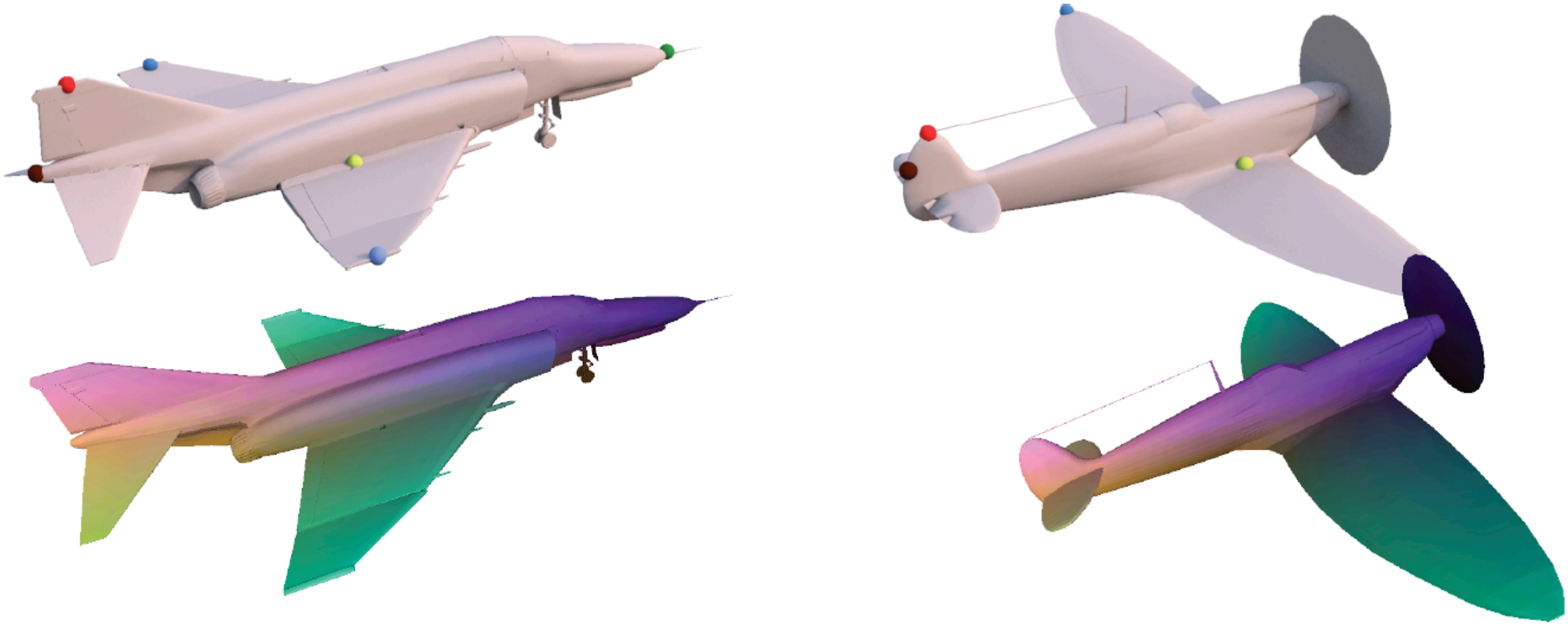
Deep Deformation Networks for 3D Geometry

Vladimir (Vova) Kim

Adobe

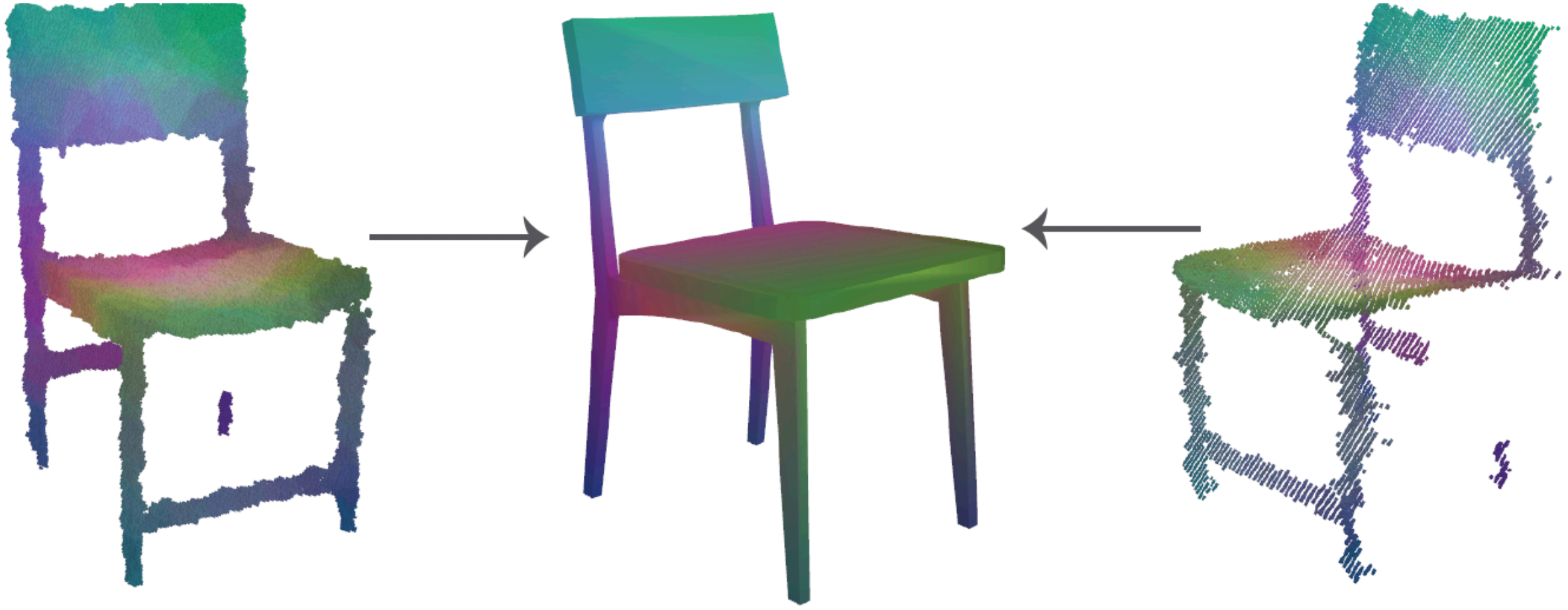
Shape Analysis

- Shape Correspondence



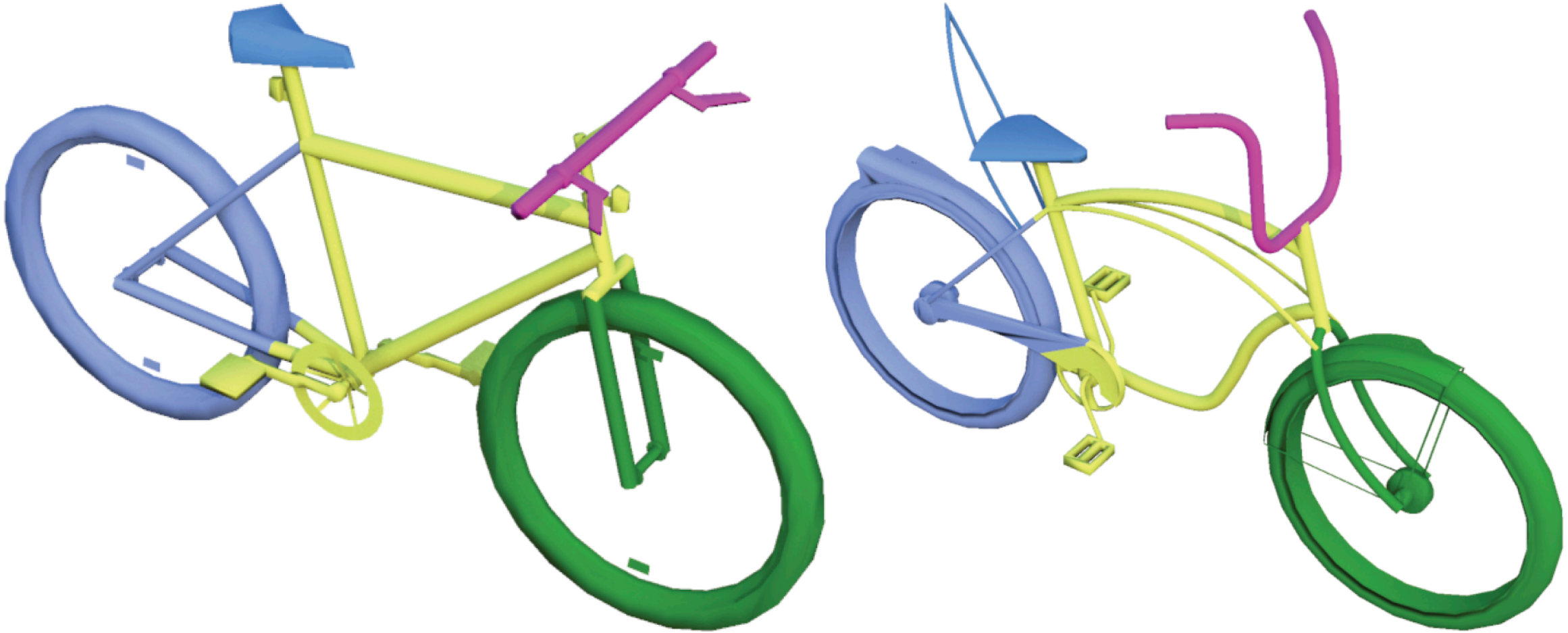
Shape Analysis

- Partial Matching



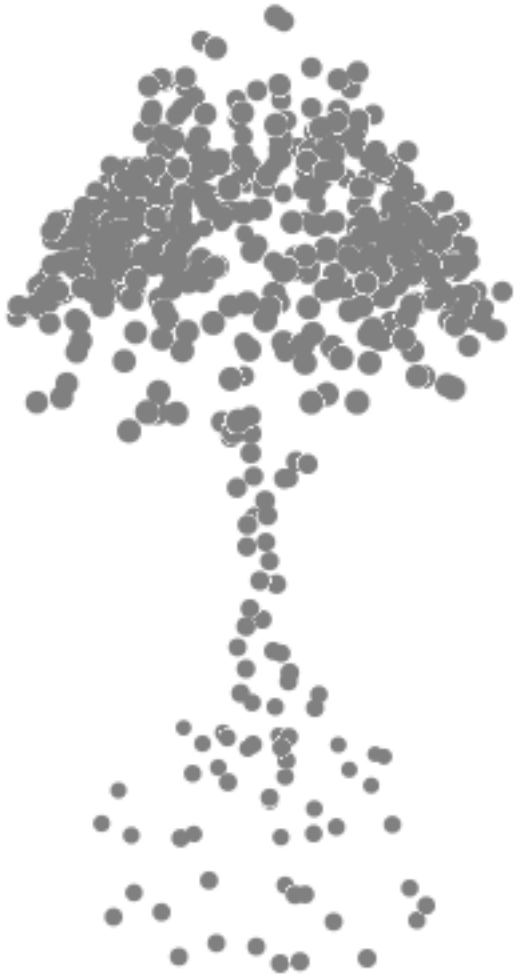
Shape Analysis

- Mesh Segmentation and Labeling



Shape Synthesis

- Point-based Reconstruction



Shape Synthesis

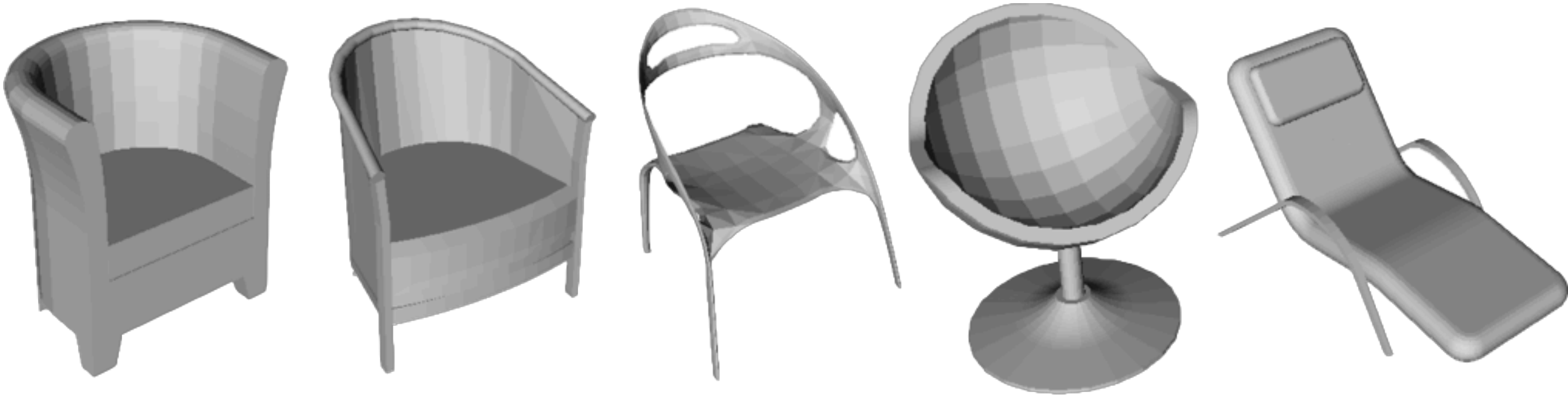
- Image-based Reconstruction

Shape Synthesis

- Interpolation

Challenges

- No structure in representation
  - No natural ordering
  - Diverse scale
  - No canonical coordinates

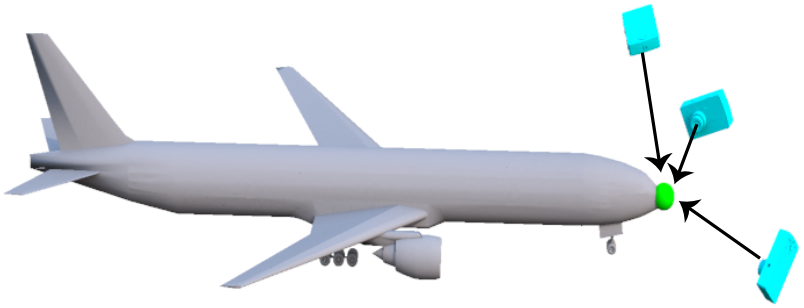


Challenges

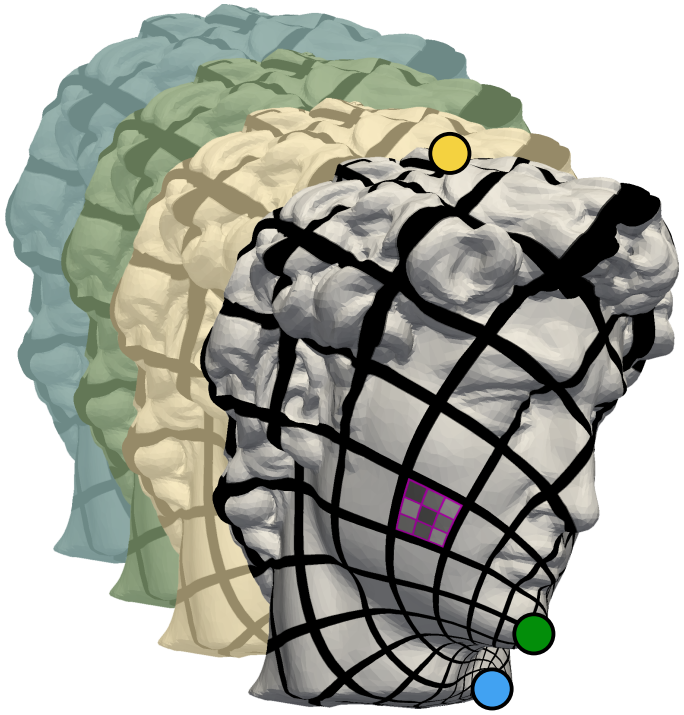
- No structure in representation
  - No natural ordering
  - Diverse scale
  - No canonical coordinates

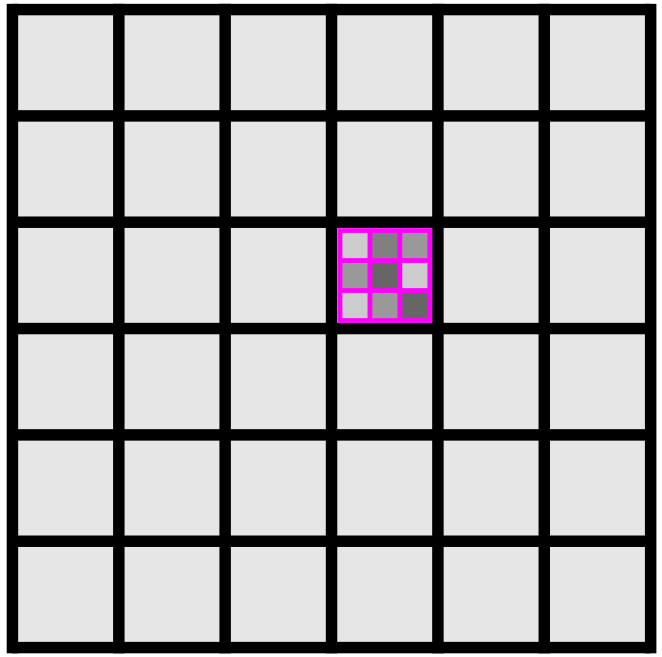
Representations

- Image-based (analysis only)



Project to an extrinsic camera



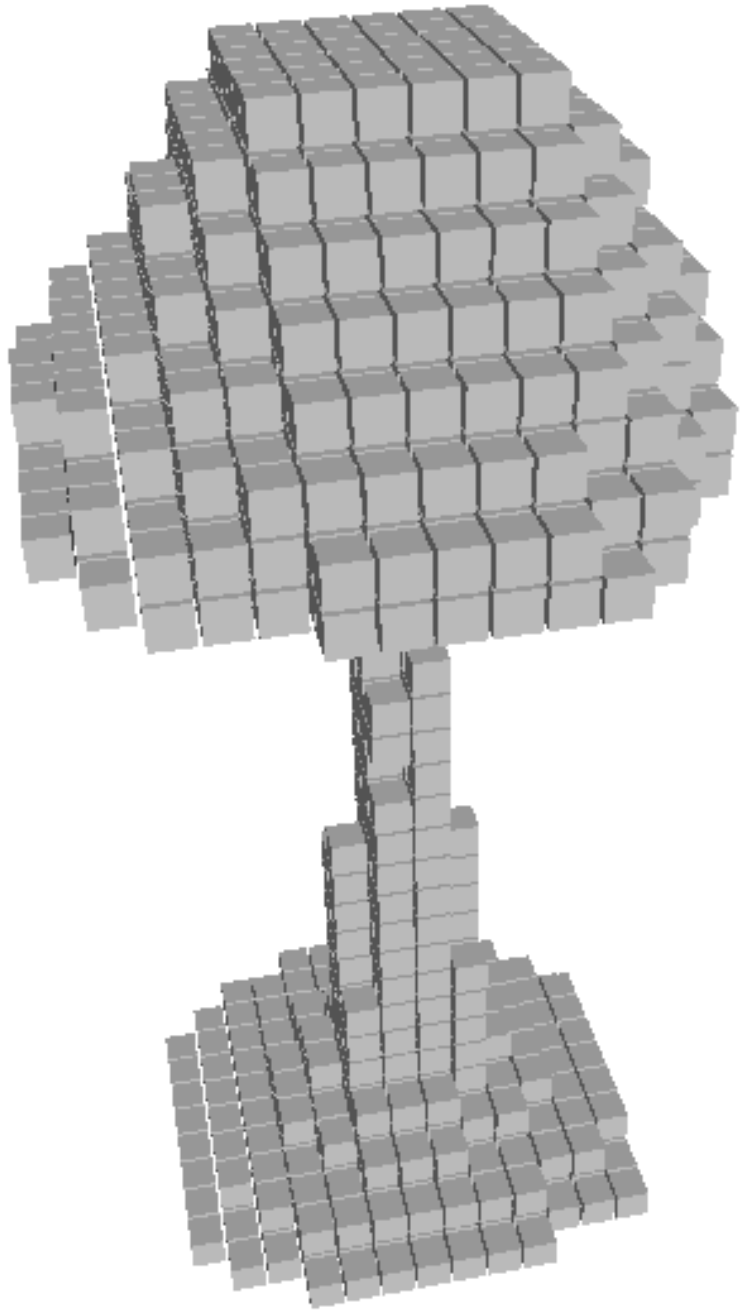


Intrinsic Parameterization

Representations

- Volume Pros:
  - Regular grid

- Volume Cons:
  - Need to normalize orientation, scale
  - Uniform resolution
  - Information is very sparse
  - Coarse generated model

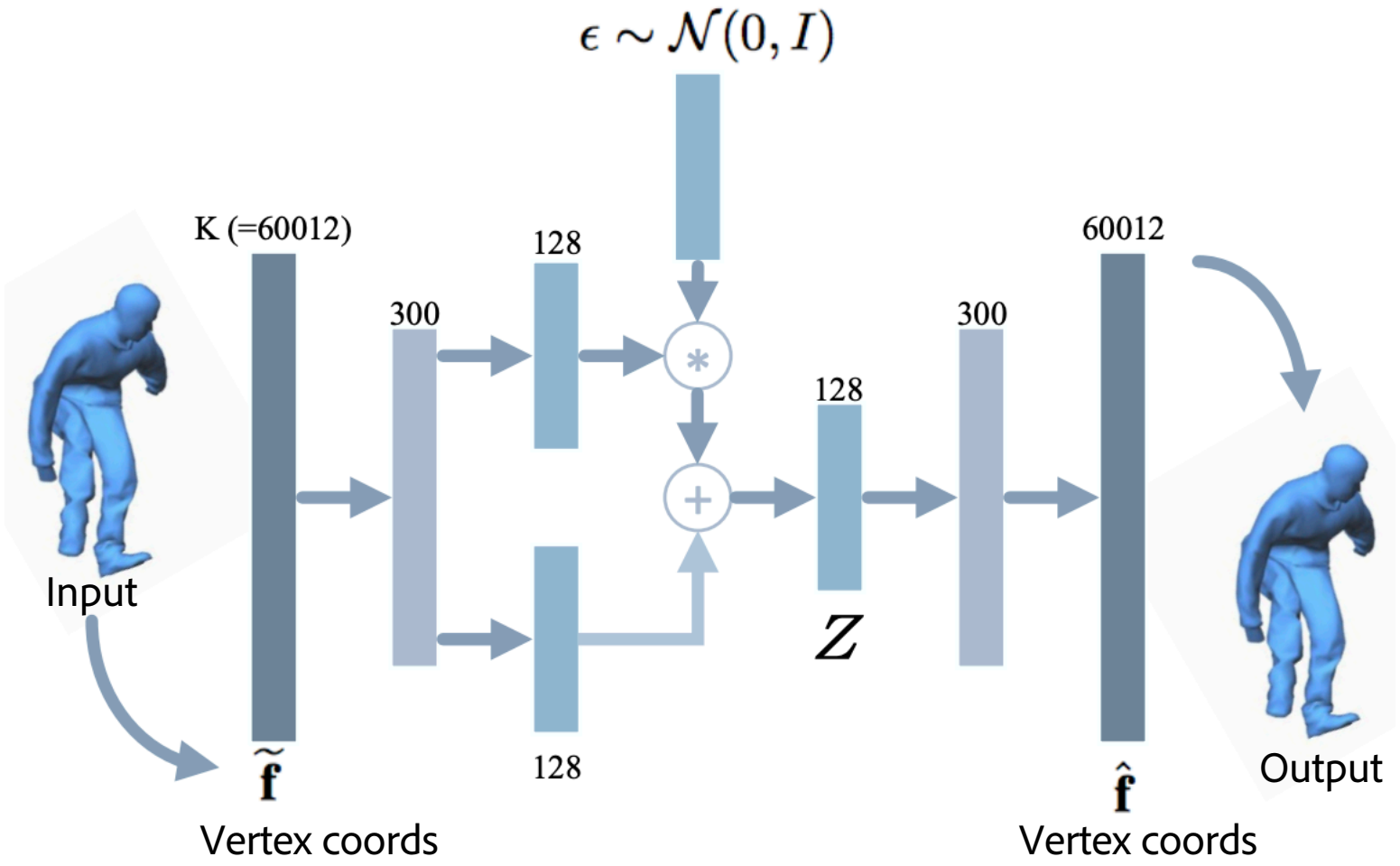


Representations

- PointNet-like architectures Pros
  - Only considers sparse input (surface points)
- PointsNet-like architectures Cons
  - Generated model is sparse

Representations

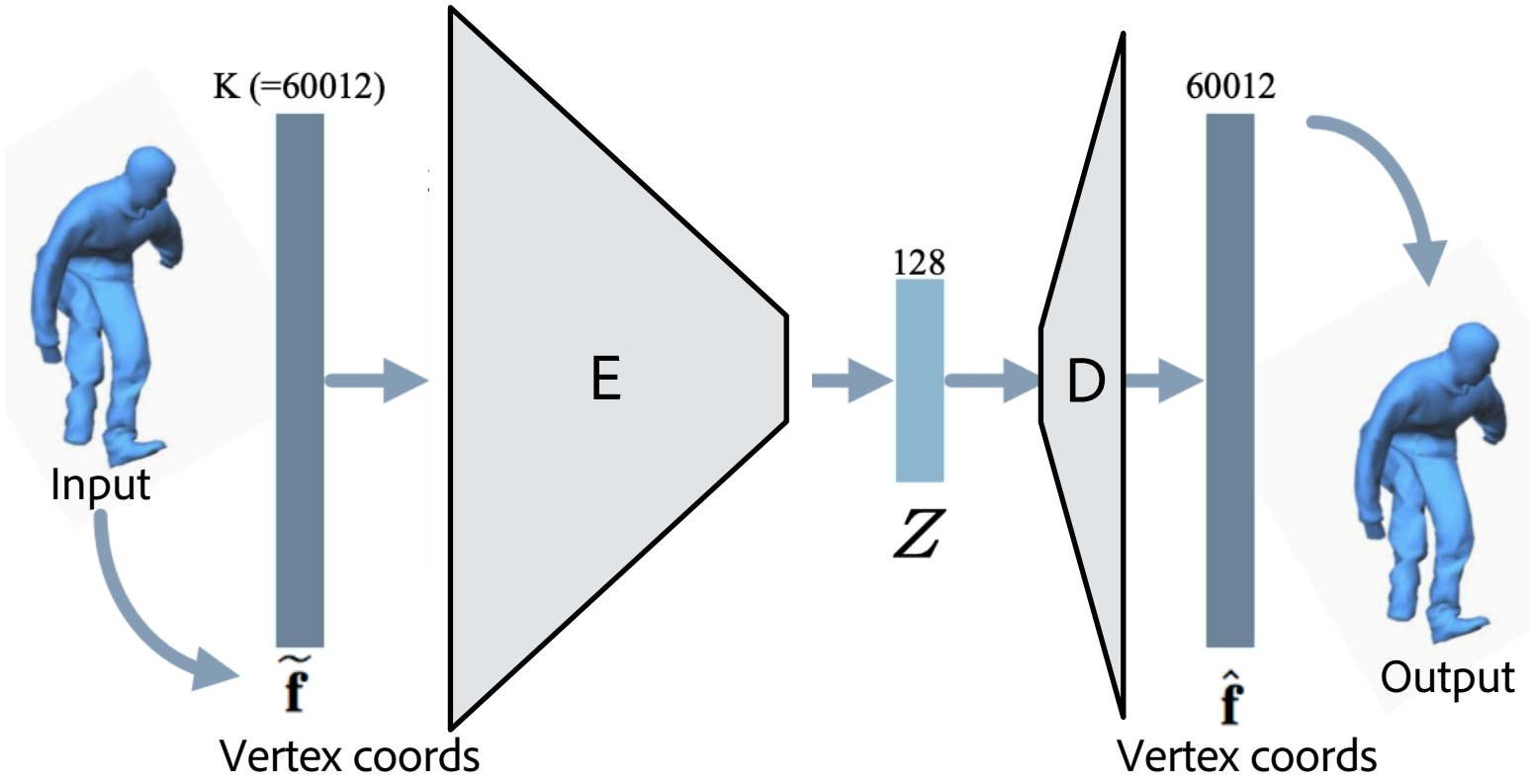
- Deformations (synthesis-only)



[Tan et al., CVPR 2018]

Representations

- Deformations (synthesis-only)
  - Sensitive to sampling
  - Input data has to be parameterized consistently with the template



[Tan et al., CVPR 2018]

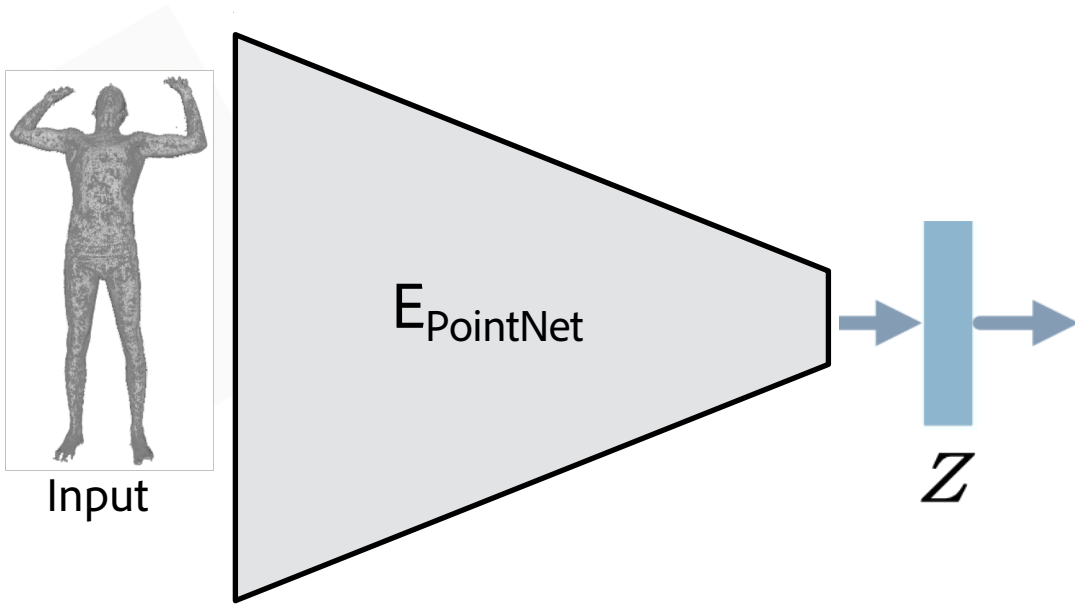
Talk Outline

- Template Fitting and Correspondence Estimation via a Deforming Neural Networks
- Template-less Modeling via Deforming Neural Networks
- Template-less Signal Transfer via Deforming Neural Networks
- Multi-view Reconstruction via Deforming Neural Networks

Thibault Groueix

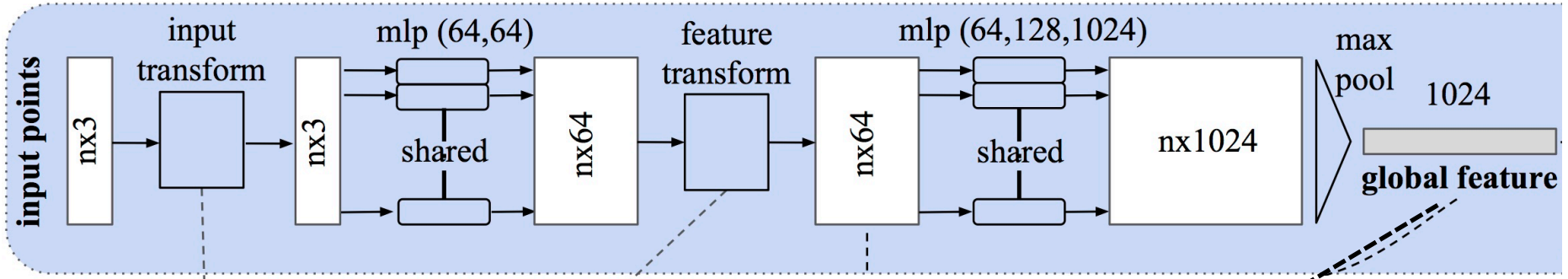
Our Architecture

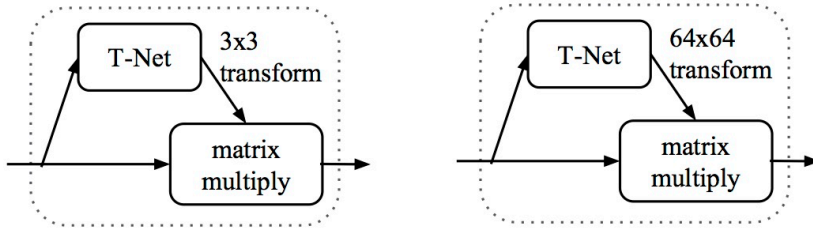
- Encoder: PointNet

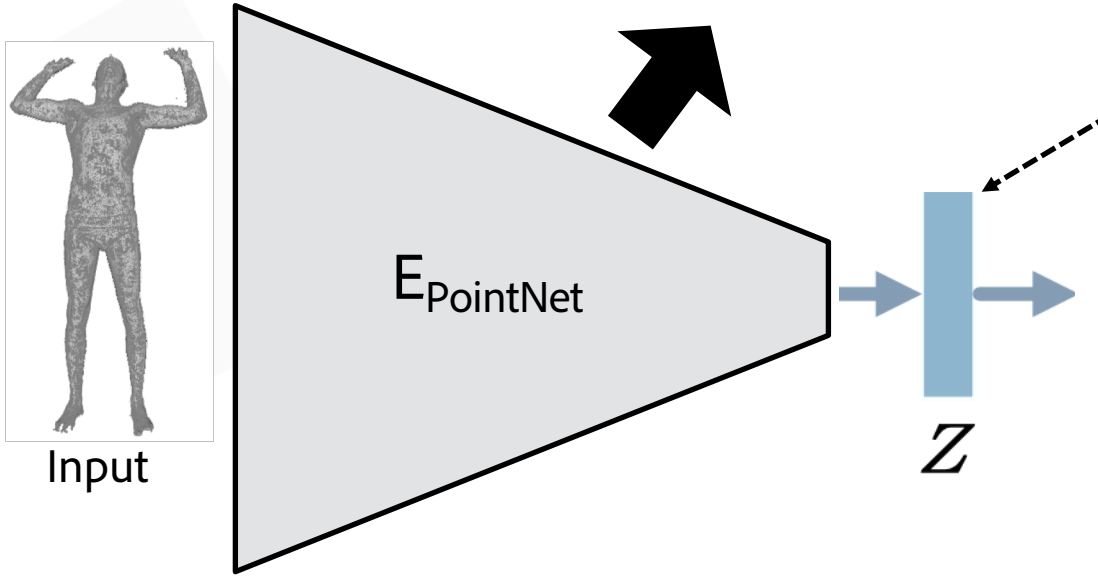


Our Architecture

- Encoder: PointNet

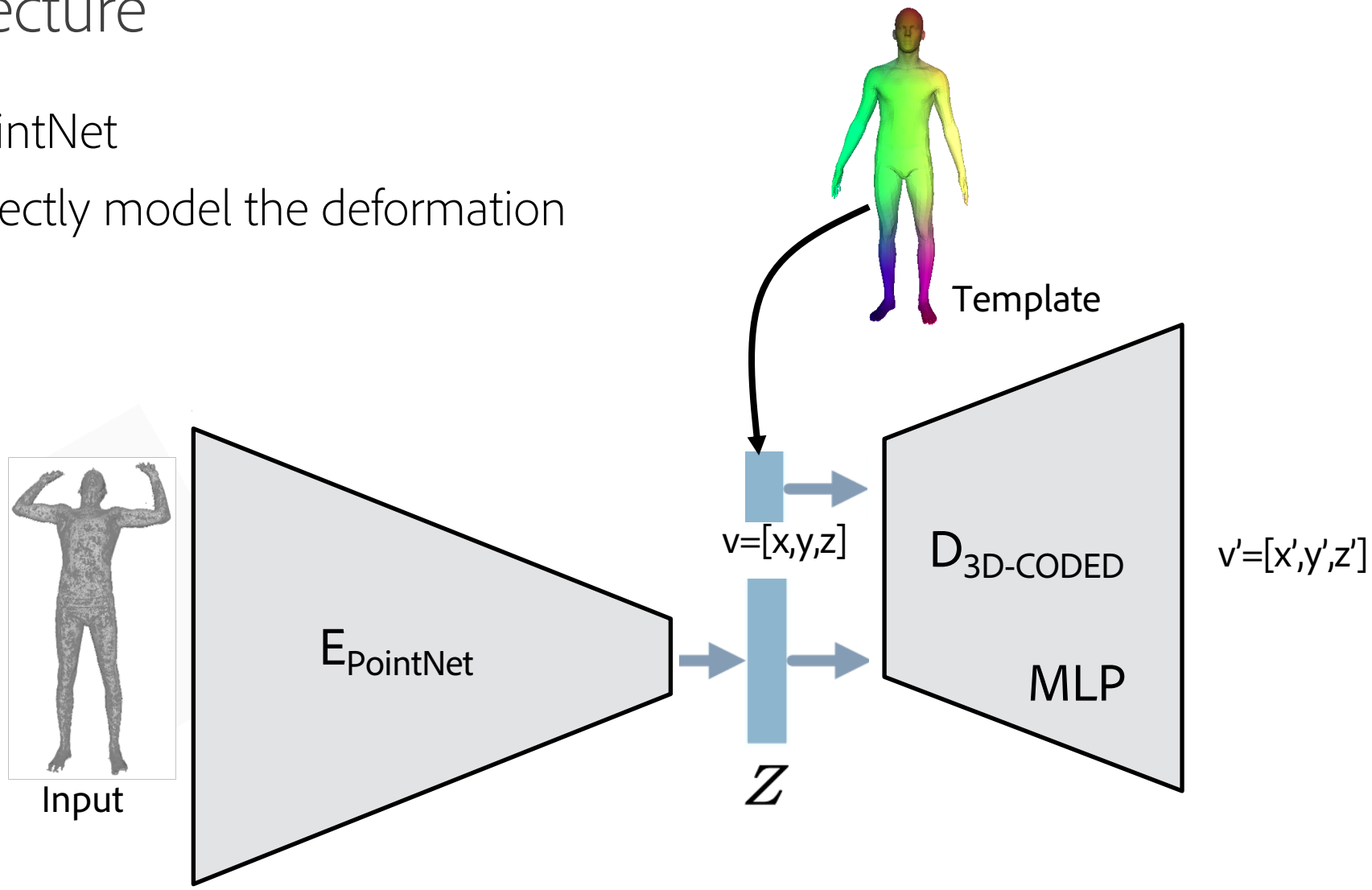






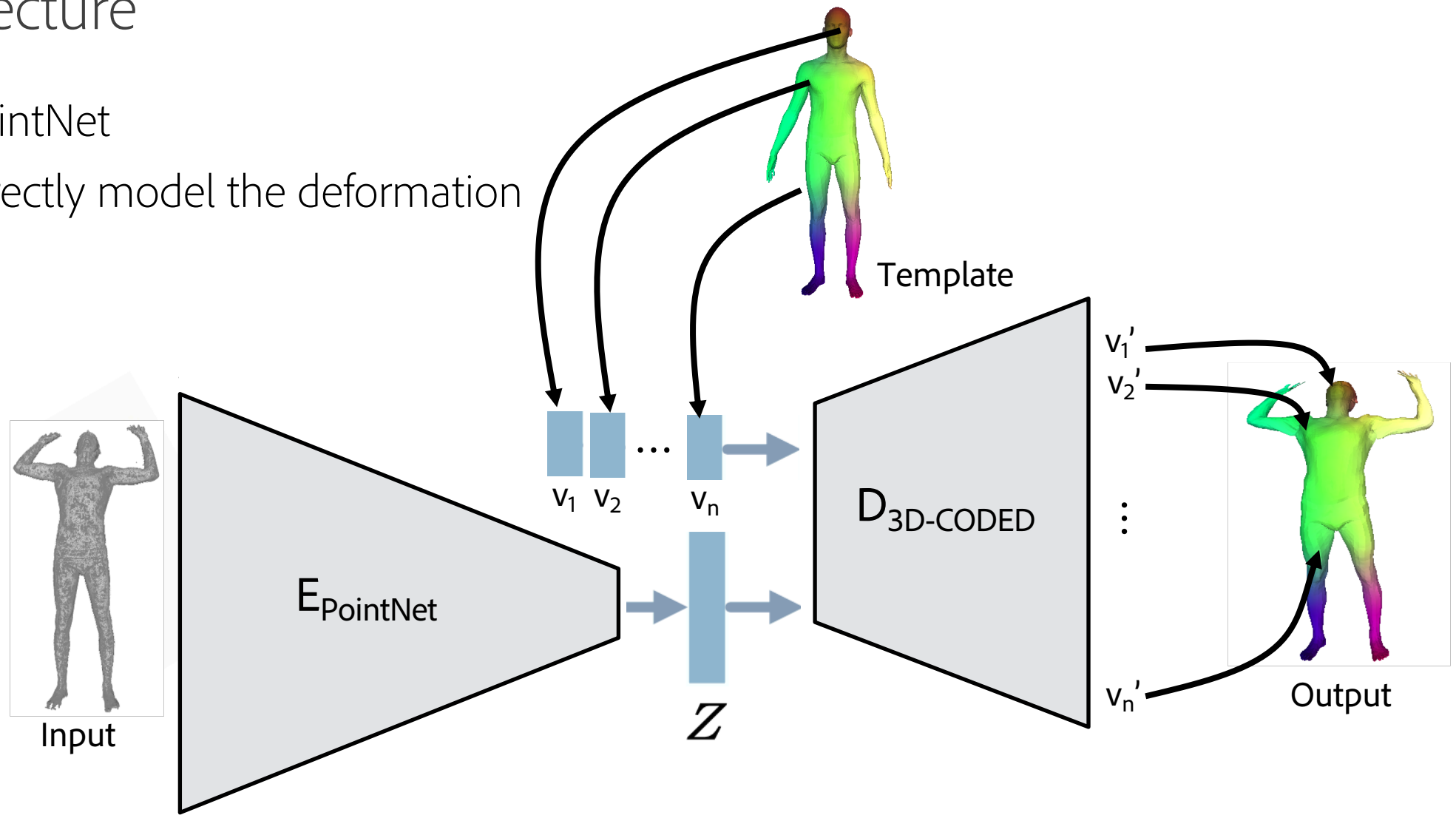
Our Architecture

- Encoder: PointNet
- Decoder: directly model the deformation



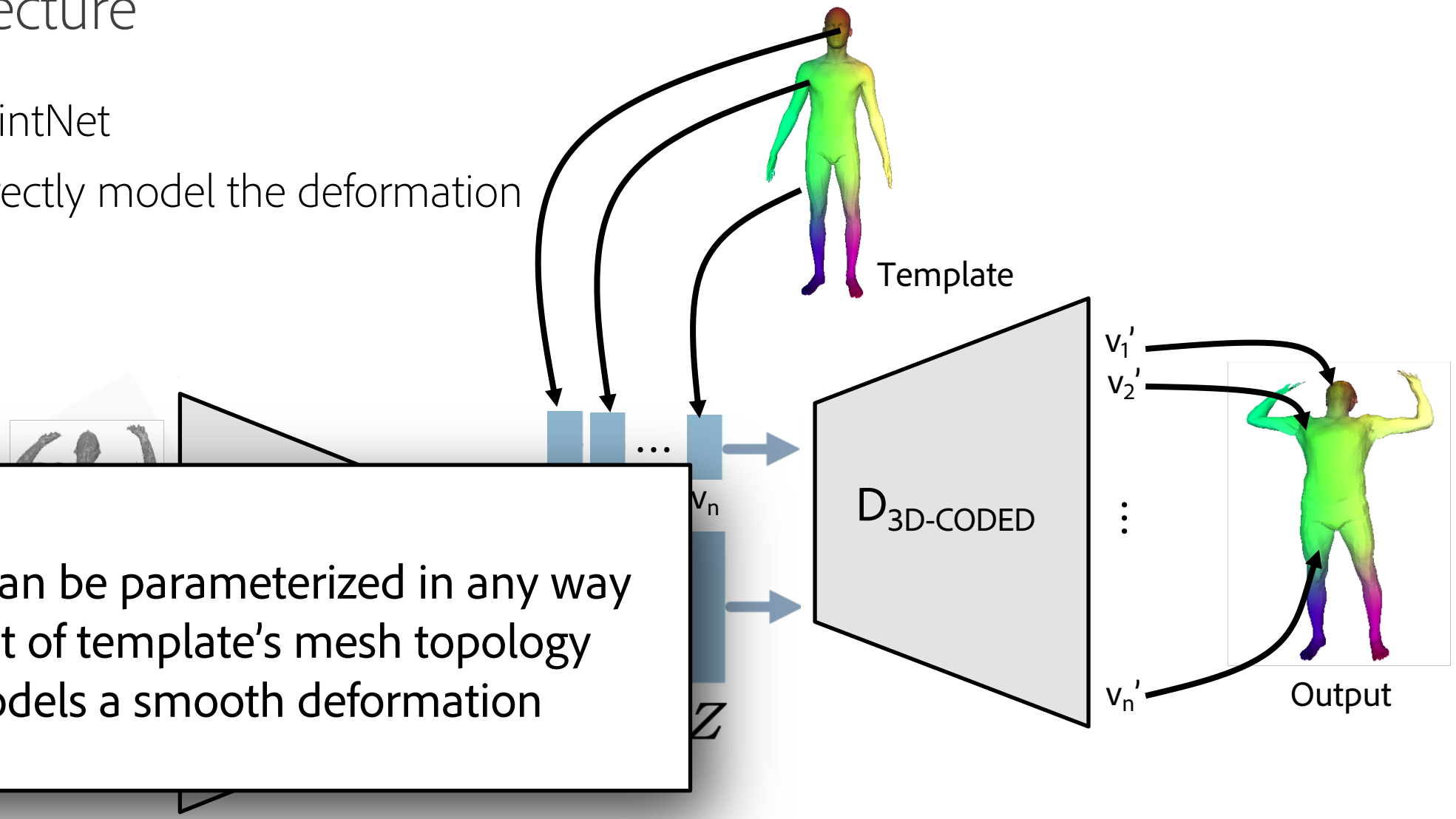
Our Architecture

- Encoder: PointNet
- Decoder: directly model the deformation



Our Architecture

- Encoder: PointNet
- Decoder: directly model the deformation

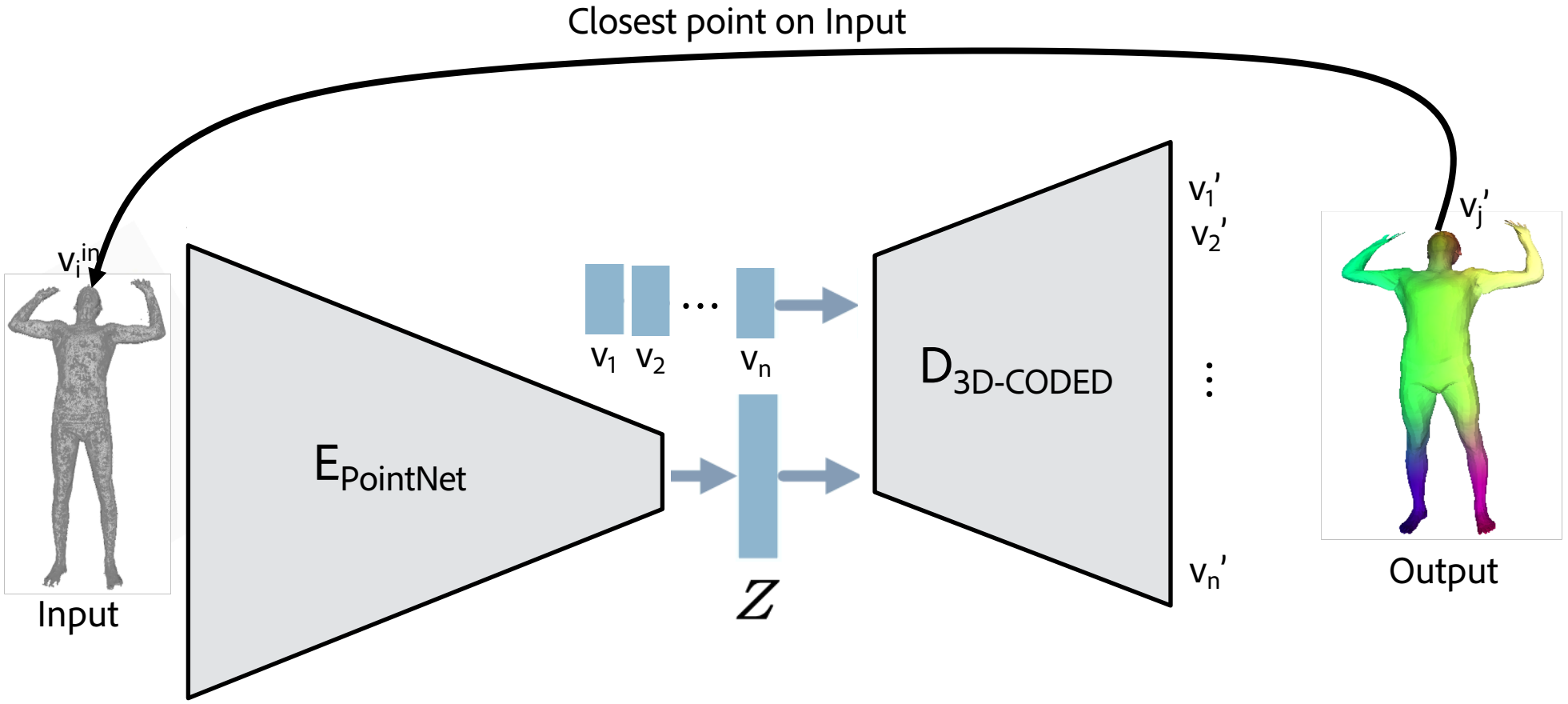


Pros:
- Input data can be parameterized in any way
- Independent of template's mesh topology
- Decoder models a smooth deformation

Training

$$\mathcal{L}^{\text{Chamfer}}(\text{Input}, \text{Output}) = \sum_{v'_j} min_{v_i^{\text{in}}} |v_i^{\text{in}} - v'_j|^2$$

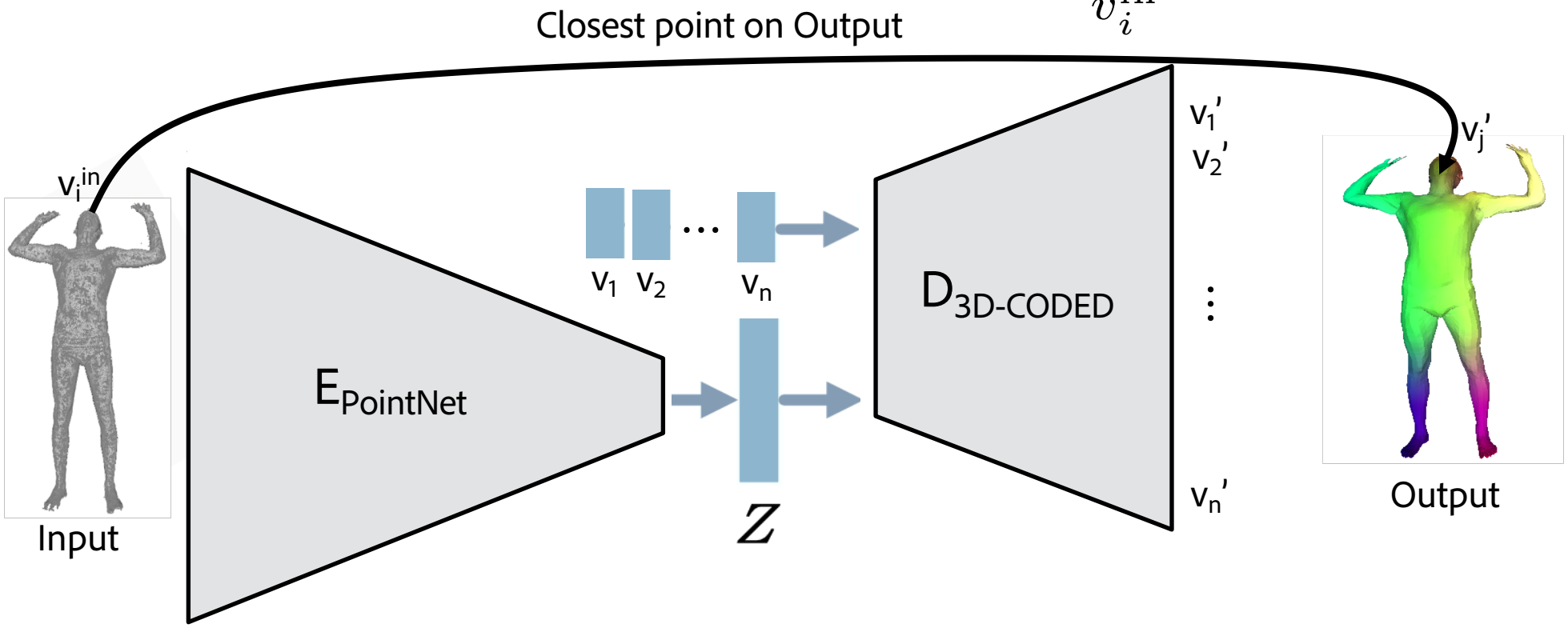
- Chamfer Distance



Training

$$\mathcal{L}^{\text{Chamfer}}(\text{Input}, \text{Output}) = \sum_{v'_j} min_{v_i^{\text{in}}} |v_i^{\text{in}} - v'_j|^2 + \sum_{v_i^{\text{in}}} min_{v'_j} |v_i^{\text{in}} - v'_j|^2$$

- Chamfer Distance

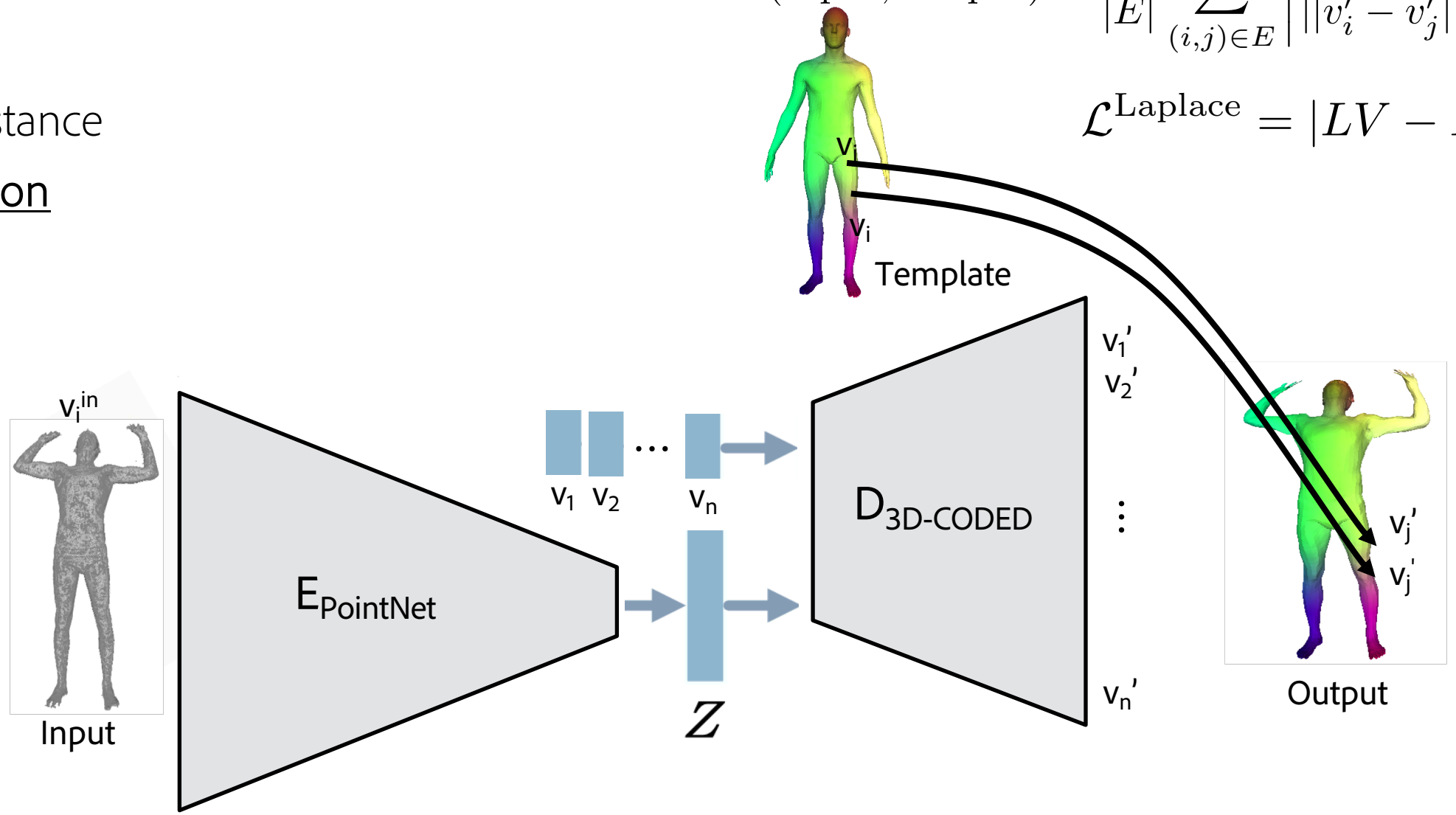


Training

- Chamfer Distance
- Regularization

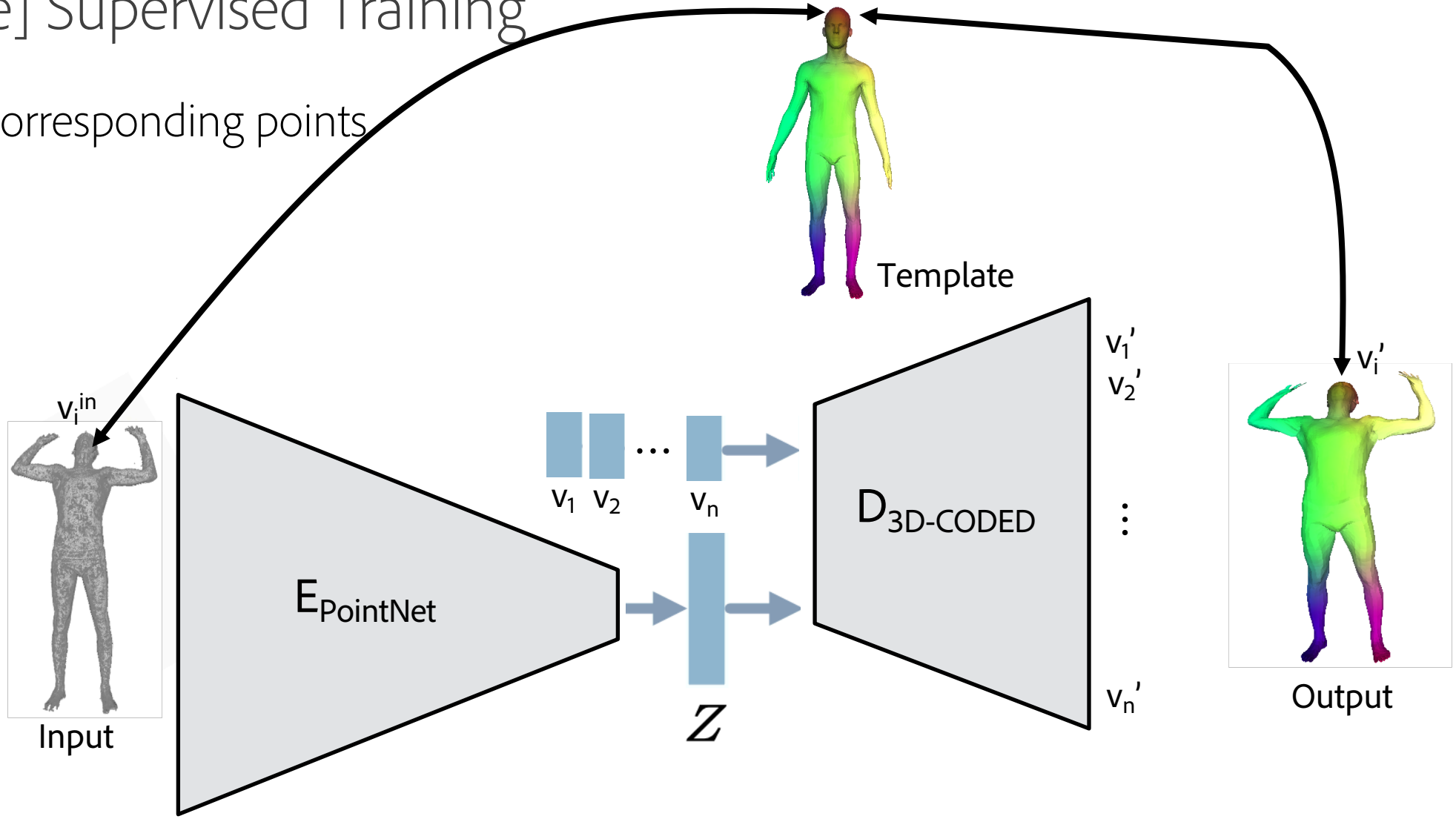
$$\mathcal{L}^{\text{isometry}}(\text{Input}, \text{Output}) = \frac{1}{|E|} \sum_{(i,j)\in E} \left| \frac{||v_i - v_j||}{||v_i' - v_j'||} - 1 \right|$$

$$\mathcal{L}^{\text{Laplace}} = |LV - LV'|$$



[Alternative] Supervised Training

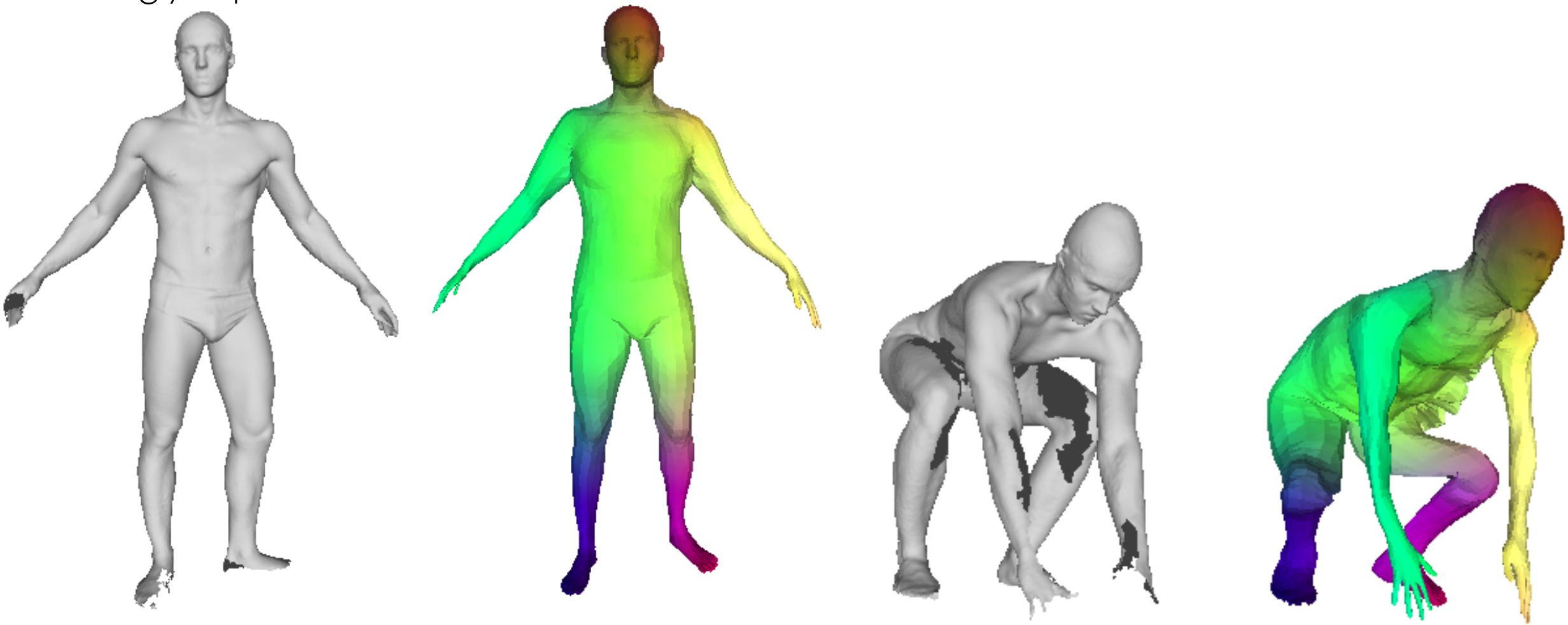
- L2 Loss on corresponding points



$$\mathcal{L}^{\text{supervised}}(\text{Input}, \text{Output}) = \sum_i |v_i^{\text{in}} - v_i'|$$

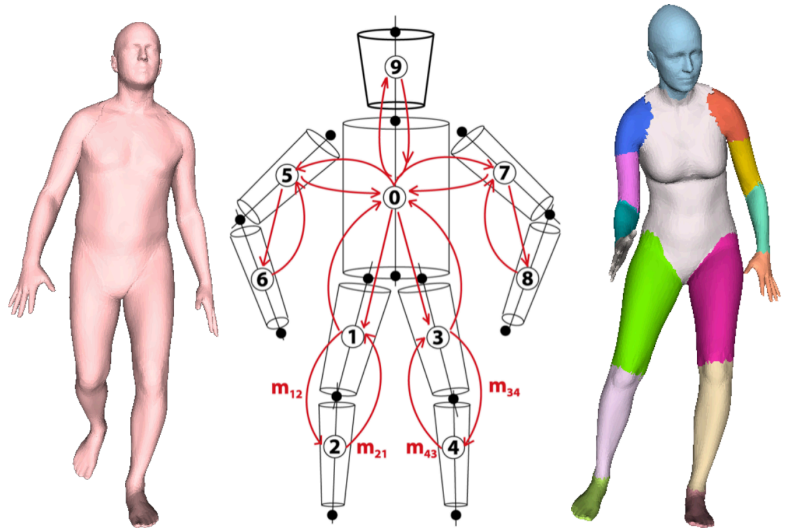
Surface Reconstruction via Template Fitting

- SCAPE dataset of human scans
- Strongly-supervised

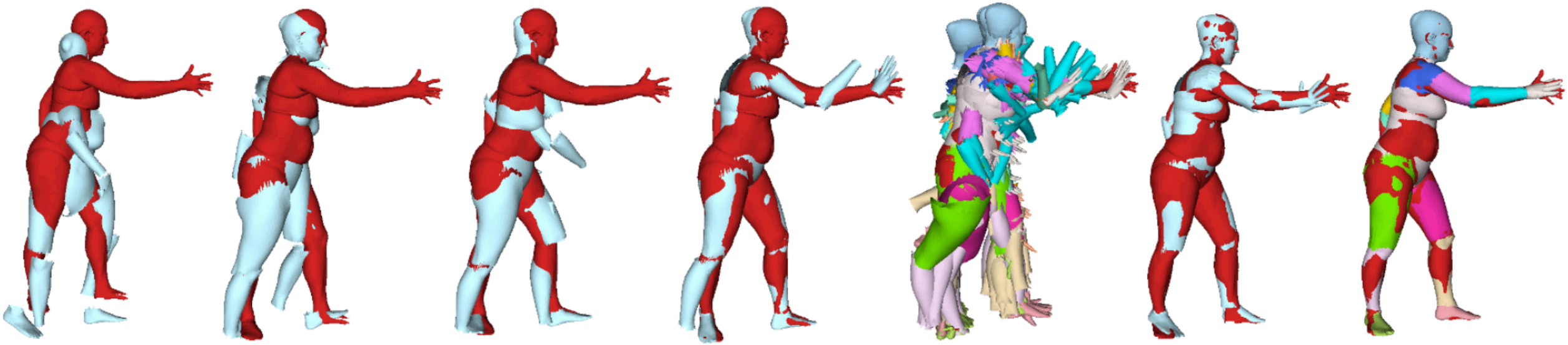


Relationship to Traditional Template Fitting

- Parameterize template deformations
- Guess initial parameters
- Gradient descent to get better parameters



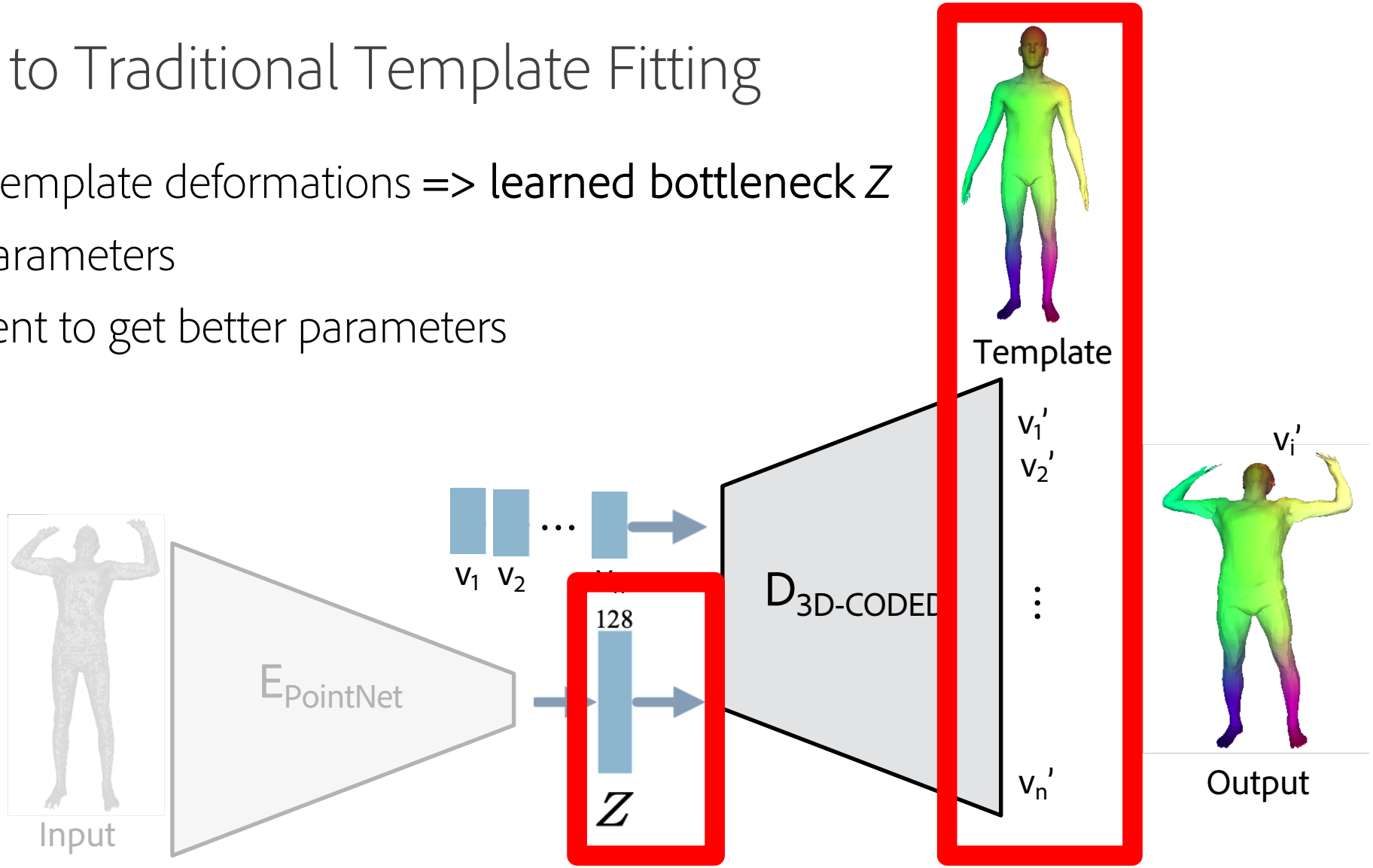
Various hand-crafted template parameterizations



E.g., Gradient descent using Stitched Puppet model by Zuffi and Black CVPR 2015

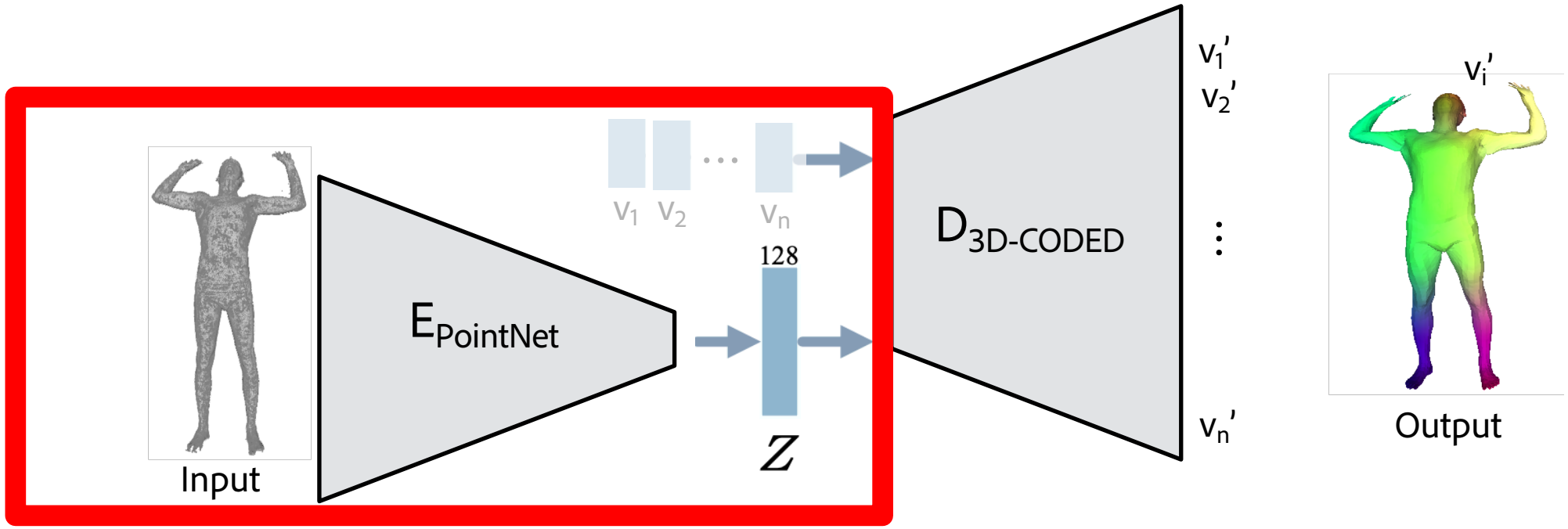
Relationship to Traditional Template Fitting

- Parameterize template deformations => learned bottleneck Z
- Guess initial parameters
- Gradient descent to get better parameters



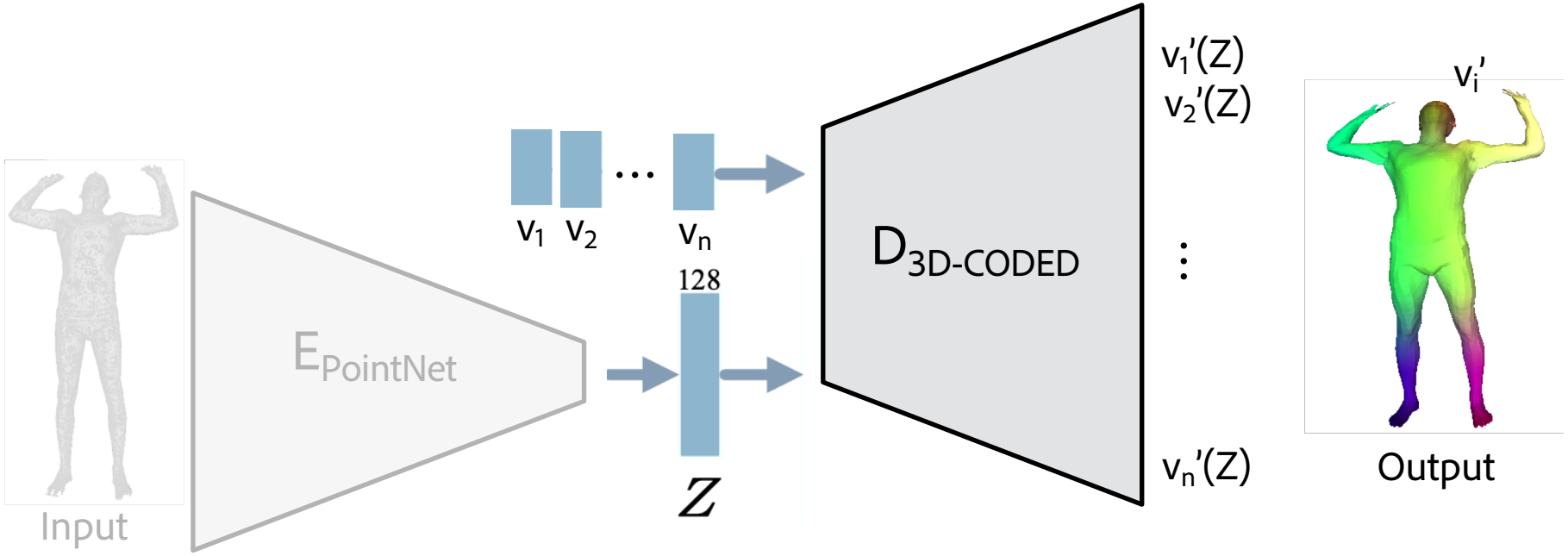
Relationship to Traditional Template Fitting

- Parameterize template deformations => learned bottleneck *Z*
- Guess initial parameters => forward pass through the neural network
- Gradient descent to get better parameters



Relationship to Traditional Template Fitting

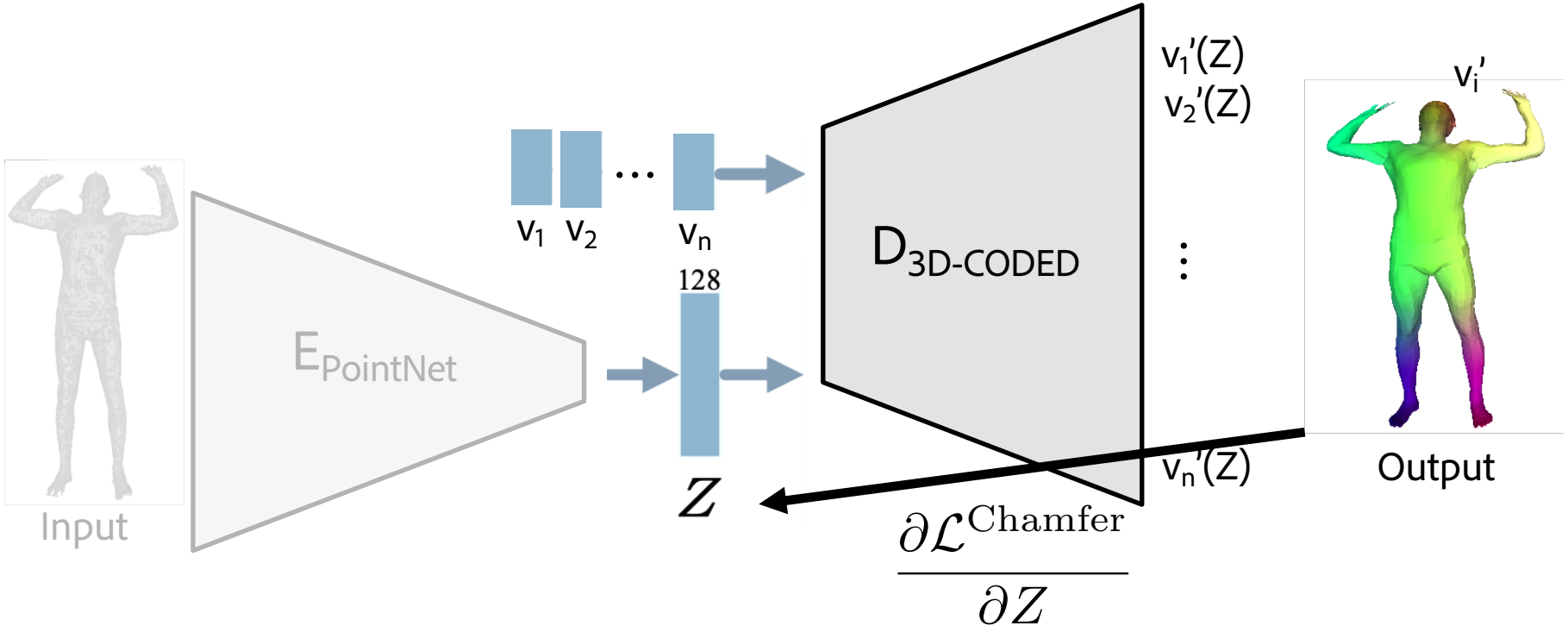
- Parameterize template deformations => learned bottleneck Z
- Guess initial parameters => forward pass through the neural network
- Gradient descent to get better parameters => backpropagate through the neural network



$$\mathcal{L}^{\text{Chamfer}}(Z) = \sum_{v'_j} min_{v_i^{\text{in}}} |v_i^{\text{in}} - v'_j(Z)|^2 + \sum_{v_i^{\text{in}}} min_{v'_j} |v_i^{\text{in}} - v'_j(Z)|^2$$

Relationship to Traditional Template Fitting

- Parameterize template deformations => learned bottleneck Z
- Guess initial parameters => forward pass through the neural network
- Gradient descent to get better parameters => backpropagate through the neural network



$$Z = argmin_{Z^*} \mathcal{L}^{\text{Chamfer}}(Z^*)$$

Surface Reconstruction via Template Fitting

Input Shape

Initial Deformation

Final Deformation

Evaluation

- Train Surreal[1] + augmented bending poses ~ 230,000 synthetic poses
- Test on FAUST[2]

Method	Faust error
Convex-Opt [3]	8.30
FMNet [4]	4.82
SP [5]	3.12
Ours Supervised	6.29
Ours Supervised + Regression	**2.87**
Ours **Unsupervised** + Regression	4.88

[1] Learning from synthetic humans, Varol et al. CVPR (2017)
[2] FAUST: Dataset and evaluation for 3D mesh registration, Bogo et al. CVPR (2014)
[3] Robust nonrigid registration by convex optimization, Chen, Koltun, ICCV (2015)
[4] Deep functional maps: Structured prediction for dense shape correspondence, Litany et al. ICCV (2017)
[5] The stitched puppet: A graphical model of 3d human shape and pose, Zuffi et al. CVPR (2015)

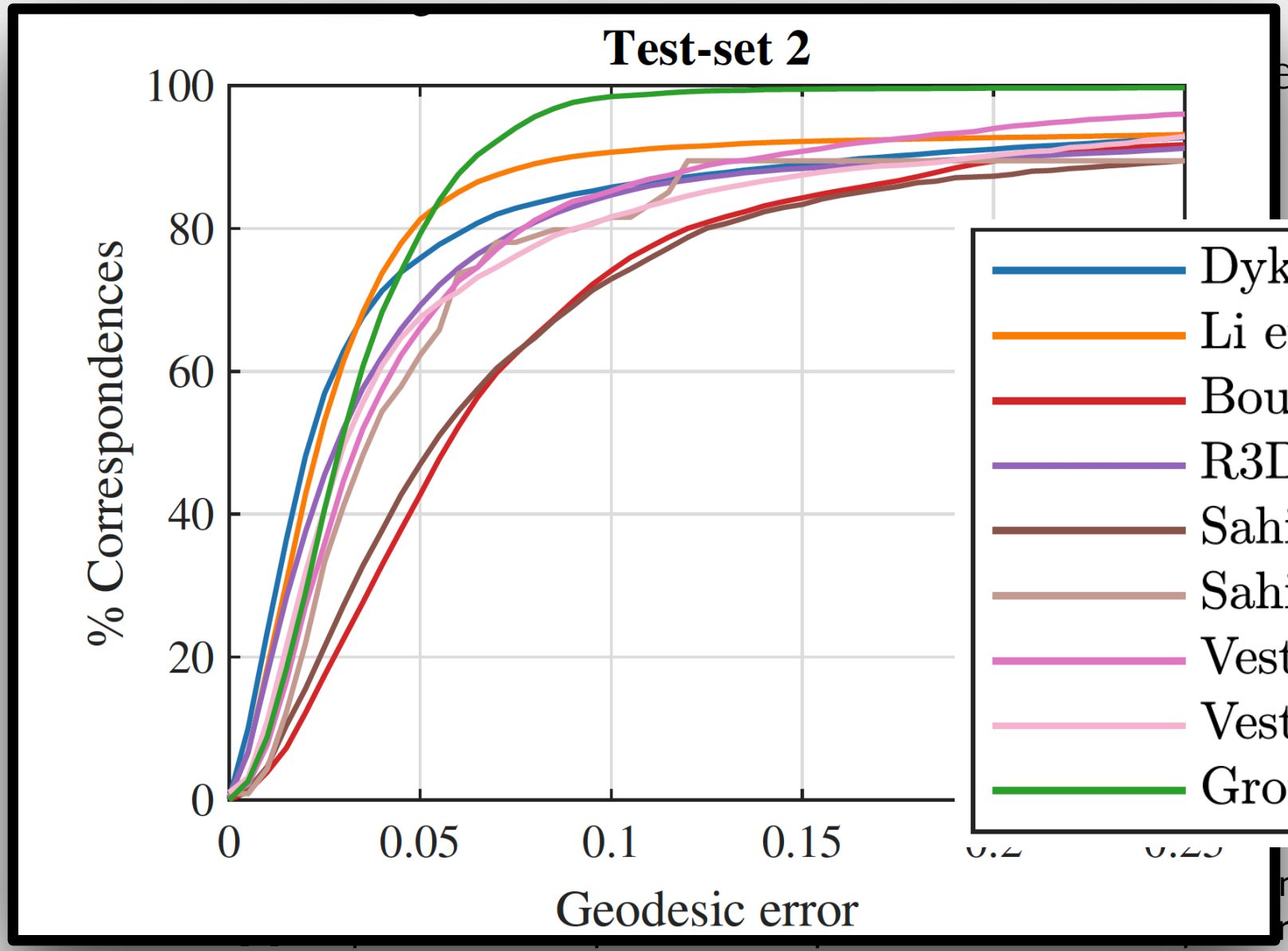
Evaluation

- Train Surreal[1] + augmented bending poses ~ 230,000 synthetic poses
- Test on FAUST[2]

Method	Faust error
Convex-Opt [3]	8.30
FMNet [4]	4.82
SP [5]	3.12
Ours Supervised	6.29
Ours Supervised + Regression	**2.87**
Ours **Unsupervised** + Regression	4.88

[1] Learning from synthetic humans, Varol et al. CVPR (2017)
[2] FAUST: Dataset and evaluation for 3D mesh registration, Bogo et al. CVPR (2014)
[3] Robust nonrigid registration by convex optimization, Chen, Koltun, ICCV (2015)
[4] Deep functional maps: Structured prediction for dense shape correspondence, Litany et al. ICCV (2017)
[5] The stitched puppet: A graphical model of 3d human shape and pose, Zuffi et al. CVPR (2015)

SHREC'19: Shape Correspondence with Isometric and Non-Isometric Deformations



etic poses

n, ICCV (2015)

respondence, Litany et al. ICCV (2017)

[5] The stitched puppet: A graphical model of 3d human shape and pose, Zuffi et al. CVPR (2015)

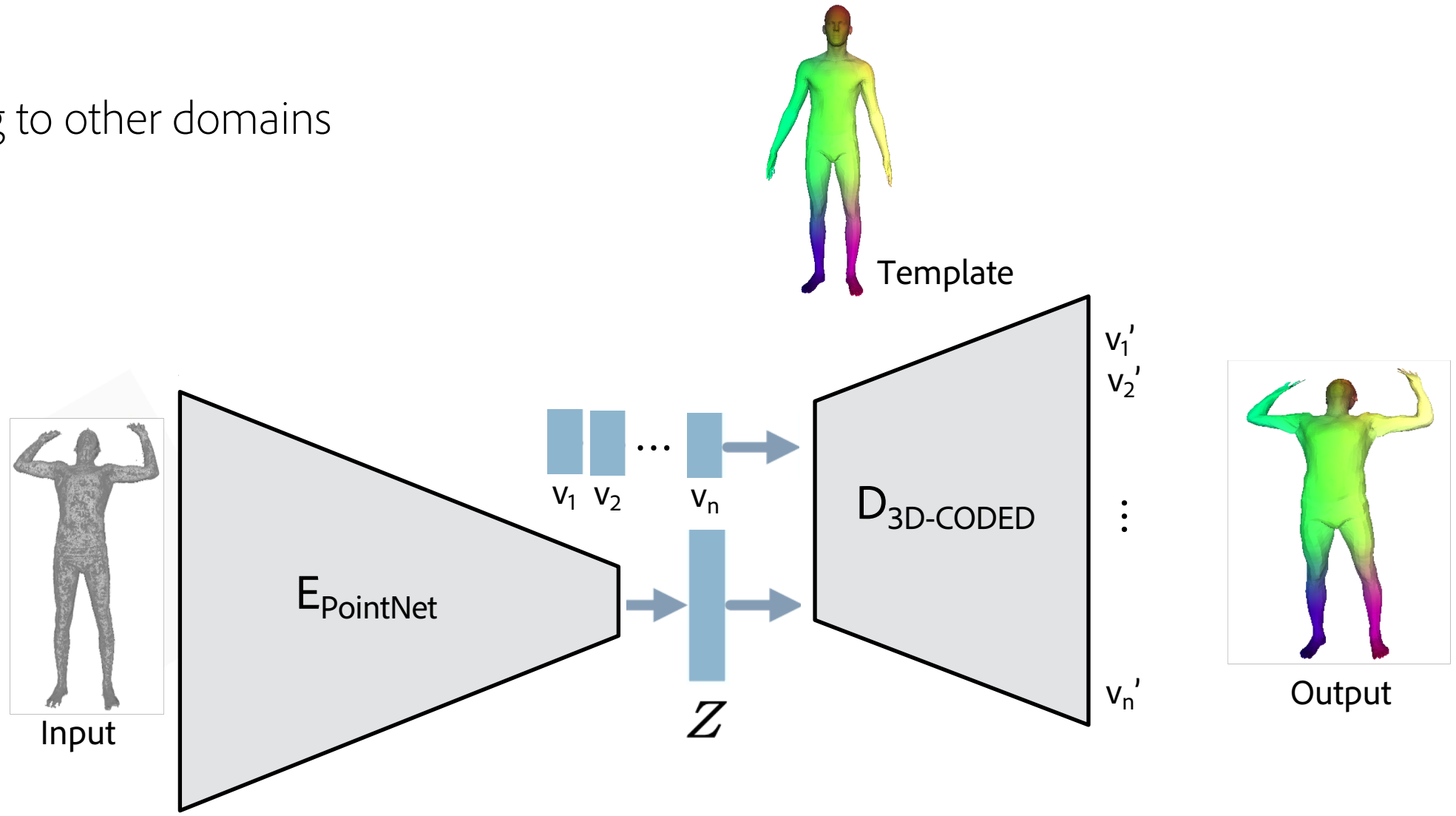
Talk Outline

- Template Fitting and Correspondence Estimation via a Deforming Neural Networks
- **Template-less Modeling via Deforming Neural Networks**
- Template-less Signal Transfer via Deforming Neural Networks
- Multi-view Reconstruction via Deforming Neural Networks

Thibault Groueix

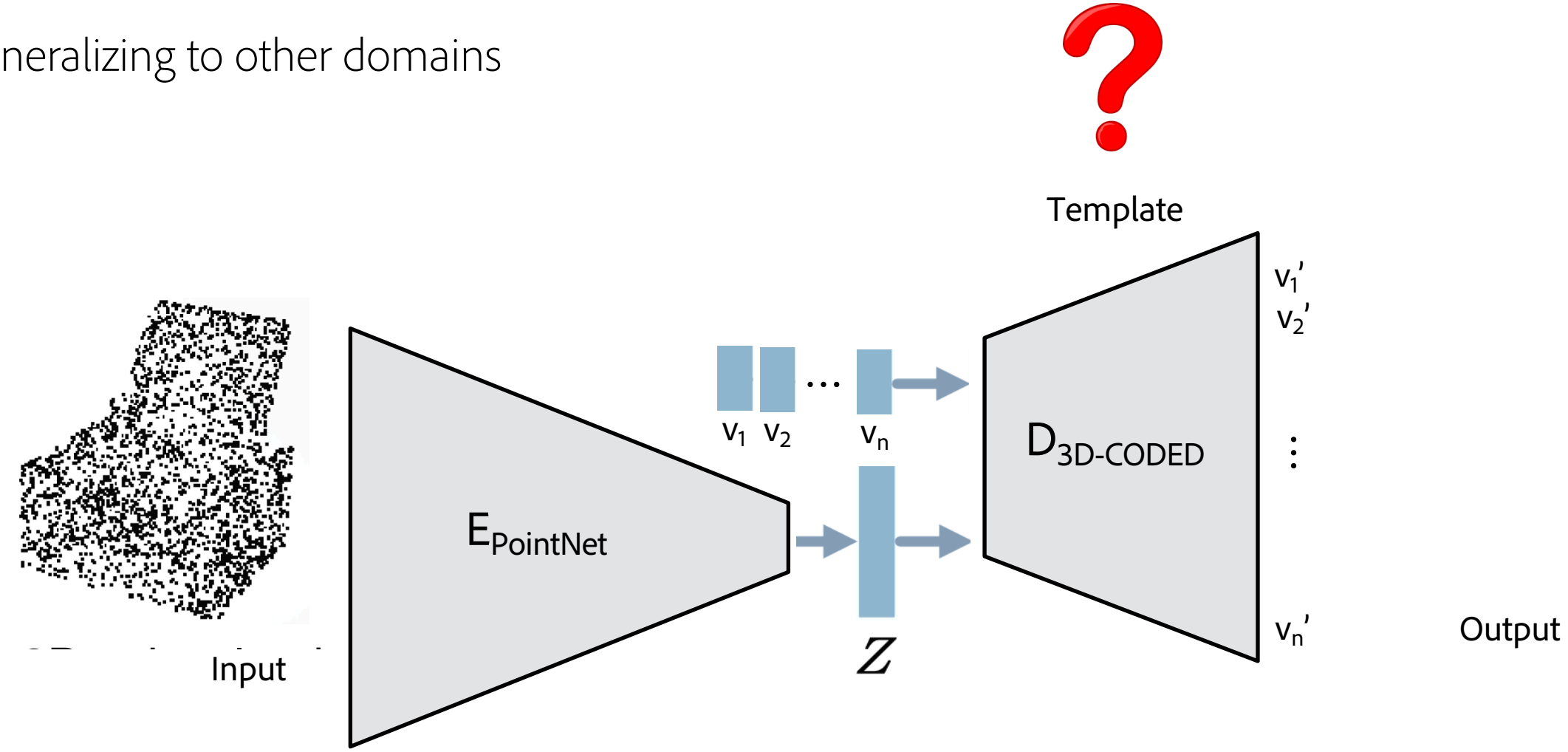
AtlasNet

- Generalizing to other domains



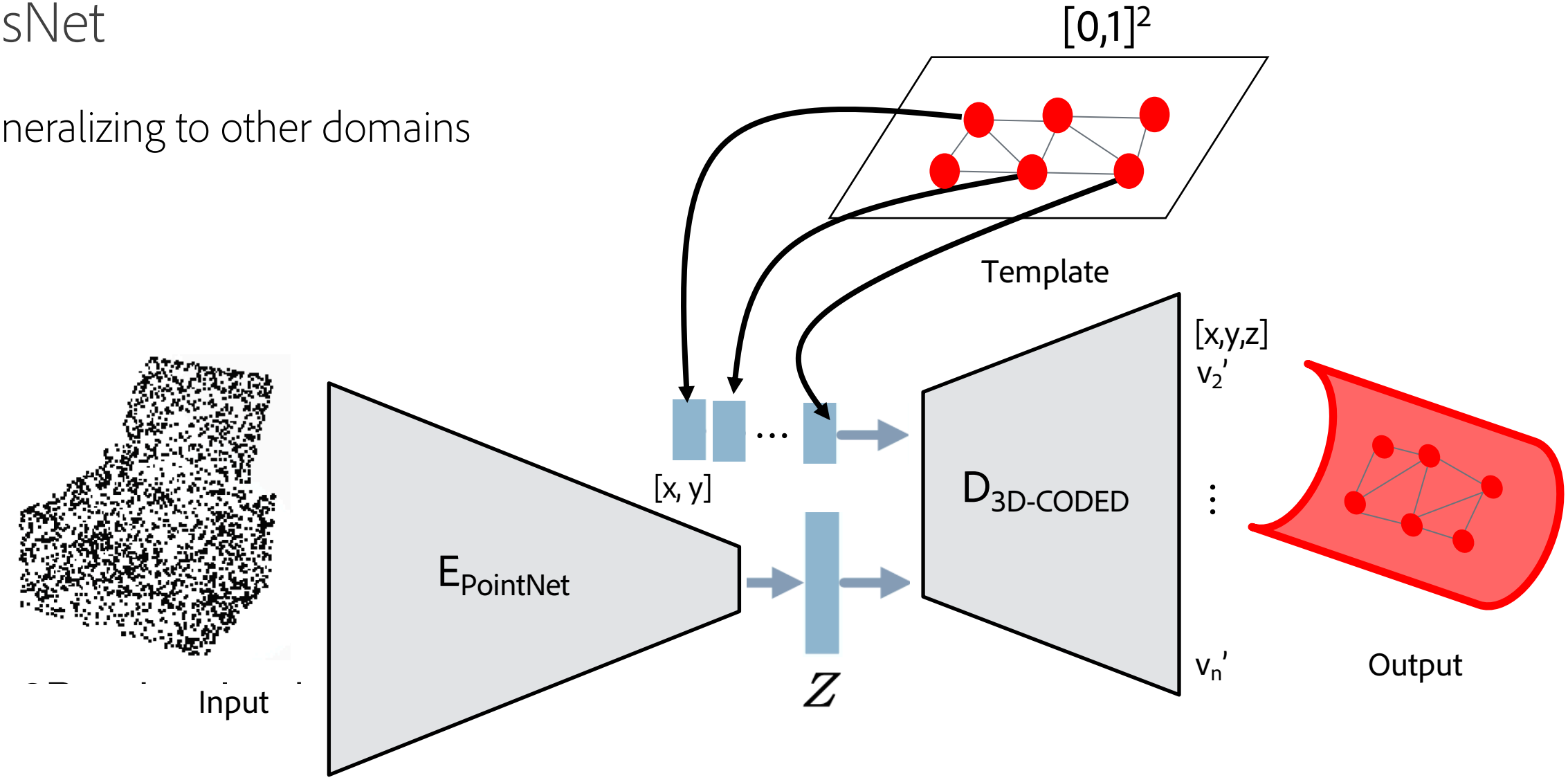
AtlasNet

- Generalizing to other domains



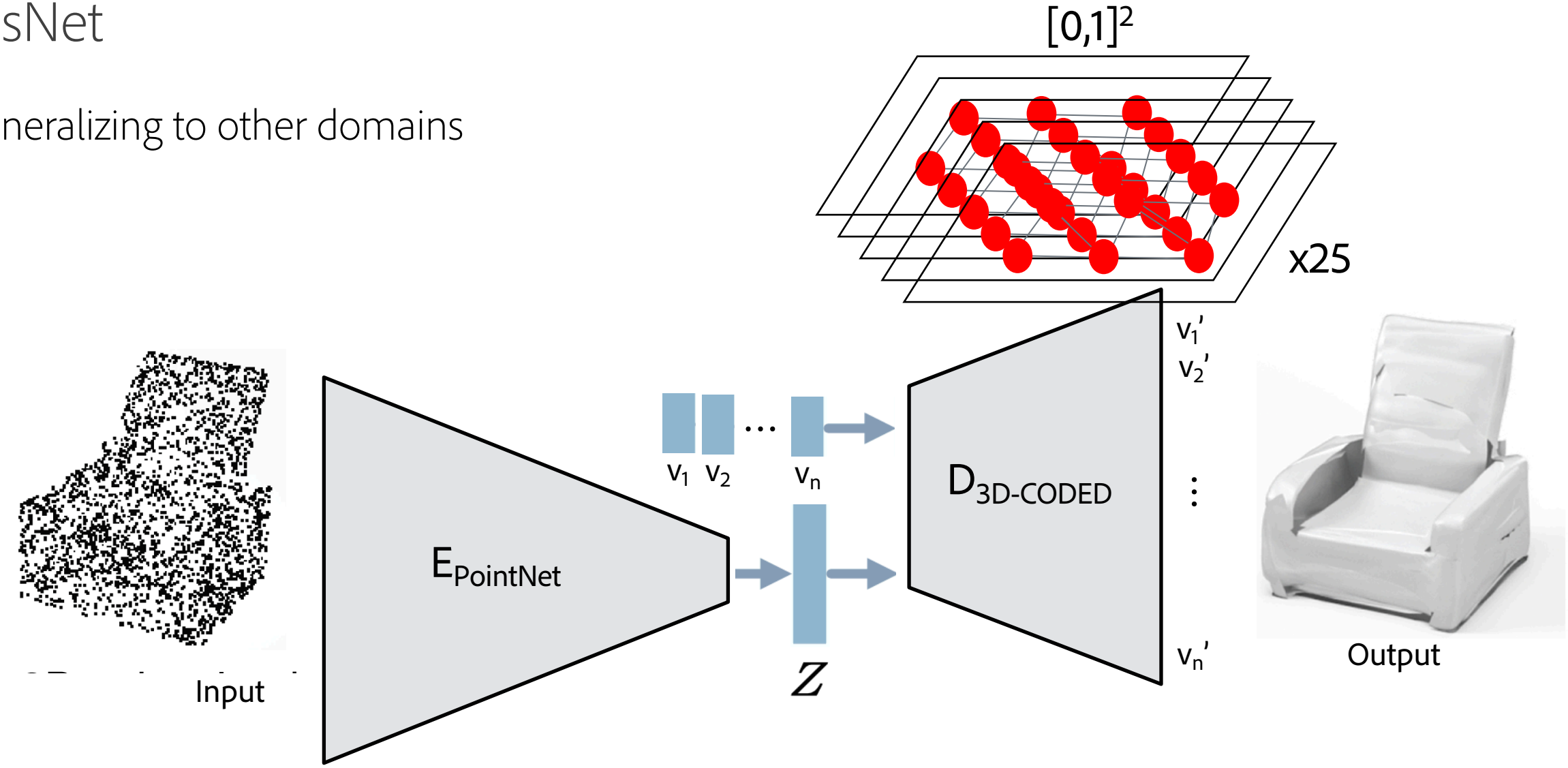
AtlasNet

- Generalizing to other domains



AtlasNet

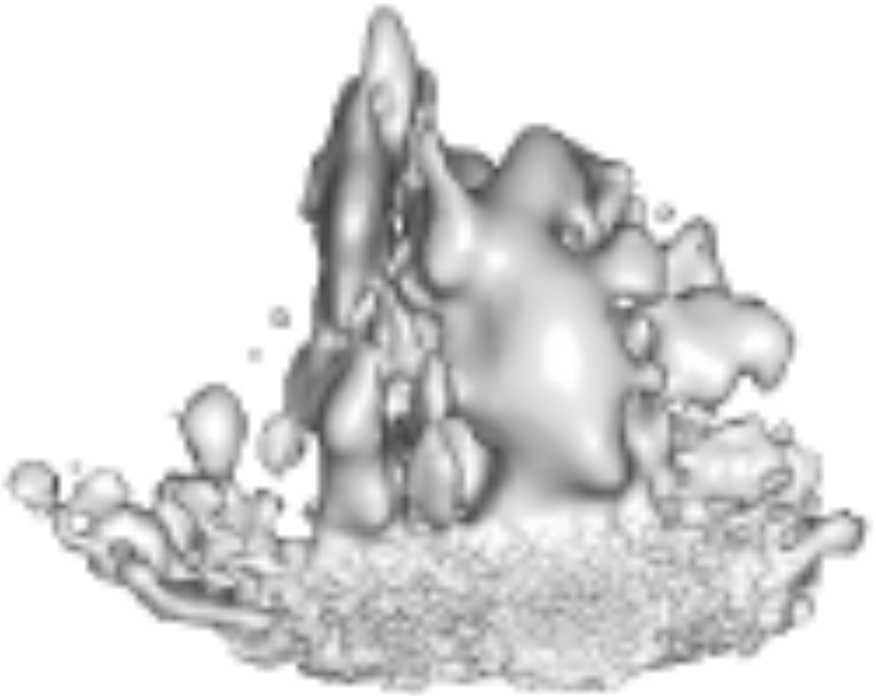
- Generalizing to other domains



Point-based Reconstructions

- Train Point-to-Atlas autoencoder with Chamfer loss

Ground Truth



Baseline
[Poisson Surface Reconstruction from Sampled Points]

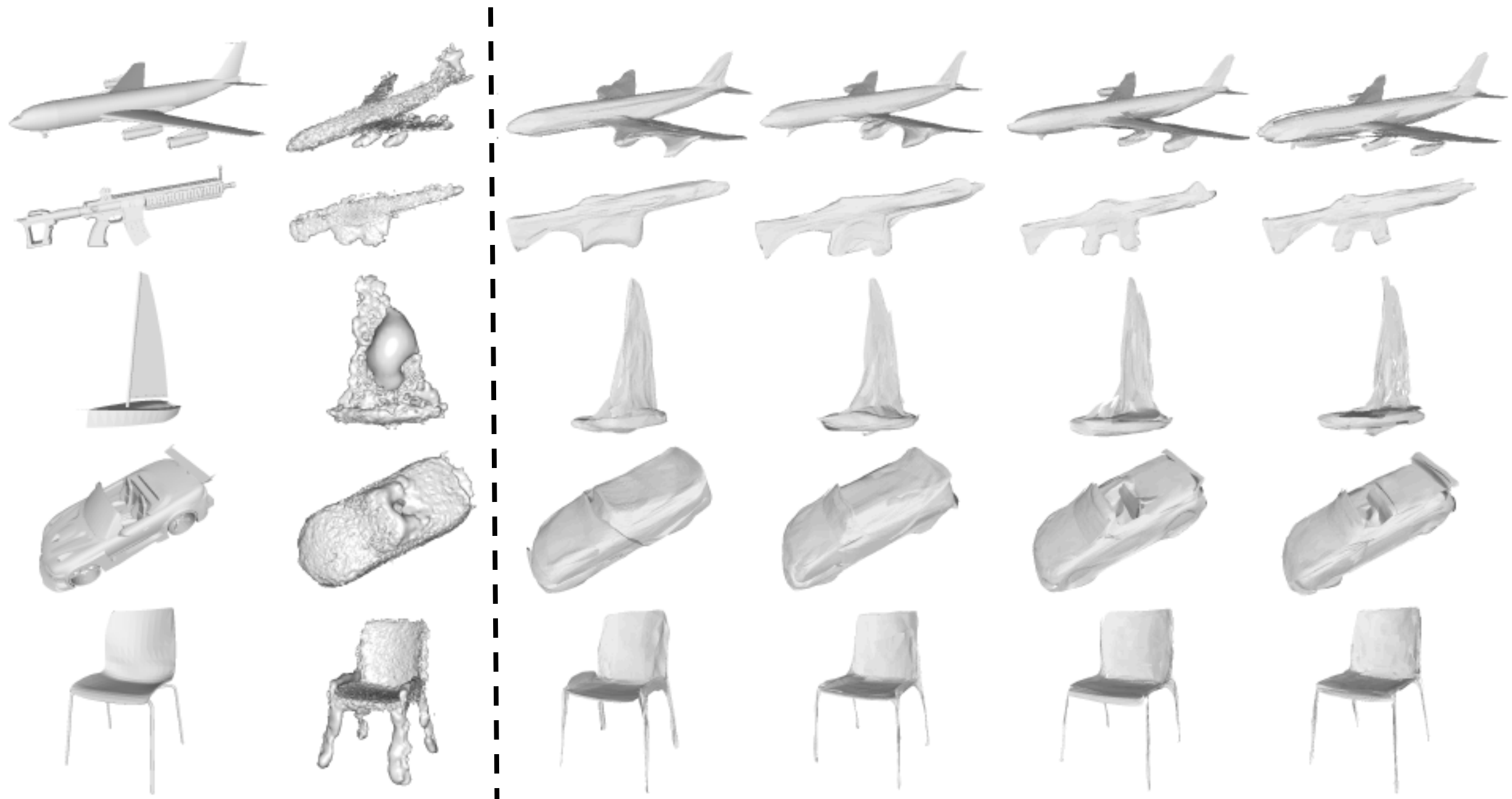
Point-based Reconstructions

- Train Point-to-Atlas autoencoder with Chamfer loss

Ground Truth

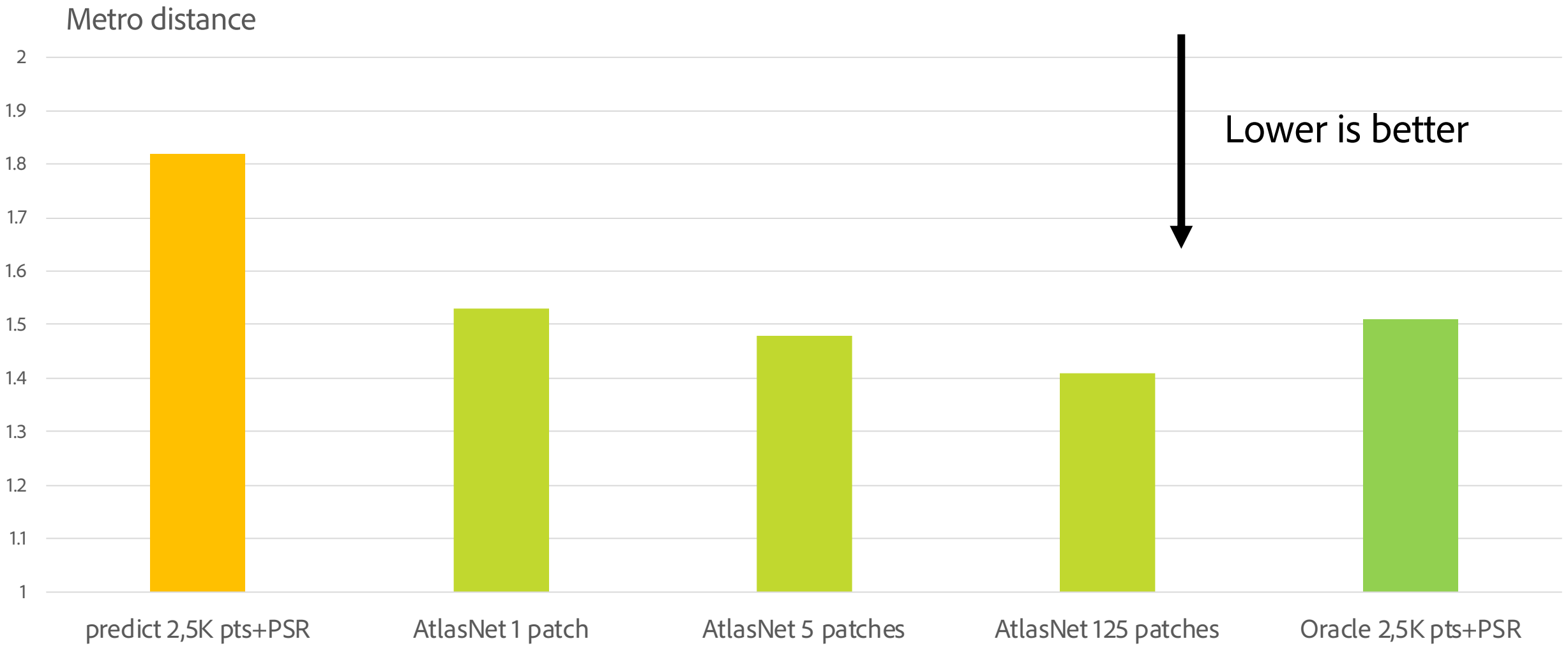
AtlastNet
with 5 patches

AtlastNet
with 25 patches

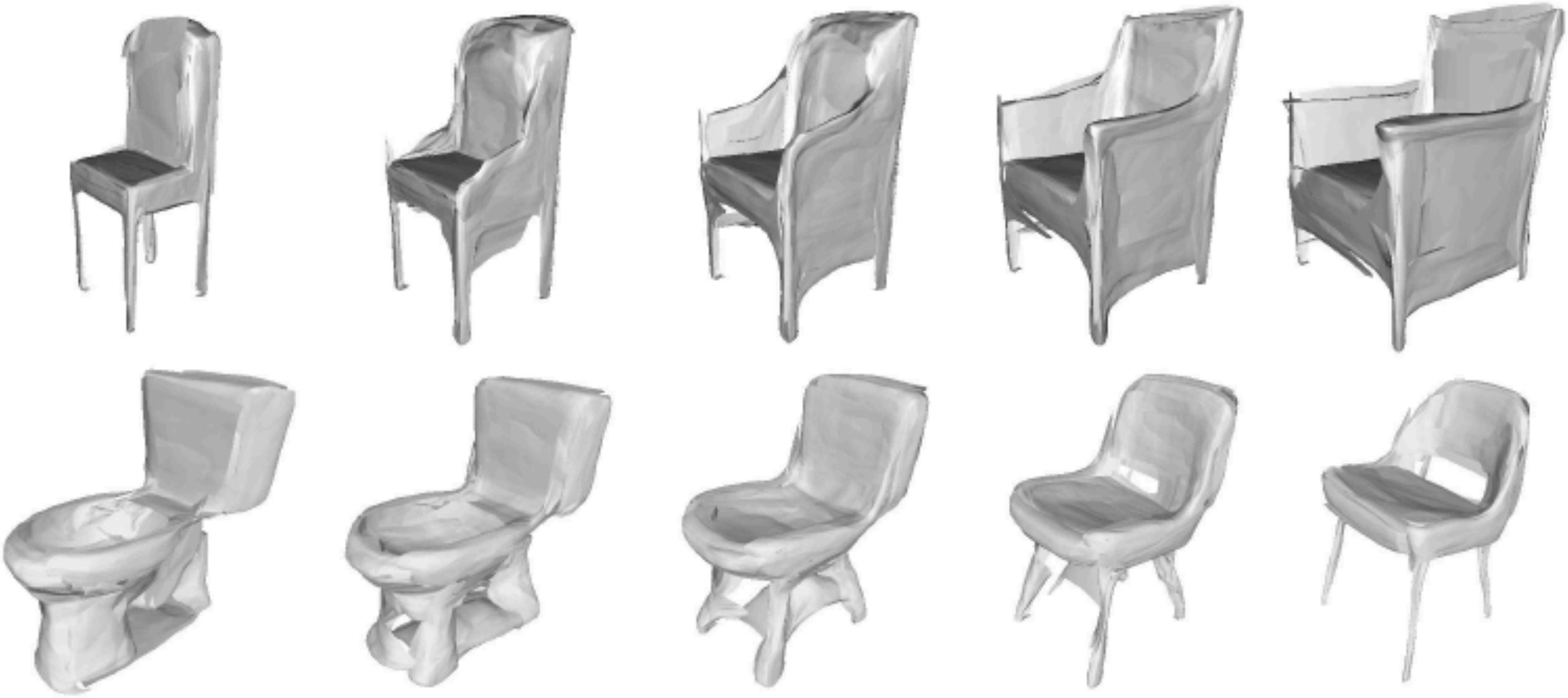


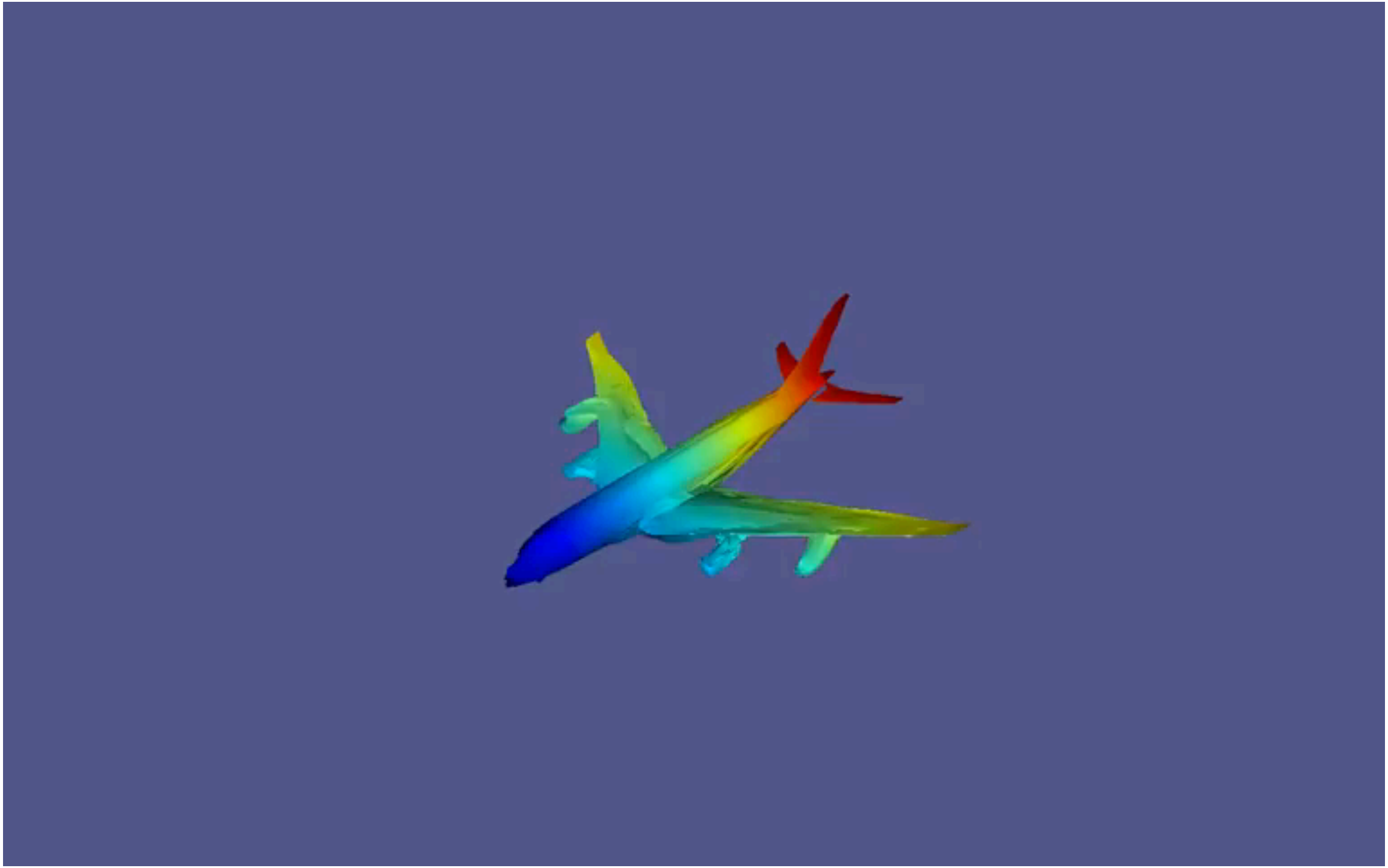
(a) Ground truth (b) Baseline (e) Ours 1 (f) Ours 5 (g) Ours 25 (h) Ours 125

Quantitative Evaluation



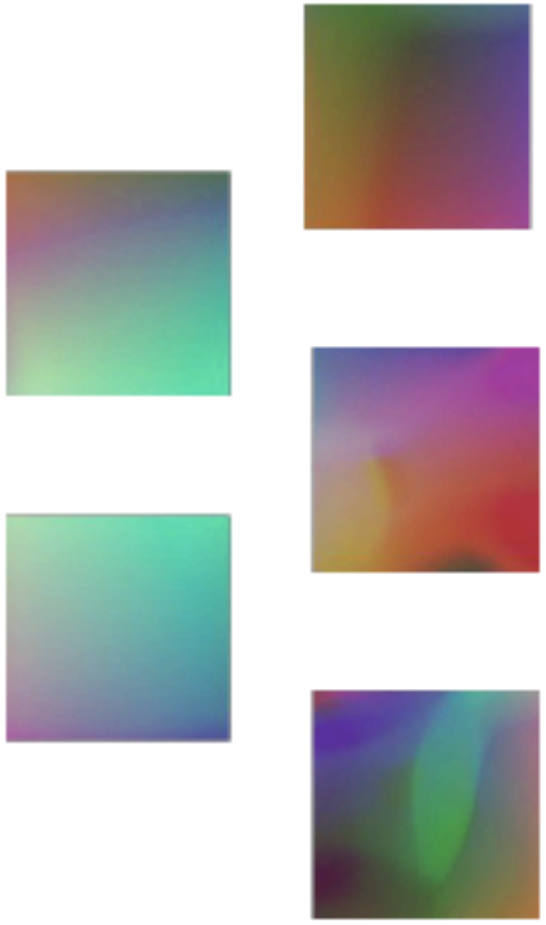
Shape Interpolation



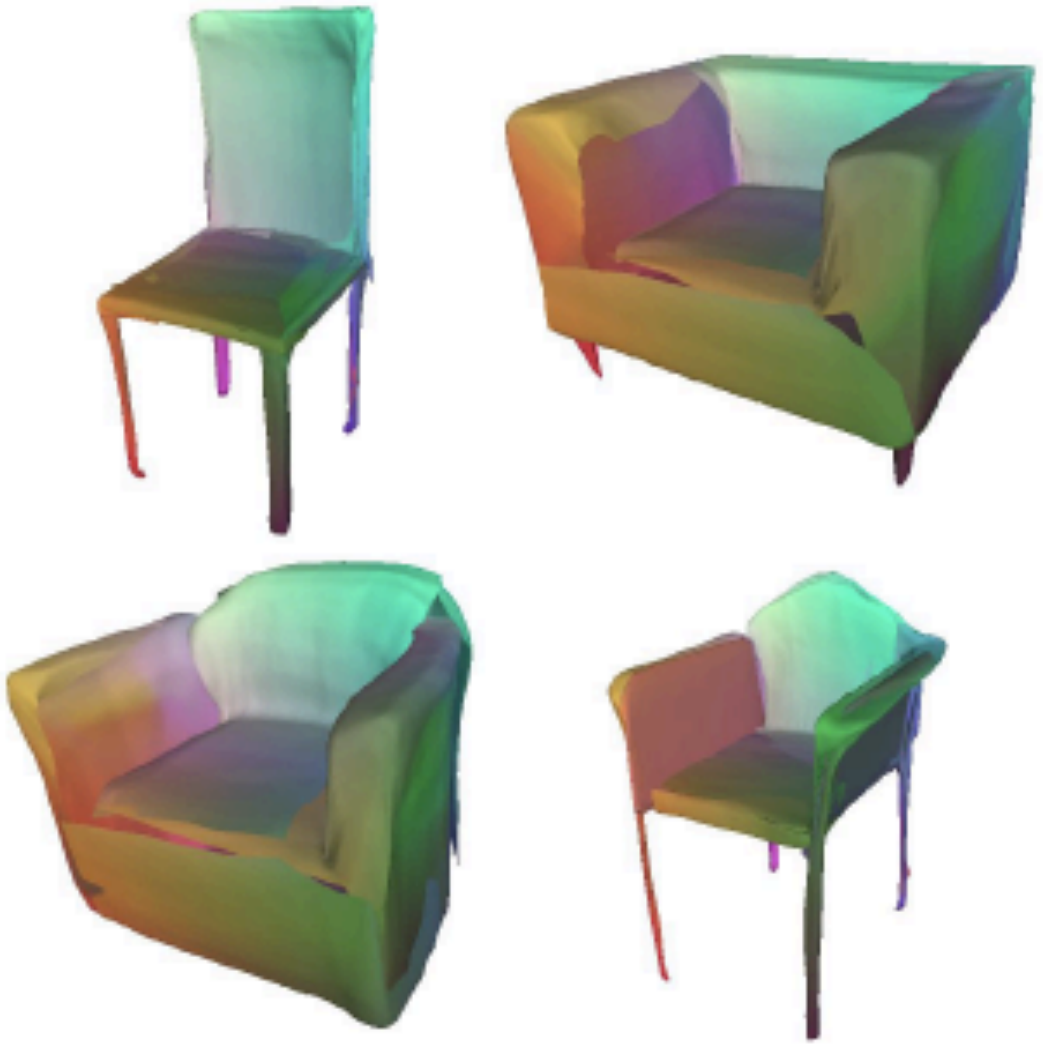


Shape Correspondences

Reference Object



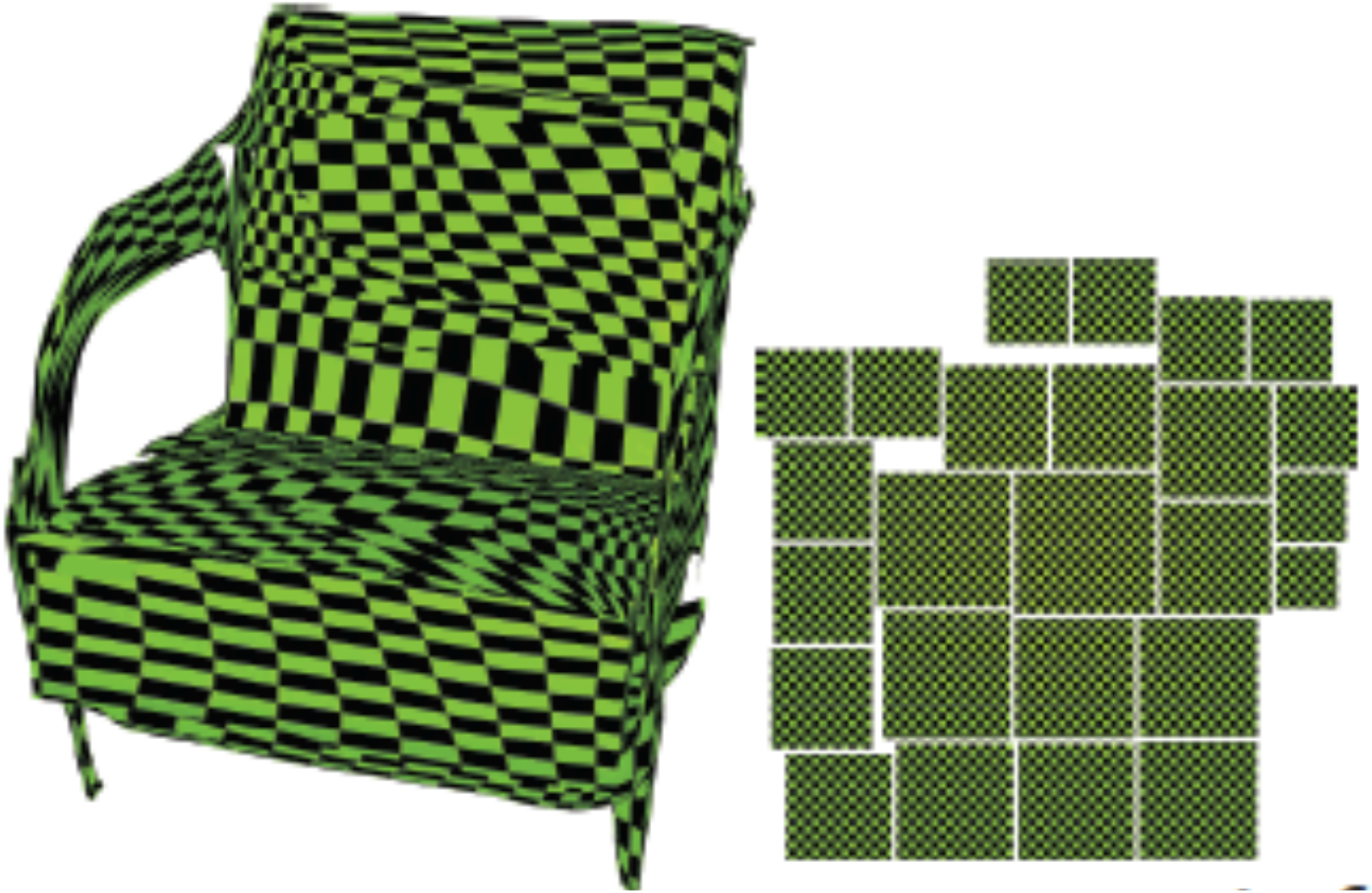
Colors Transferred
to the Inferred Atlas



Colors Transferred
to Other Objects

Taking advantage of UV parameterizations

- Bijective map, but with high distortion



Taking advantage of UV parameterizations

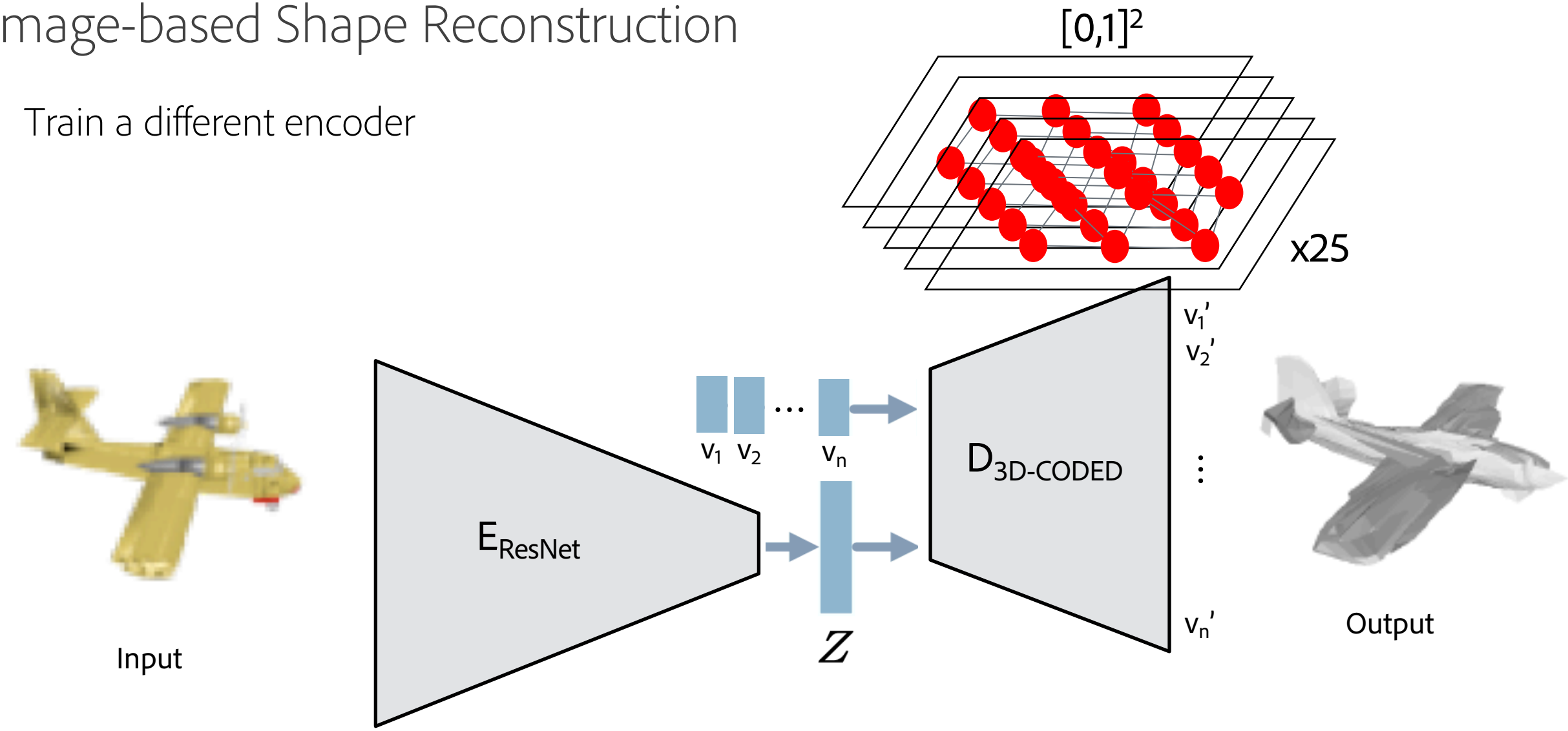
- Bijective map, but with high distortion

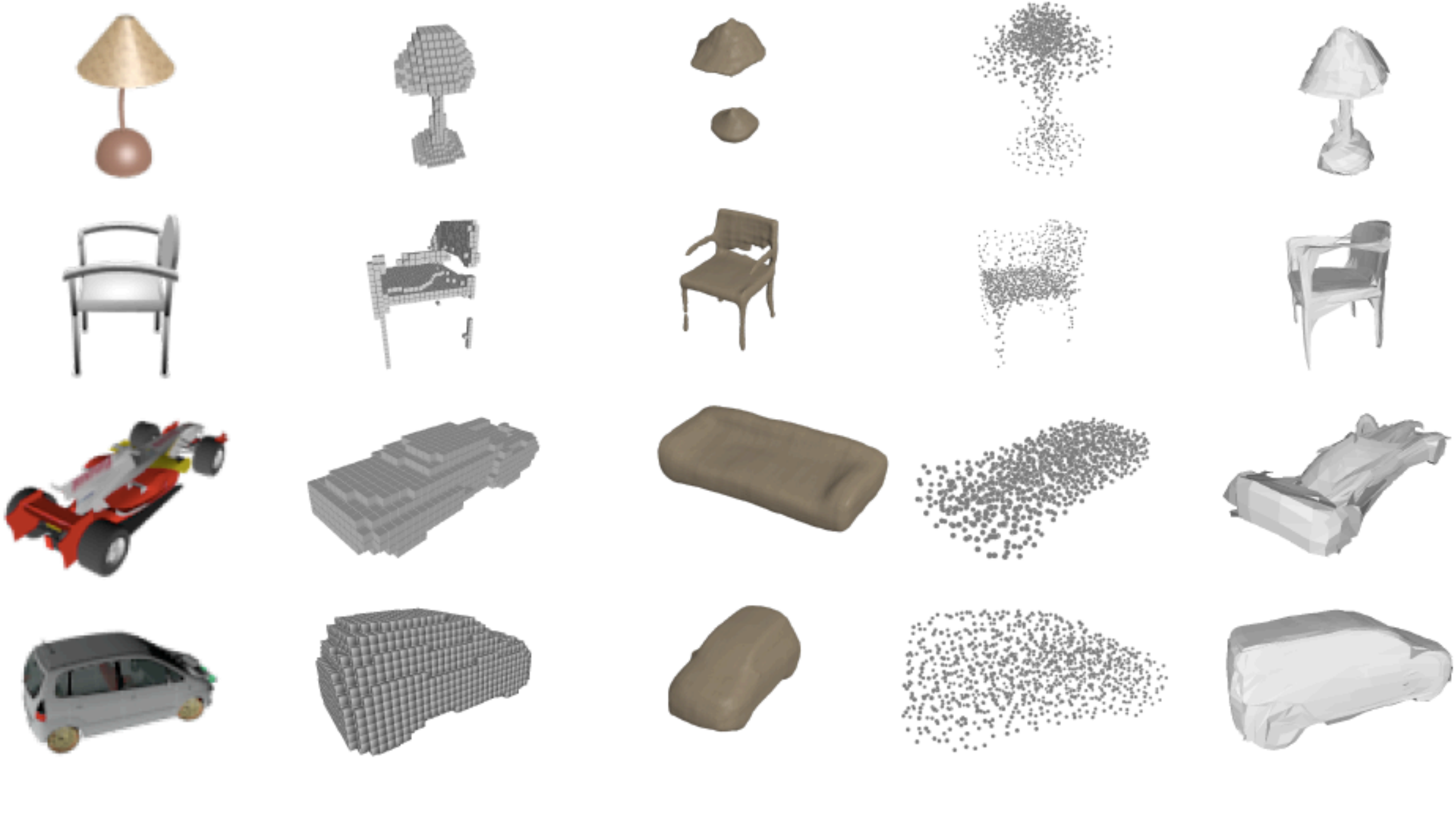
Optimized Atlas

Taking advantage of UV parameterizations

Image-based Shape Reconstruction

- Train a different encoder




(a) Input
(b) 3D-R2N2
(c) HSP
(d) PSG
(e) Ours

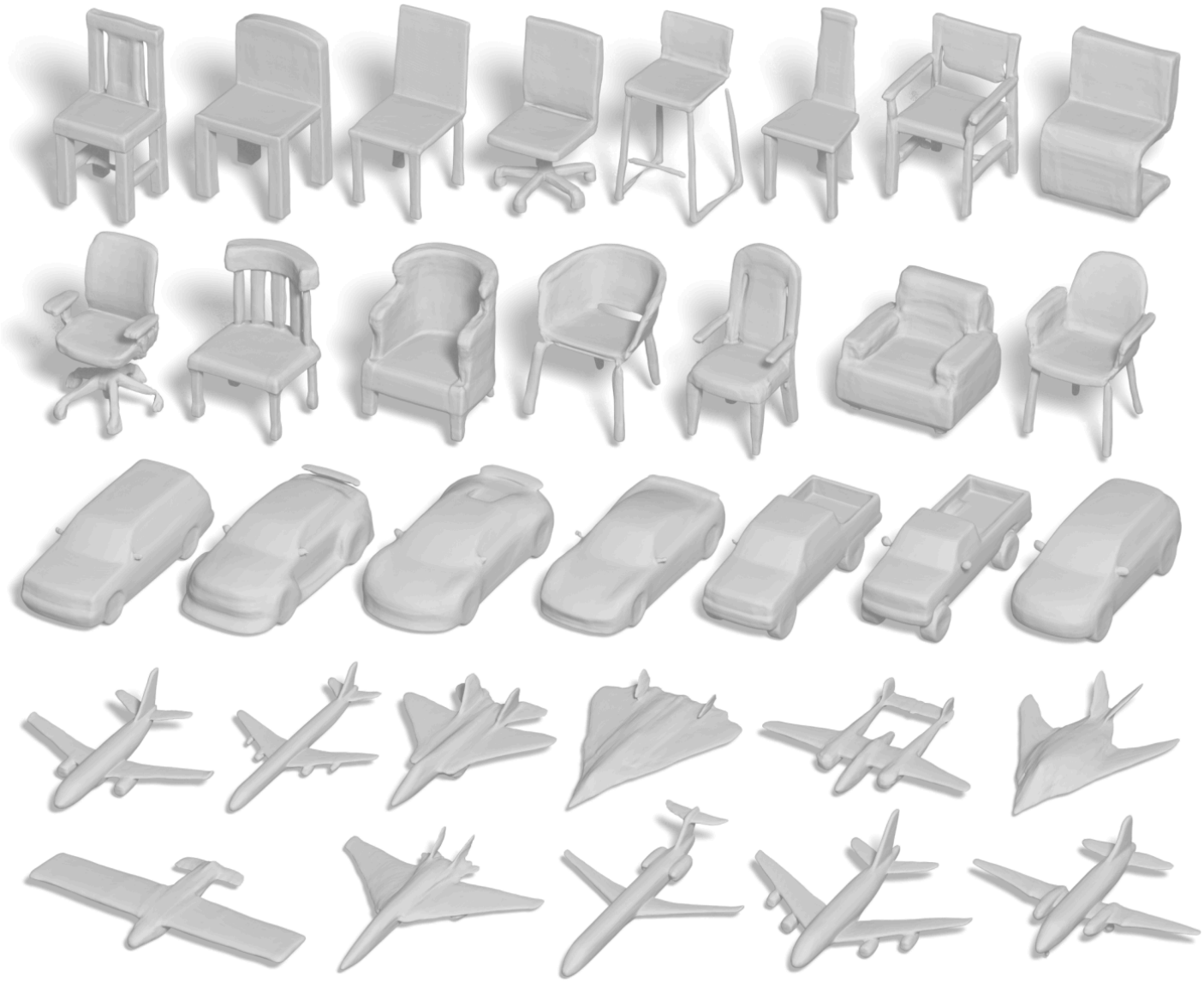
Try It Yourself!

- Some examples I tried recently

https://github.com/ThibaultGROUEIX/AtlasNet

Implicit Functions vs Deformation-based Models

- Possible to infer deformation directly [3rd section of this talk]
- Meshes are easier to plug into existing Graphics and Vision pipelines [4th section of this talk]
- Implicit surfaces look much smoother: loss function or representation? [future challenges]

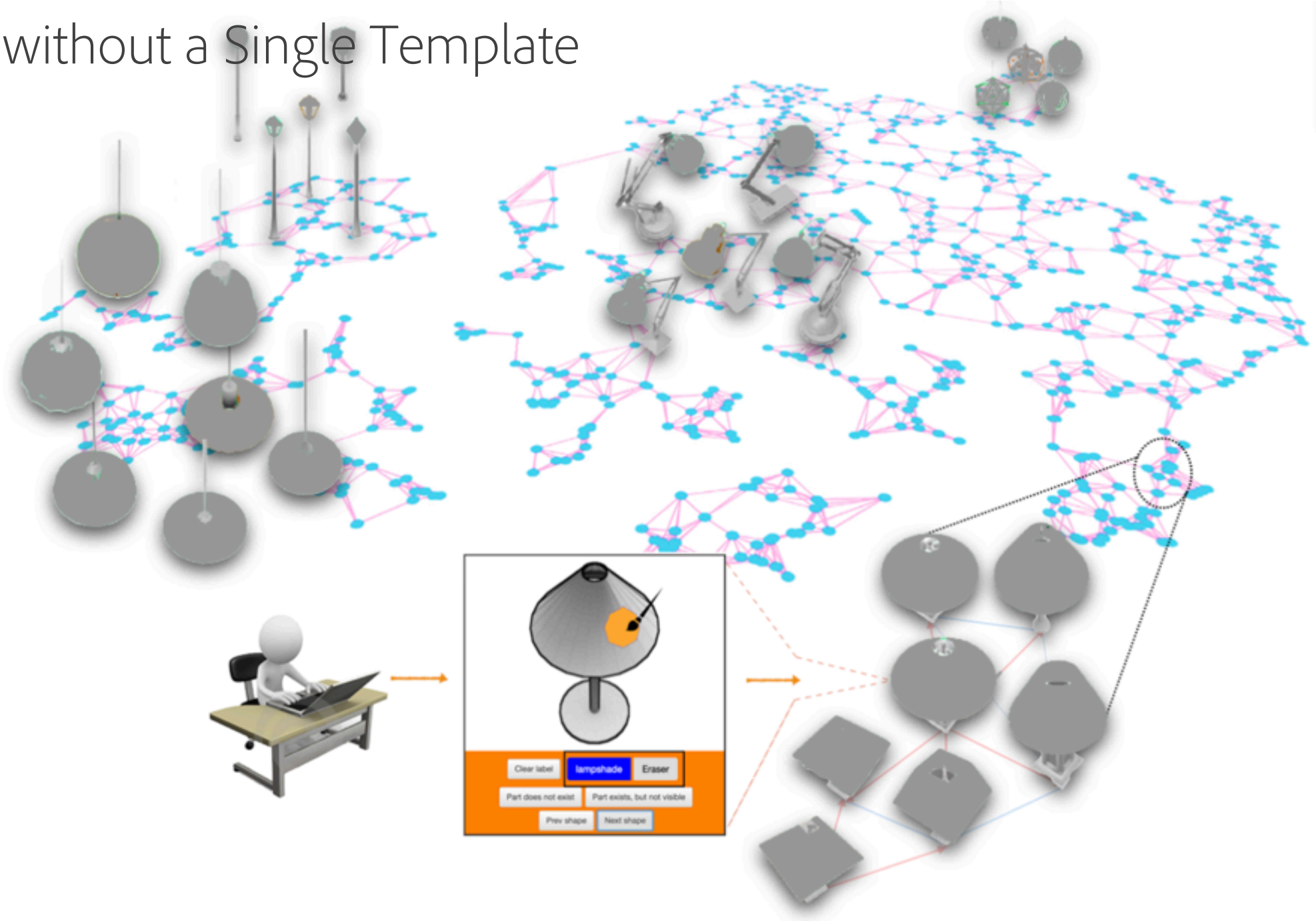


Talk Outline

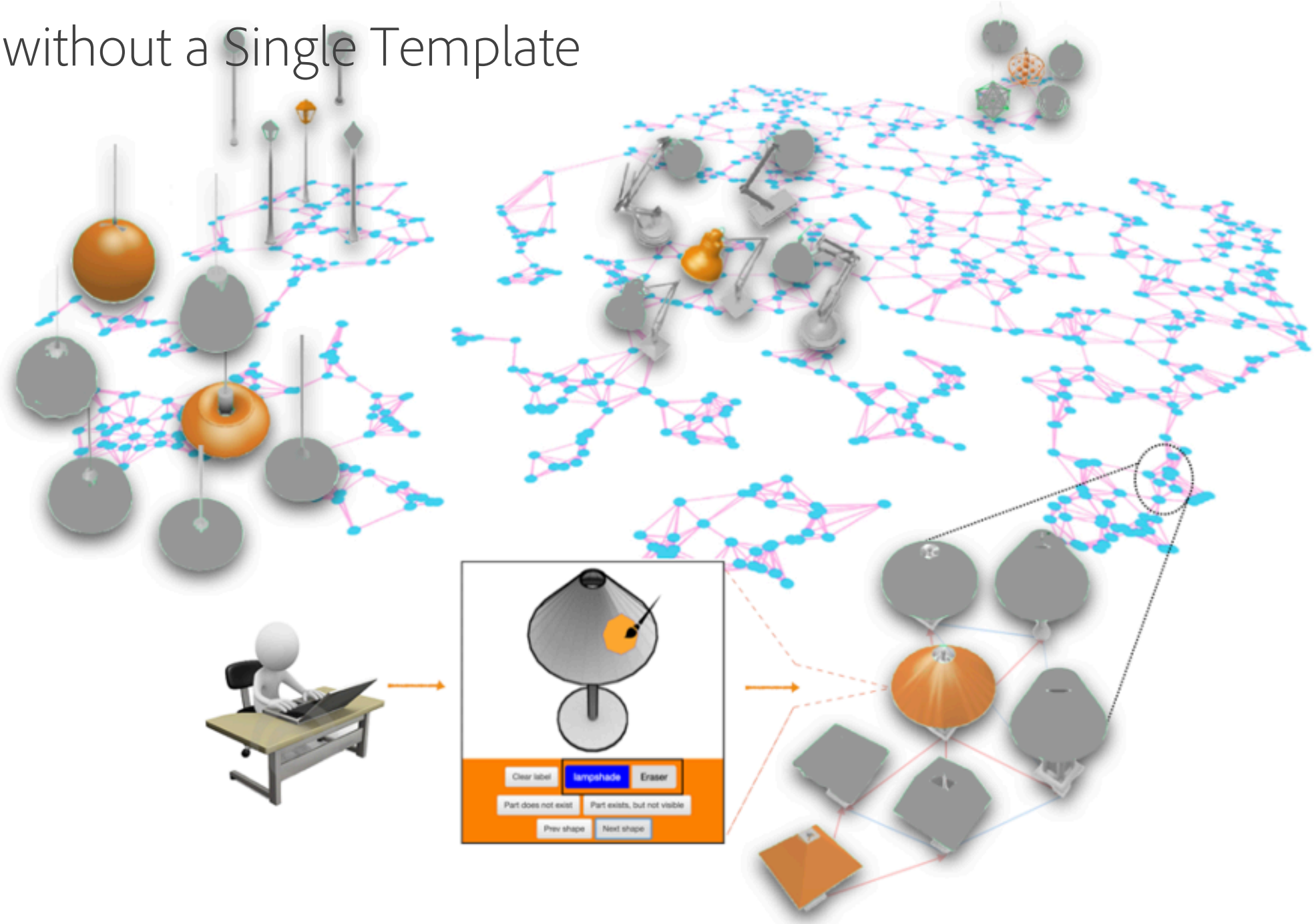
- Template Fitting and Correspondence Estimation via a Deforming Neural Networks
- Template-less Modeling via Deforming Neural Networks
- Template-less Signal Transfer via Deforming Neural Networks
- Multi-view Reconstruction via Deforming Neural Networks

Thibault Groueix

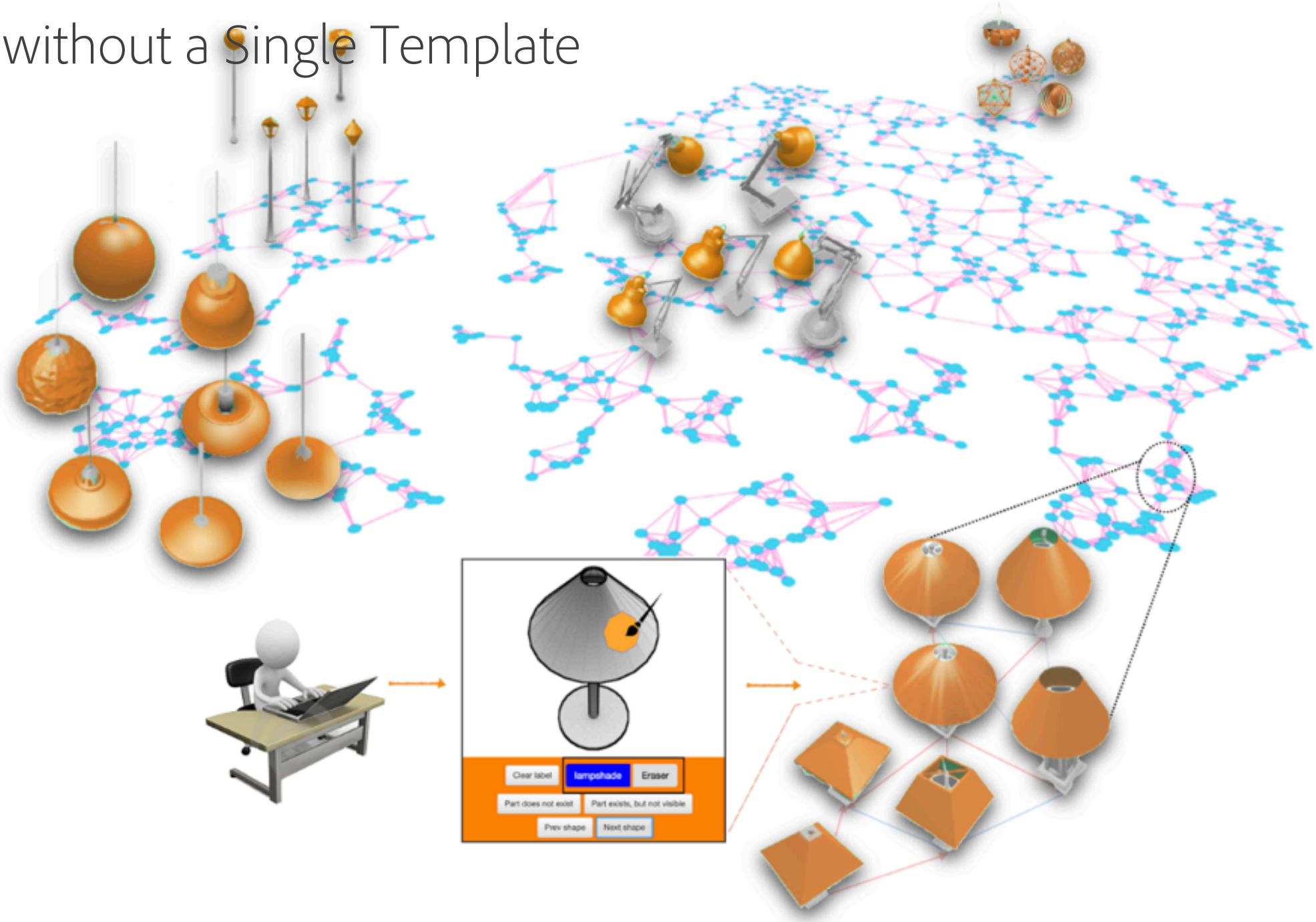
Signal Transfer without a Single Template



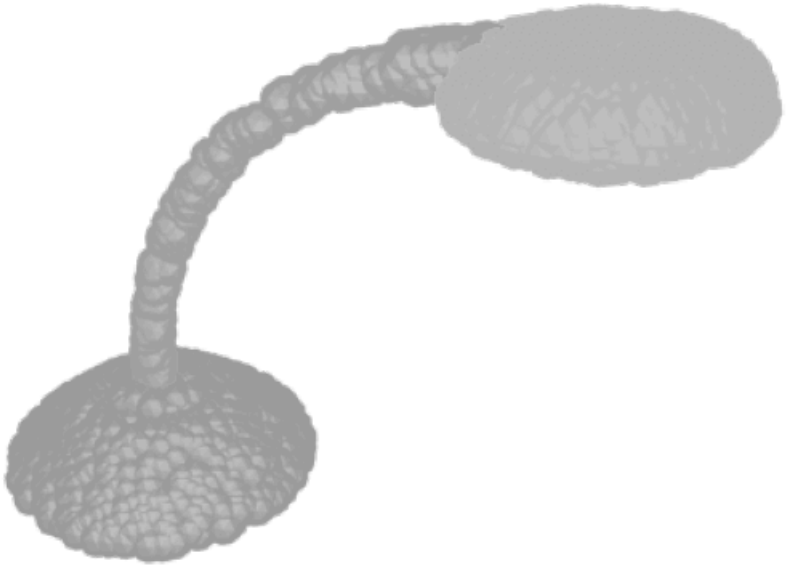
Signal Transfer without a Single Template



Signal Transfer without a Single Template



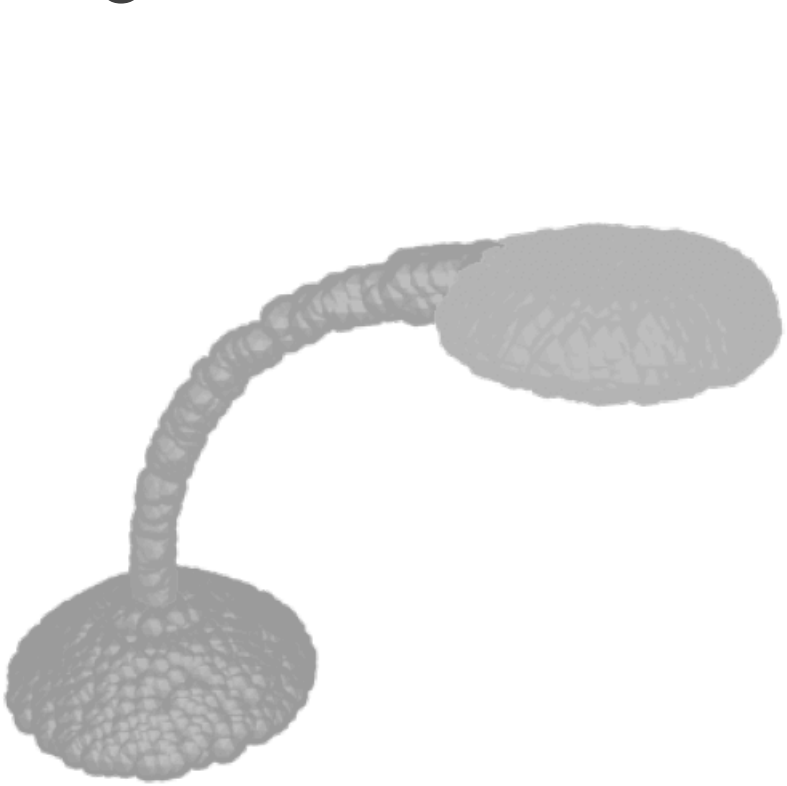
Signal Transfer from a Labeled Source



Input Shape

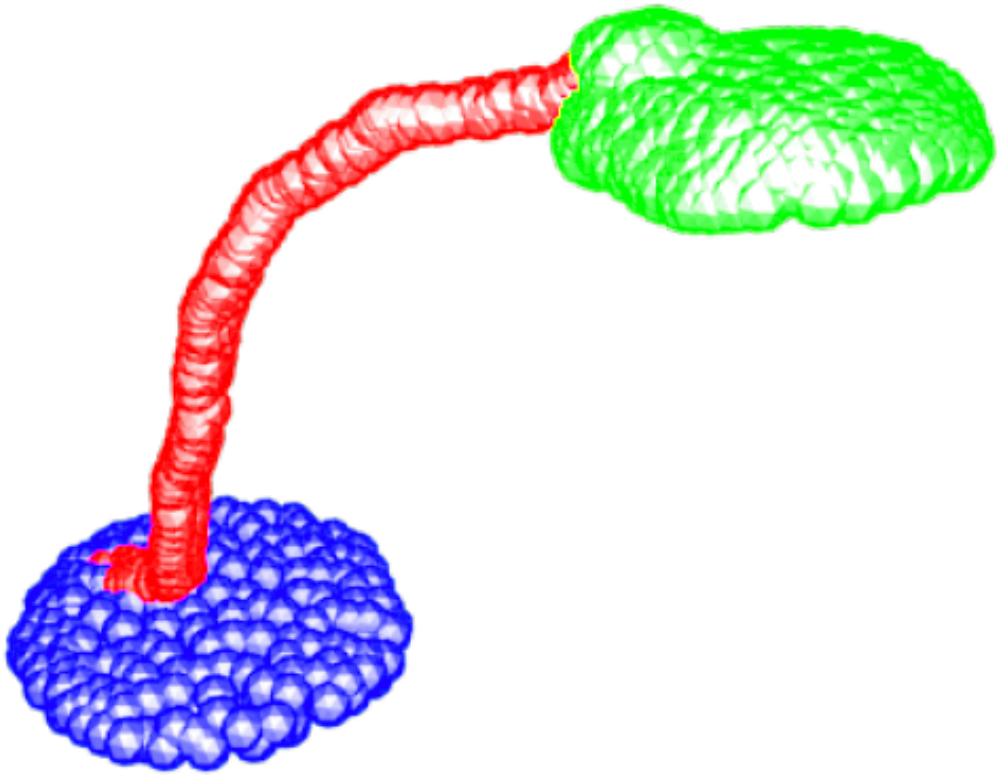
Labeled Source

Signal Transfer from a Labeled Source



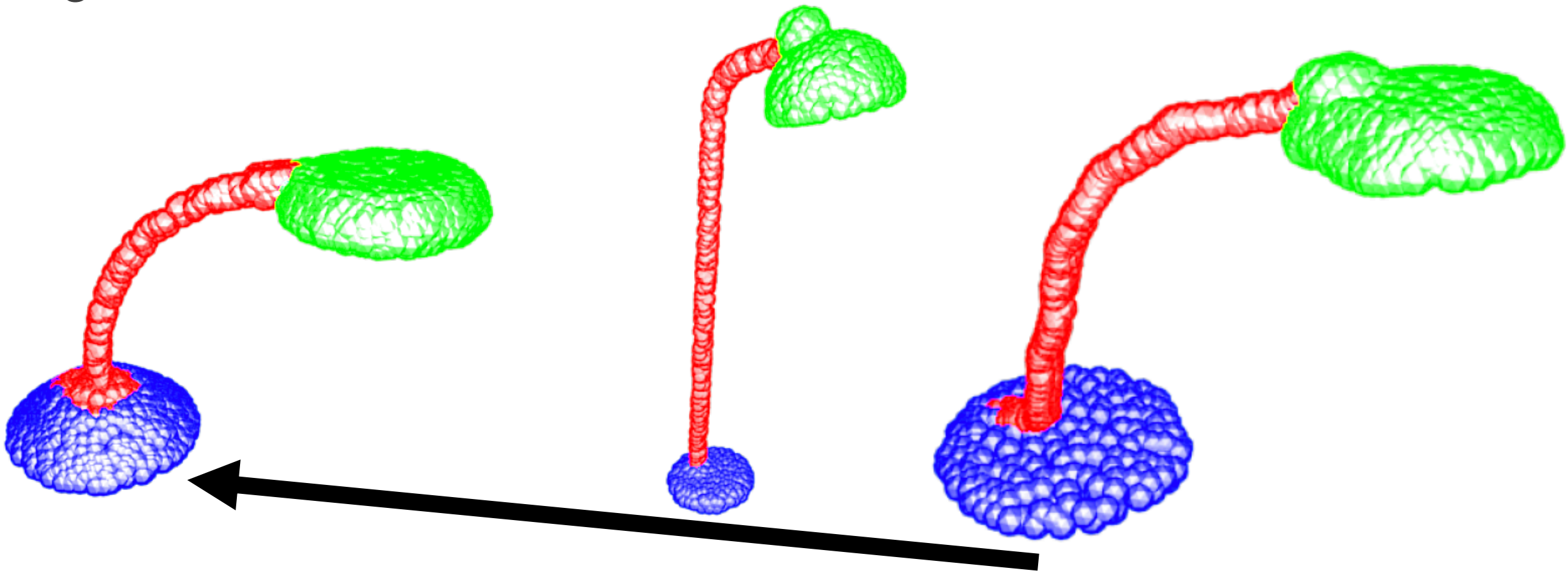
Input Shape

Labeled Source



Deformed Source

Signal Transfer from a Labeled Source

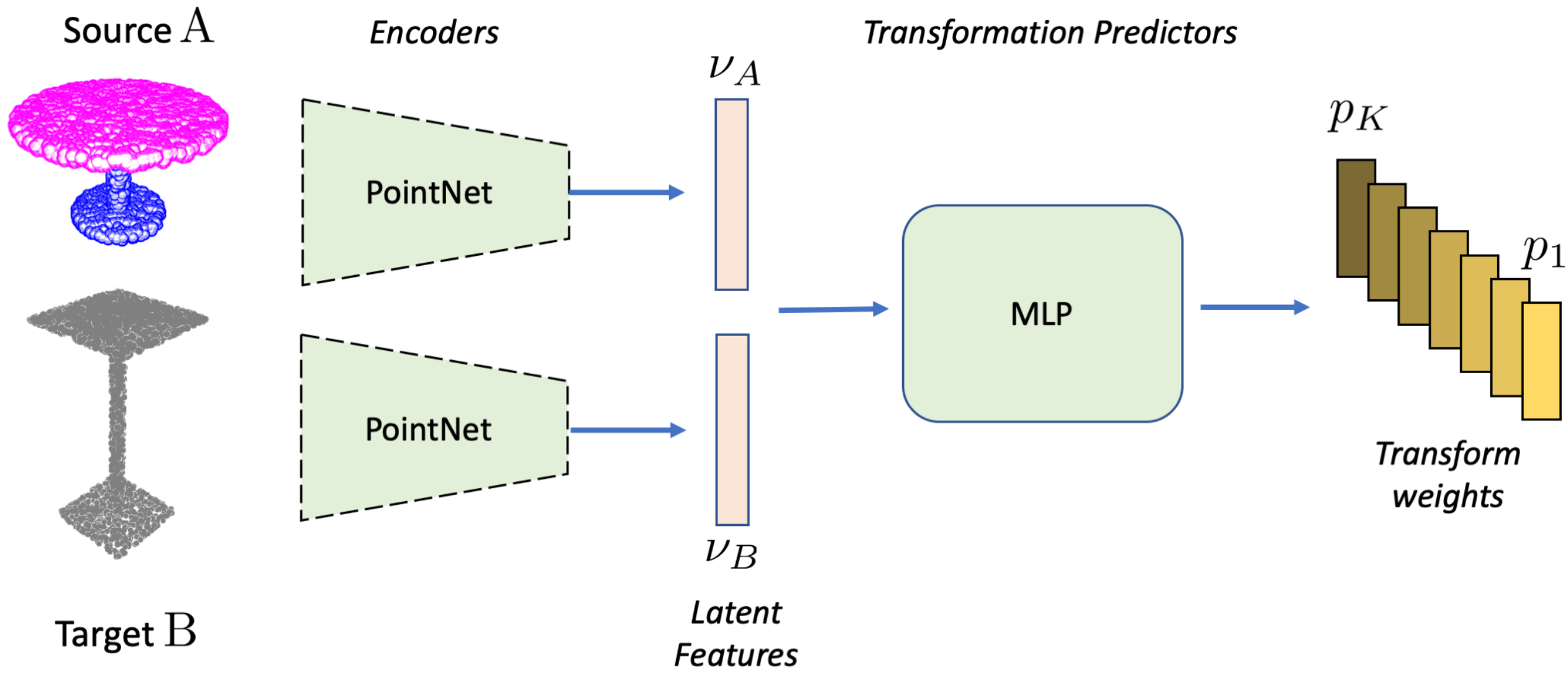


Input Shape

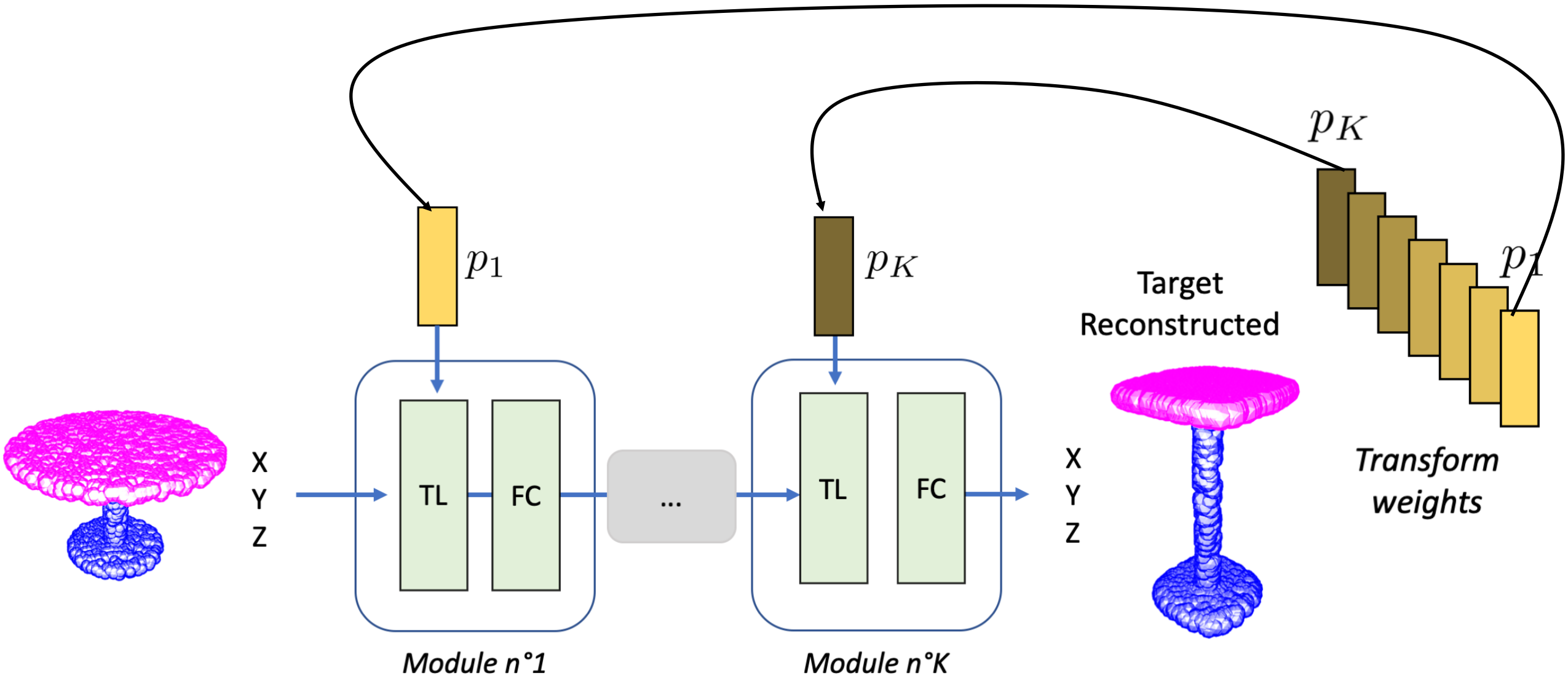
Labeled Source

Deformed Source

Template-less Deformation Network

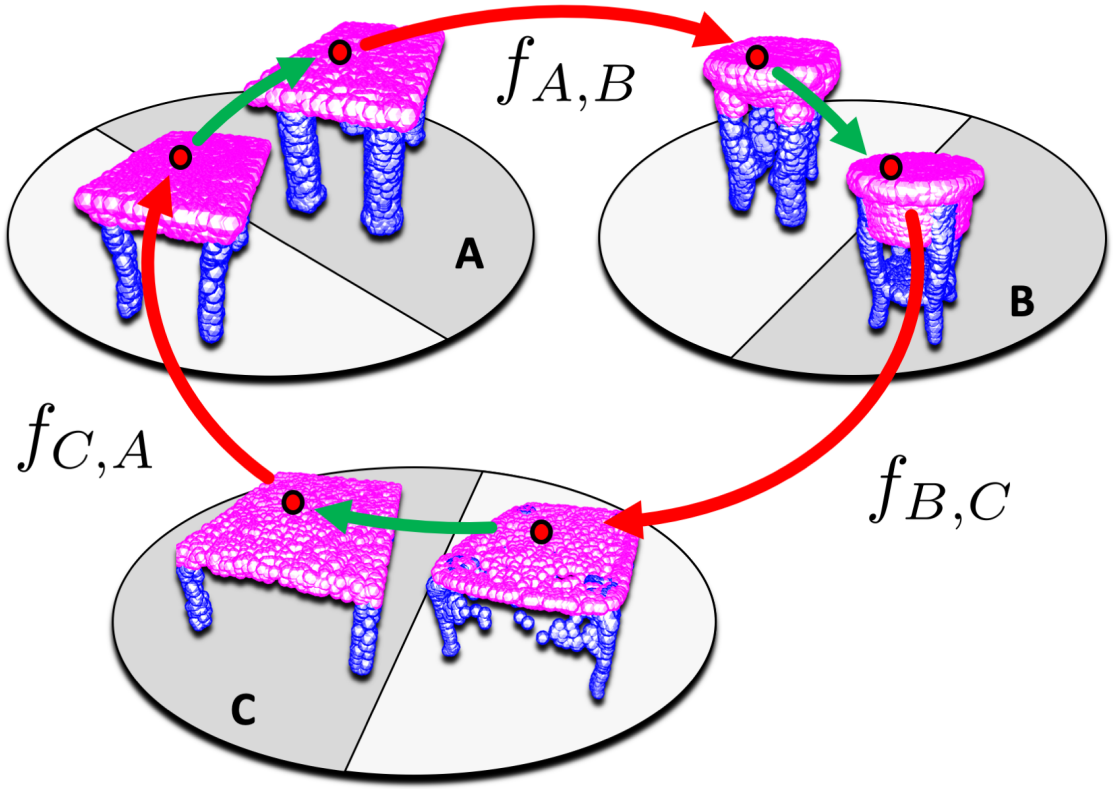


Template-less Deformation Network



Cycle-consistent Deformation Network

- Self-reconstruction loss
- Asymmetric Chamfer loss
- Cycle-consistency loss

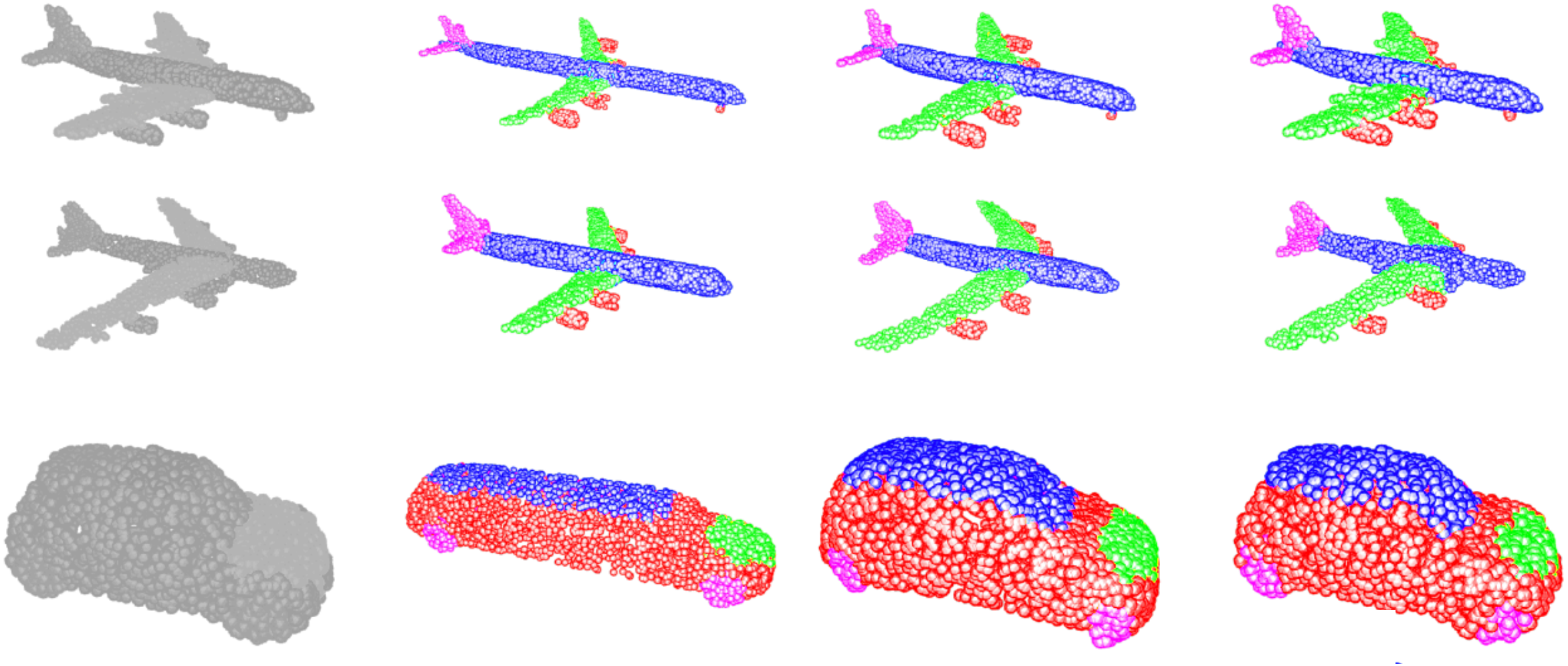


$$Cy_2(X,Y) = \frac{1}{|X|}\sum_{\mathbf{p}\in X}\left|\mathbf{p} - f_{Y,X}\circ\pi_Y\circ f_{X,Y}(\mathbf{p})\right|_2$$

$$Cy_3(X,Y,Z) = \frac{1}{|X|}\sum_{\mathbf{p}\in X}\left|\mathbf{p} - f_{Z,X}\circ\pi_Z\circ f_{Y,Z}\circ\pi_Y\circ f_{X,Y}(\mathbf{p})\right|_2$$

Segmentation Transfer

- Segmentation accuracy with only 10 labeled examples



Segmentation Transfer

- Segmentation accuracy with only 10 labeled examples

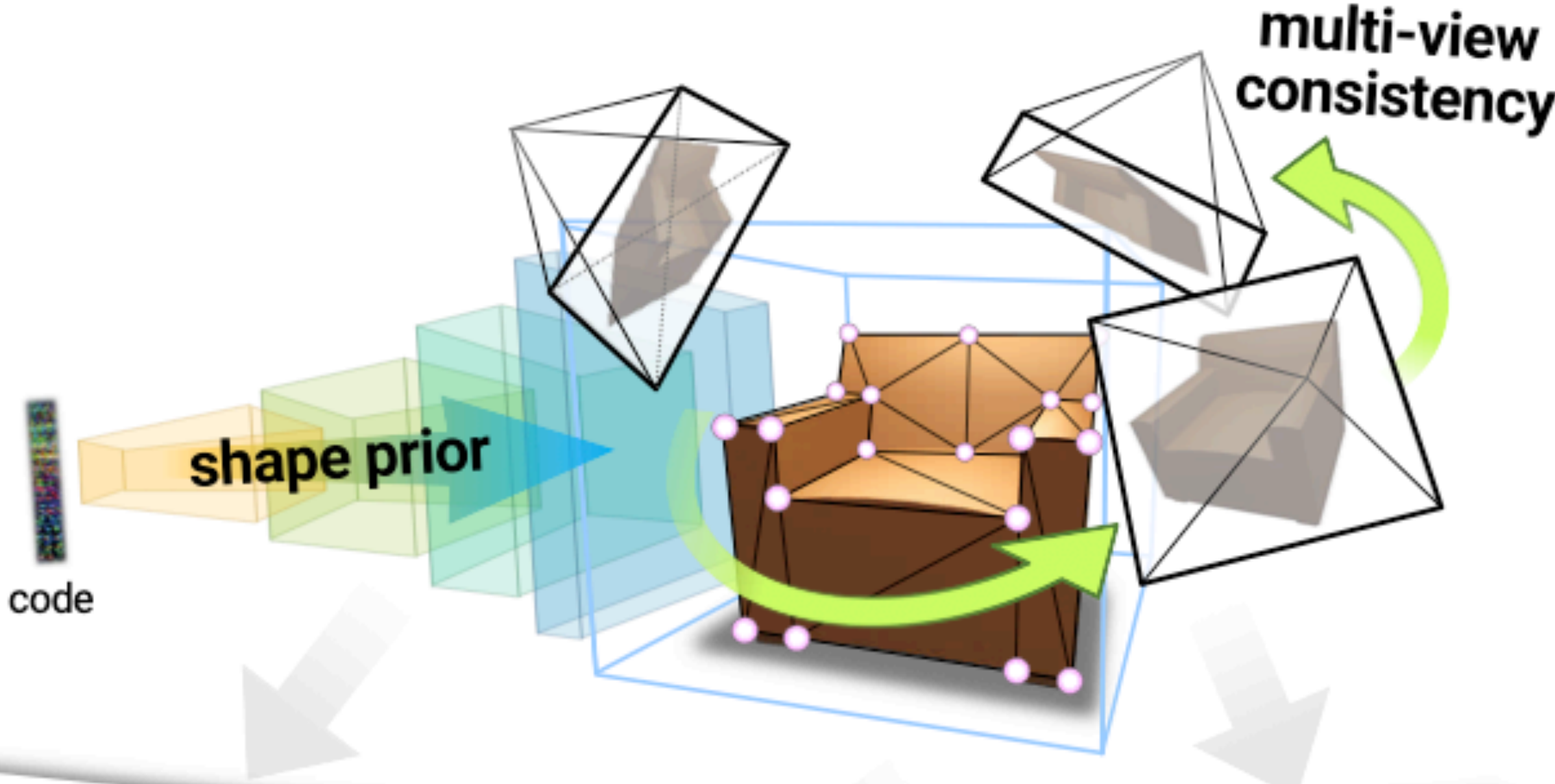
10 shots	Airplane	Car	Chair	Lamp	Table
(a) Pointnet	14.0 ± 8.0	11.7 ± 10.4	21.1 ± 13.1	26.0 ± 13.2	43.5 ± 15.5
(b) Atlasnet Patch	62.6 ± 2.4	52.3 ± 9.1	72.1 ± 1.2	62.8 ± 2.2	61.6 ± 3.7
(c) Atlasnet Sphere	62.2 ± 2.2	52.9 ± 9.1	70.2 ± 1.2	59.3 ± 1.8	60.0 ± 5.1
(d) ICP	65.5 ± 3.1	61.3 ± 1.1	75.8 ± 1.2	64.8 ± 5.0	64.9 ± 3.9
(e) Ours	$\mathbf{67.1 \pm 2.9}$	$\mathbf{61.4 \pm 1.1}$	$\mathbf{78.9 \pm 1.1}$	$\mathbf{65.8 \pm 5.2}$	$\mathbf{66.1 \pm 4.5}$

Talk Outline

- Template Fitting and Correspondence Estimation via a Deforming Neural Networks
- Template-less Modeling via Deforming Neural Networks
- Template-less Signal Transfer via Deforming Neural Networks
- Multi-view Reconstruction via Deforming Neural Networks

Chen-Hsuan Lin

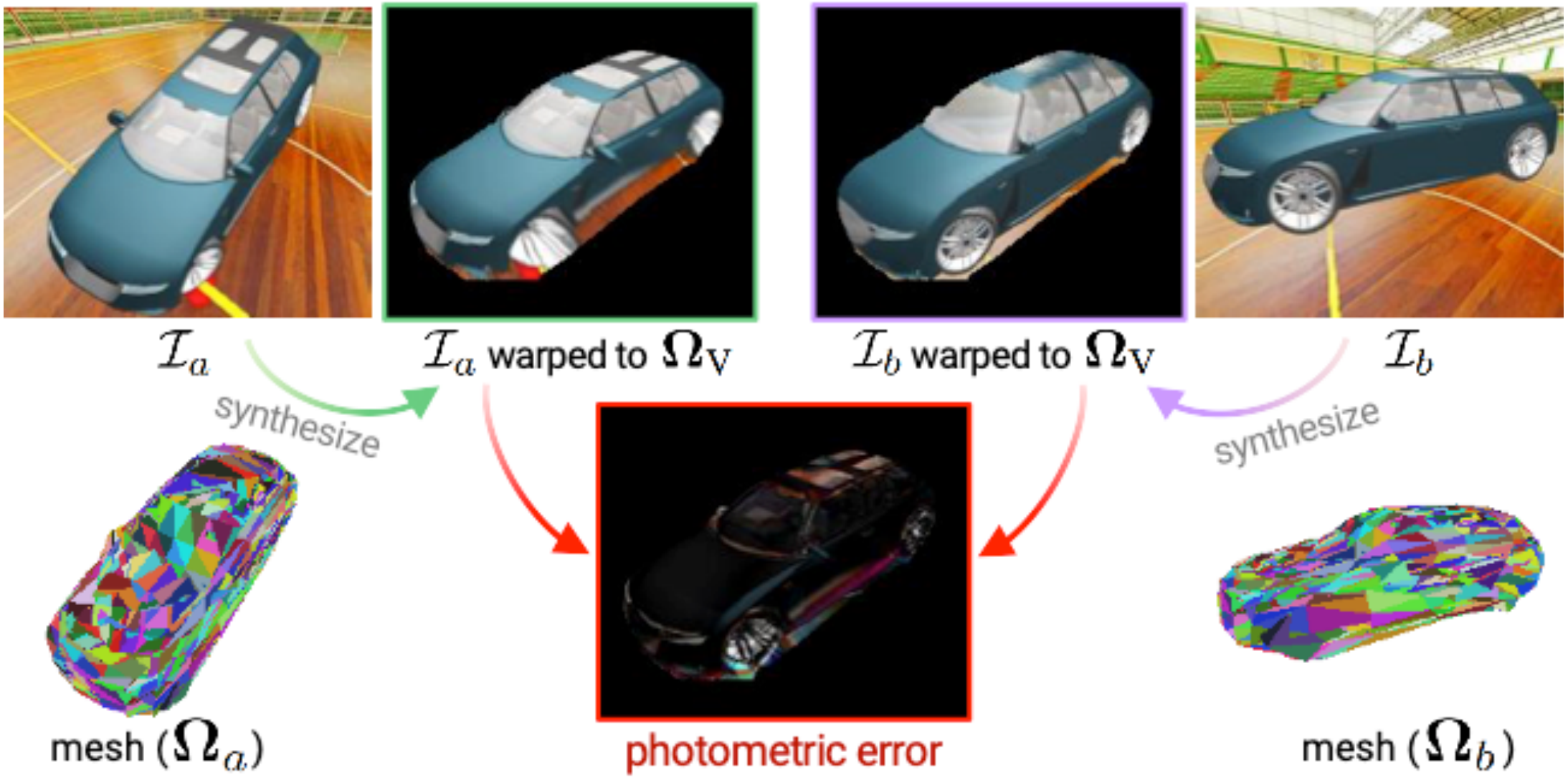
Video-based Shape Reconstruction



Video-based Shape Reconstruction

Loss Functions

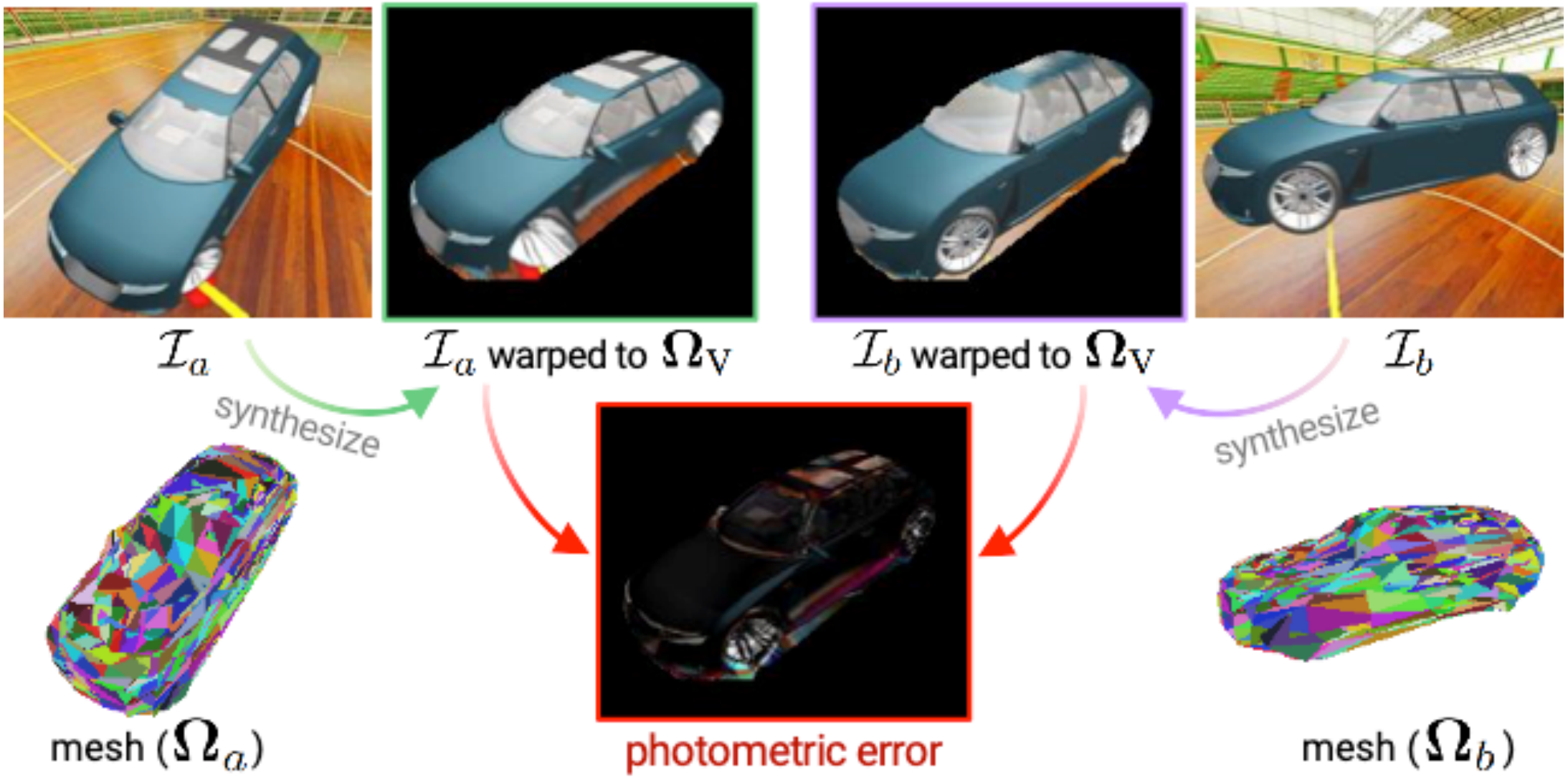
- Photometric Loss



Loss Functions

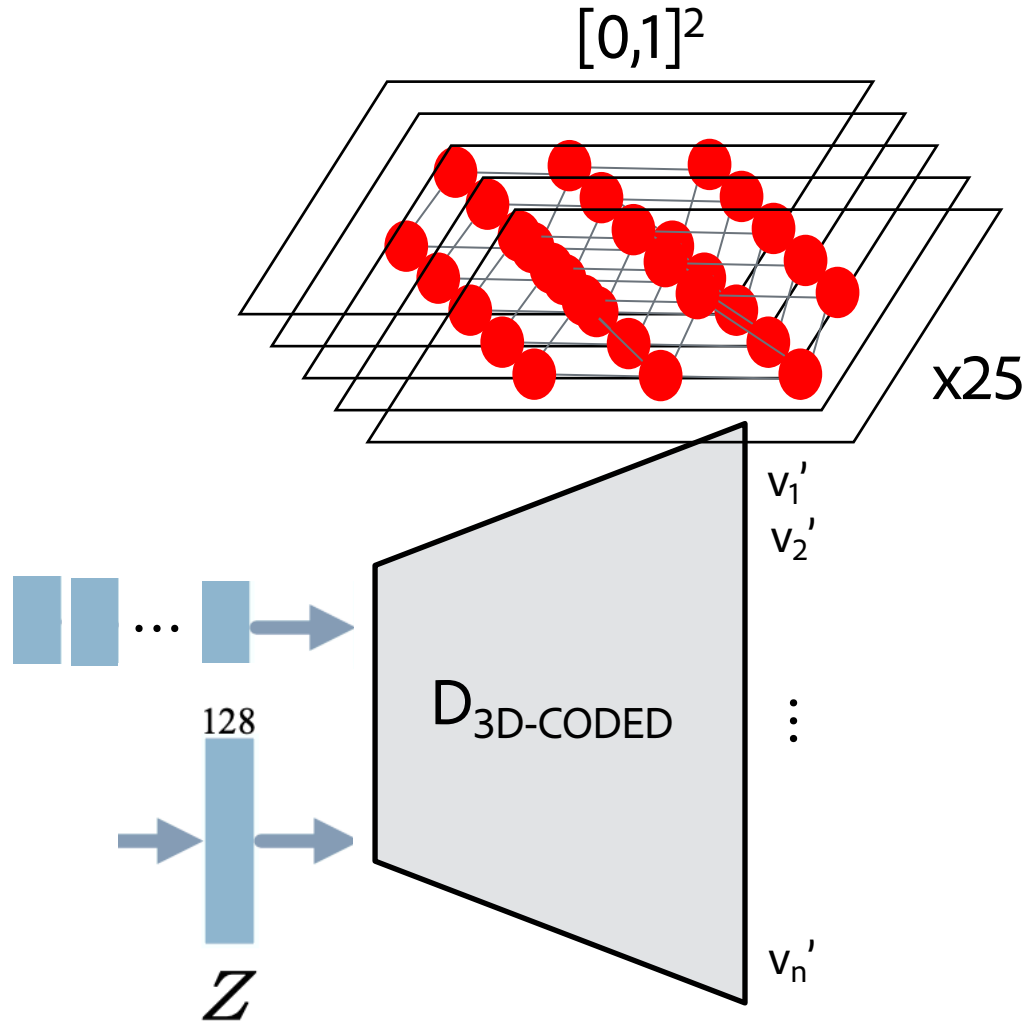
- Photometric Loss

$$\mathcal{L}_{\text{photo}}(\mathcal{I}_a, \mathcal{I}_b, \boldsymbol{\Omega}_a, \boldsymbol{\Omega}_b; \mathcal{G}(\mathbf{z}))$$
$$= \sum_j \sum_{i:\mathbf{p}_i \in \mathcal{P}_j} \left\| \mathcal{I}_a\left(\pi\left(\mathbf{p}_i(\mathbf{z}); \boldsymbol{\Omega}_a\right)\right) - \mathcal{I}_b\left(\pi\left(\mathbf{p}_i(\mathbf{z}); \boldsymbol{\Omega}_b\right)\right) \right\|_1$$



Loss Functions

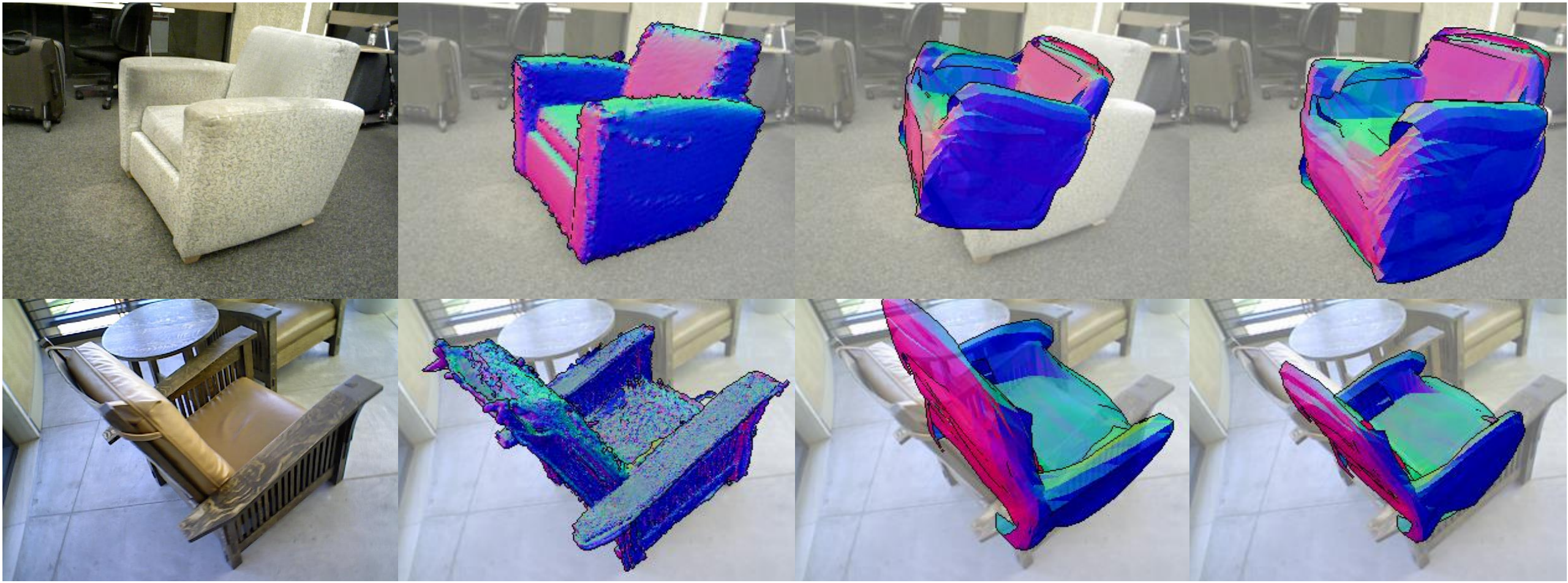
- AtasNet code loss



$$\mathcal{L}^{\text{reg}} = |z - z'|^2$$

Video-based Shape Reconstruction Results

RGB sequence | Traditional pipeline | AtlasNet (init.) | Ours



A Large Dataset of Object Scans, Choi et al. 2015

3D Challenges

- Representations
- Training with strong and weak supervision
- Hierarchical and multi-resolution approaches
- Materials
- Human-object interactions

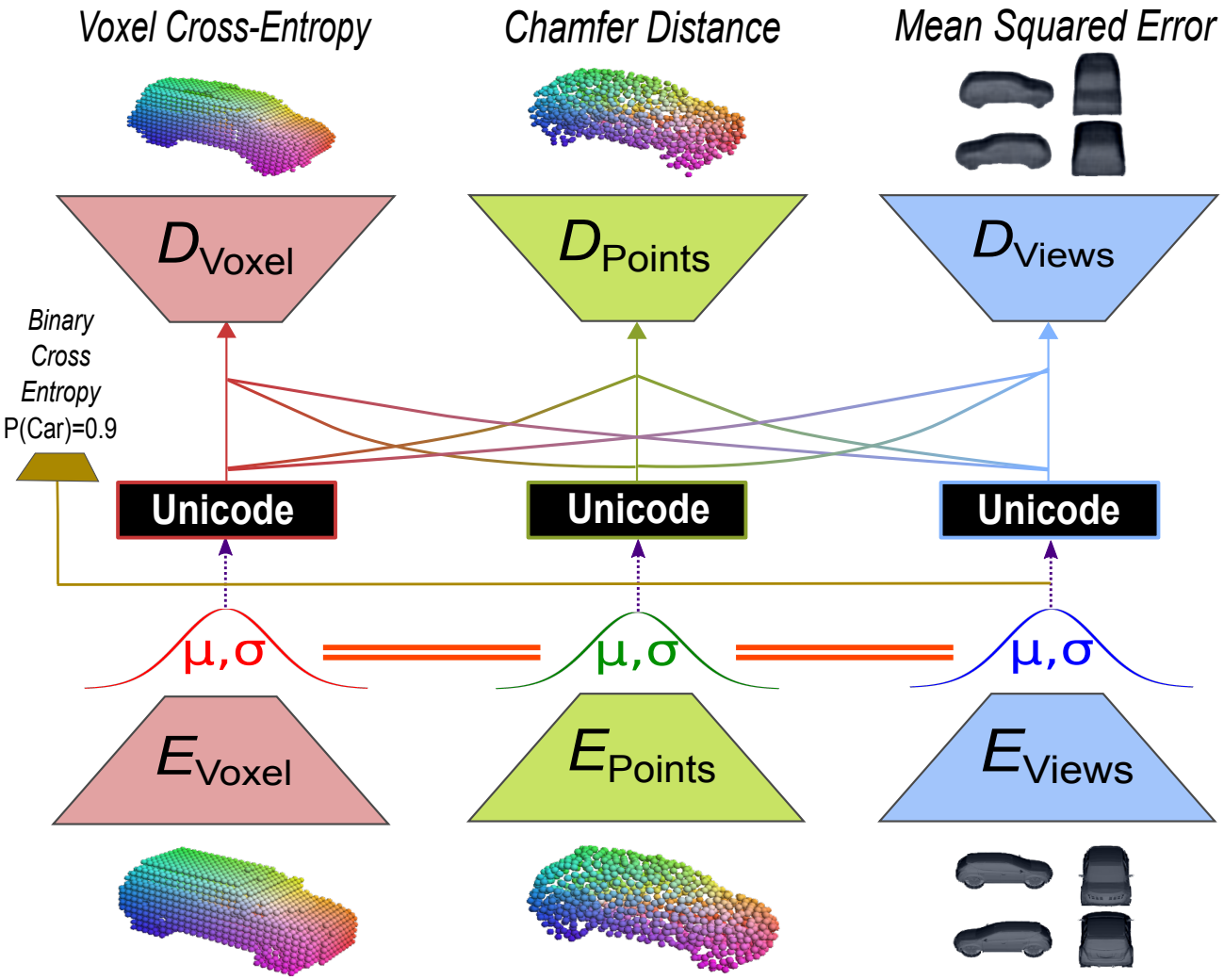
Unicode Representation

- Any encoder should be decoded with any decoder

Sanjeev Muralikrishnan

Unicode Representation

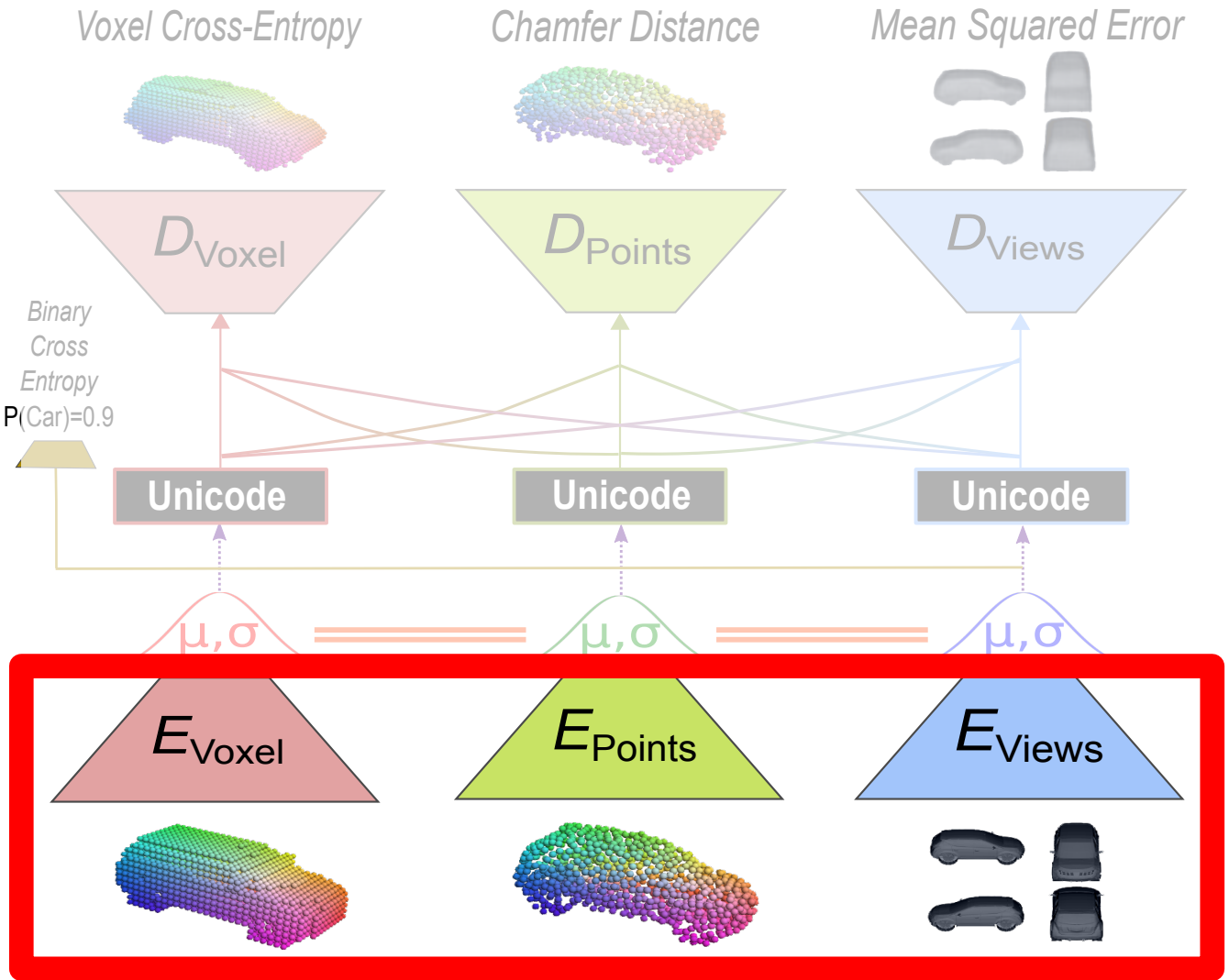
- Any encoder should be decoded with any decoder



[Muralikrishnan et al., CVPR 2019]

Unicode Representation

- Any encoder should be decoded with any decoder

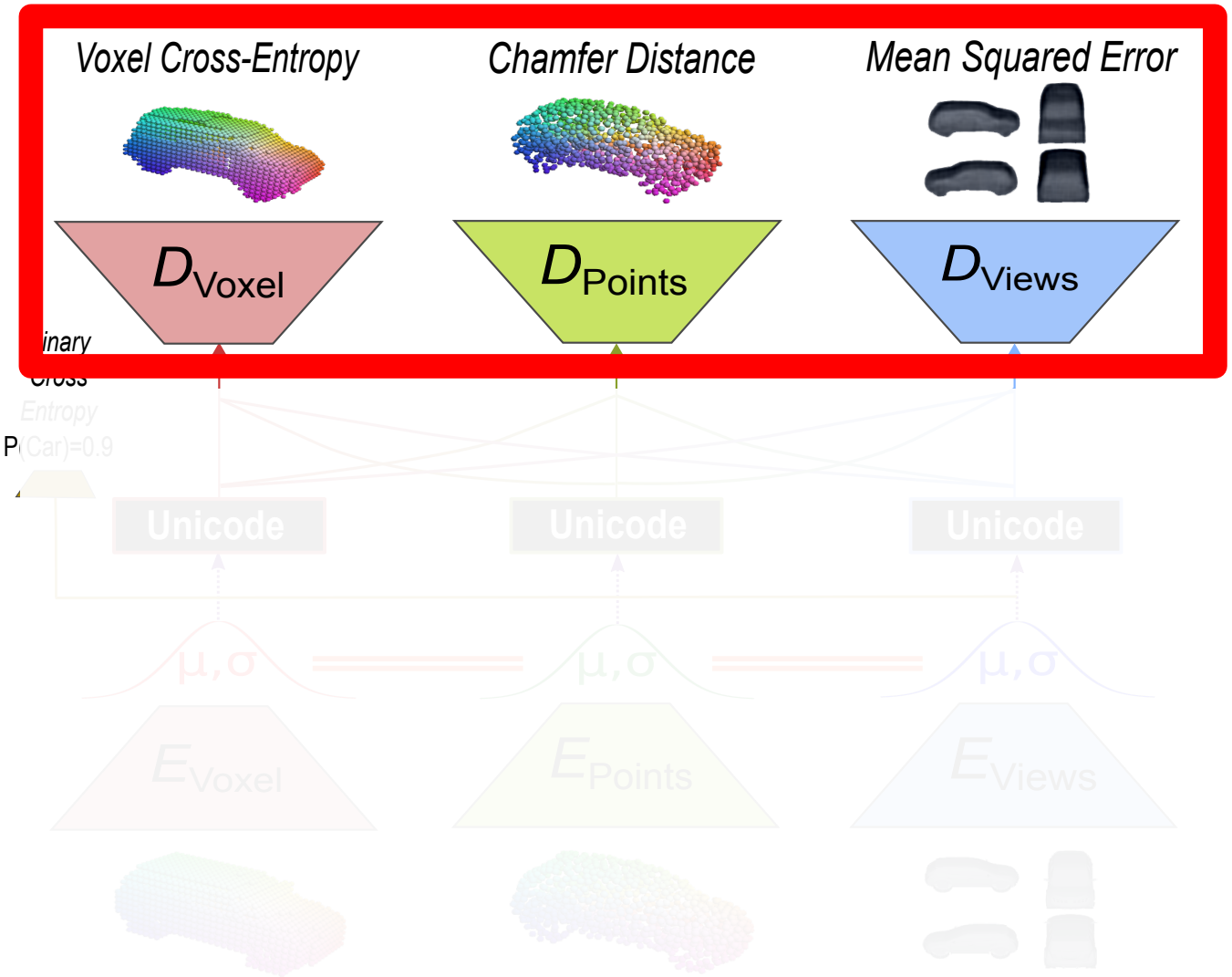


$E_x(S_x)$ - encoders

[Muralikrishnan et al., CVPR 2019]

Unicode Representation

- Any encoder should be decoded with any decoder



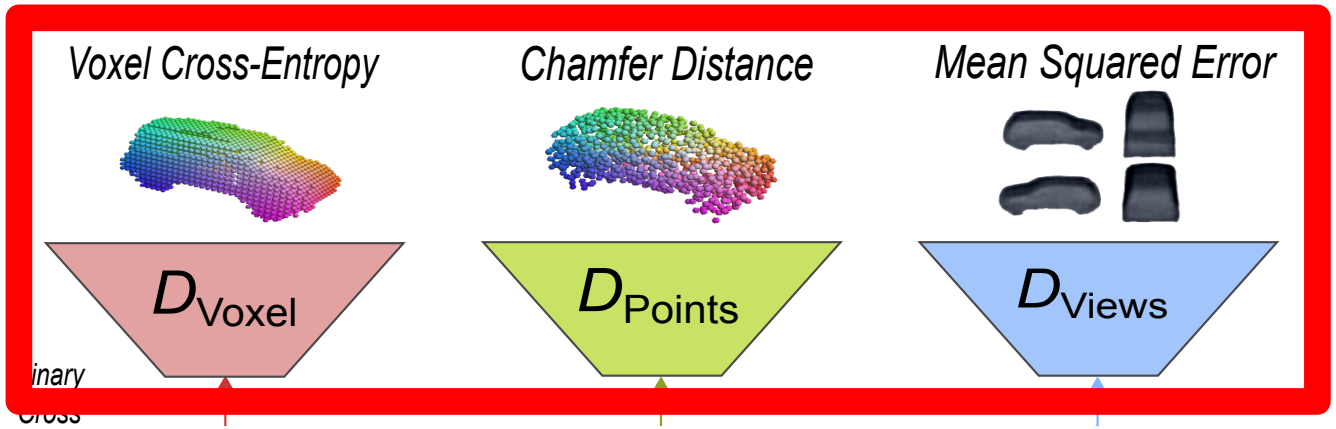
$E_x(S_x)$ - encoders

$D_y(\text{unicode})$ - decoders

[Muralikrishnan et al., CVPR 2019]

Unicode Representation

- Any encoder should be decoded with any decoder



$E_x(S_x)$ - encoders

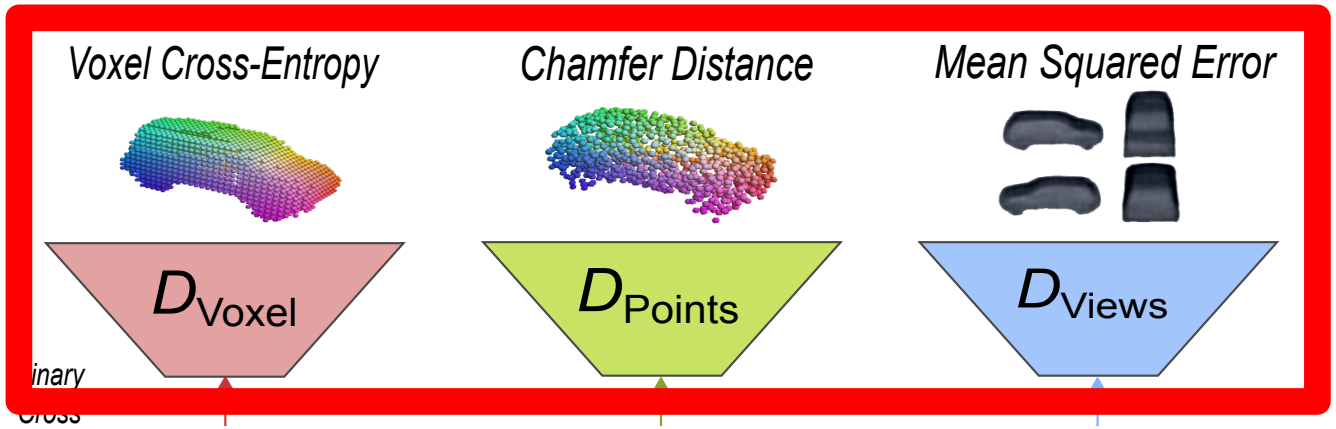
$D_y(\text{unicode})$ - decoders

$$\mathcal{L}^{\text{reconstruction}} = \sum_{x \in \mathcal{R}} \sum_{y \in \mathcal{R}} w_{x \to y} \text{Dist}_y(D_y(E_x(S_x)), S_y),$$

[Muralikrishnan et al., CVPR 2019]

Unicode Representation

- Any encoder should be decoded with any decoder



$E_x(S_x)$ - encoders

$D_y(\text{unicode})$ - decoders

$$\mathcal{L}^{\text{reconstruction}} = \sum_{x \in \mathcal{R}} \sum_{y \in \mathcal{R}} w_{x \to y} \text{Dist}_y(D_y(E_x(S_x)), S_y),$$

$\text{Dist}_{\text{Voxel}}$: Per voxel cross-entropy

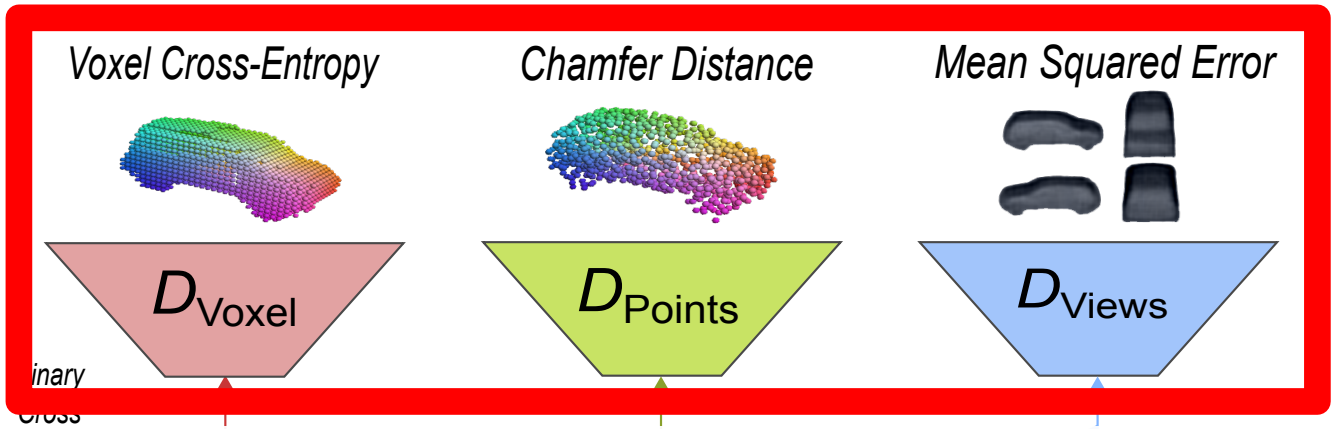
$\text{Dist}_{\text{Points}}$: Chamfer distance

$\text{Dist}_{\text{Views}}$: L_2

[Muralikrishnan et al., CVPR 2019]

Unicode Representation

- Any encoder should be decoded with any decoder



$E_x(S_x)$ - encoders

$D_y(\text{unicode})$ - decoders

Normalized in mini batches

$$\mathcal{L}^{\text{reconstruction}} = \sum_{x \in \mathcal{R}} \sum_{y \in \mathcal{R}} w_{x \to y} \text{Dist}_y(D_y(E_x(S_x)), S_y),$$

$\text{Dist}_{\text{Voxel}}$: Per voxel cross-entropy

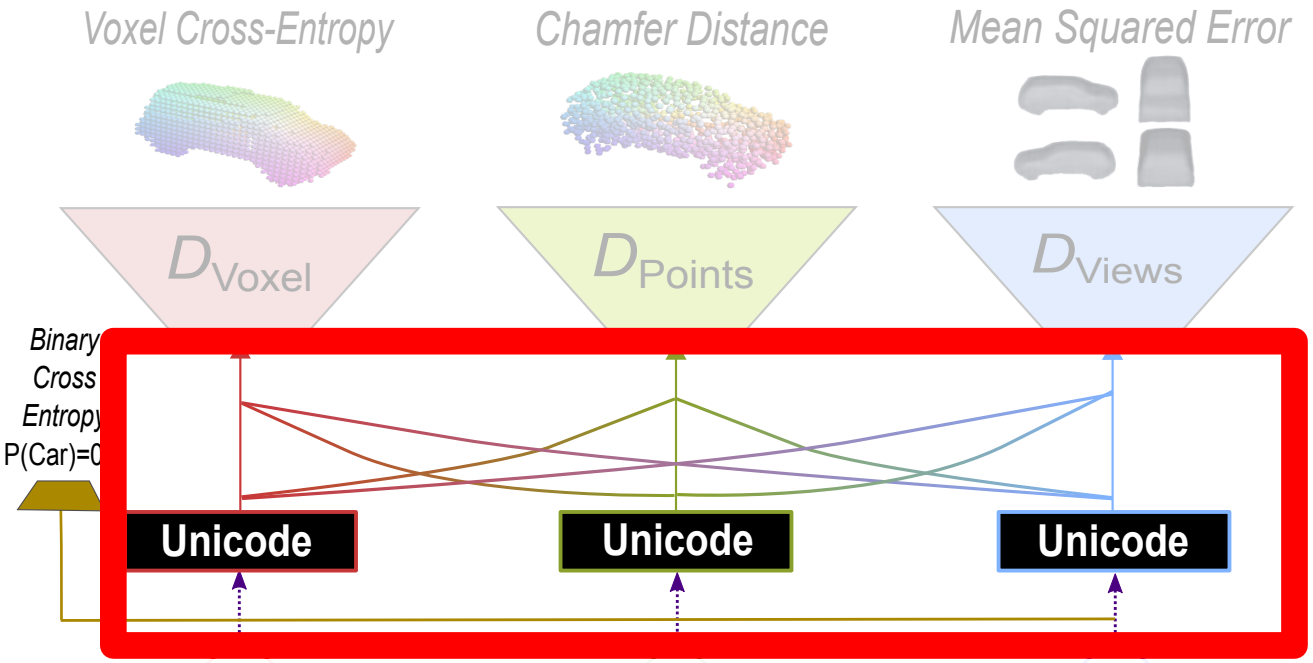
$\text{Dist}_{\text{Points}}$: Chamfer distance

$\text{Dist}_{\text{Views}}$: L_2

[Muralikrishnan et al., CVPR 2019]

Unicode Representation

- Any encoder should be decoded with any decoder

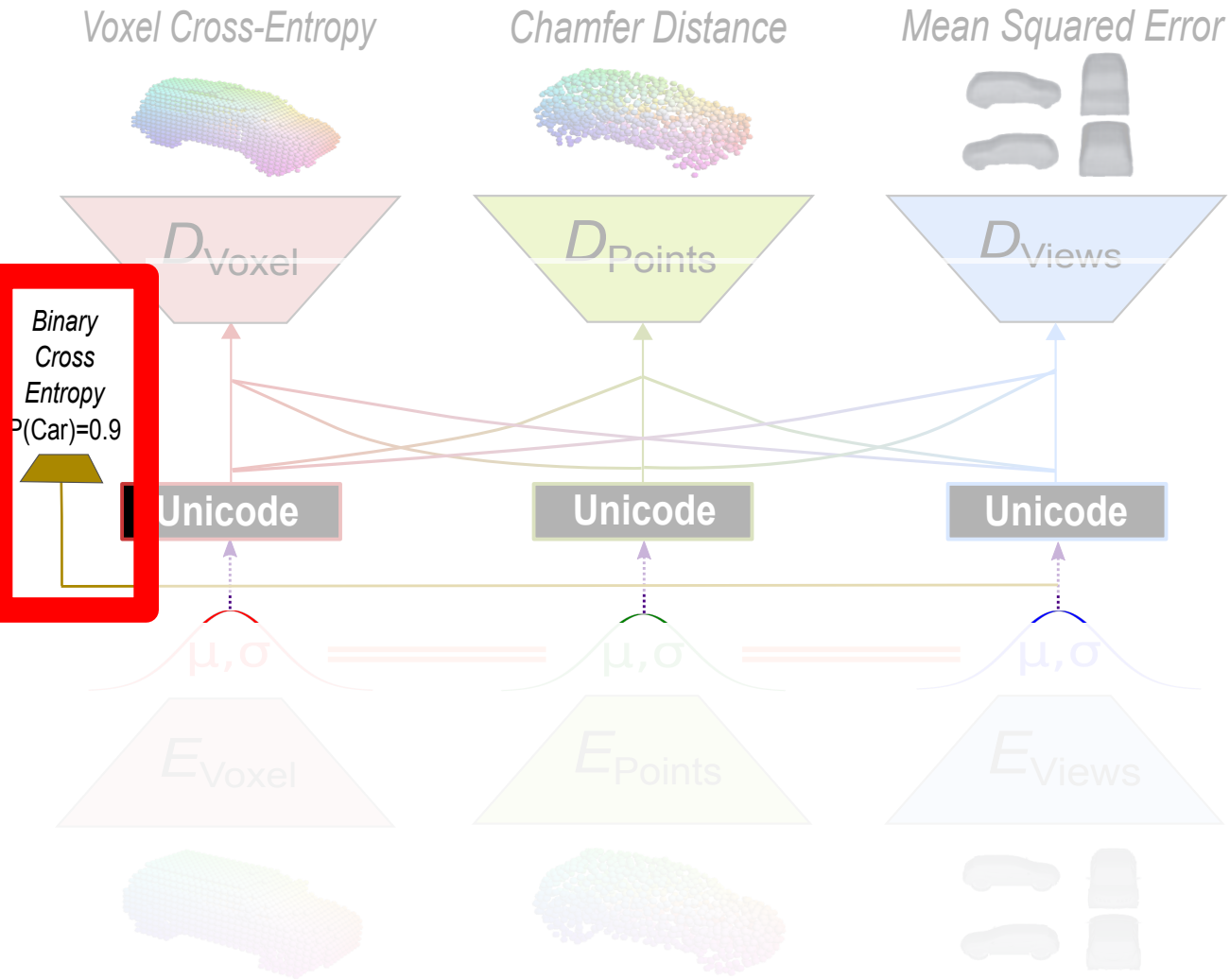


$$\mathcal{L}^{\text{embedding}} = \sum_{x \in \mathcal{R}} \sum_{y \in (\mathcal{R} \setminus x)} | E_x - E_y |_1$$

[Muralikrishnan et al., CVPR 2019]

Unicode Representation

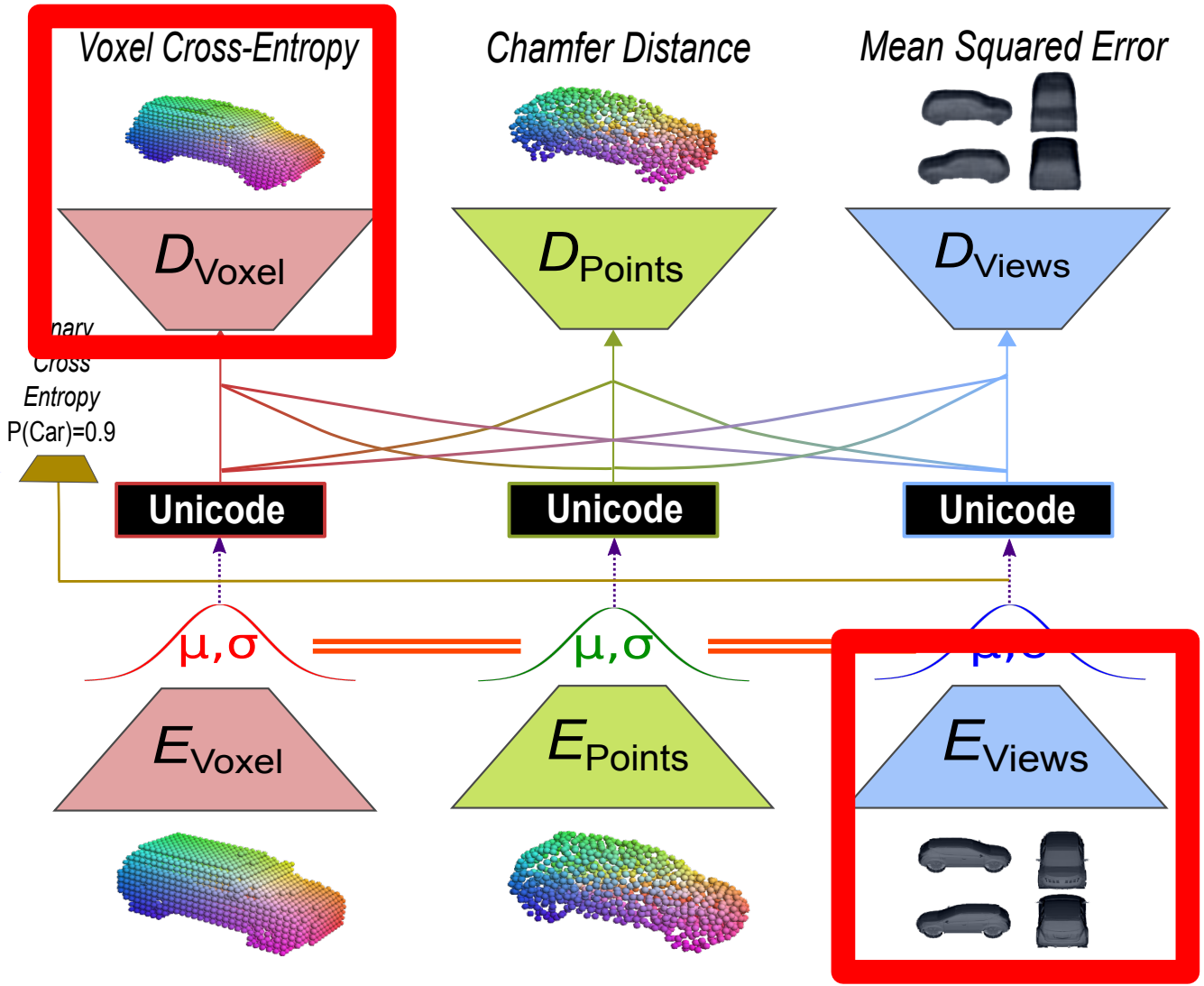
- Any encoder should be decoded with any decoder



[Muralikrishnan et al., CVPR 2019]

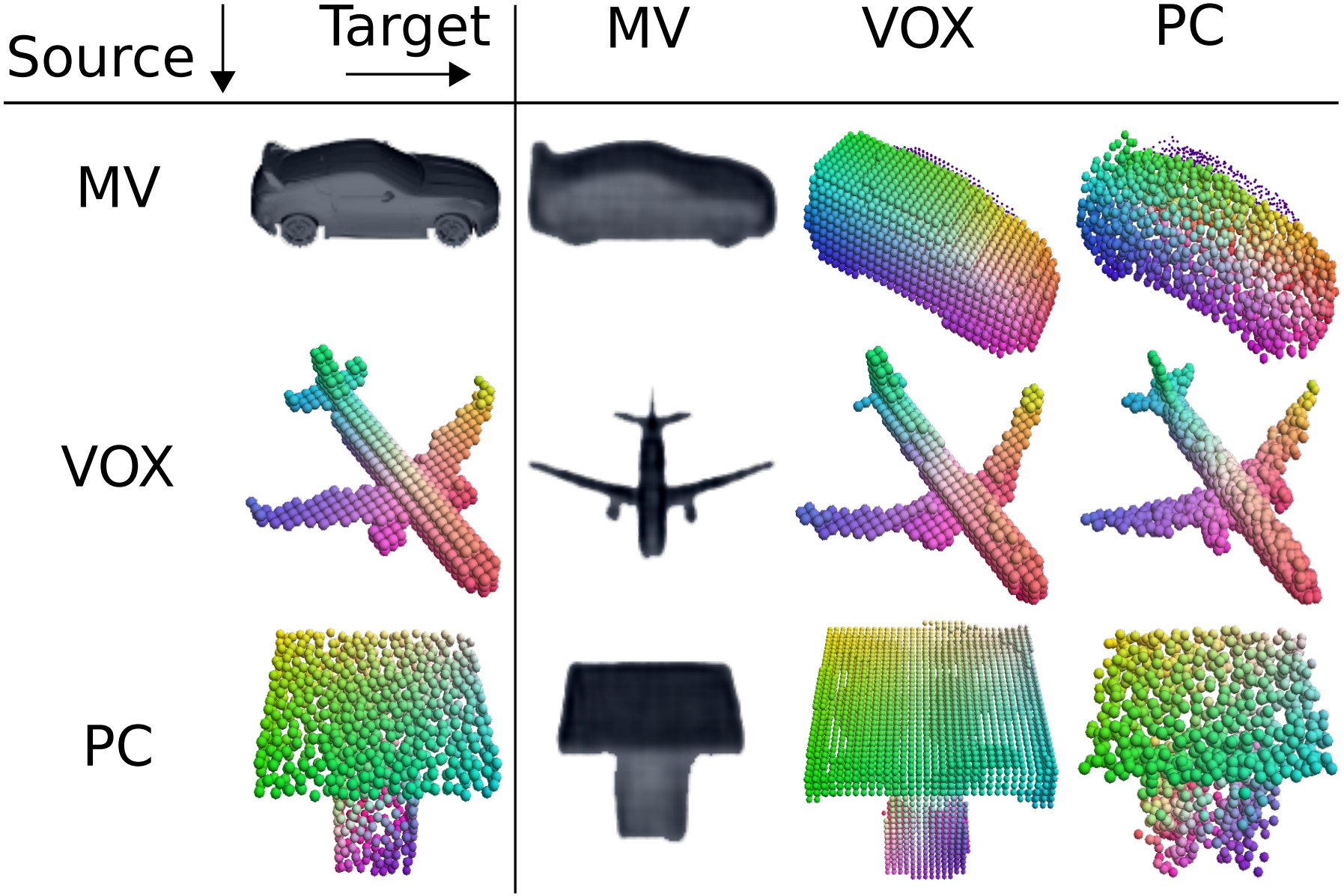
Unicode Representation

- Any encoder should be decoded with any decoder



[Muralikrishnan et al., CVPR 2019]

Translation Results



[Muralikrishnan et al., CVPR 2019]

Retrieval and Classification Results

- Only one representation is needed at test time
- Useful if training data is not in uniform format

Representation	Solo Autoencoder	Unicode Autoencoder
Points	0.73	0.73
Voxels	0.81	0.71
Multi-view	0.72	0.72

F1 Retrieval Score

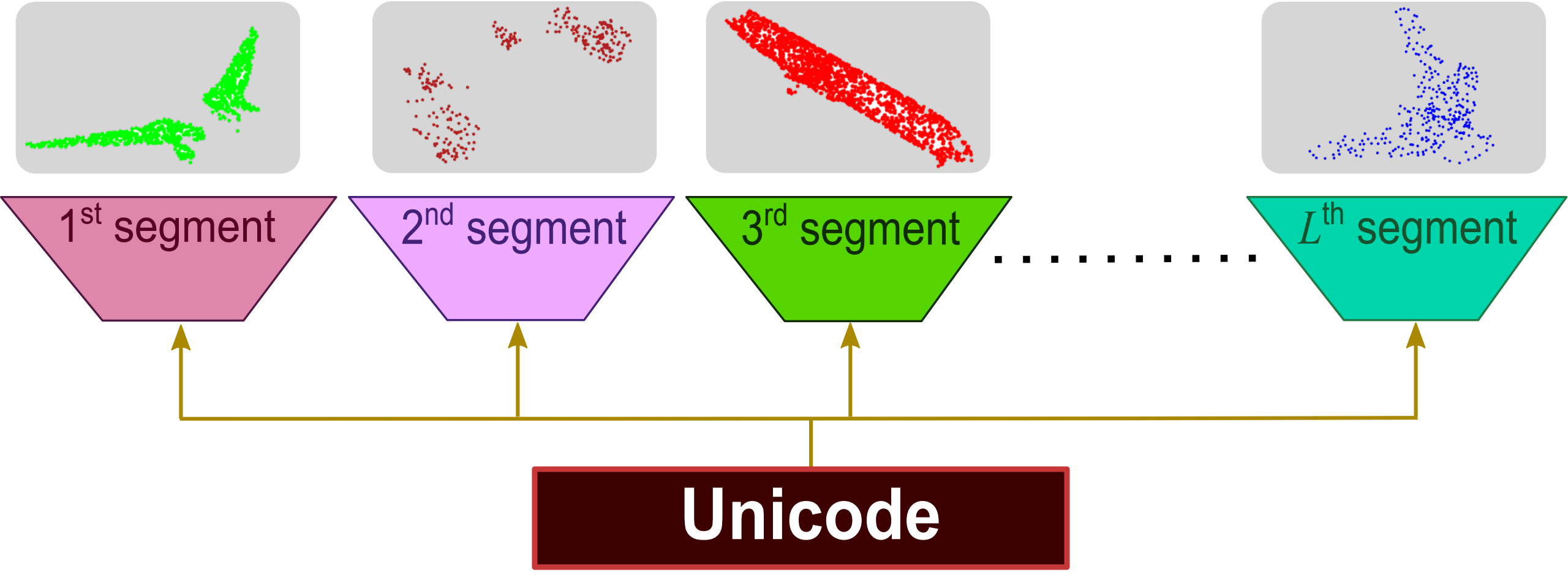
Representation	Solo Autoencoder	Unicode Autoencoder
Points	84.07	84.23
Voxels	80.76	82.48
Multi-view	83.58	83.38

Classification Accuracy

[Muralikrishnan et al., CVPR 2019]

Shape Segmentation

- Decode Segments



[Muralikrishnan et al., CVPR 2019]

Shape Segmentation

- Decode Segments

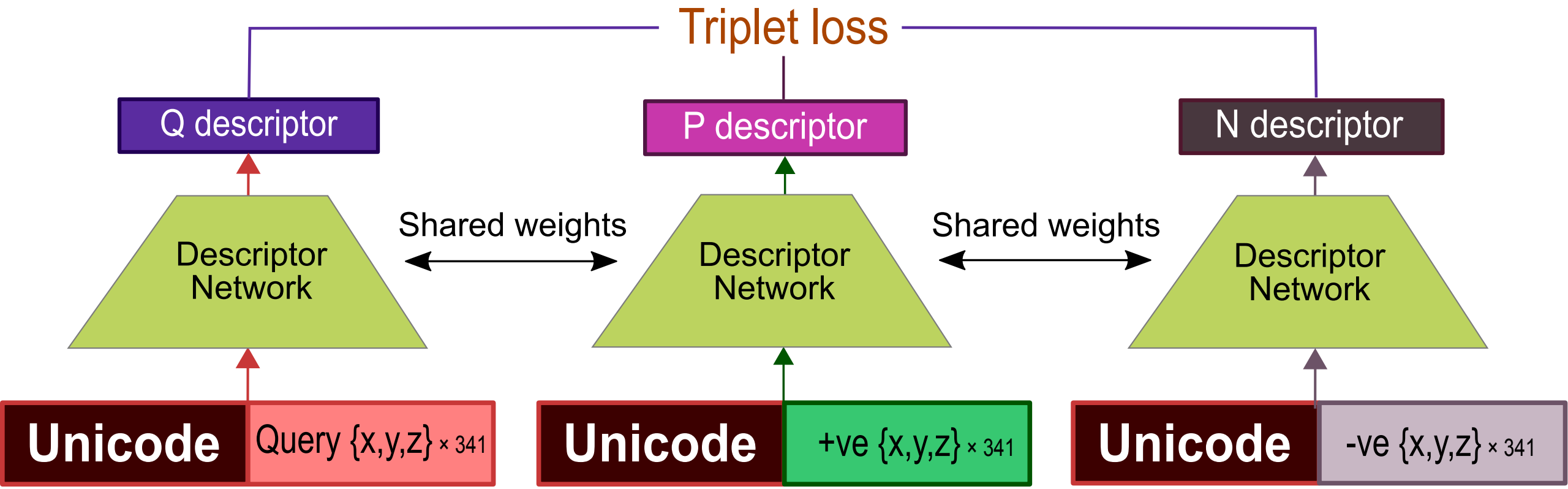
Representation	Solo Autoencoder	Unicode Autoencoder
Points	85.73	86.73
Voxels	85.05	87.26
Multi-view	-	86.79

Segmentation Accuracy

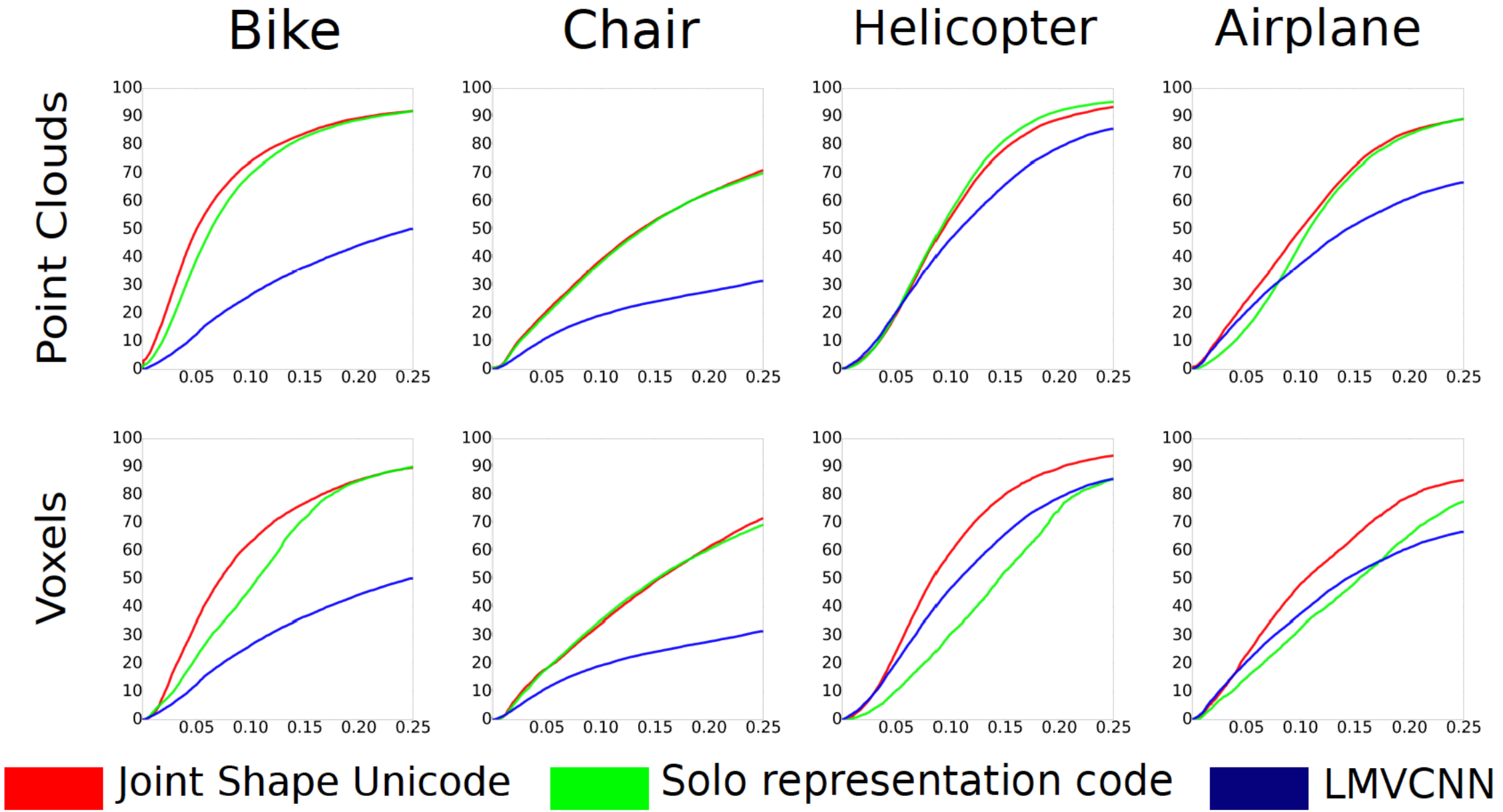
[Muralikrishnan et al., CVPR 2019]

Shape Correspondences

- Learning Descriptors



[Muralikrishnan et al., CVPR 2019]

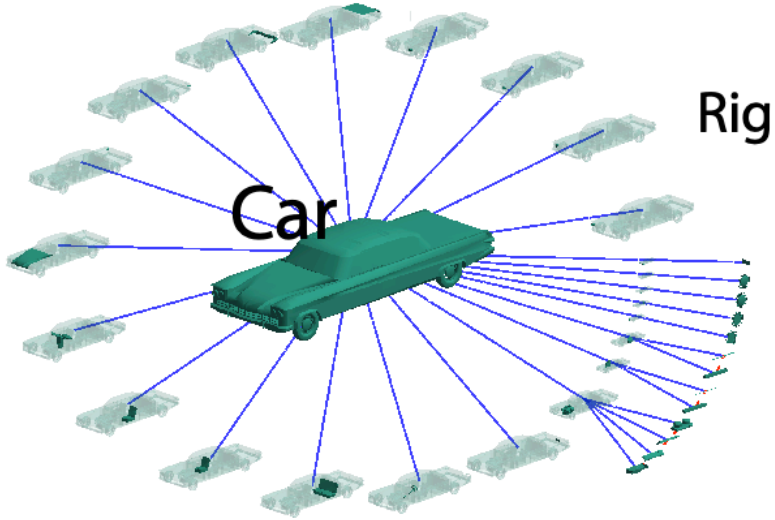


[Muralikrishnan et al., CVPR 2019]

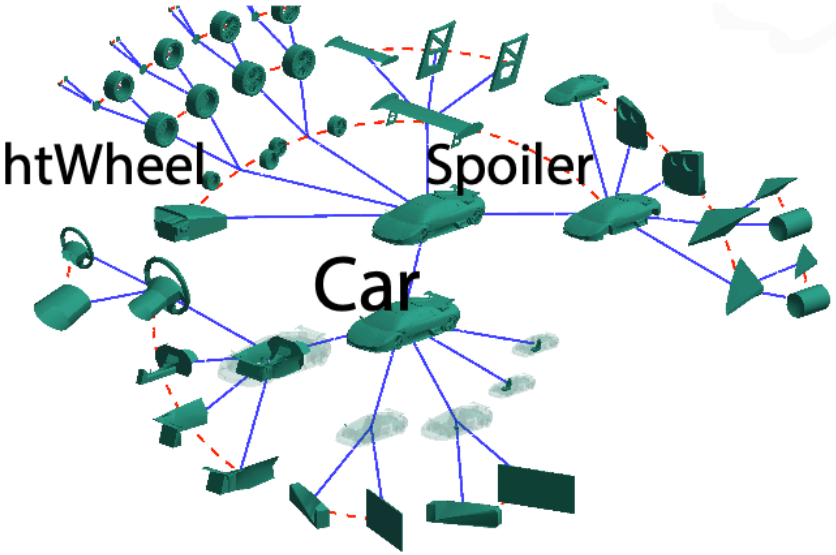
Training with Strong and Weak Supervision

Tags

[Muralikrishnan et al., CVPR 2018]

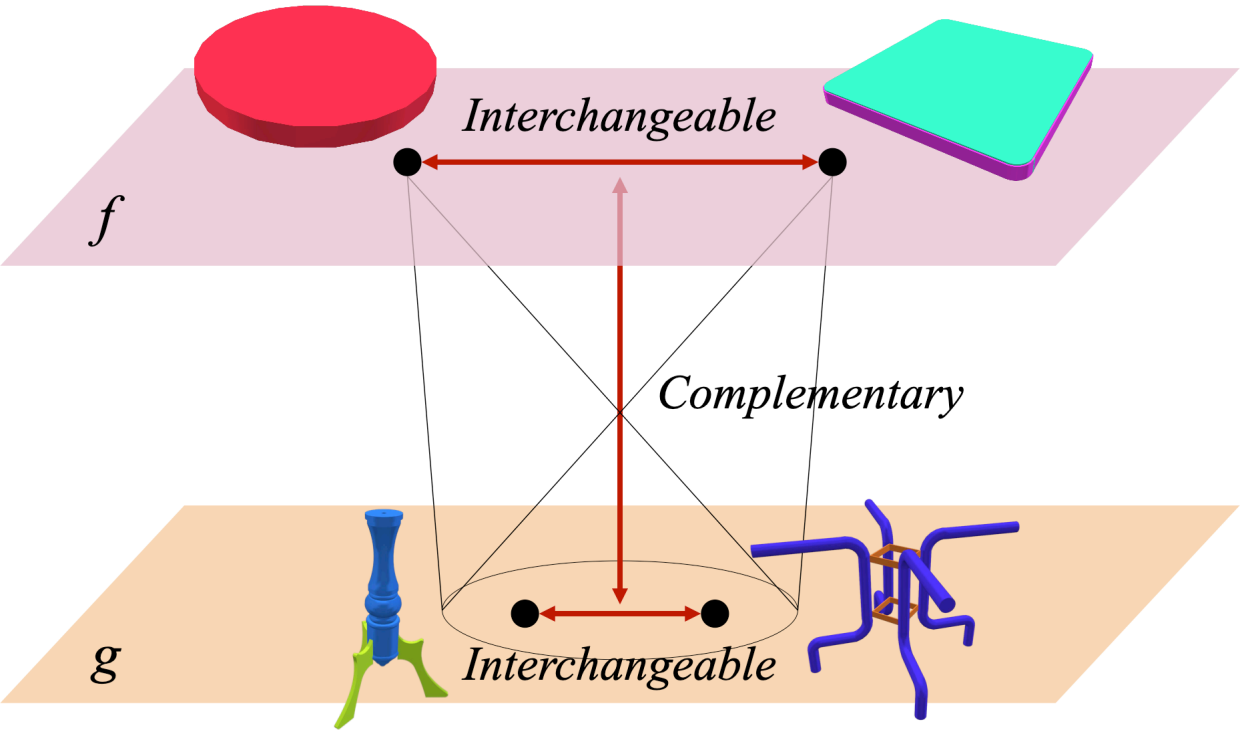


Scene Graphs



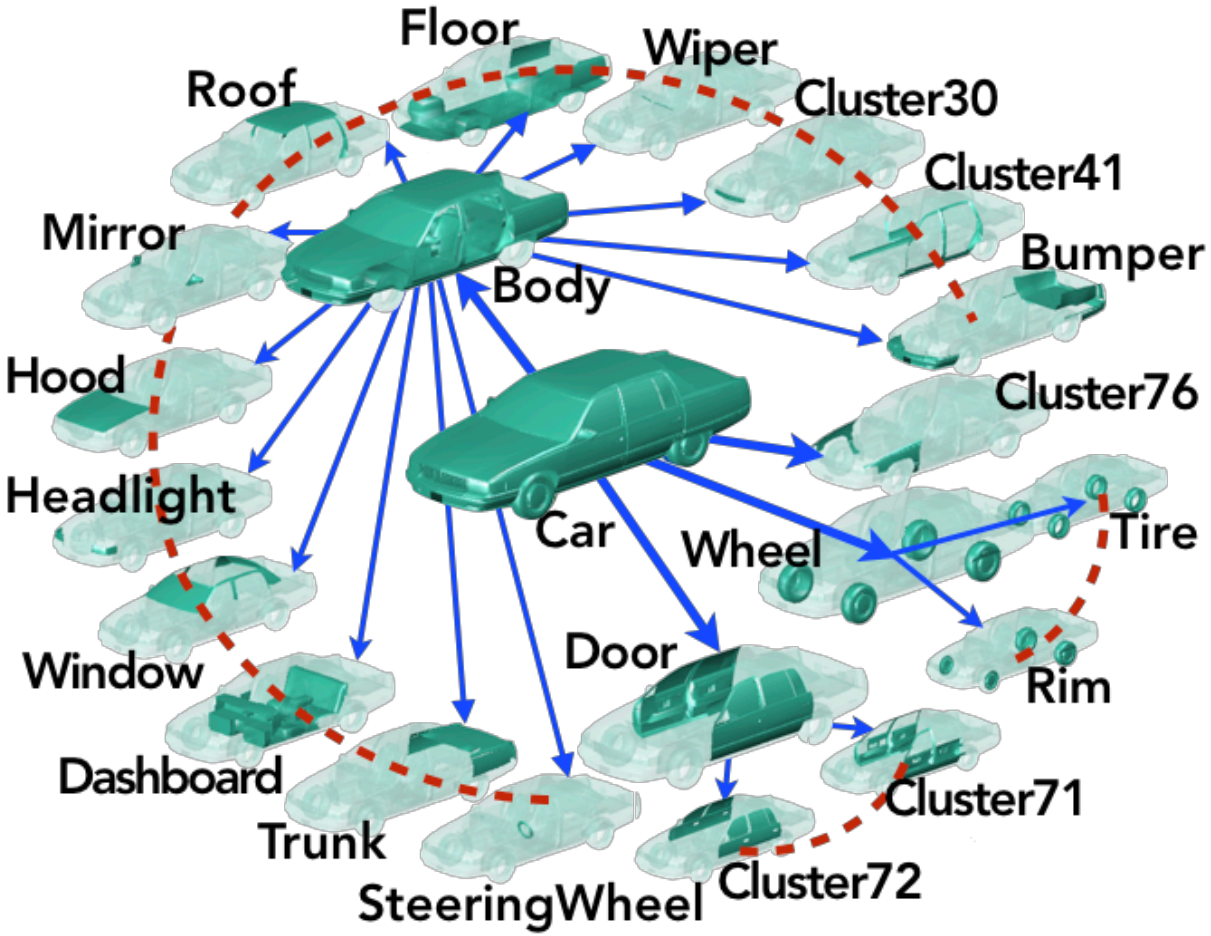
[Li et al., SIGGRAPH 2017]

Hierarchical and Multi-Resolution Approaches



Part Relations

[Sung et al., SIGGRAPH 2017, SGP 2018]

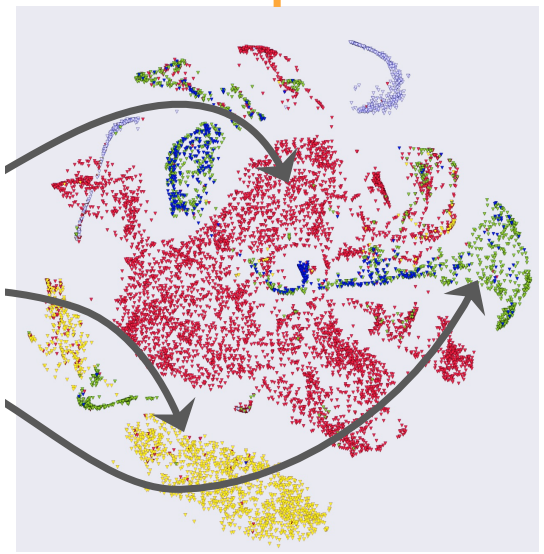


Part Hierarchies

[Li et al., SIGGRAPH 2017]

Materials

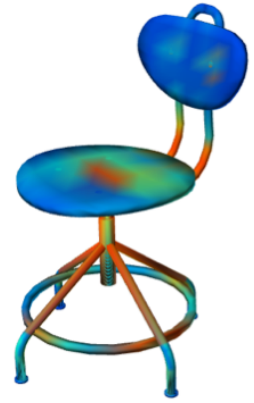
- Appearance: relate images and shapes
- Physical objects: novel applications



Material-Aware
Descriptor Space

Material-Aware
Auto-Texturing

Material-Aware
Part Retrieval

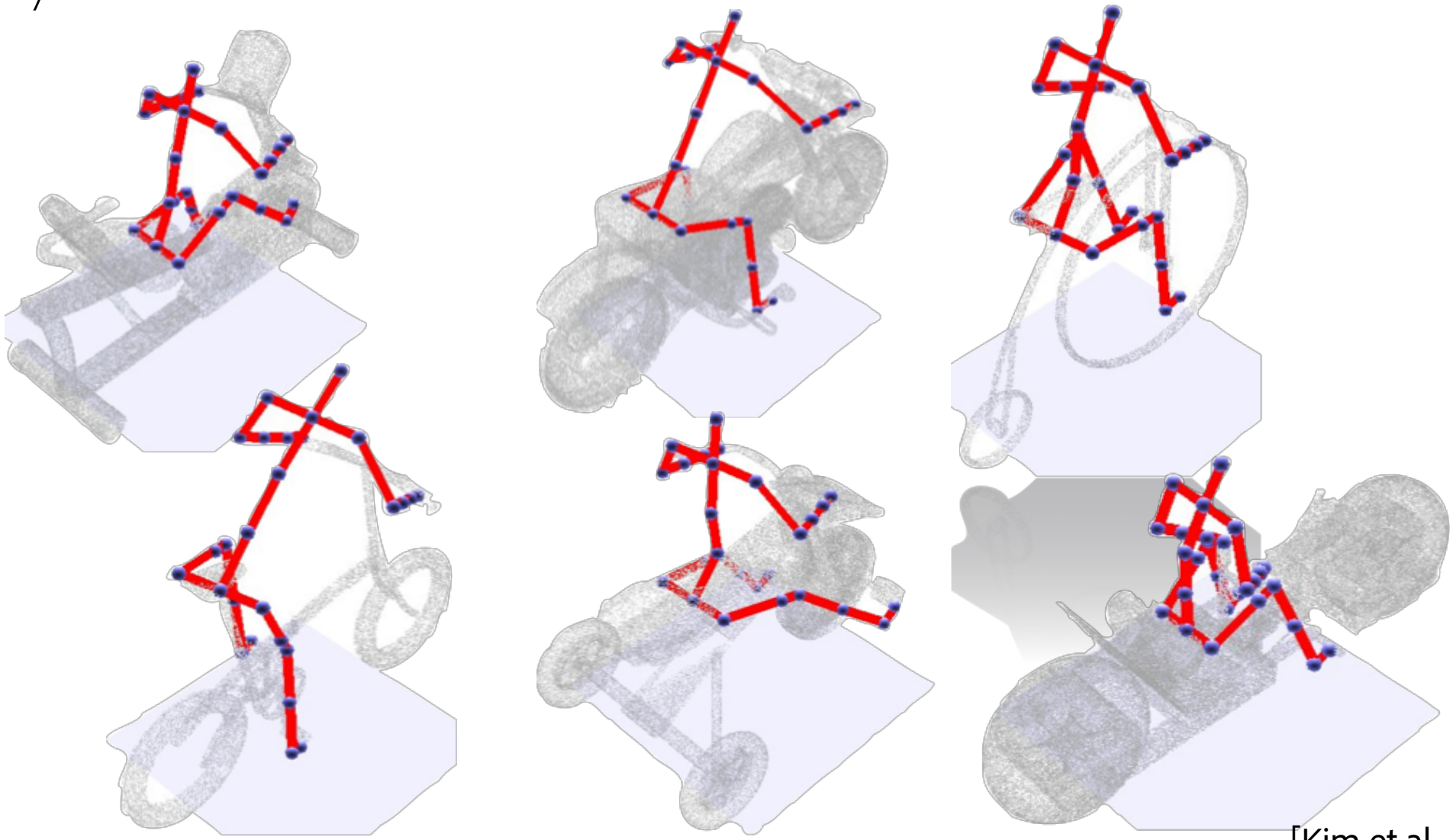


Material-Aware
Simulation

[Lin et al., 3DV 2018]

Human-Object Interactions

- Geometry & Interactions



[Kim et al., SIGGRAPH 2013]

Thank you!

- AtlasNet,
  3D-CODED,
  Signal Transfer

- Photometric
  Optimization

- Unicode


Thibault Groueix


Mathieu Aubry


Matt Fisher


Bryan Russell


Chen-Hsuan Lin


Simon Lucey


Oliver Wang


Eli Shechtman

Matt Fisher

Bryan Russell


Sanjeev Muralikrishnan


Siddhartha Chaudhuri

Matt Fisher