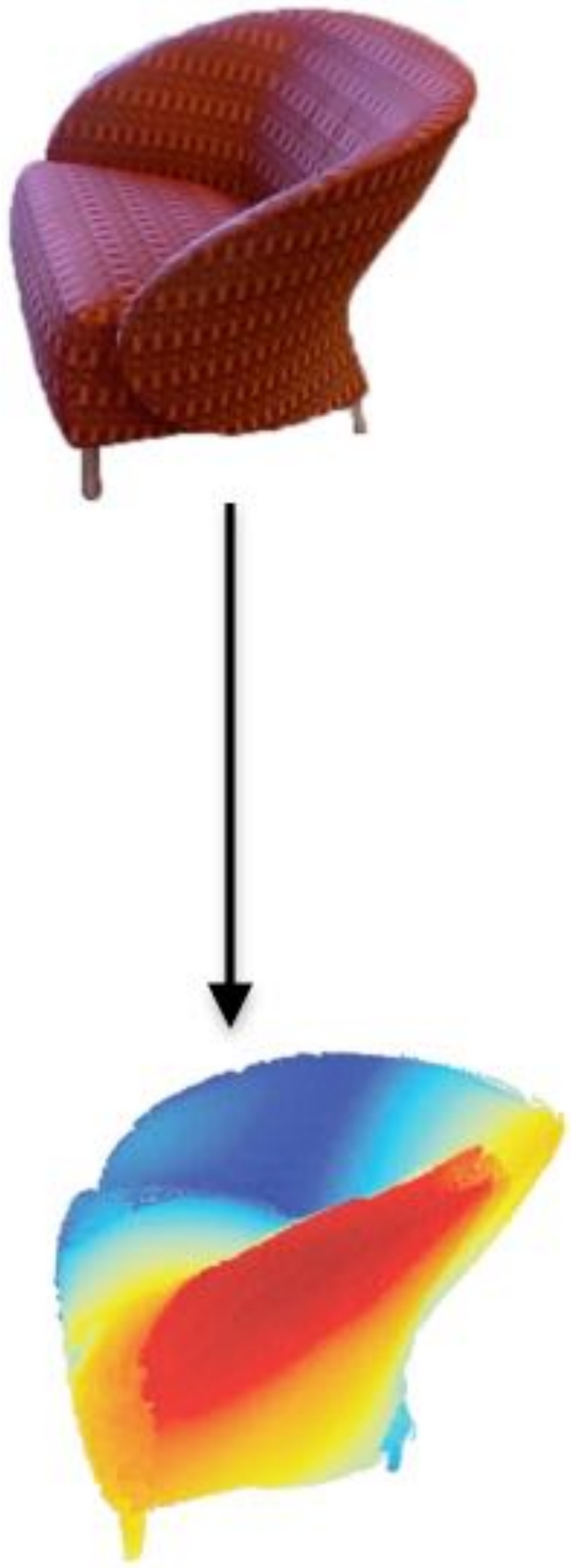
motion

symmetry

part

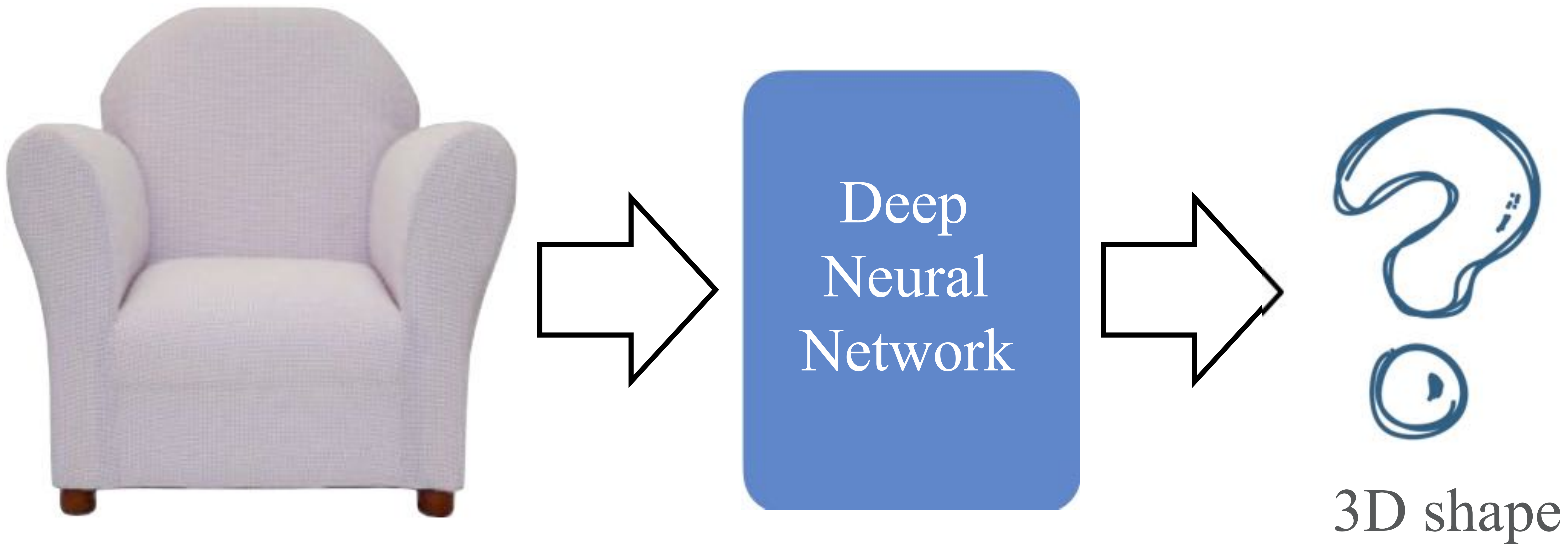
category-specific 3D knowledge

……



Slides credit: Hao Su

Learning-based 3D Reconstruction



Outline

- Task
- ***Synthesis-for-Learning Pipeline***
- Single-image to Point Cloud
- Single-image to Mesh

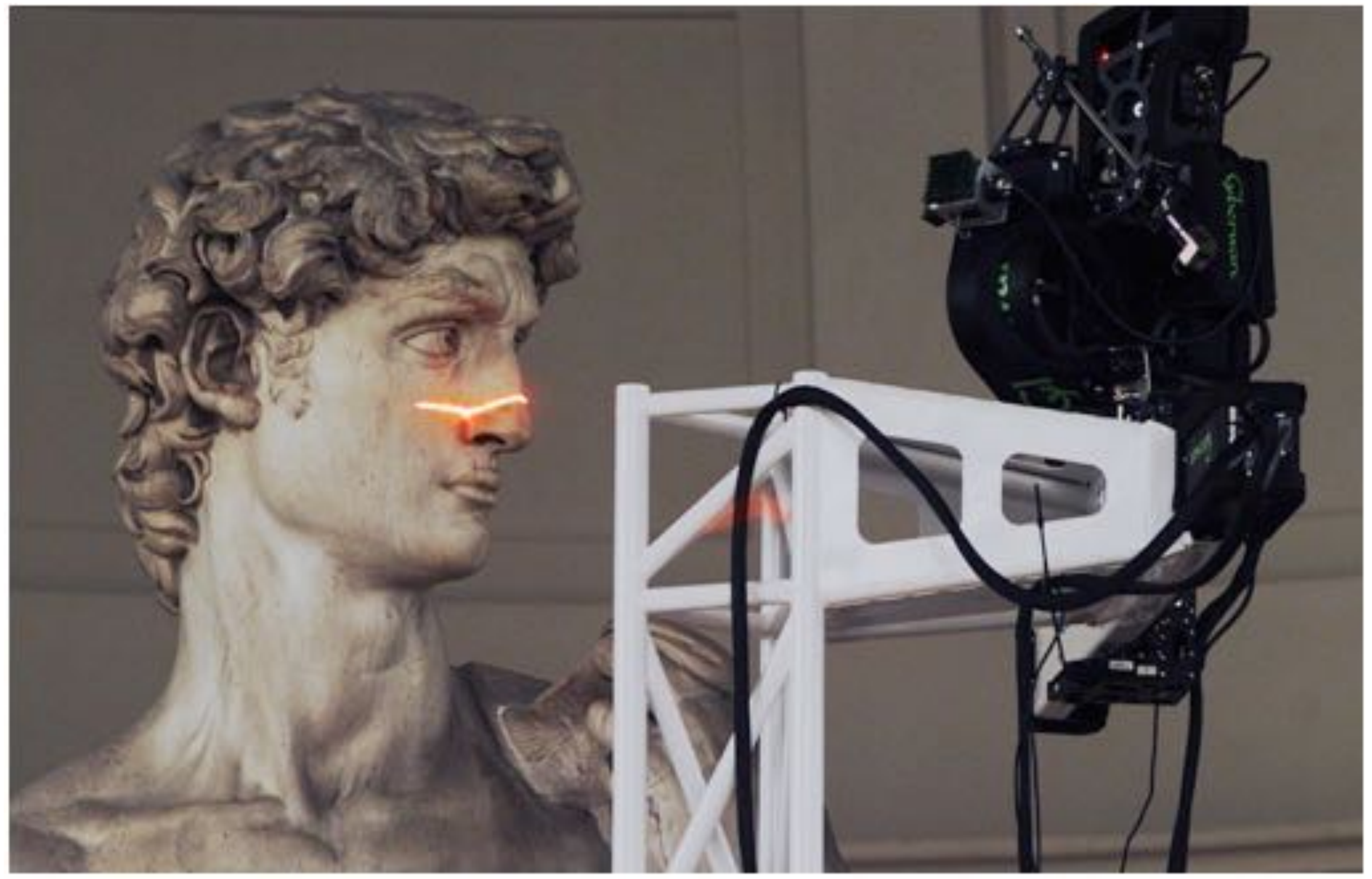
Where Are My Training Data?

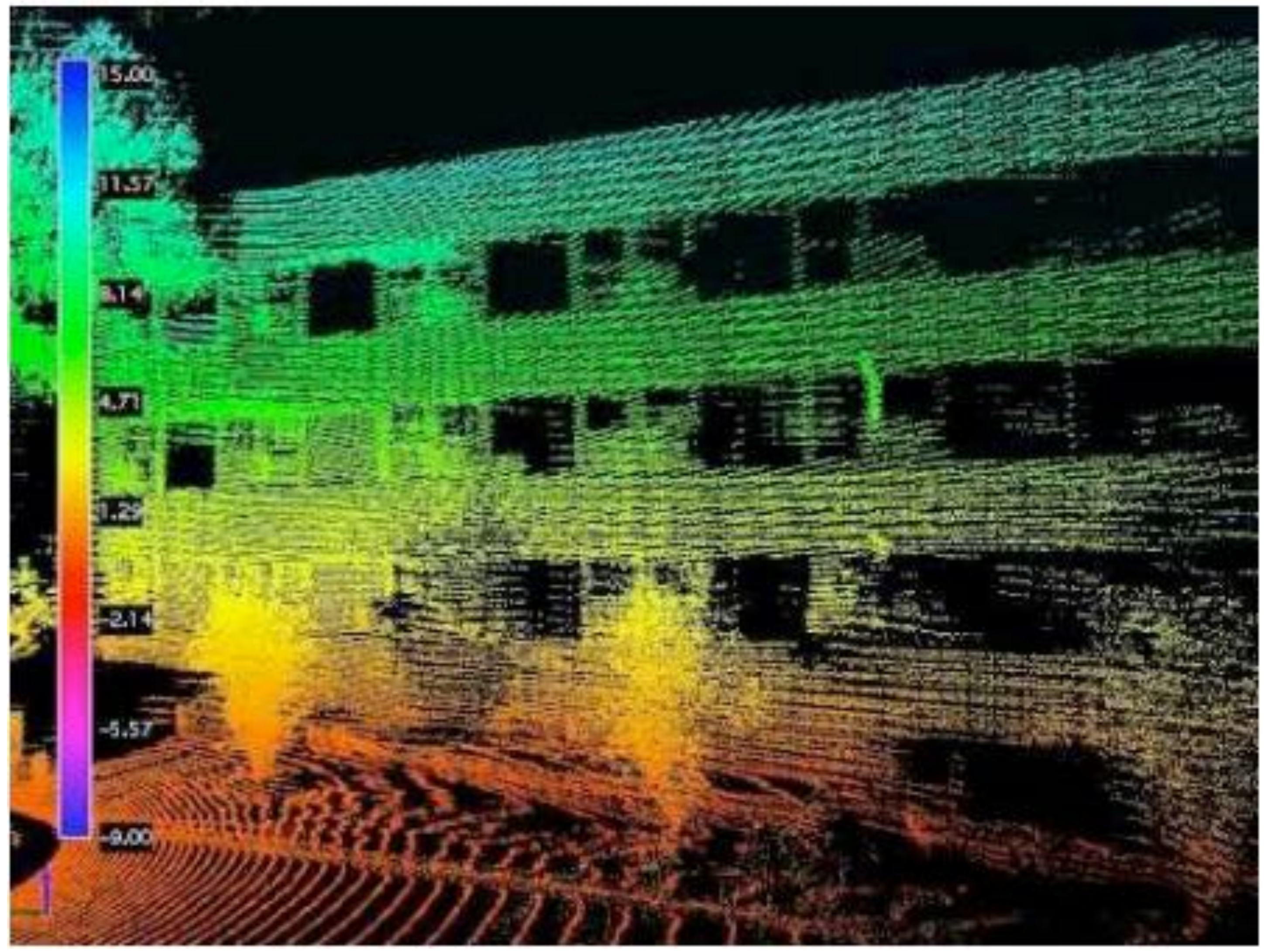
- In general, training deep networks needs a lot of data with labels!
- In our case, we need many image-3D shape pairs…
- Before talking about learning algorithms, obtaining training data is already a challenge!

Slides credit: Hao Su

Source I: Real Data

- Many techniques
  - Indoor: ToF or stereo sensors (Kinect, RealSense, …)
  - Outdoor: LiDAR

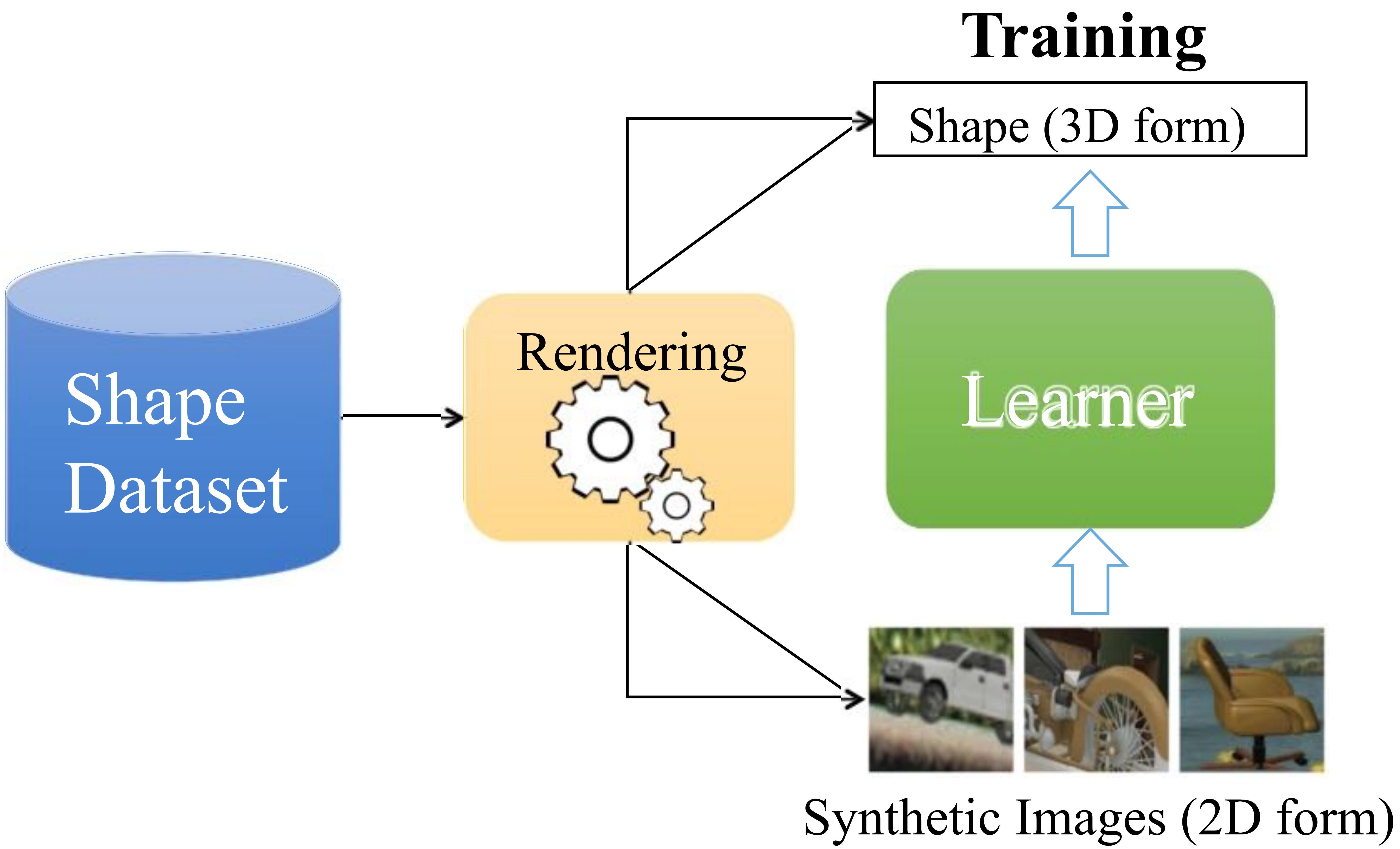




- The amount of real data is increasing quickly

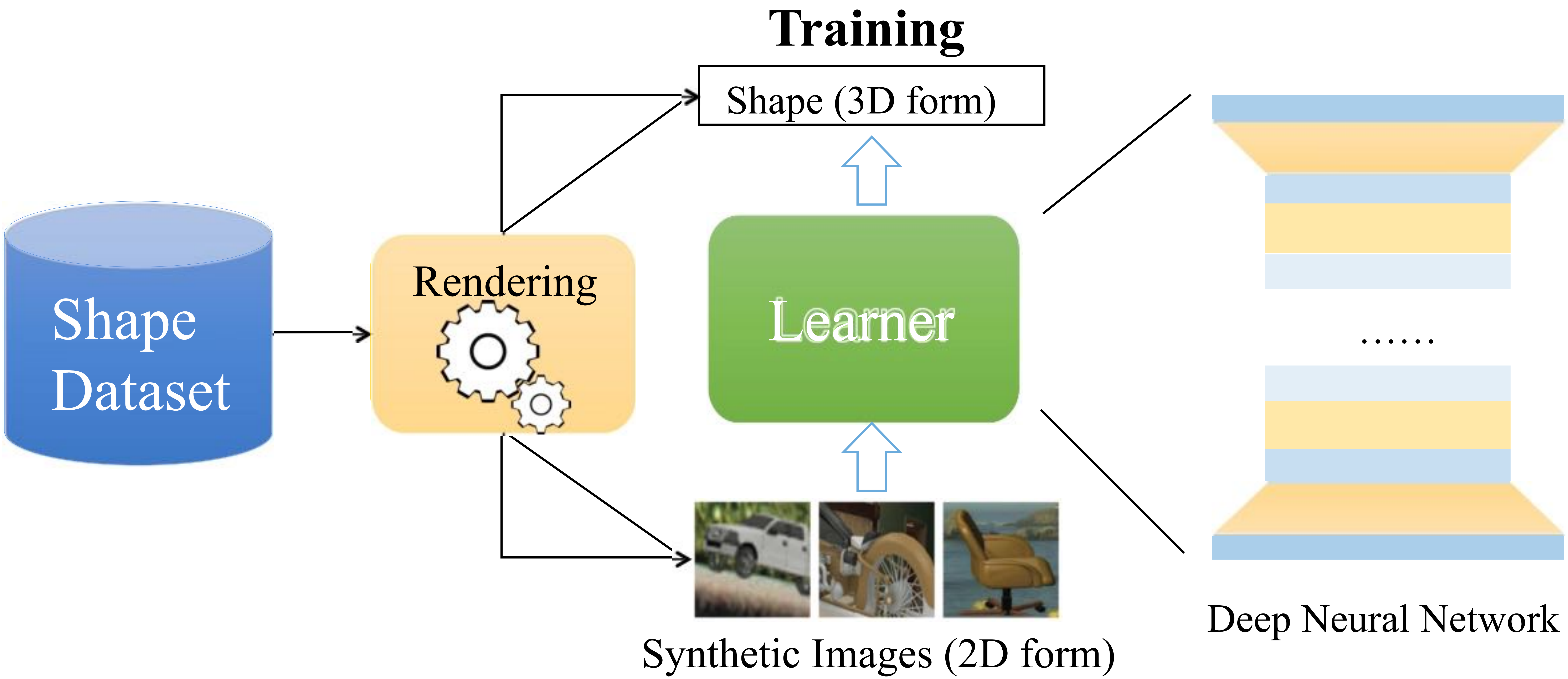
Slides credit: Hao Su

Source II: Synthesis for Learning



[Su et al., RenderForCNN, ICCV15]

Source II: Synthesis for Learning

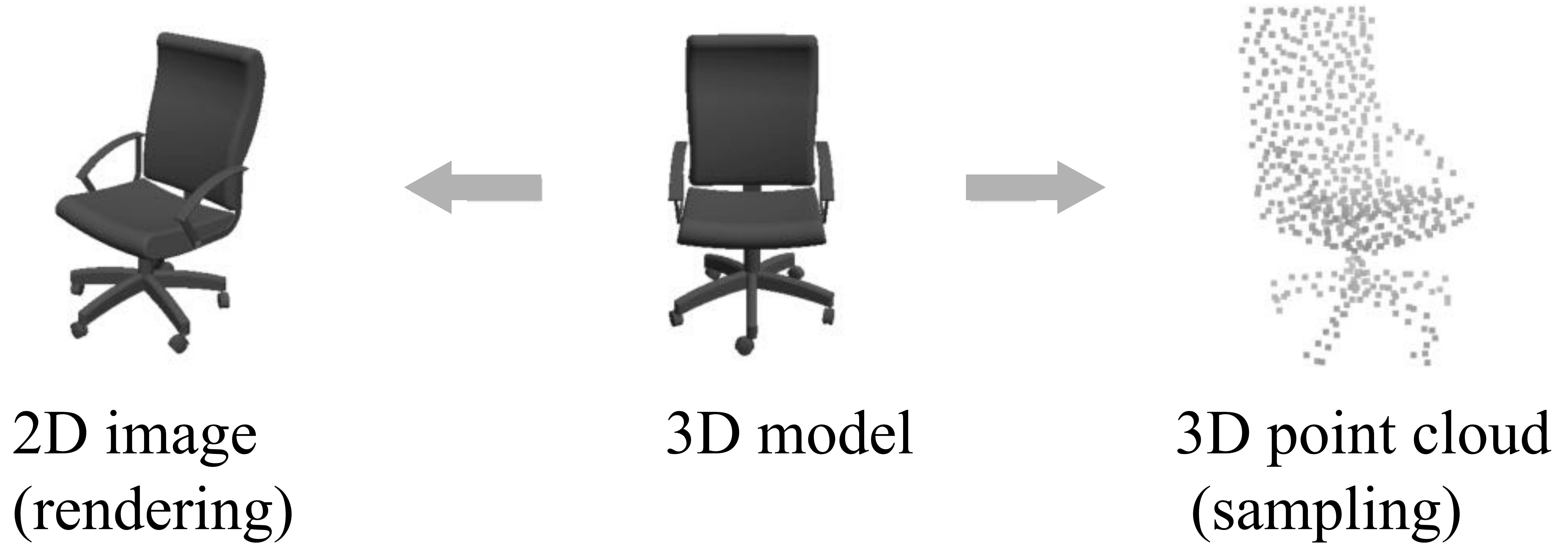


[Su et al., RenderForCNN, ICCV15]

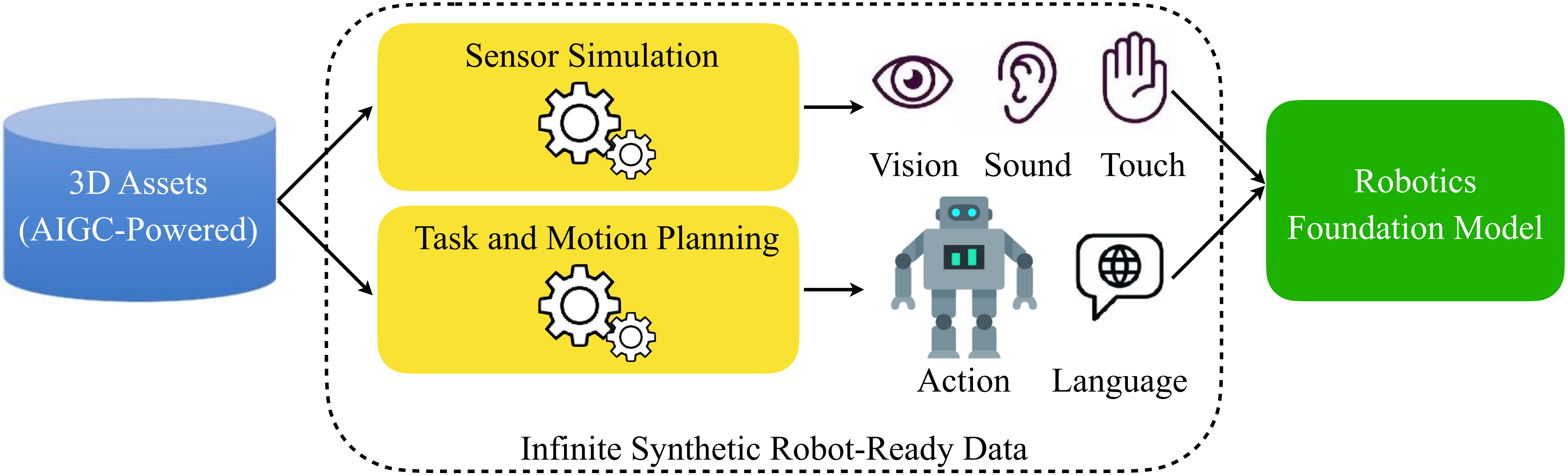
Slides credit: Hao Su

Source II: Synthesis for Learning

- For example, image ⟶ point cloud



Synthesis for Learning Beyond 3D Reconstruction



Large-Scale Synthetic 3D Dataset

- For example,
  - ShapeNet: http://www.shapenet.org

…

Large-Scale Synthetic 3D Dataset

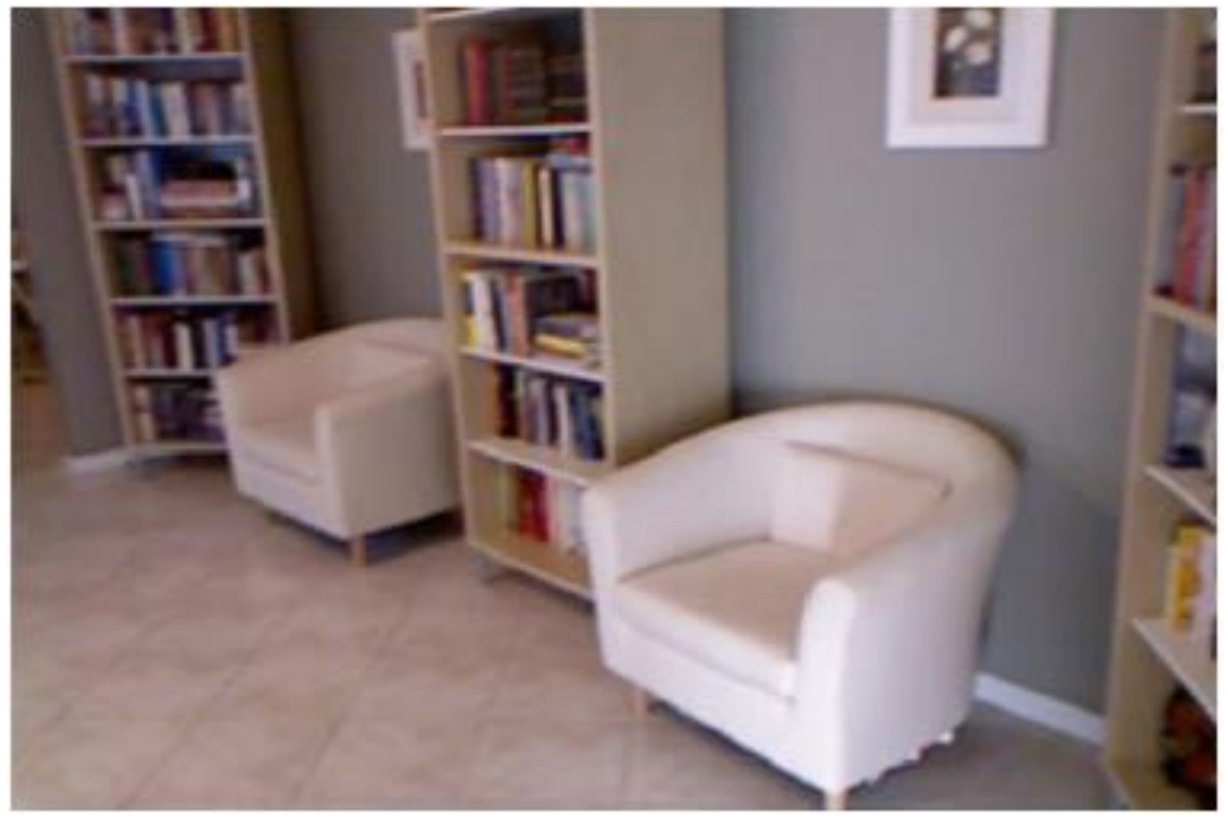
- For example,
  - Objaverse-XL (10M CAD): https://objaverse.allenai.org/

…

A Very Coarse Literature Review

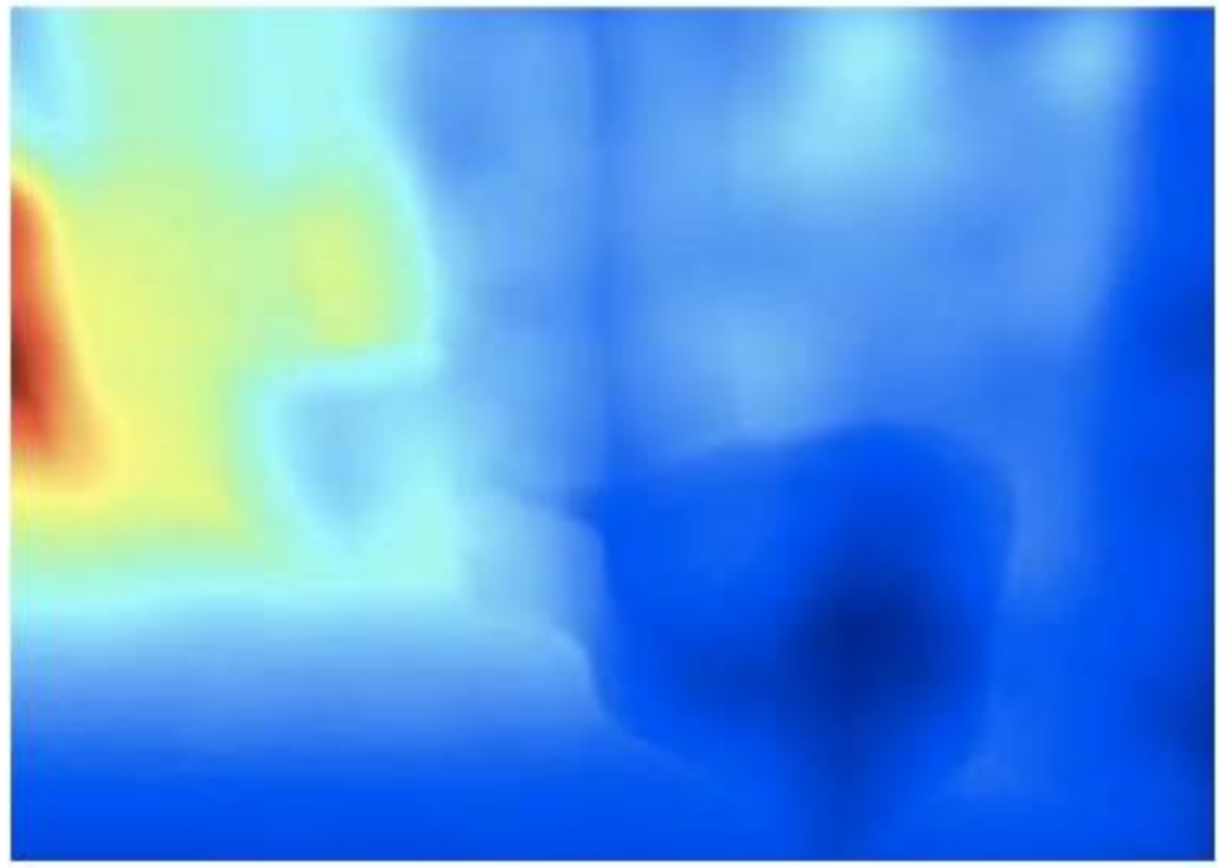
Literature: to Depth Map

- Fully-convolutional



Input image

(a) Normal map



(b) Depth map

Qi et al., " **GeoNet: Geometric Neural Network for Joint Depth and Surface Normal Estimation**", *CVPR 2018*

Recall: Issue of L_p Depth Loss

Prediction Groundtruth

- Common strategy: Depth-Normal consistency
- Review lecture 5
- Limitation: partial 3D info from camera view

Qi et al., " **GeoNet: Geometric Neural Network for Joint Depth and Surface Normal Estimation**", *CVPR 2018*

Slides credit: Hao Su

Literature: to Point Cloud

- From a single image to 3D point cloud generation.

Input image

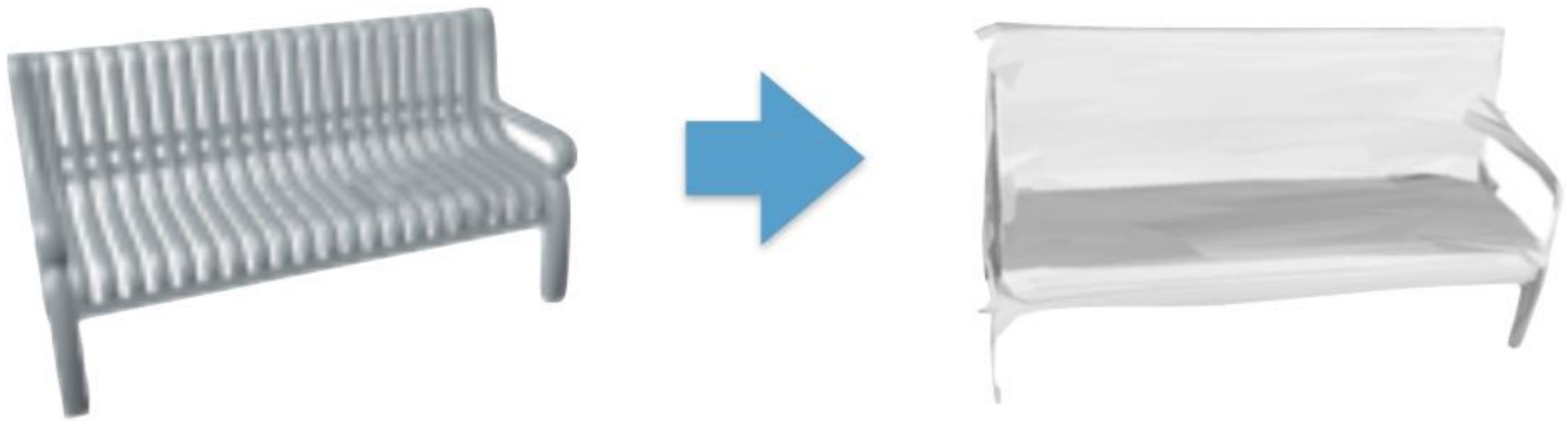
Reconstructed 3D point cloud

Fan et al., " **A Point Set Generation Network for 3D Object Reconstruction from a Single Image** ", *CVPR 2017*

Explain with Details Later

Literature: to Mesh

- From a single image to mesh surface.



Input image

Reconstructed 3D mesh

Groueix et al, **AtlasNet:A papier-mâché approach to learning 3d surface generation**, CVPR 2018

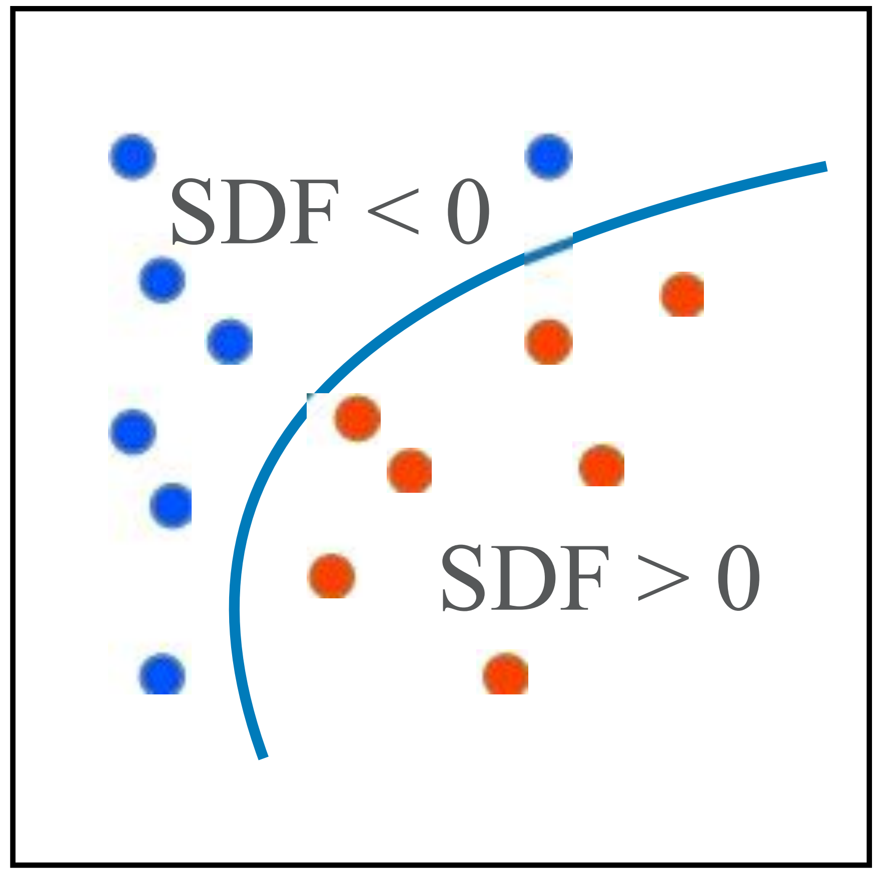
Explain with Details Later

Slides credit: Hao Su

Literature: to Implicit Field Function

- From a single image to implicit field function.

Input image



Implicit field function

$$F(x) = 0$$

Mescheder et al., "**Occupancy networks: Learning 3d reconstruction in function space** ", *CVPR 2019*

Check by Yourself

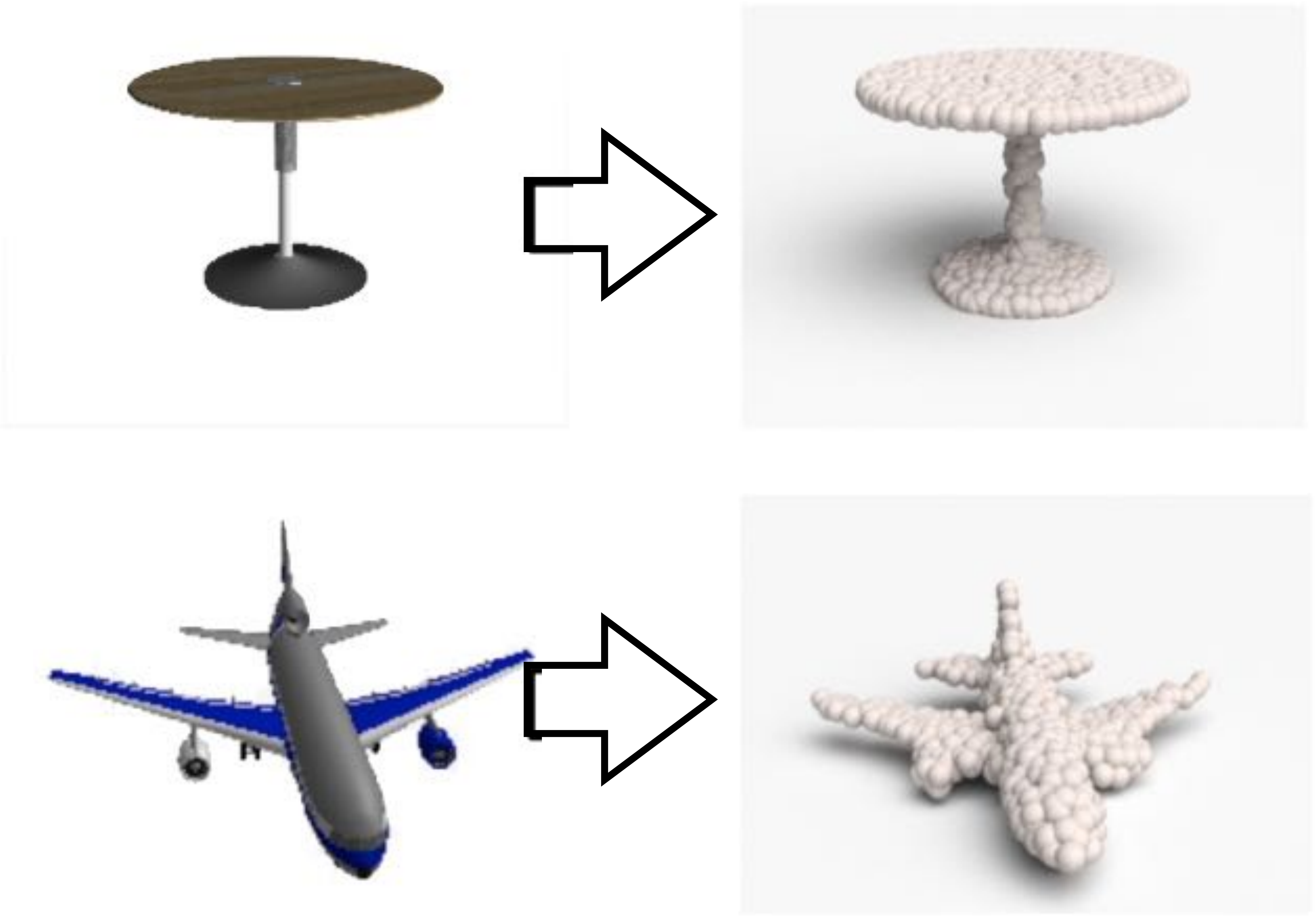
Slides credit: Hao Su

Outline

- Task
- Synthesis-for-Learning Pipeline
- ***Single-image to Point Cloud***
- Single-image to Mesh

Why Point Representation?

- Previous depth map covers only visible area.
- A flexible representation
  - A few thousands of points can model a great variety of shapes.



Slides credit: Hao Su

Point Cloud as a Set

3D mesh

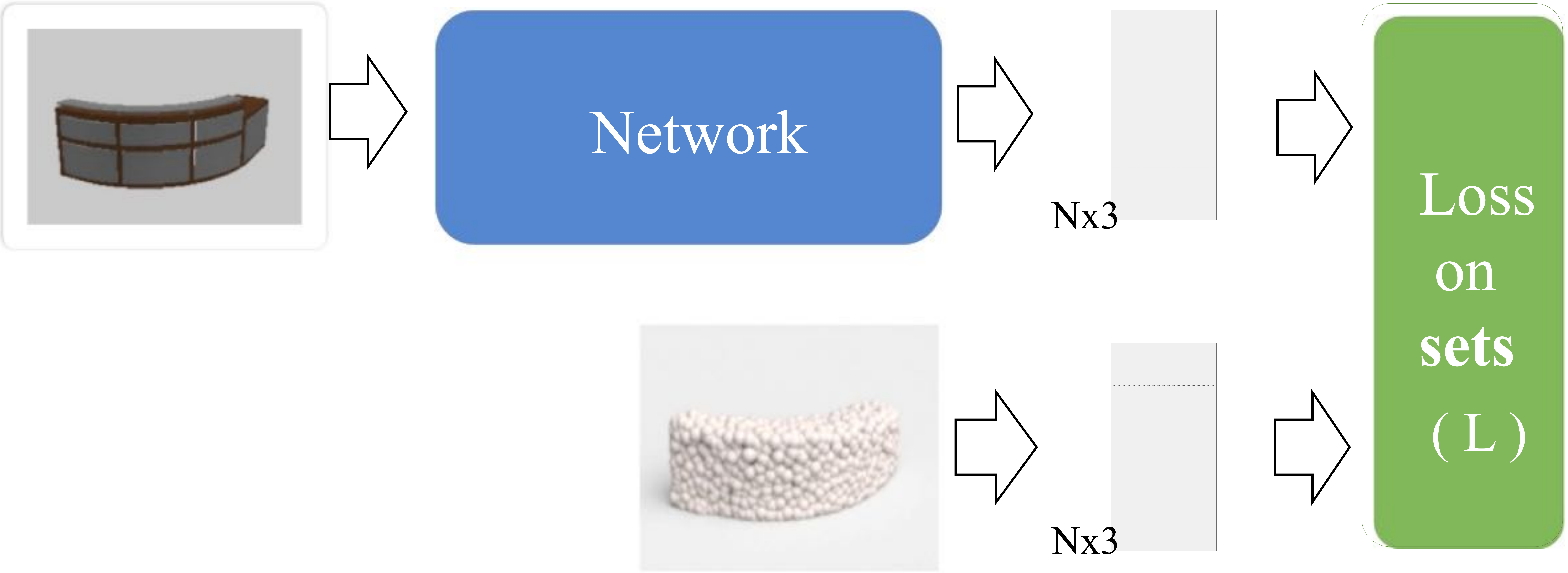
sampling

as

$$\left\{ \begin{array}{c} (x_1, y_1, z_1) \\ (x_2, y_2, z_2) \\ \dots \\ (x_n, y_n, z_n) \end{array} \right\}$$

Slides credit: Hao Su

Pipeline



Fan et al., " **A Point Set Generation Network for 3D Object Reconstruction from a Single Image** ", *CVPR 2017*

Slides credit: Hao Su

Real-world Results

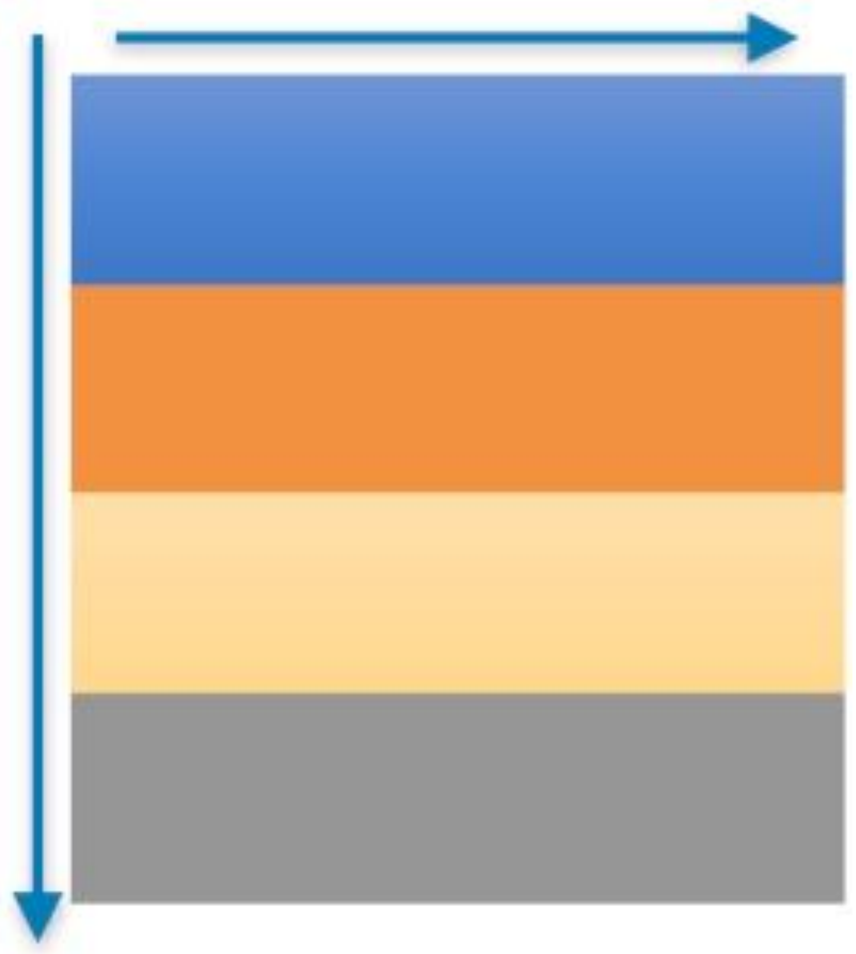
Some results

input observed view 90° input observed view 90°

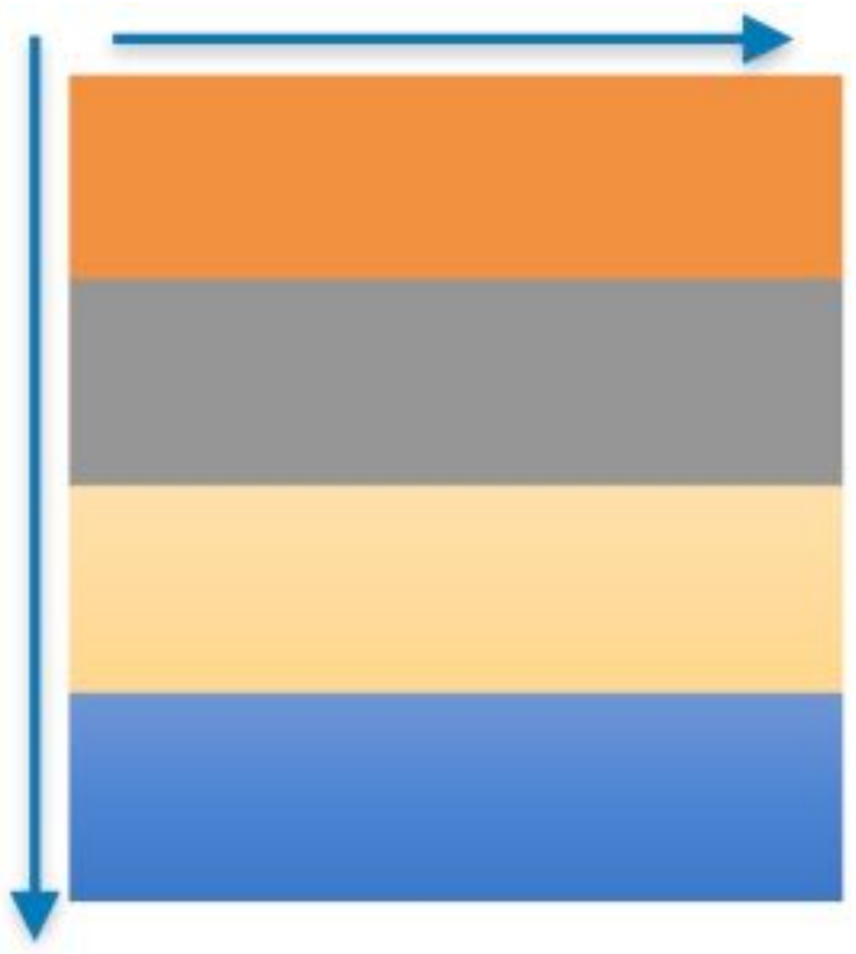
Differentiable Loss for Point Clouds

Permutation Invariance

- Point cloud: N orderless points, each represented by a D dim vector



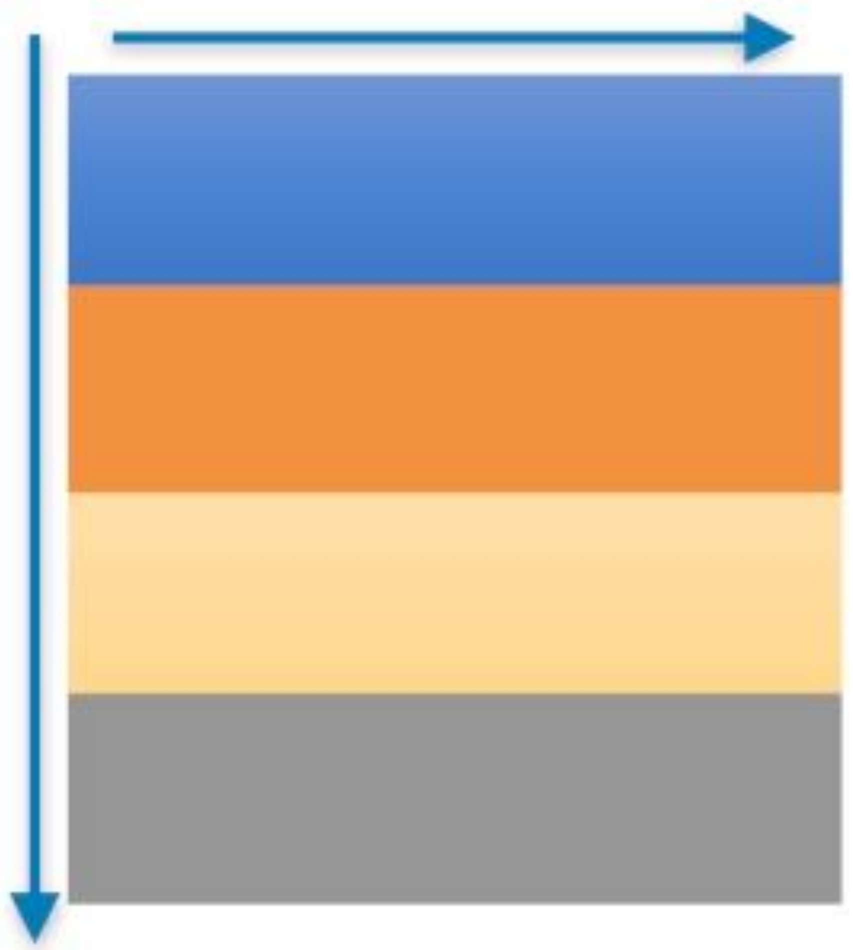
represents the same **set** as



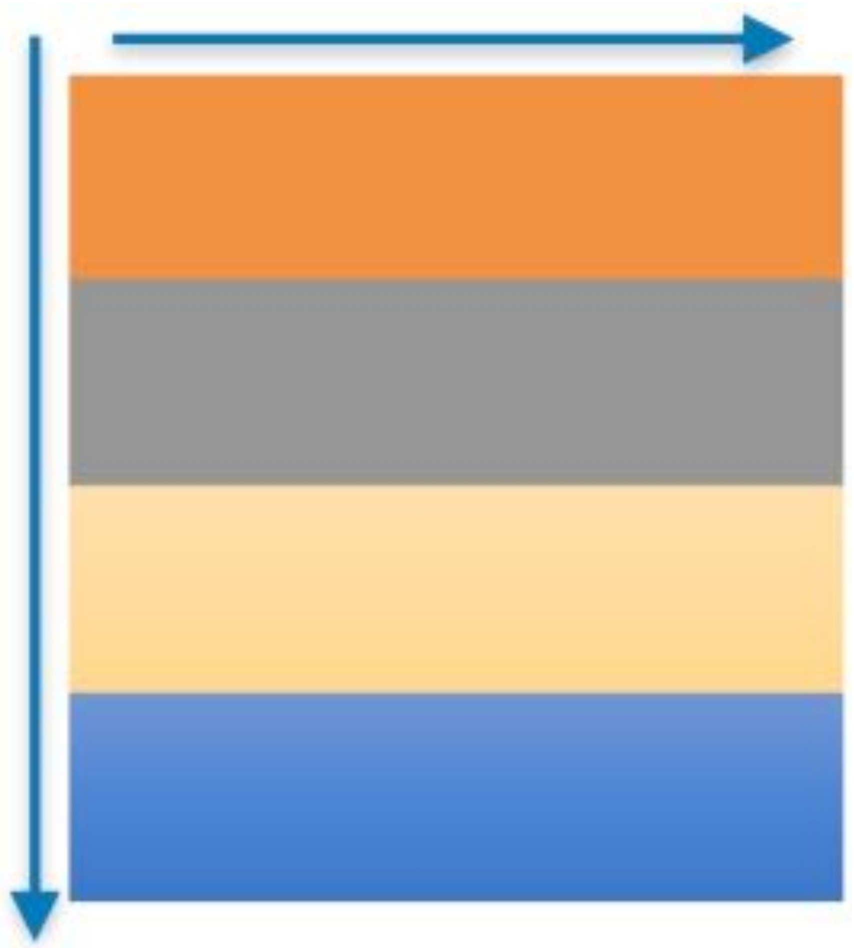
Slides credit: Hao Su

Permutation Invariance

- Point cloud: N orderless points, each represented by a D dim vector



represents the same **set** as



Loss needs to be invariant to ordering of points!

Slides credit: Hao Su

Metric for Point Clouds

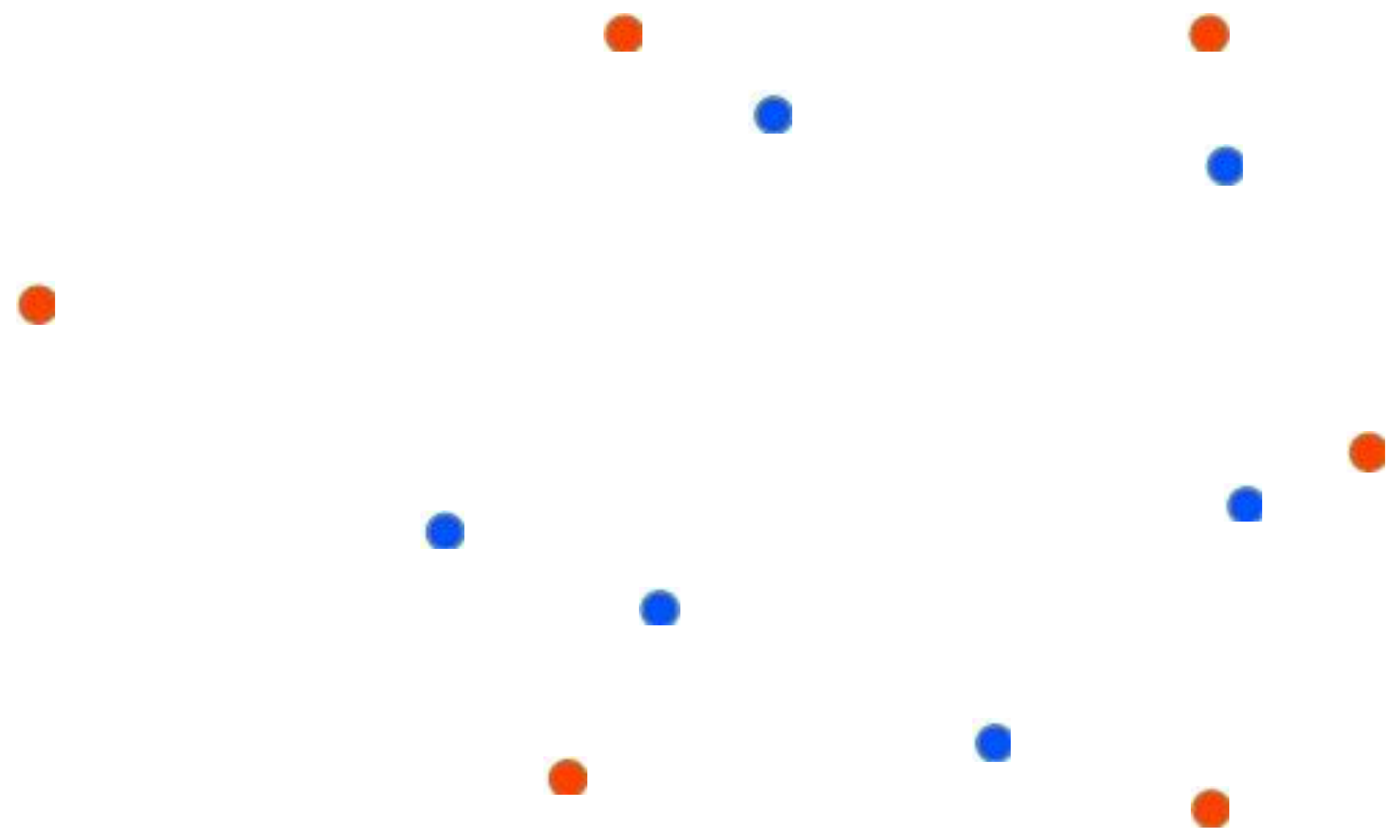
- L_2 loss does not work for point cloud.

- Need a metric to measure distance between two point sets

- Two popular choices
  - Earth Mover's Distance
  - Chamfer Distance

Earth Mover's Distance

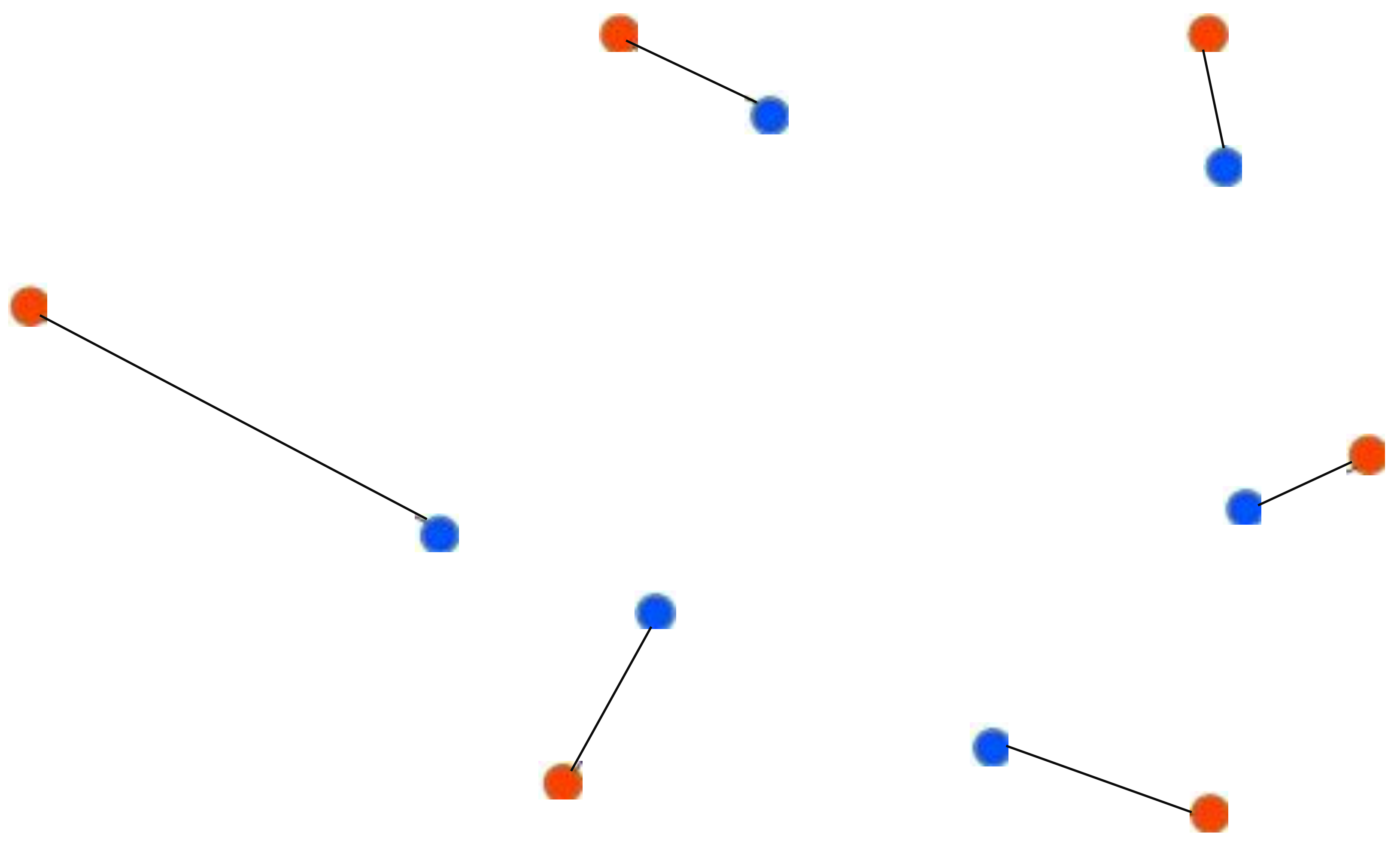
- Find a 1-1 correspondence between point sets



Slides credit: Hao Su

Earth Mover's Distance

- Find a 1-1 correspondence between point sets



$$d_{EMD}(S_1, S_2) = \min_{\phi: S_1 \to S_2} \sum_{x \in S_1} \|x - \phi(x)\|_2$$

where $\phi : S_1 \to S_2$ is a bijection

Earth Mover's Distance

$$d_{EMD}(S_1, S_2) = \min_{\phi: S_1 \to S_2} \sum_{x \in S_1} \|x - \phi(x)\|_2$$

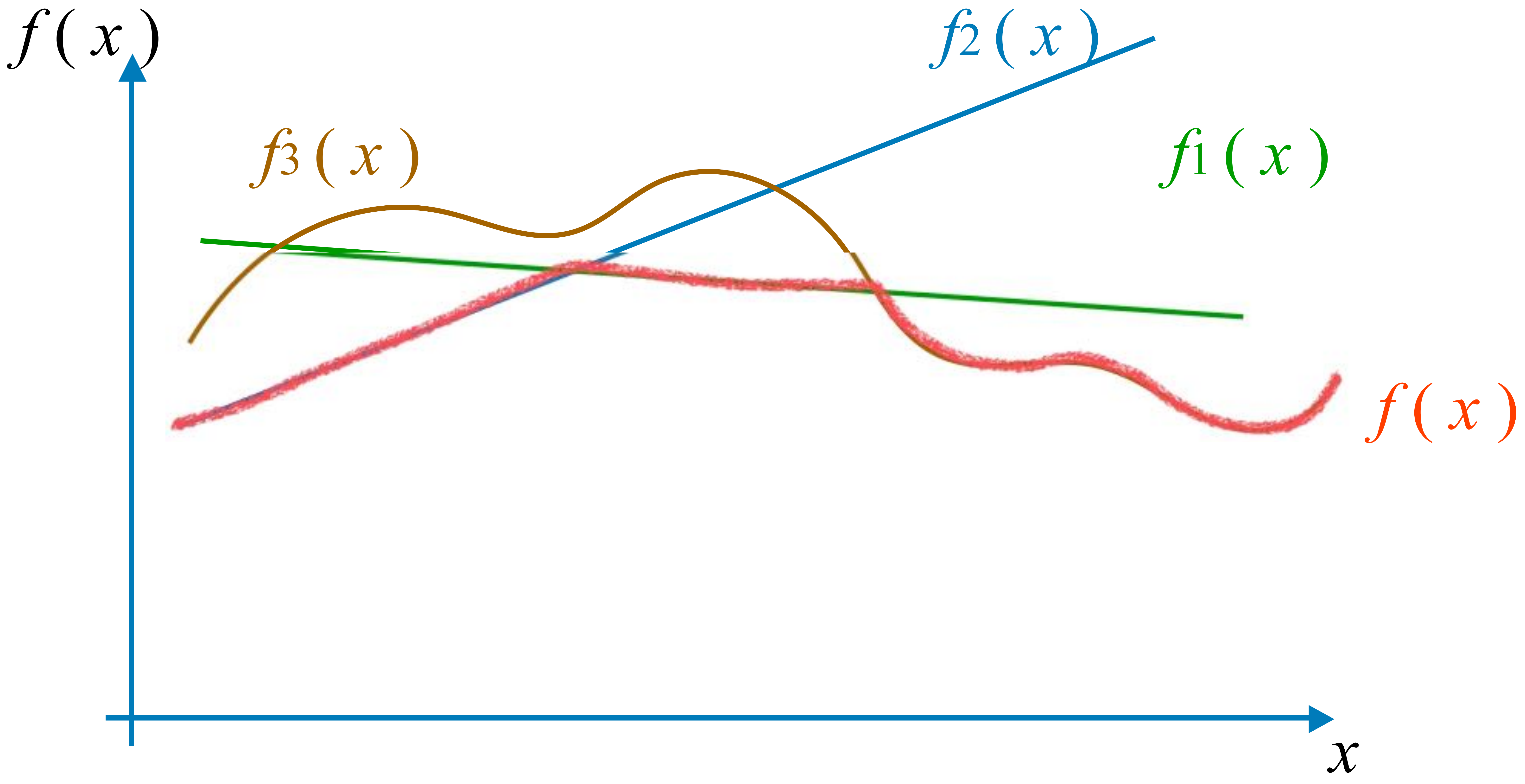
where $\phi : S_1 \to S_2$ is a bijection

Question:

Viewing $d_{EMD}(S_1, S_2)$ as a function of point coordinates in S_1 , is this function **continuous** ?

Lemma

- For a family of continuous functions $\{f_i(x)\}$, the point-wise minimum $f(x) = \min_i \{f_i(x)\}$ is continuous.



Slides credit: Hao Su

Continuity of $d_{EMD}(S_1, S_2)$

$$d_{EMD}(S_1, S_2) = \min_{\phi: S_1 \to S_2} \sum_{x \in S_1} \|x - \phi(x)\|_2$$

where $\phi : S_1 \to S_2$ is a bijection

- $\phi(x)$ defines a point-wise correspondence ($n!$ possibilities, $n =$ size of S_1).
- For a fixed ϕ, define $f_\phi(S_1) = \sum_{x \in S_1} \|x - \phi(x)\|_2$, and $f_\phi(S_1)$ is obviously continuous
- $d_{EMD}(S_1, S_2) = \min_{\phi: S_1 \to S_2} f_\phi(S_1)$ is thus continuous!

Slides credit: Hao Su

Differentiable?

$$d_{EMD}(S_1, S_2) = \min_{\phi:S_1 \to S_2} \sum_{x \in S_1} \|x - \phi(x)\|_2$$

where $\phi : S_1 \to S_2$ is a bijection

- From the example, we see that $d_{EMD}(S_1, S_2)$ can be constructed in a piece-wise manner
- Inside each piece, it is $f_{\phi_i}(S_1)$ by some ϕ_i, which is obviously differentiable (as $\phi_i(x)$ is a constant)
- $d_{EMD}(S_1, S_2)$ is differentiable except for zero-measure set!

Implementation

- Many algorithmic study on fast EMD computation (a specific bipartite matching problem)

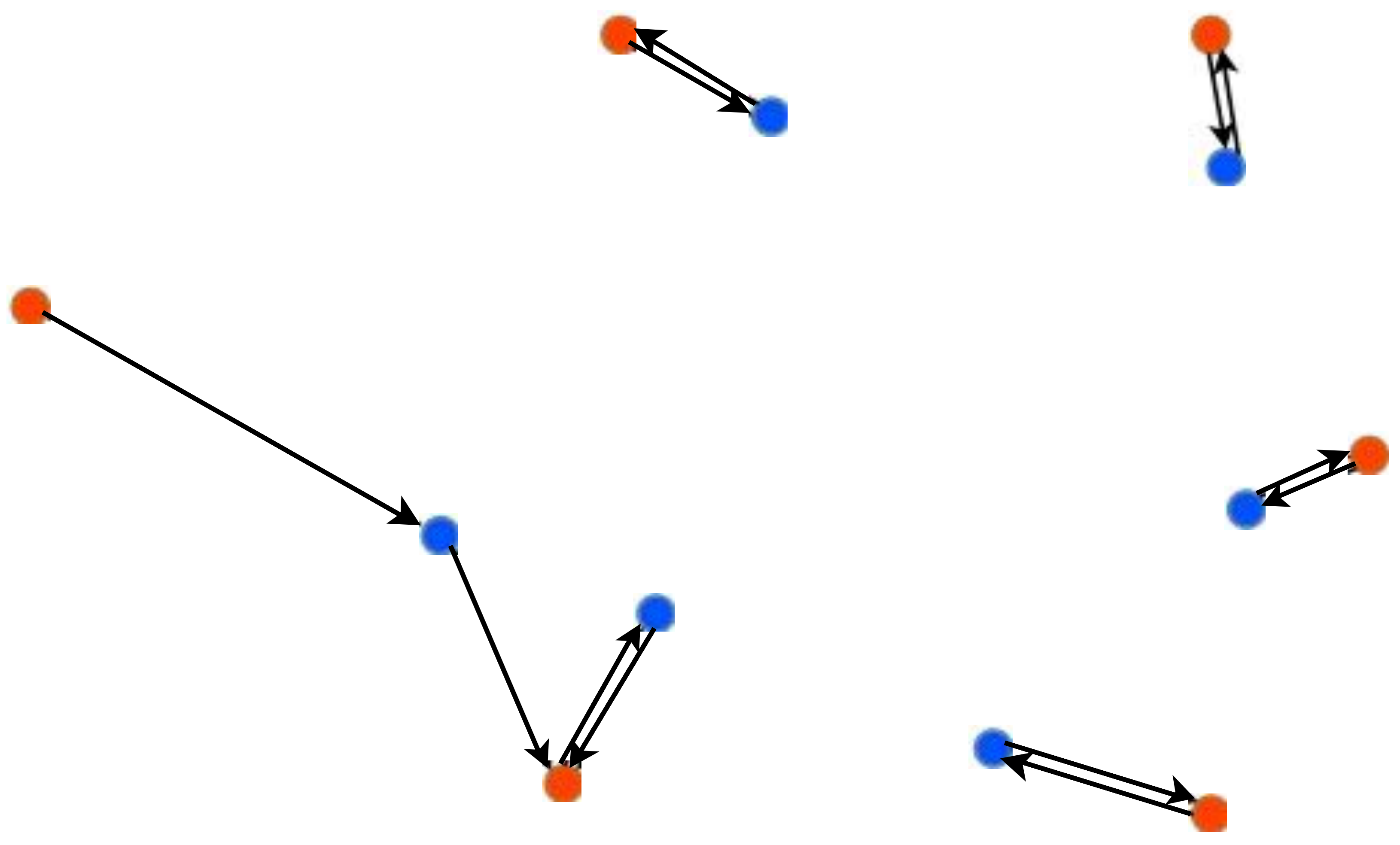
- There exists parallelizable implementation of EMD on CUDA

- A fast implementation (approximated EMD): https://github.com/Colin97/MSN-Point-Cloud-Completion

Slides credit: Hao Su

Chamfer Distance

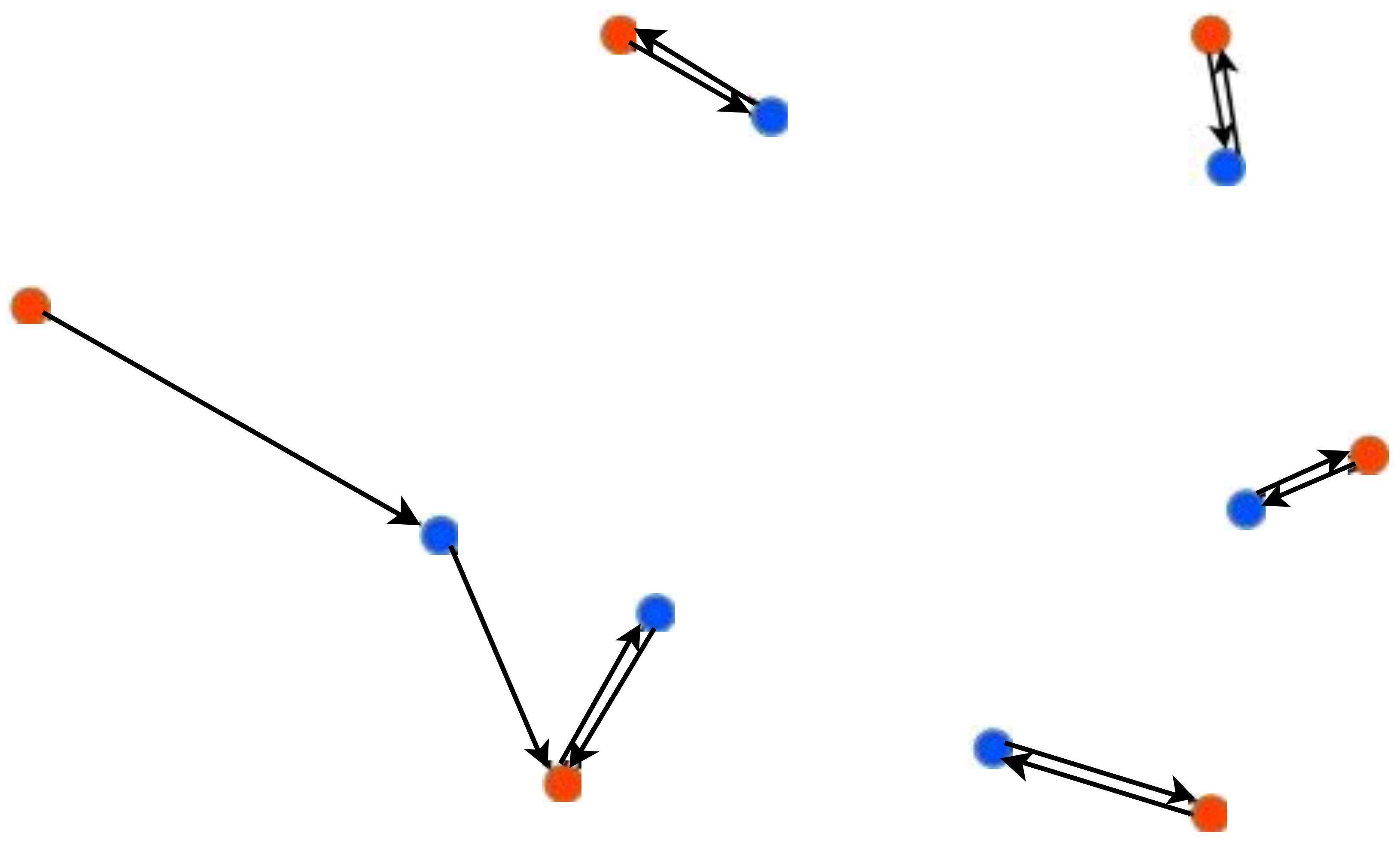
- Nearest neighbor correspondence for each point



Slides credit: Hao Su

Chamfer Distance

- Nearest neighbor correspondence for each point

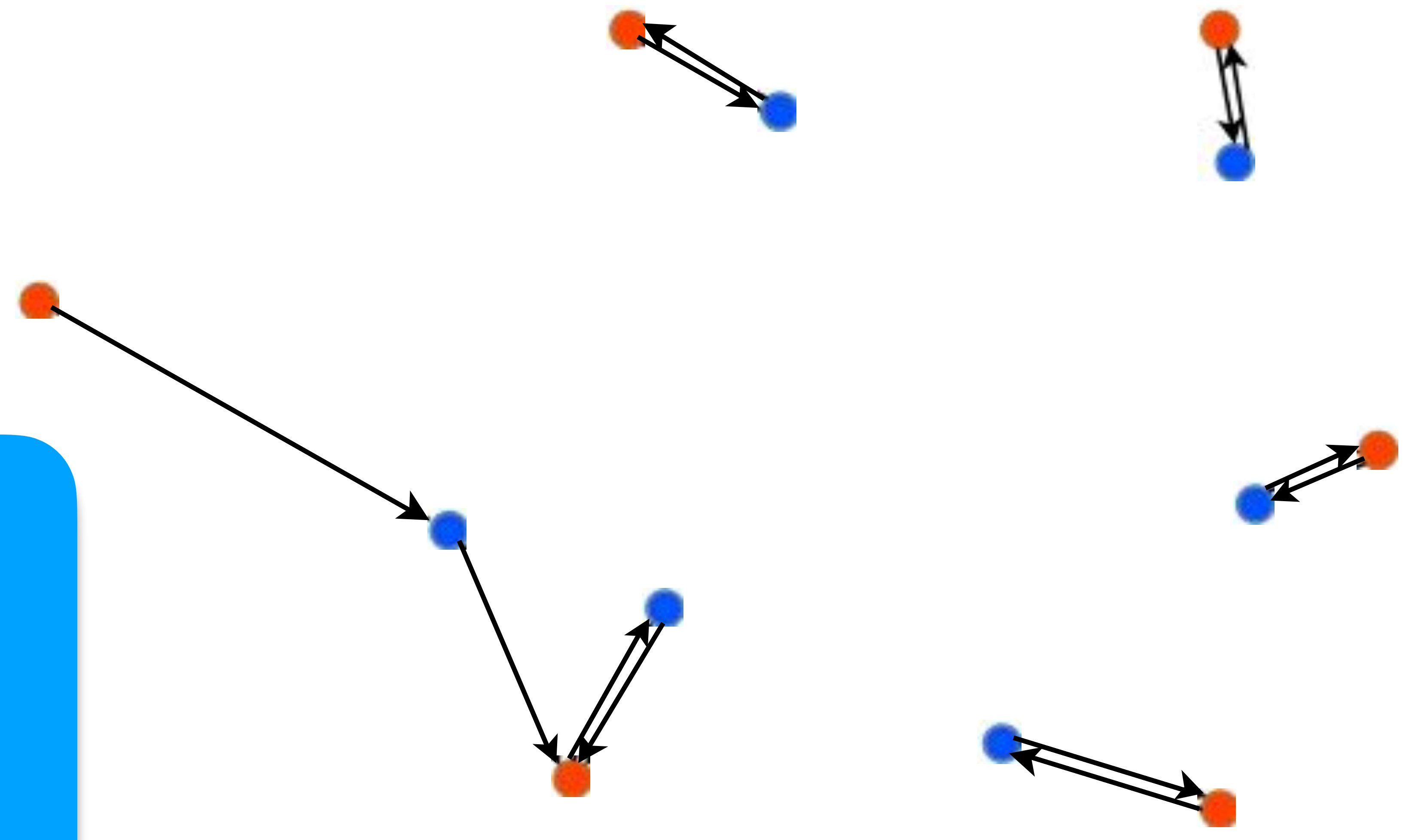


$$d_{CD}(S_1, S_2) = \sum_{x \in S_1} \min_{y \in S_2} \|x - y\|_2^2 + \sum_{y \in S_2} \min_{x \in s_1} \|x - y\|_2^2$$

Slides credit: Hao Su

Chamfer Distance

- Nearest neighbor correspondence for each point



Differentiable?

$$d_{CD}(S_1, S_2) = \sum_{x \in S_1} \min_{y \in S_2} \|x - y\|_2^2 + \sum_{y \in S_2} \min_{x \in s_1} \|x - y\|_2^2$$

How Distance Metric Affect Learning?

- A fundamental issue: inherent ambiguity in 2D-3D dimension lifting.

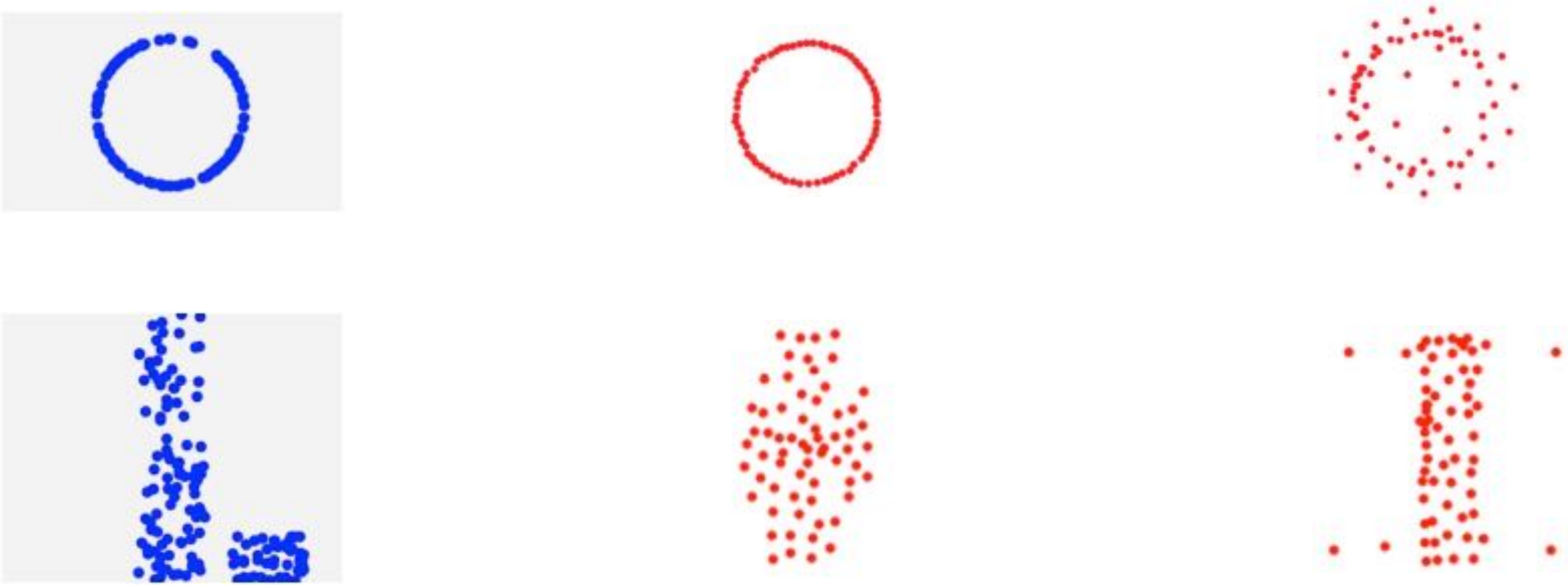
How Distance Metric Affect Learning?

- A fundamental issue: inherent ambiguity in 2D-3D dimension lifting.

- By loss minimization, the network tends to predict a " **mean shape**" that averages out uncertainty

Distance Metrics Affect Mean Shapes

- The mean shape carries characteristics of the distance metric.



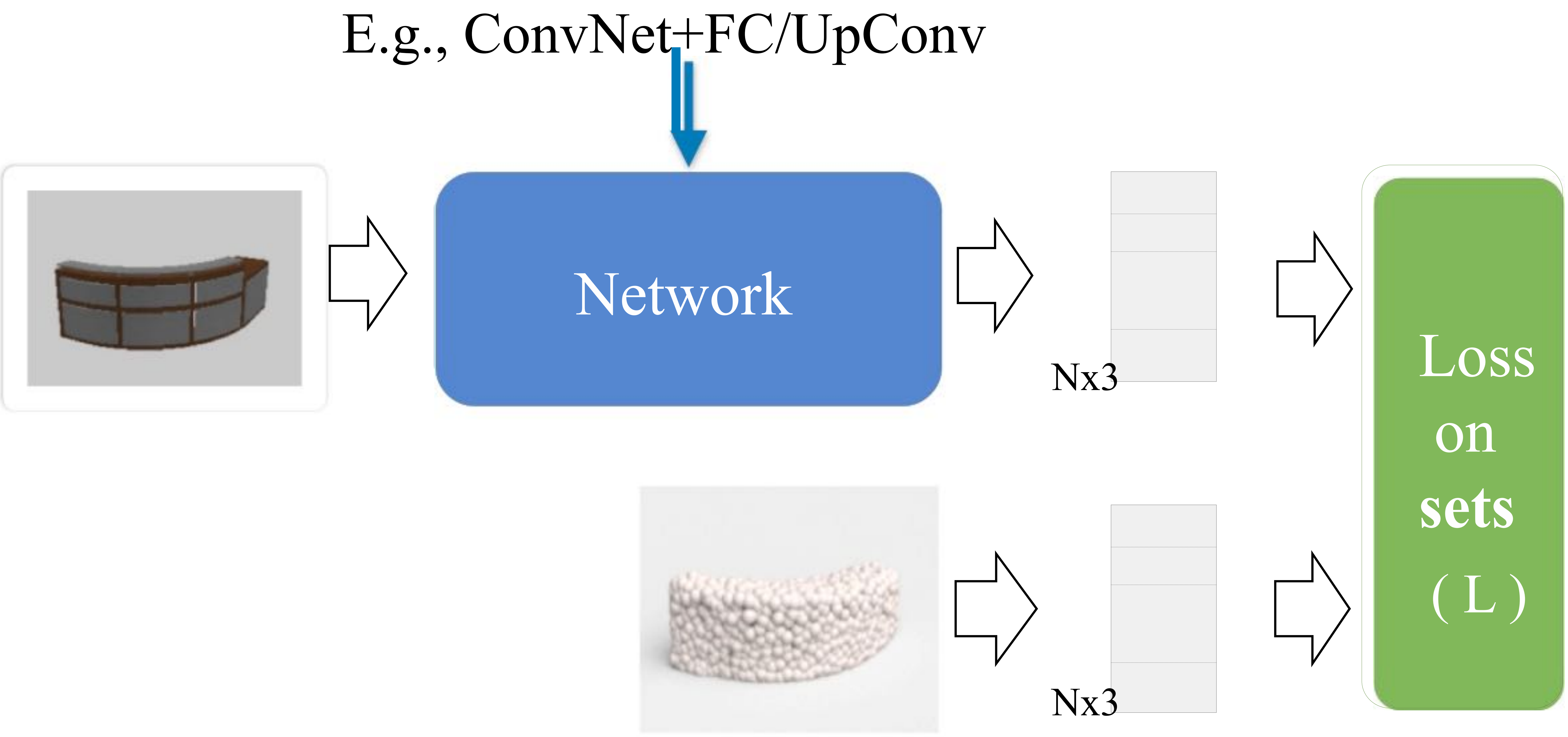
Continuous hidden variable (radius)

Discrete hidden variable (add-on location)

Input | EMD mean | CD mean

Slides credit: Hao Su

Network Choice: Certain Tricks



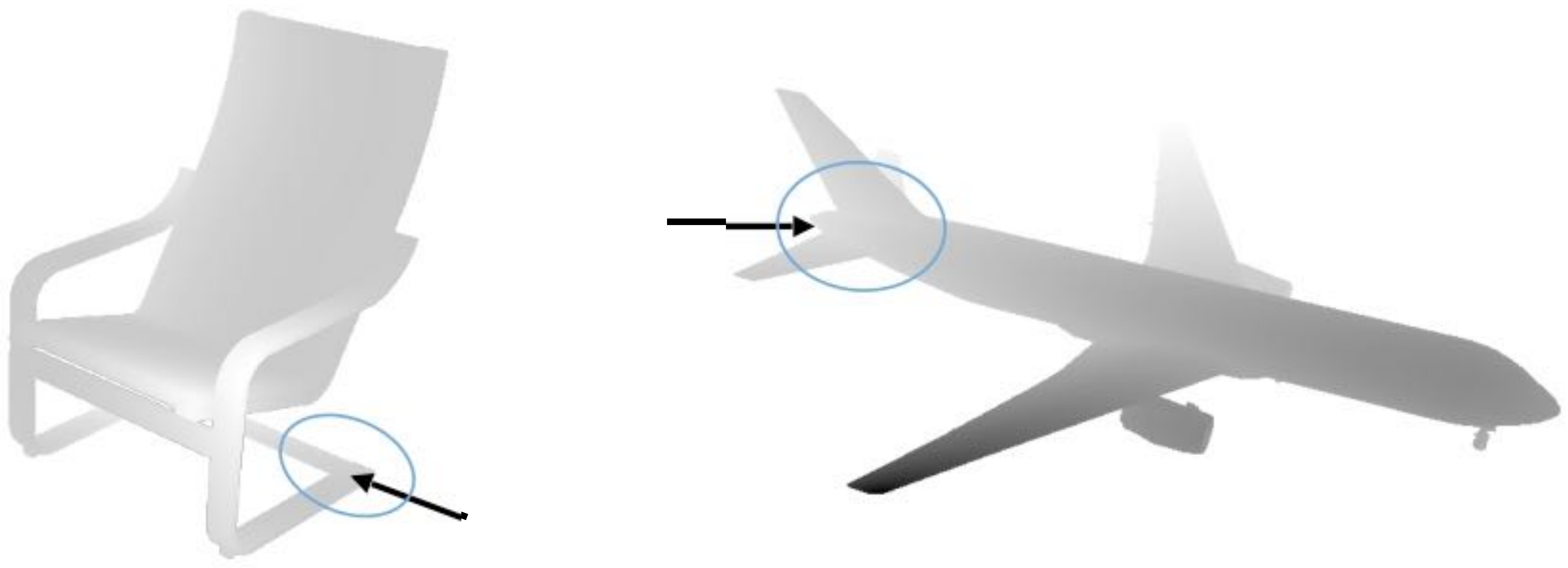
Fan et al., " **A Point Set Generation Network for 3D Object Reconstruction from a Single Image** ", *CVPR 2017*

Network Design: Respect Natural Statistics of Geometry

- Many local structures are common

Fan et al., " **A Point Set Generation Network for 3D Object Reconstruction from a Single Image** ", *CVPR 2017*

Network Design: Respect Natural Statistics of Geometry

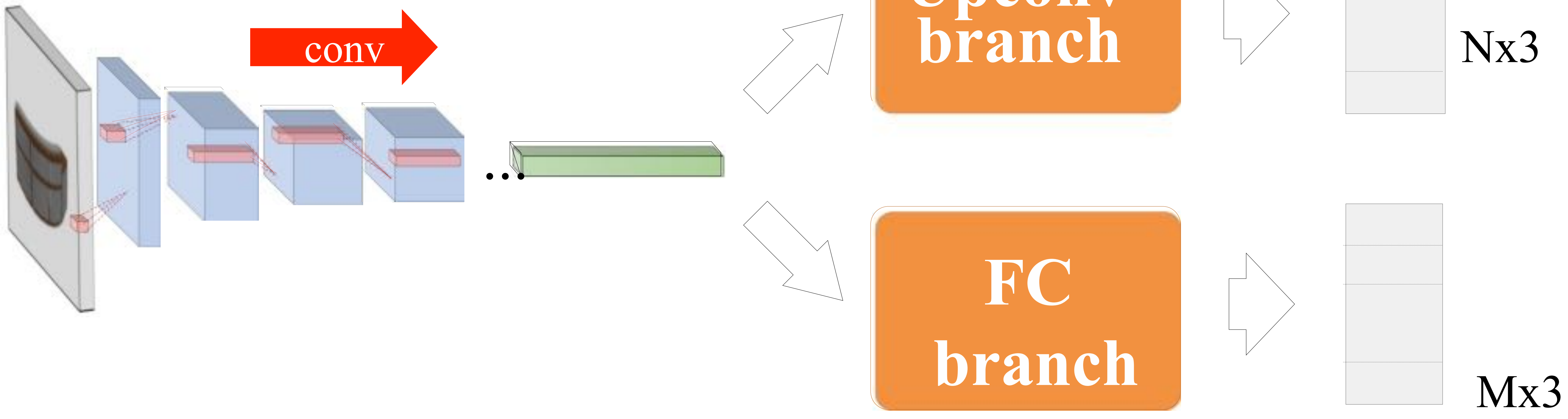


- Many local structures are common

- Also some intricate structures

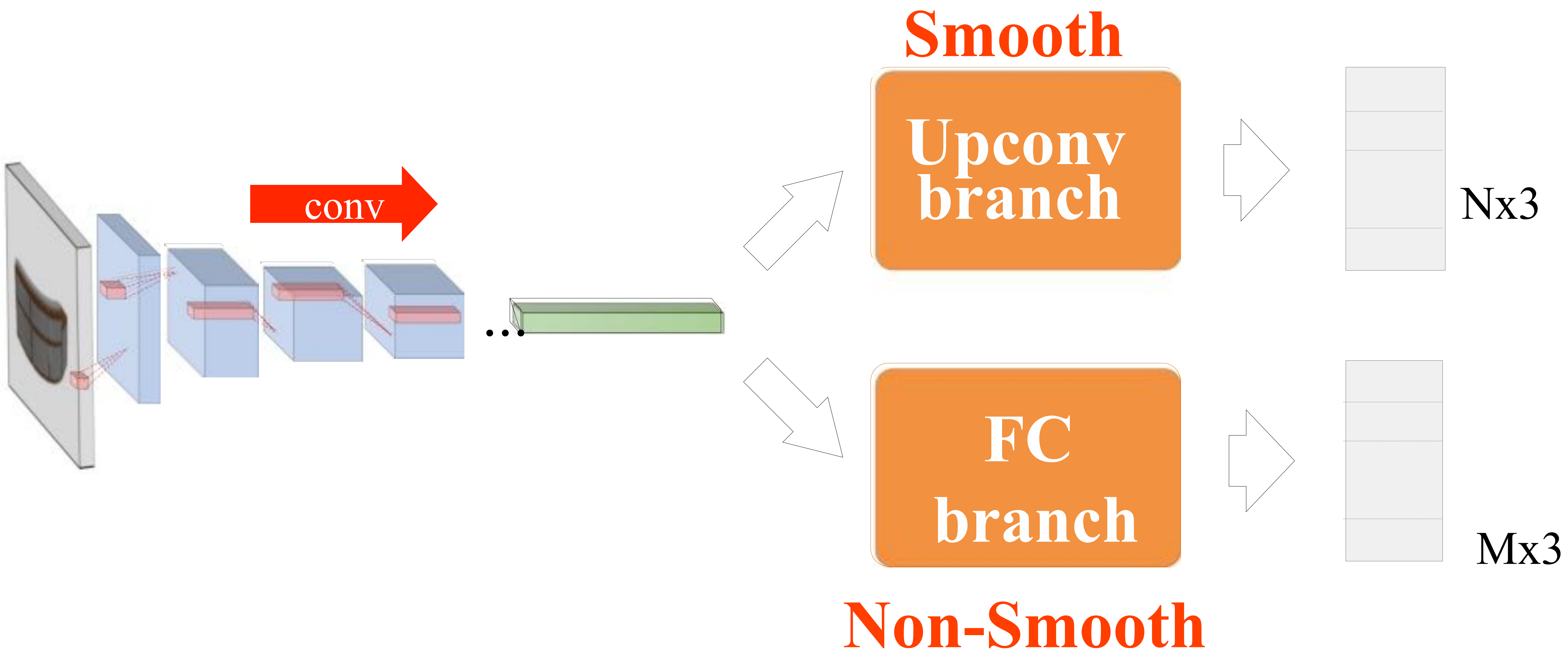
Fan et al., " **A Point Set Generation Network for 3D Object Reconstruction from a Single Image** ", *CVPR 2017*

Two-Branch Architecture



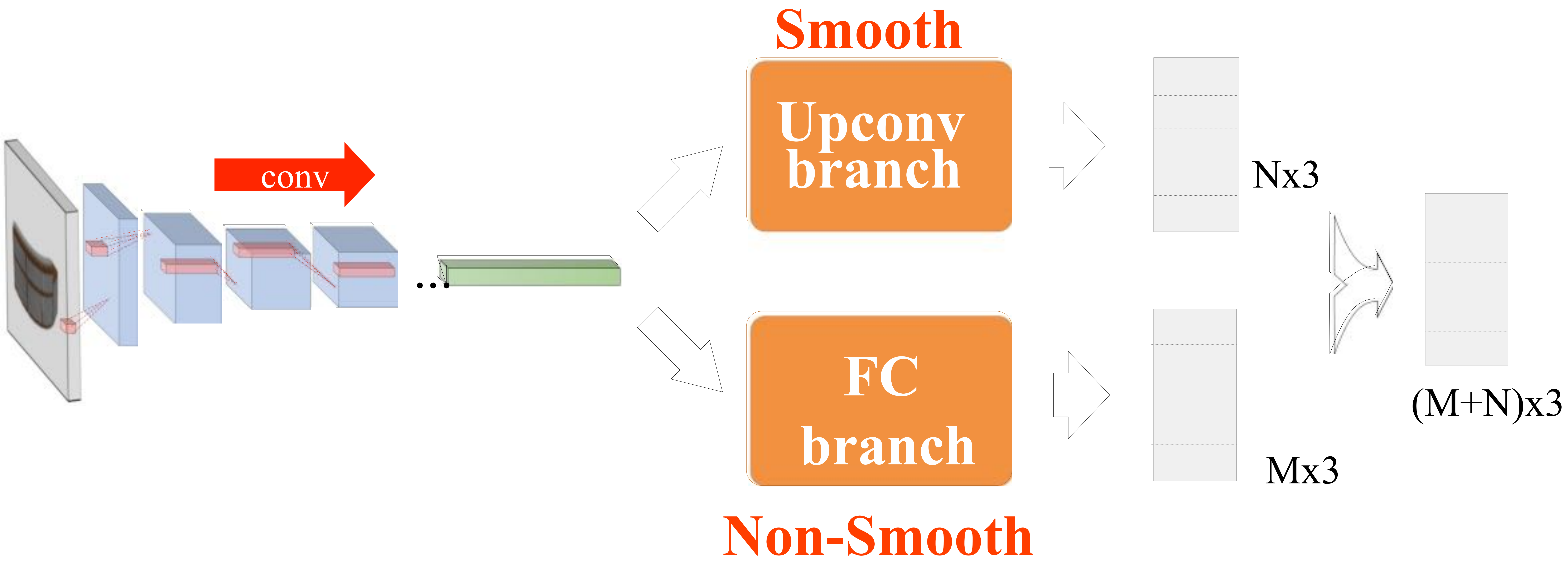
Fan et al., " **A Point Set Generation Network for 3D Object Reconstruction from a Single Image** ", *CVPR 2017*

Two-Branch Architecture



Fan et al., " **A Point Set Generation Network for 3D Object Reconstruction from a Single Image** ", *CVPR 2017*

Two-Branch Architecture



Set union by array concatenation

Fan et al., " **A Point Set Generation Network for 3D Object Reconstruction from a Single Image** ", *CVPR 2017*

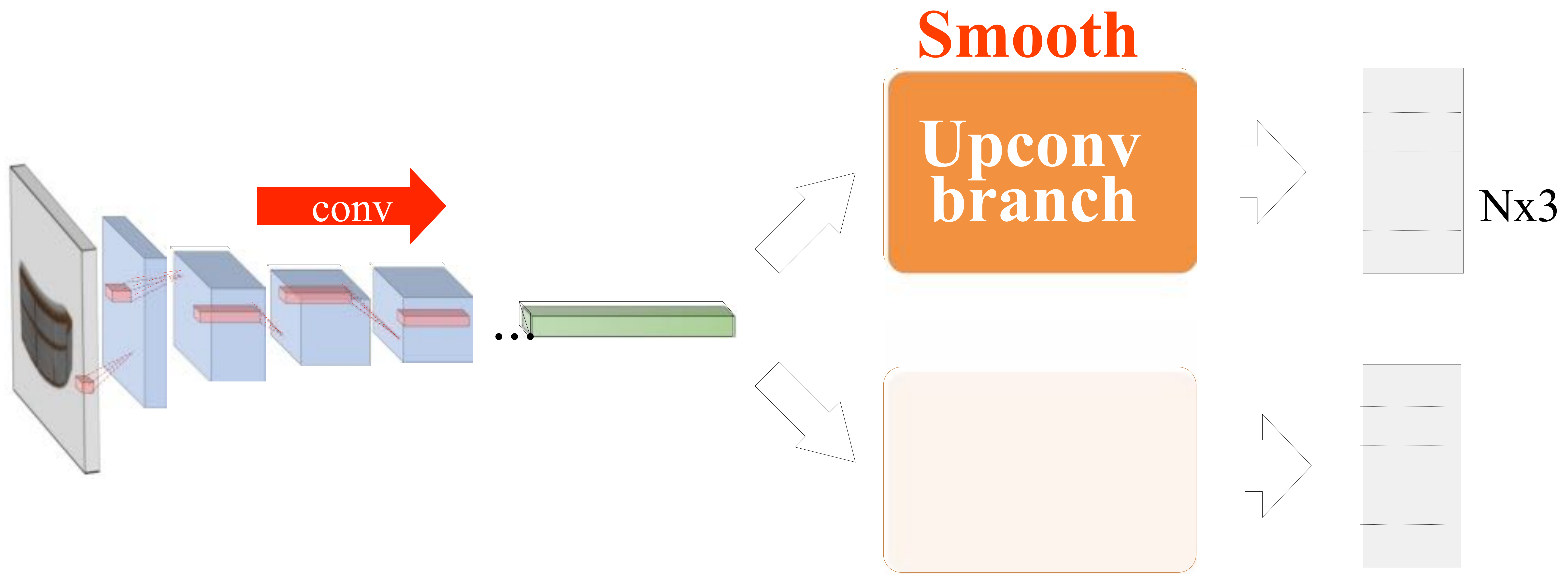
Slides credit: Hao Su

Which Color Corresponds to the Upconv Branch? FC Branch?

Fan et al., " **A Point Set Generation Network for 3D Object Reconstruction from a Single Image** ", *CVPR 2017*

Slides credit: Hao Su

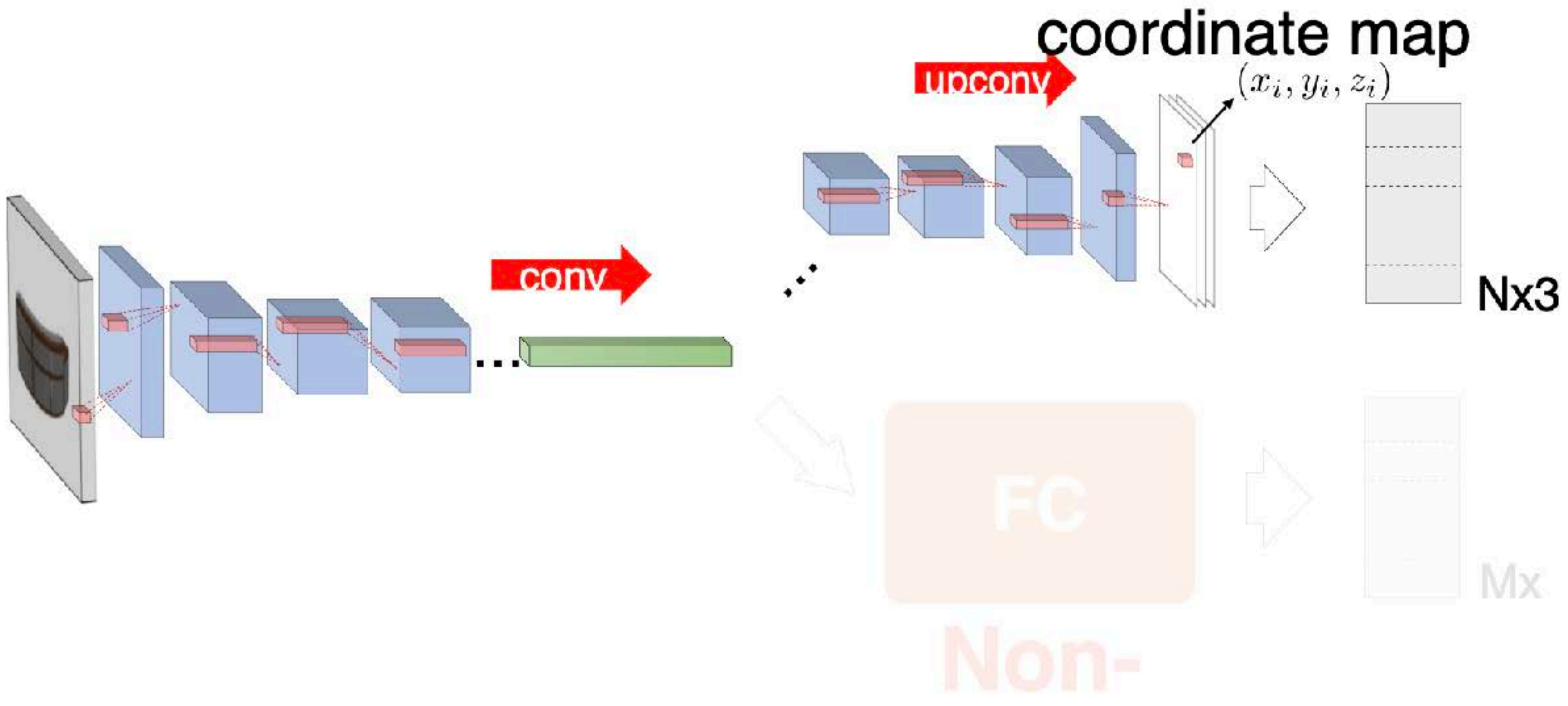
Design of Upconvolution Branch



Set union by array concatenation

Fan et al., " **A Point Set Generation Network for 3D Object Reconstruction from a Single Image** ", *CVPR 2017*

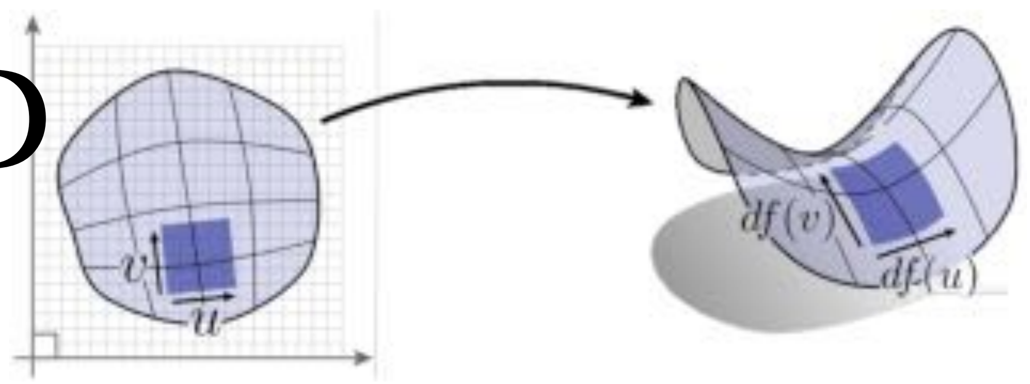
Design of Upconvolution Branch



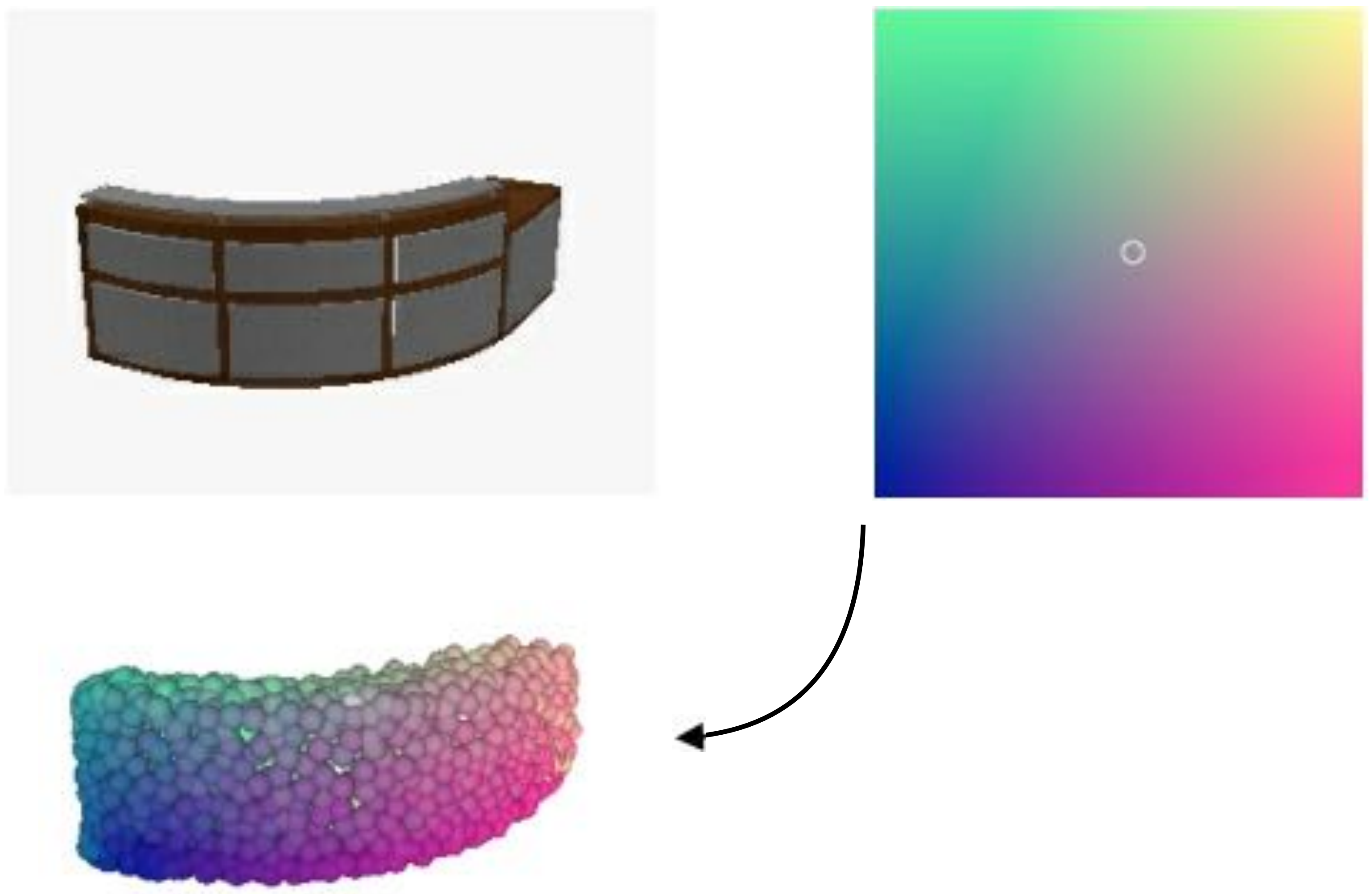
Fan et al., " **A Point Set Generation Network for 3D Object Reconstruction from a Single Image** ", *CVPR 2017*

Learns a Surface Parameterization

Smooth parameterization from 2D to 3D



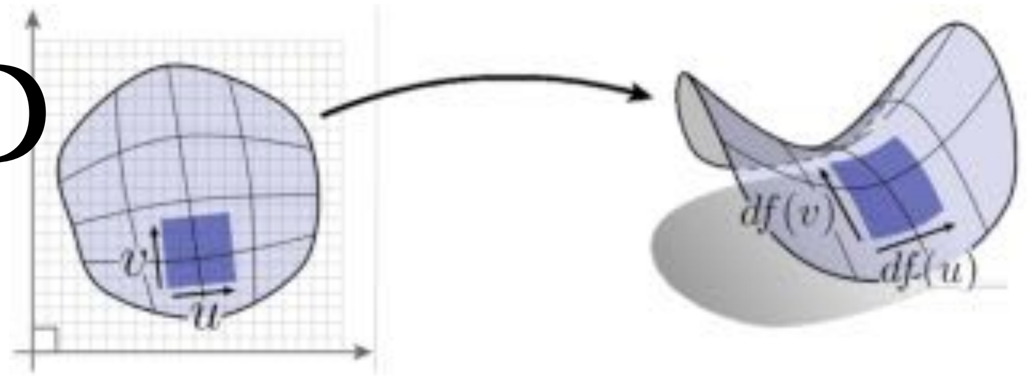
[image credit: Keenan Crane]



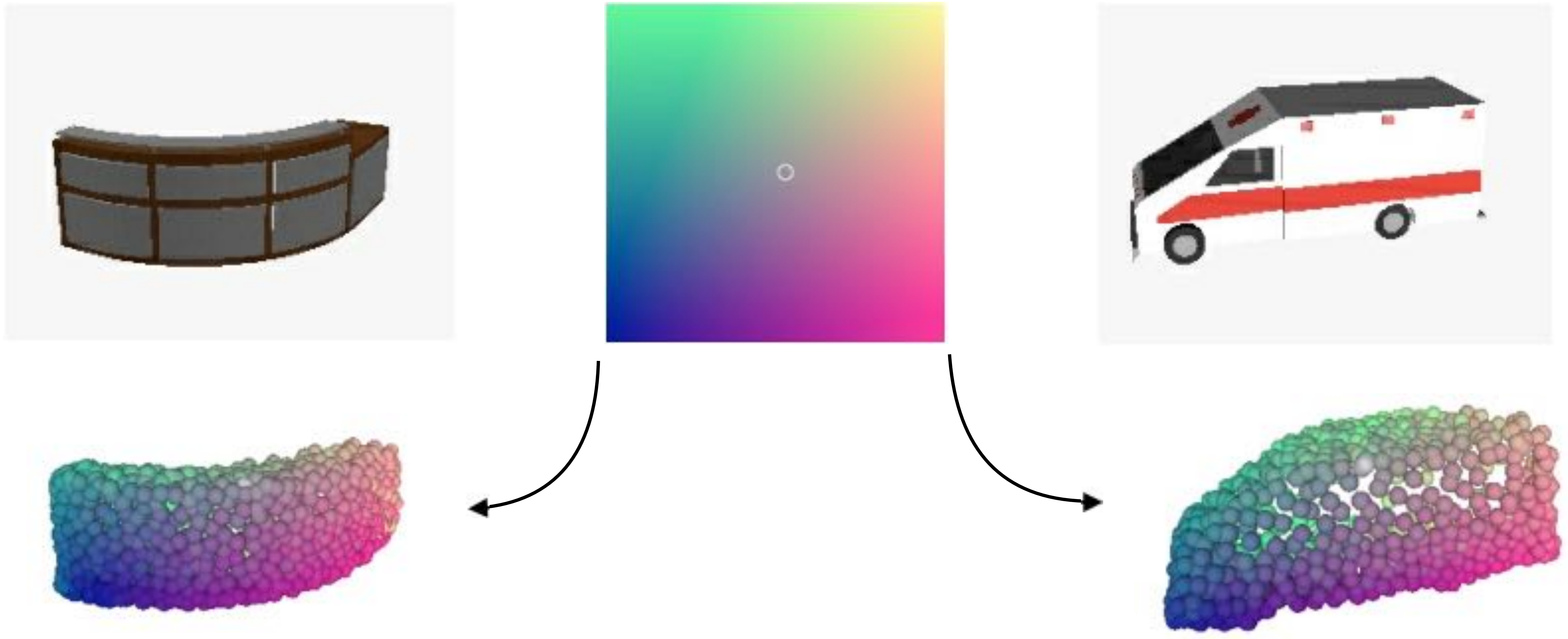
Learns a Surface Parameterization

Smooth parameterization from 2D to 3D

Consistent across objects



[image credit: Keenan Crane]



Slides credit: Hao Su

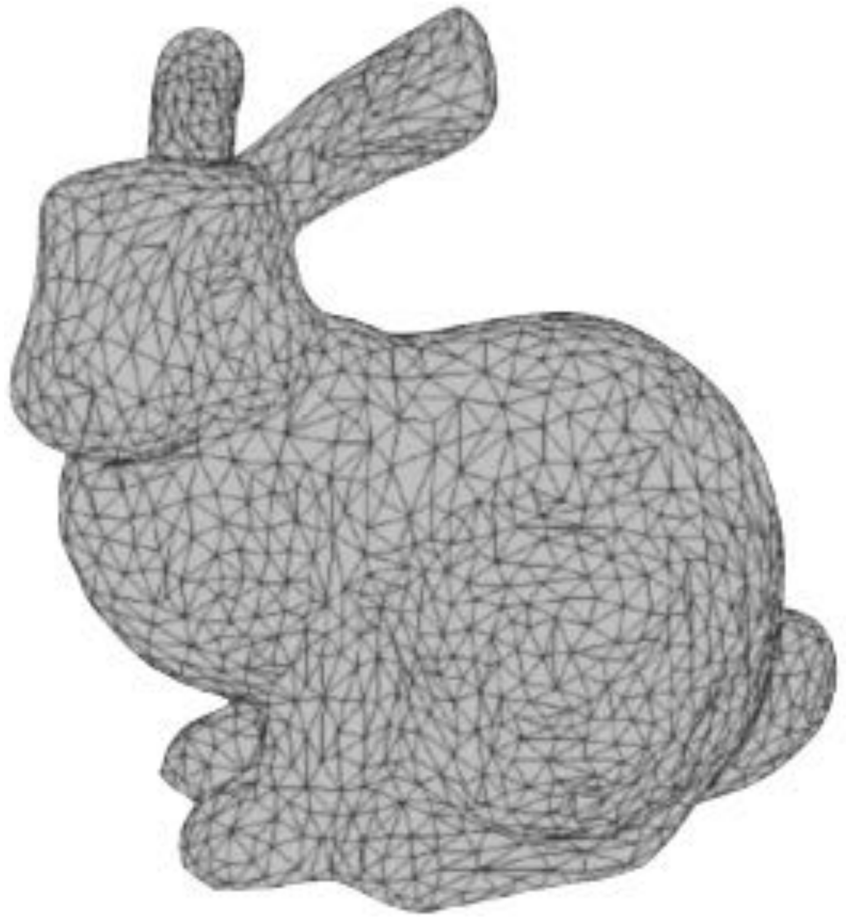
Outline

- Task
- Synthesis-for-Learning Pipeline
- Single-image to Point Cloud
- ***Single-image to Mesh***

Mesh Representation

- Previous point representation predicts only geometry without point connectivity.

- Mesh elements include mesh connectivity and mesh geometry $G = (V, E)$.



Mesh

Slides credit: Hao Su

Topology Ambiguity

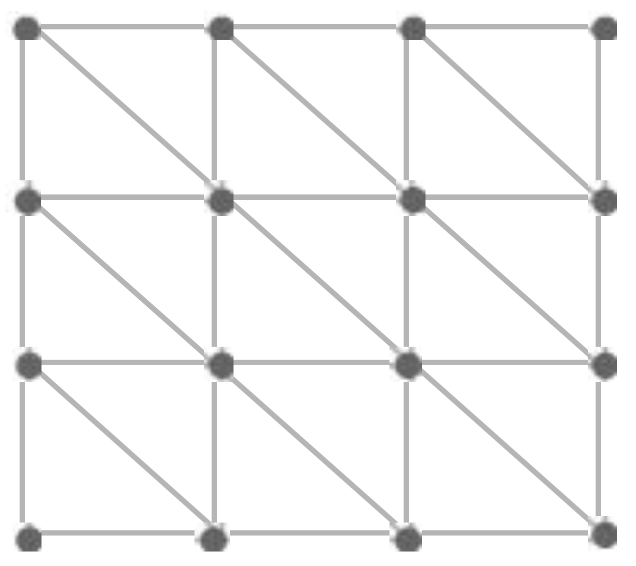
· Can we regress the vertices **and edges** from neural network?

- Estimate vertices as a set of points. ✓

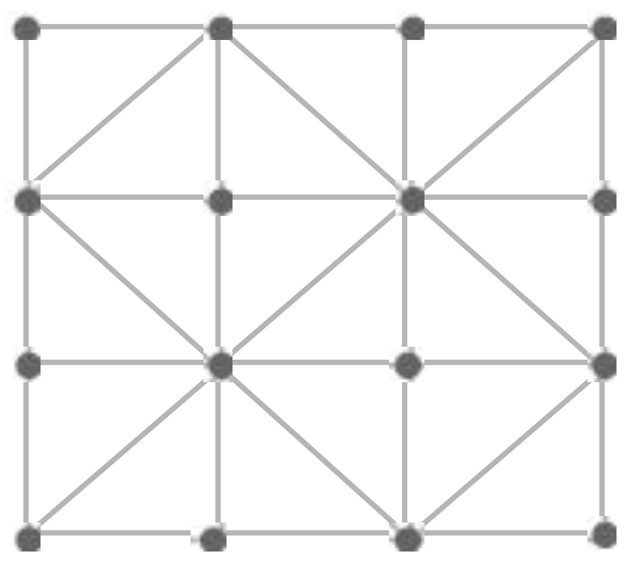
- Estimate edges?

Designing Loss for Edge Prediction is Hard

- **Key observation**: given vertices, there are many possible ways to connect them and represent the same underlying surface:



$G = (V, E)$



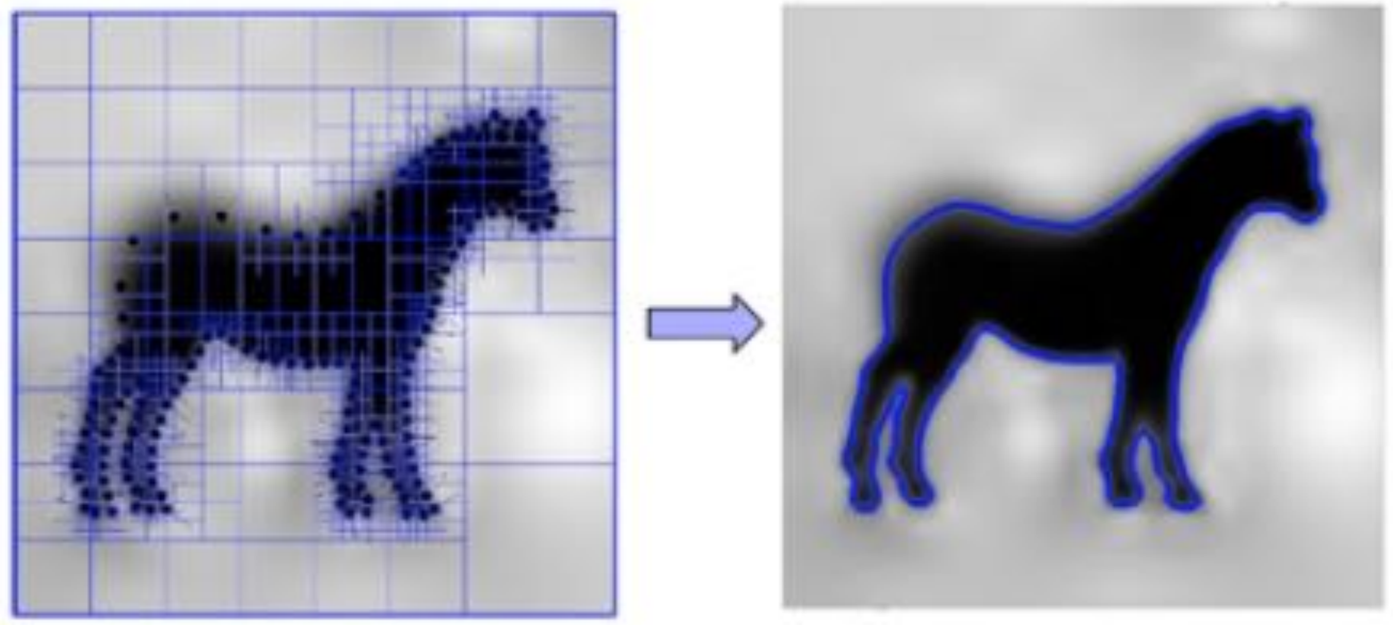
$G = (V, E')$

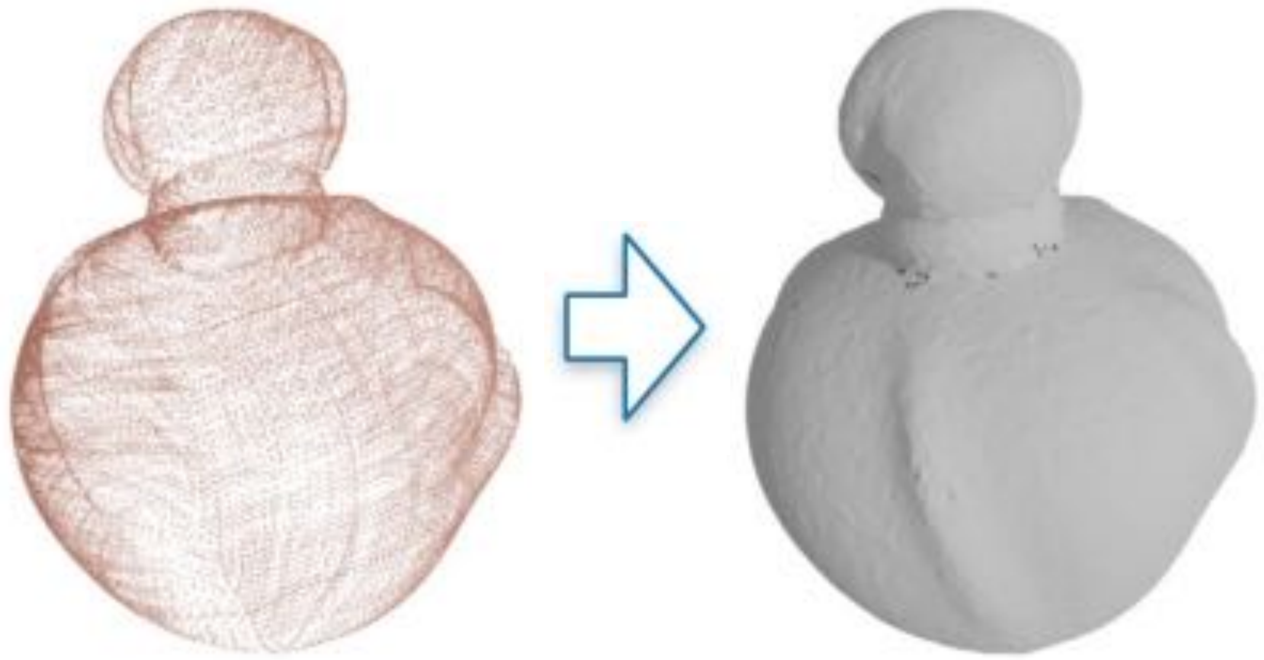
Slides credit: Hao Su

Image → Intermediate Repr. → Mesh

- One option is to first build a high-resolution intermediate representation, and then convert the point cloud to mesh

- Intermediate representations:
  - Voxel
  - Implicit surface
  - Point cloud





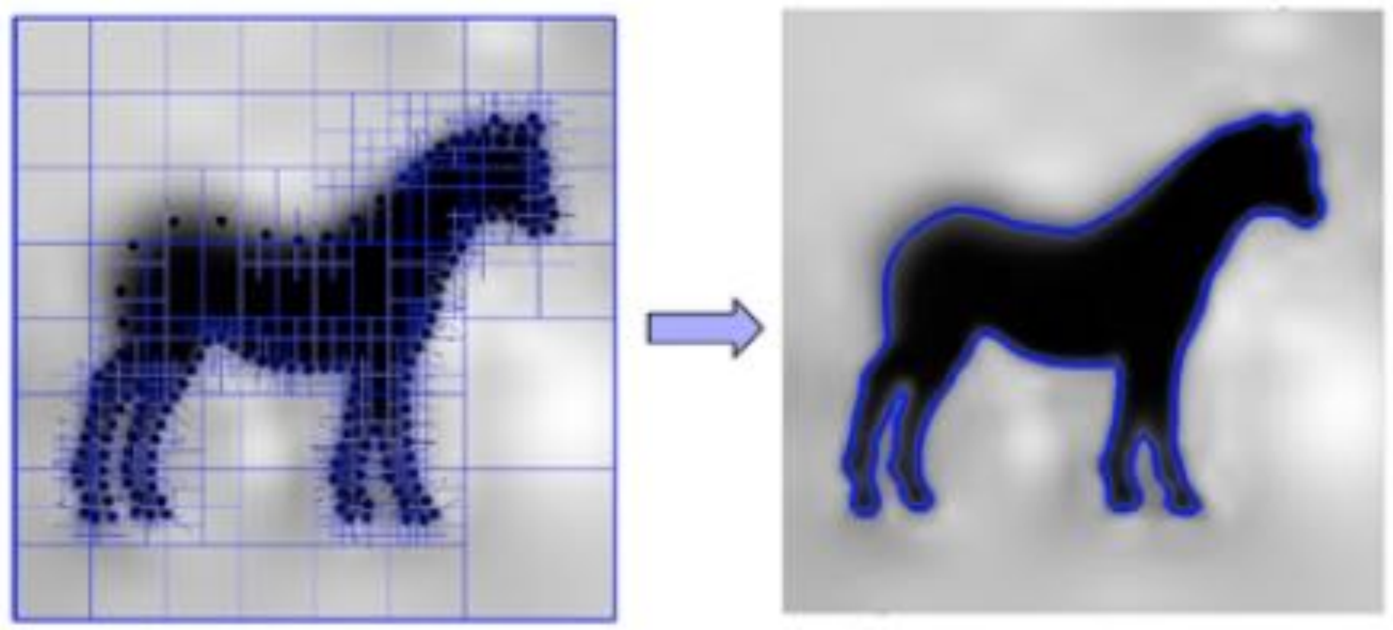
Slides credit: Hao Su

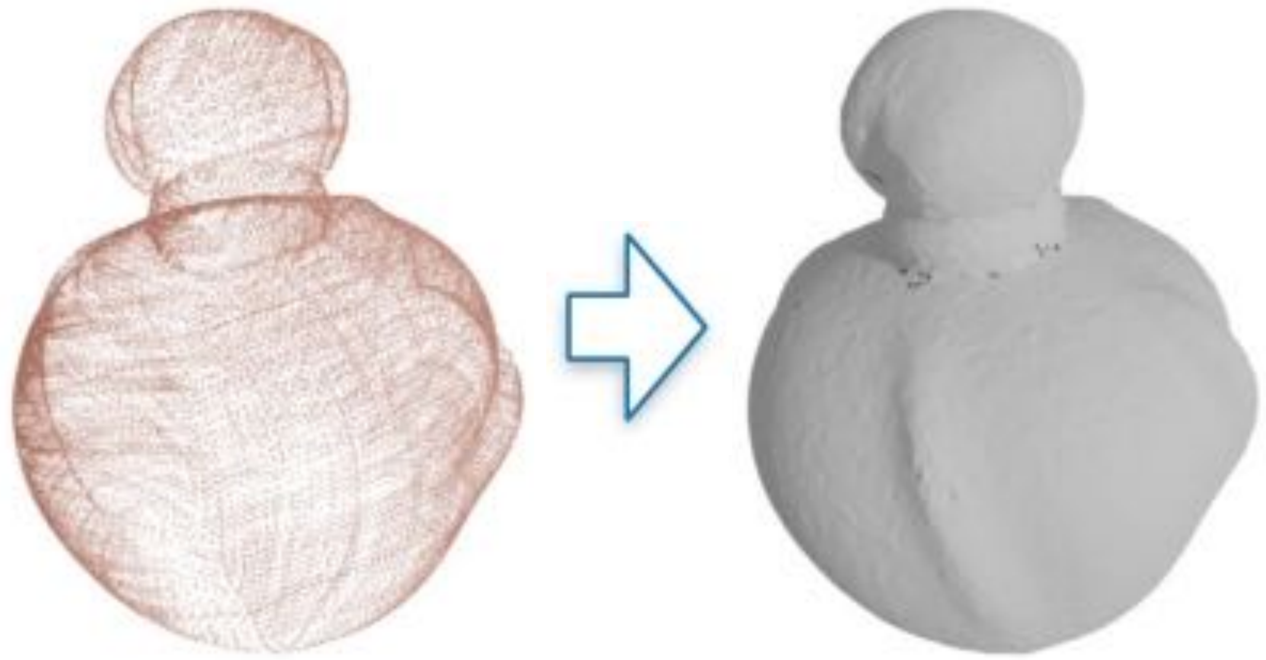
Image → Intermediate Repr. → Mesh

- One option is to first build a high-resolution intermediate representation, and then convert the point cloud to mesh

- Intermediate representations:
  - Voxel
  - Implicit surface
  - Point cloud

Defer to a later lecture!





Slides credit: Hao Su

Editing-based Mesh Modeling

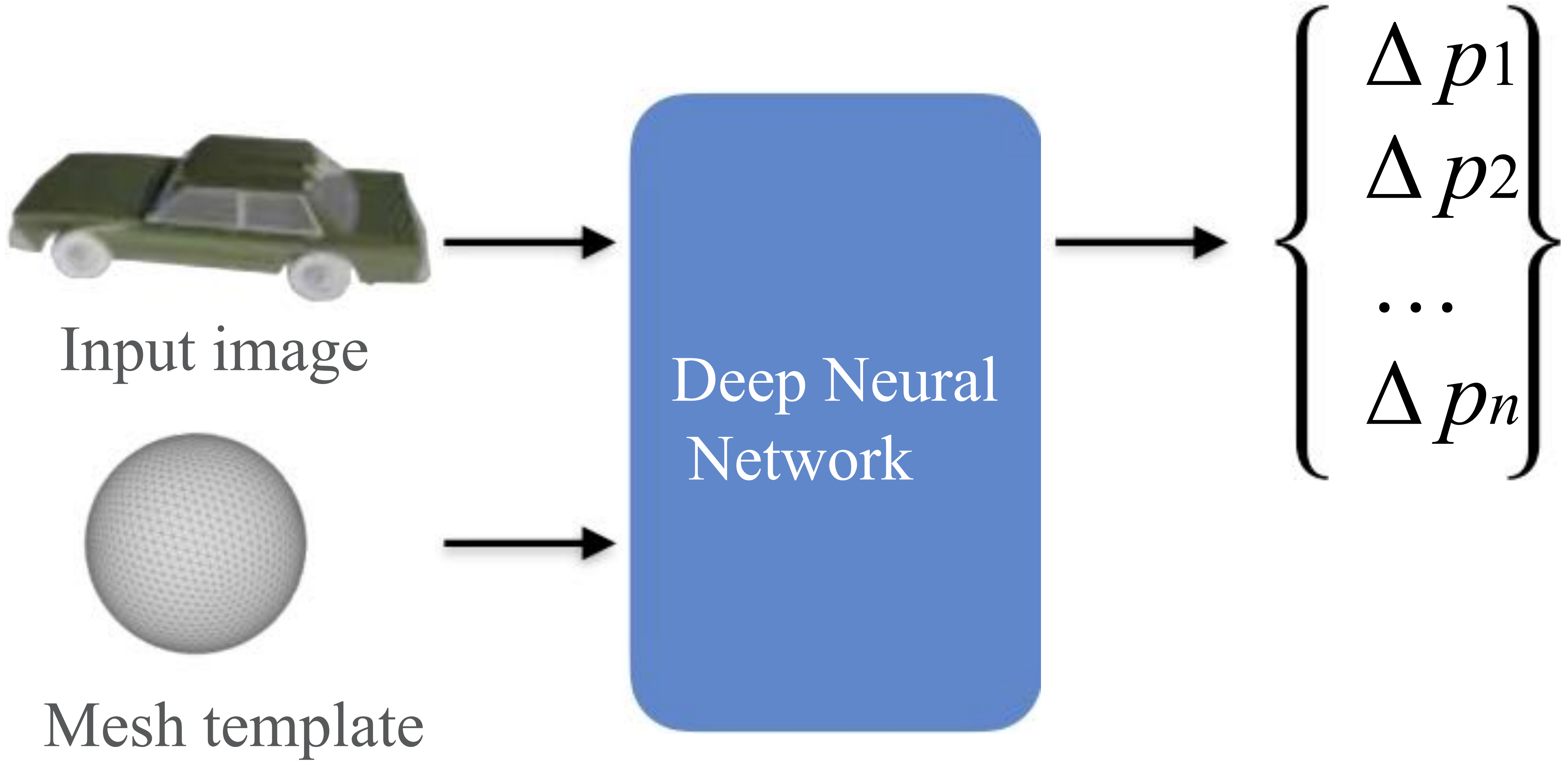
- Can we model mesh without predicting edges?

Mesh Editing-based Methods

Slides credit: Hao Su

Editing-based Mesh Modeling

- Key idea: starting from an established mesh and modify it to become the target shape

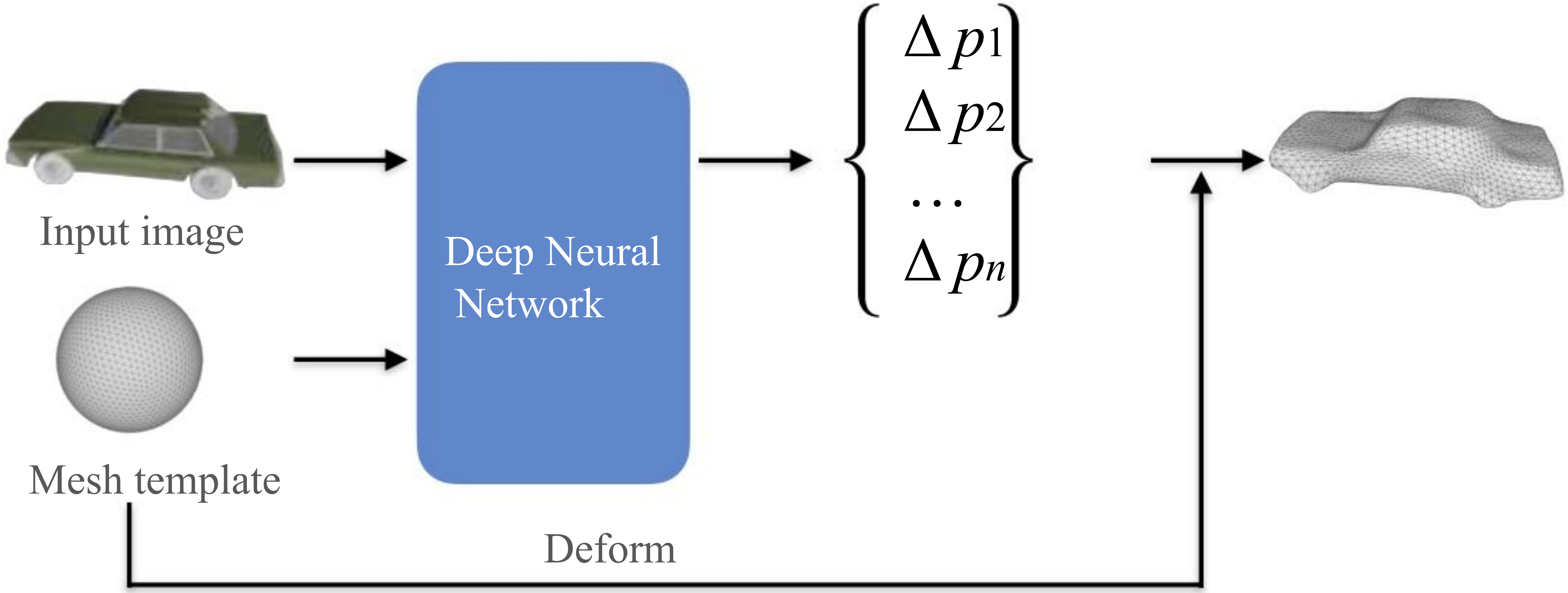


Slides credit: Hao Su

Editing-based Mesh Modeling

- Key idea: starting from an established mesh and modify it to become the target shape

For example, deformation-based modeling:



Slides credit: Hao Su

Some Example Losses for Mesh Editing

- Vertices distance.
  - Vertices point set distance.

- Uniform vertices distribution.
  - Edge length regularizer.

- Mesh surface smoothness.

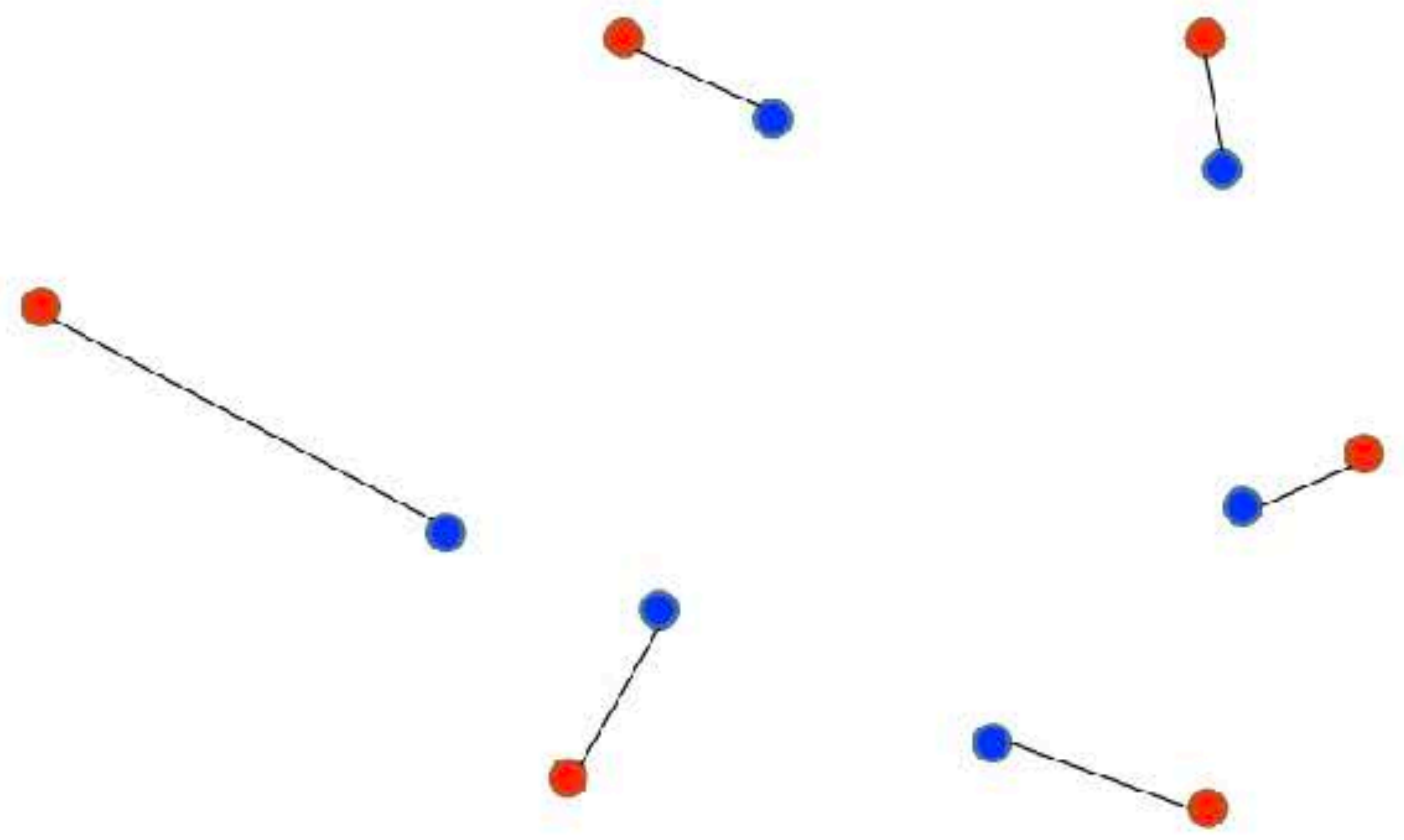
- Normal Loss.

Slides credit: Hao Su

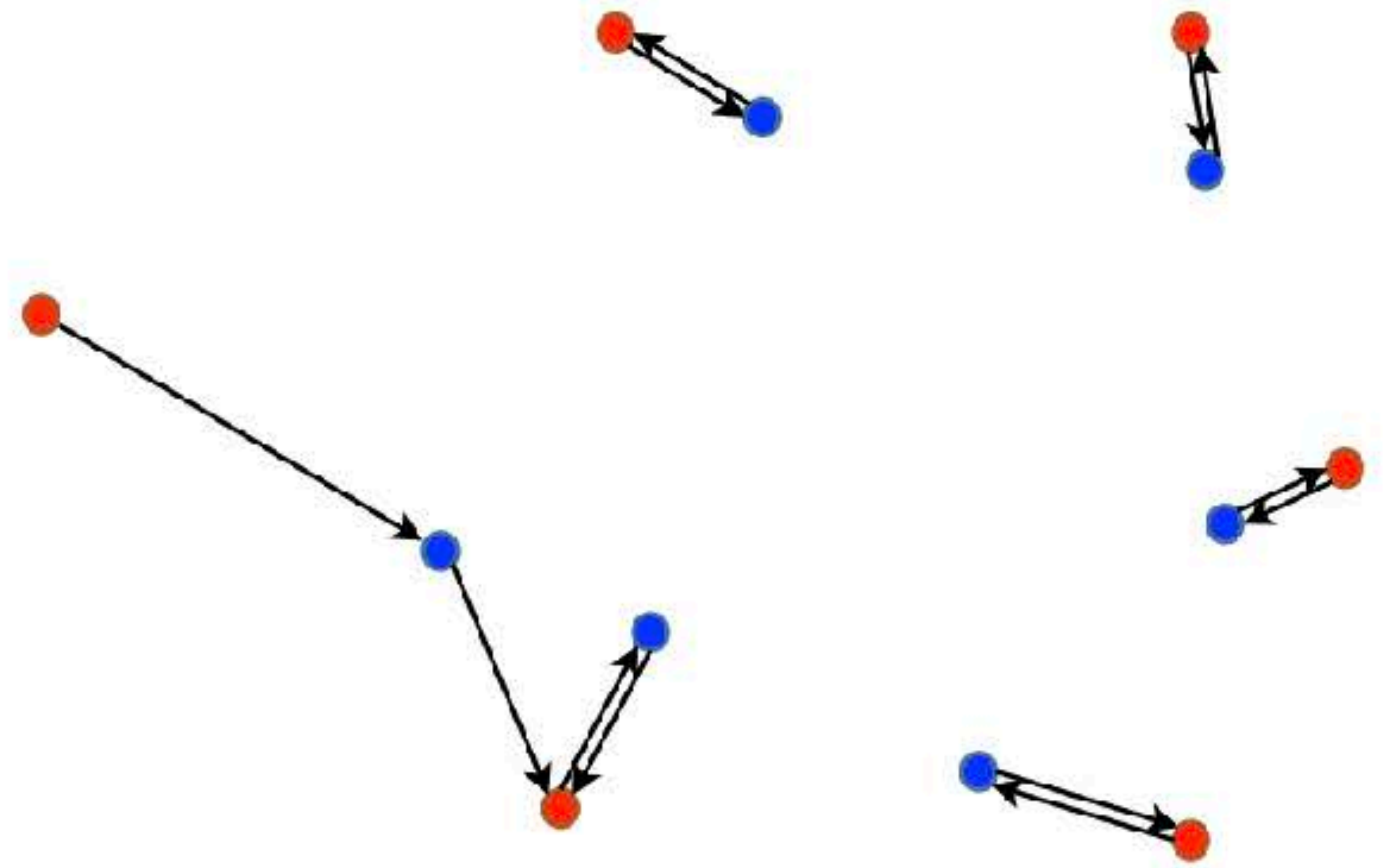
Loss I: Set Distance between Vertices

- Vertices are a set of points

- Recall the metrics for point clouds



Earth Mover's distance



Chamfer distance

$$d_{EMD}(S_1, S_2) = \min_{\phi: S_1 \rightarrow S_2} \sum_{x \in S_1} \|x - \phi(x)\|_2$$

$$d_{CD}(S_1, S_2) = \sum_{x \in S_1} \min_{y \in S_2} \|x - y\|_2^2 + \sum_{y \in S_2} \min_{x \in S_1} \|x - y\|_2^2$$

Slides credit: Hao Su

Loss II: Uniform Vertices Distribution

- Penalizes the flying vertices and overlong edges to guarantee the high quality of recovered 3D geometry
- Encourage equal edge length between vertices

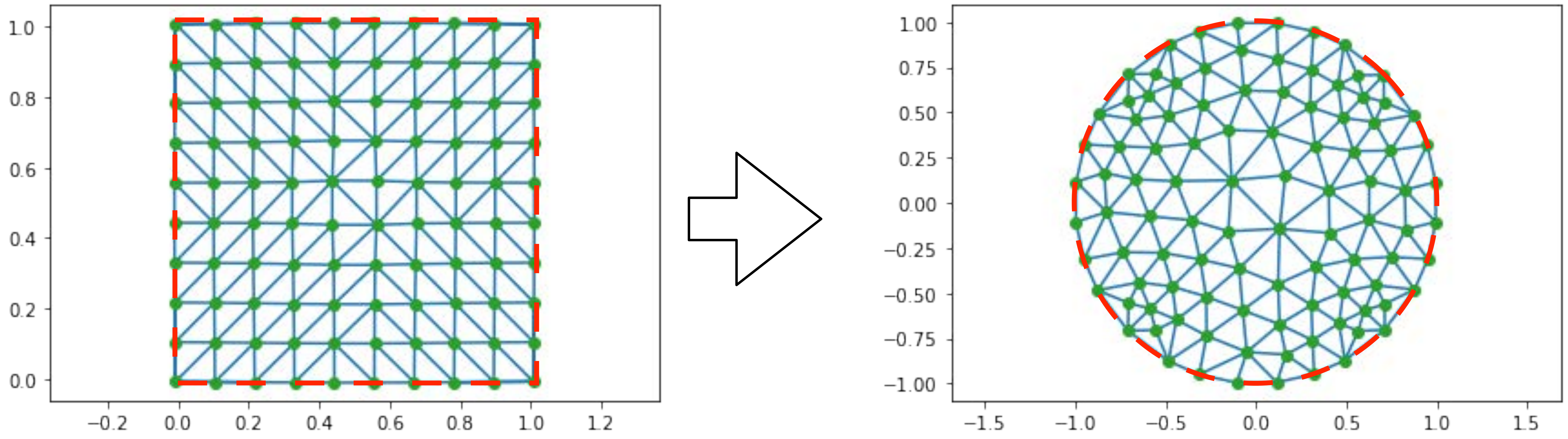
$$L_{\text{unif}} = \sum_{p} \sum_{k \in N(p)} \|p - k\|_2^2$$

Slides credit: Hao Su

Loss II: Uniform Vertices Distribution

$$L_{\text{unif}} = \sum_{p} \sum_{k \in N(p)} \|p - k\|_2^2$$

Effect of minimizing l when fixing topology and setting boundary points to the new positions

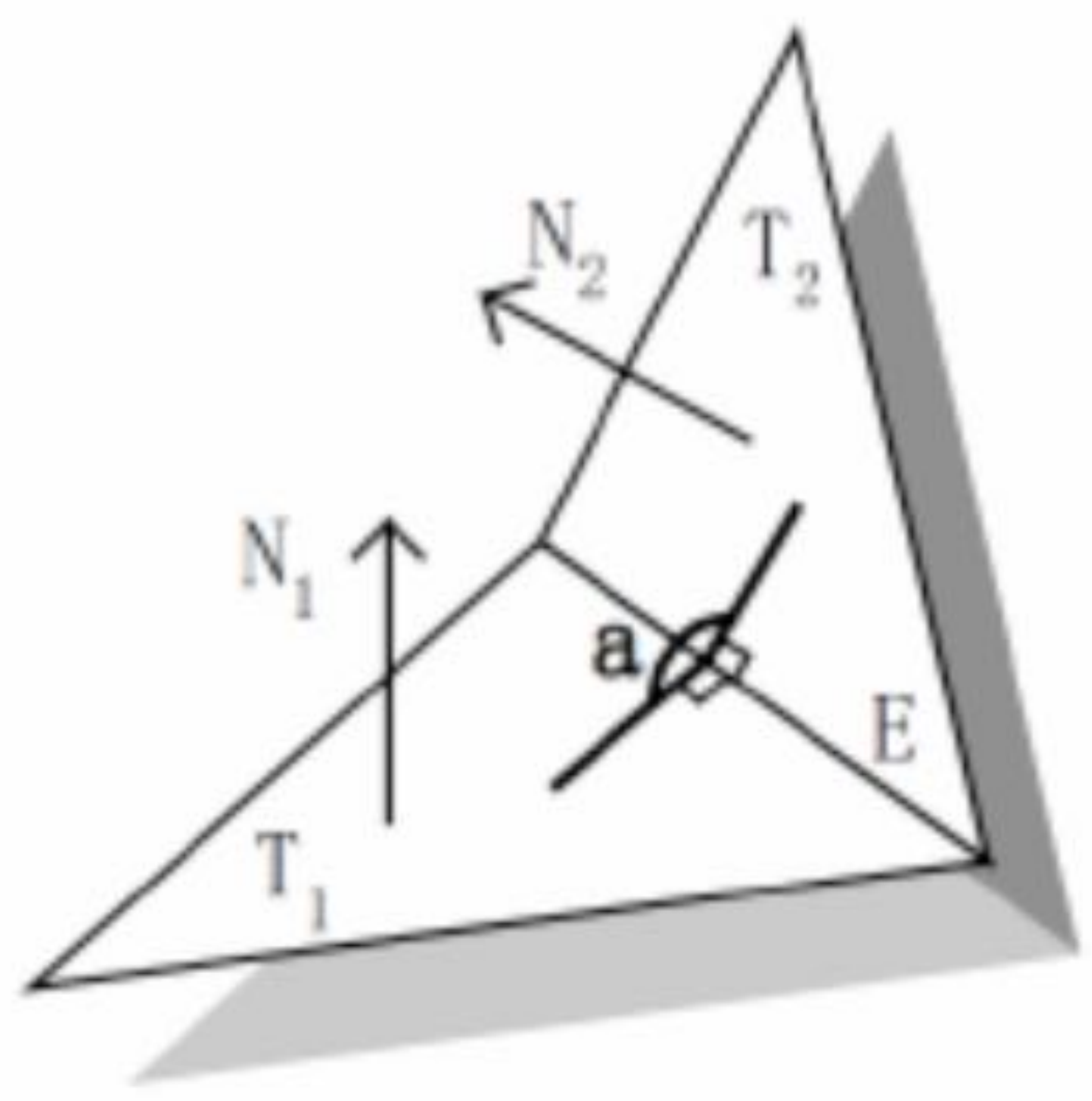


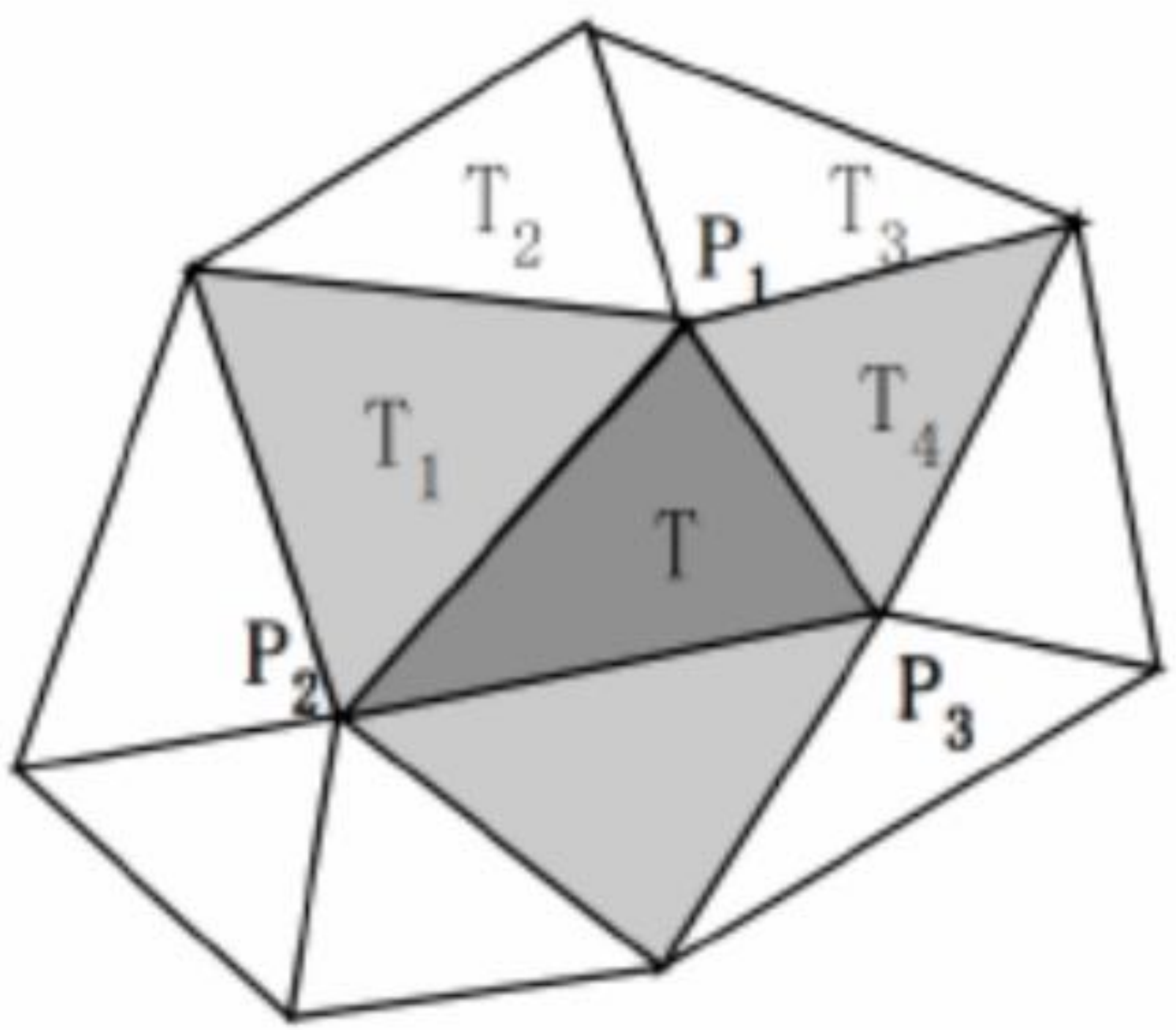
Slides credit: Hao Su

Loss III: Mesh Smoothness

- L_{smooth} encourages that intersection angles of faces are close to 180 degrees.

$$L_{smooth} = \sum_i (\cos \theta_i + 1)^2$$



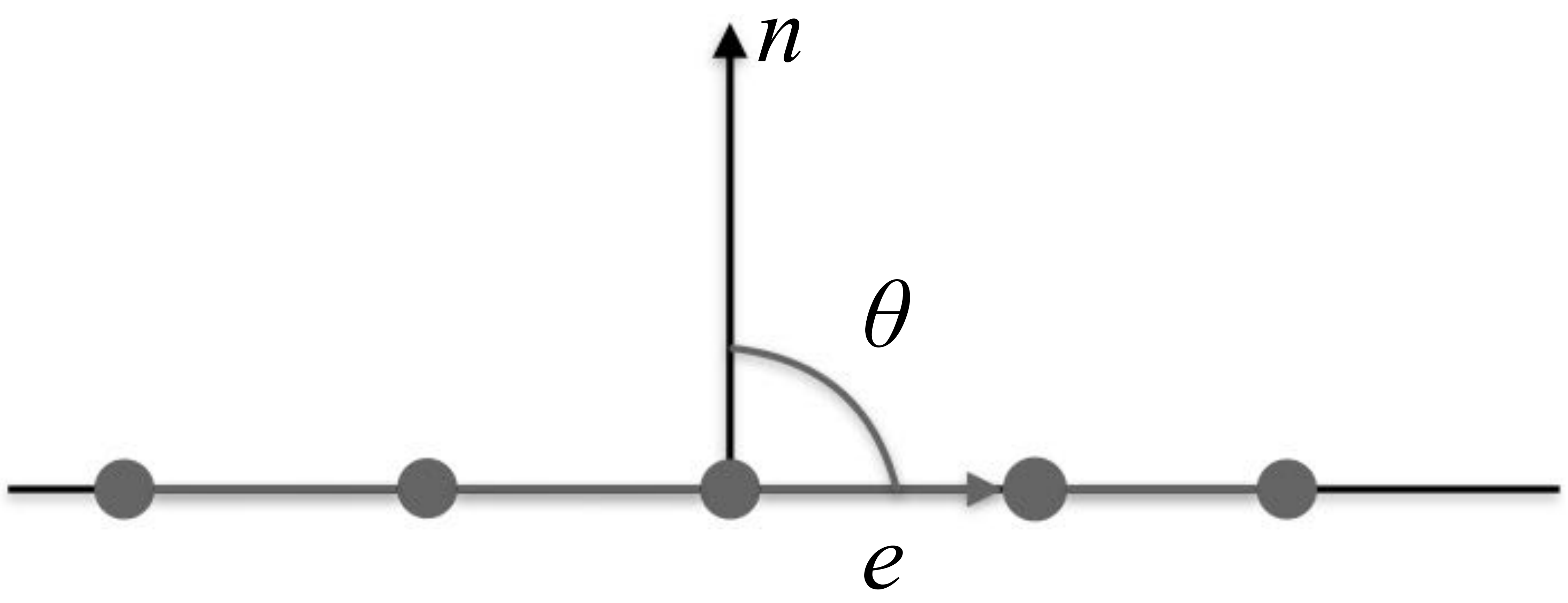


Slides credit: Hao Su

Loss IV: Normal Loss

- **Key assumption**: vertices within a local neighborhood lie on the same tangent plane.

- Regularize edge to be perpendicular to the underlying groundtruth vertex normal



Slides credit: Hao Su

Loss IV: Normal Loss

- But how to find the vertices normal?

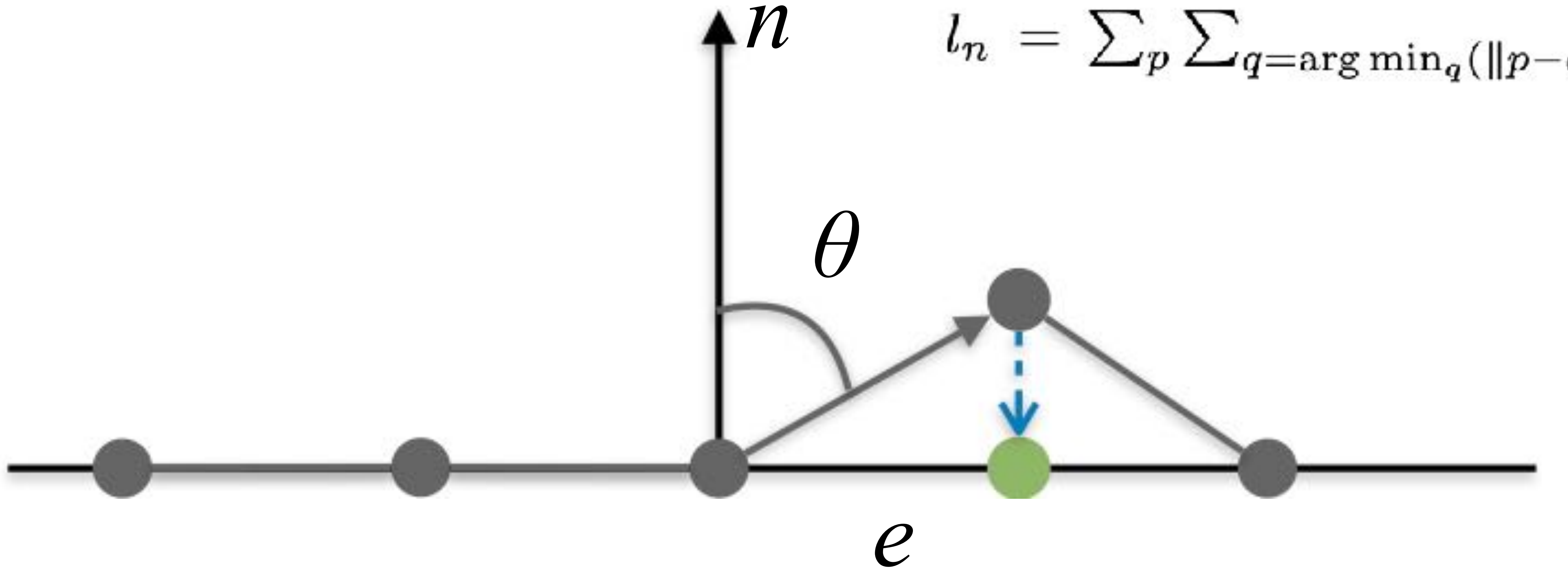
- One approach: use the nearest ground truth point normal as current vertex normal.

Loss IV: Normal Loss

· But how to find the vertices normal?

· One approach: use the nearest ground truth point normal as current vertex normal.

· Penalize the edge direction to perpendicular to vertex normal.

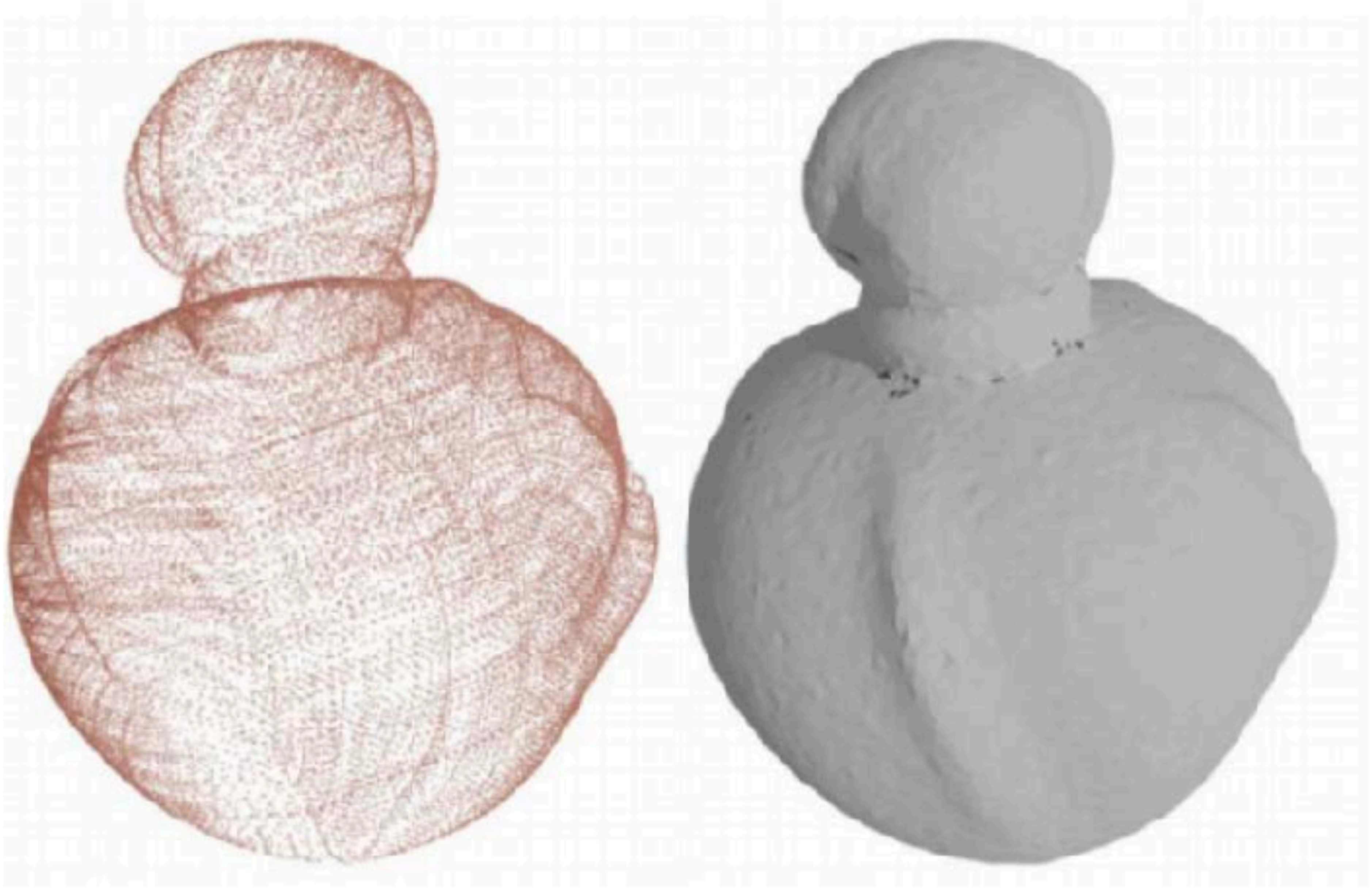


$$l_n = \sum_p \sum_{q=\arg\min_q(\|p-q\|_2^2)} \|\langle p-k, \mathbf{n}_q\rangle\|_2^2, \text{ s.t. } k \in \mathcal{N}(p)$$

Summary

- Synthesis-for-learning pipeline leverages easy-to-obtain synthetic data for challenging 3D visual understanding tasks
- Single image to 3D point cloud is possible with properly defined set metric (EMD and CD)
- Natural ambiguity in single image to 3D
- Single image to mesh can be achieved through template deformation
- Mesh reconstruction requires more regularizations

Next Time



Surface Reconstruction