Single View 3D Object Shape

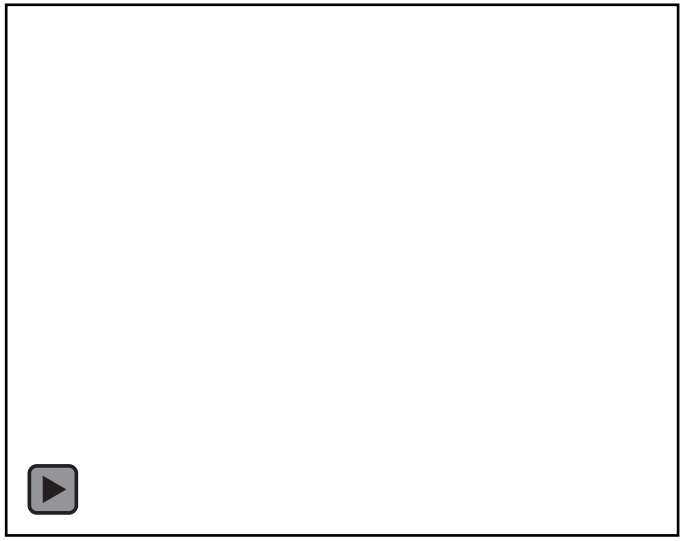
3D Vision

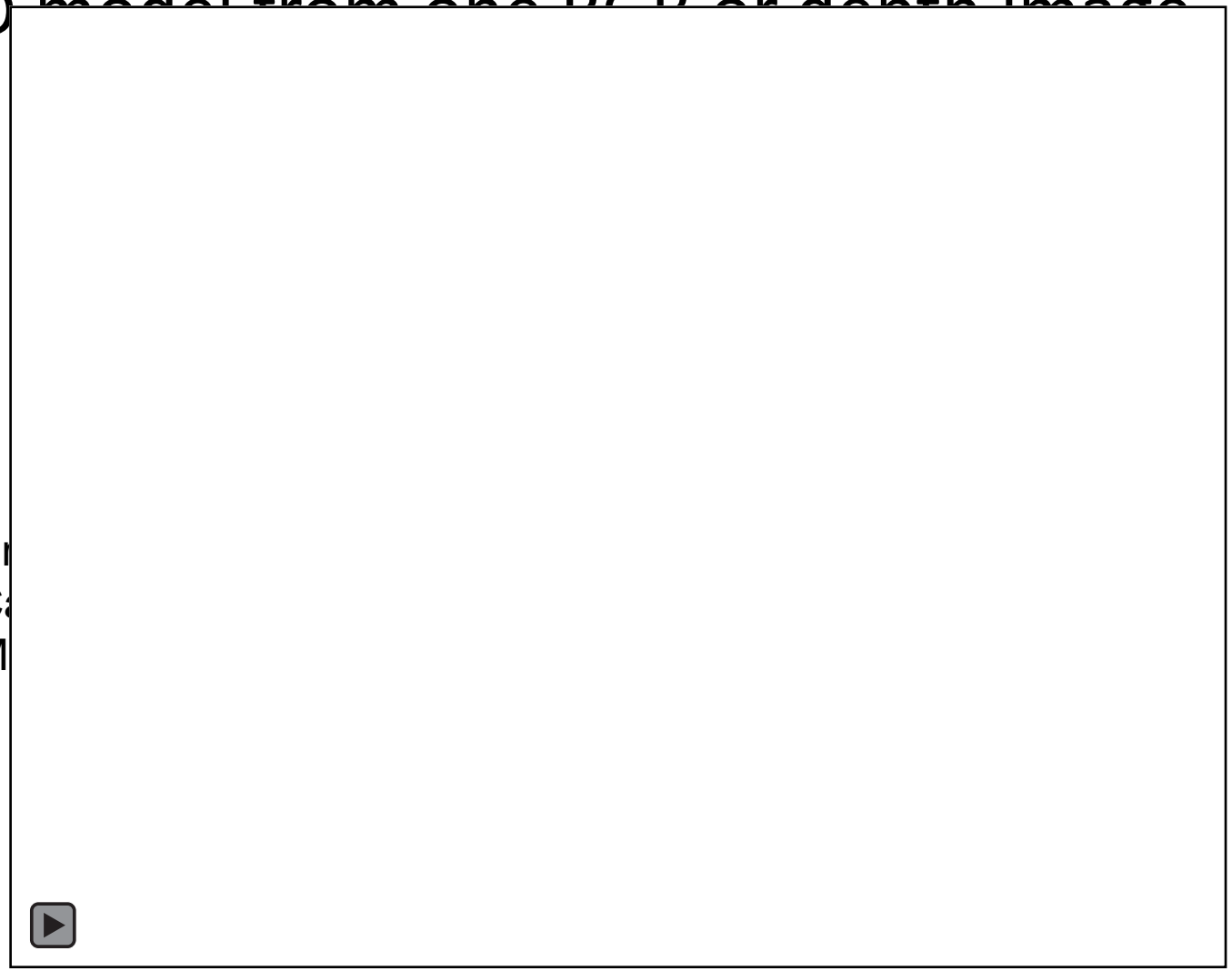
University of Illinois

Derek Hoiem

Goal: produce a complete 3D model from one RGB or depth image

Input Depth Image





Challenges

- Most of shape is not observed in single view
- Many different possible 3D shapes
  - Simple structural priors are not sufficient
  - Parameterizing shape is difficult
- Want it to work for any category

Problem and solution design decisions

- Reference frame: Object centric vs. viewer centric
- Shape output: points, depths, mesh, voxels
- Generalization: category-specific or category agnostic
- Cues: RGB appearance, depth, boundaries, surface normals, symmetry

Early approaches to recover 3D object shape

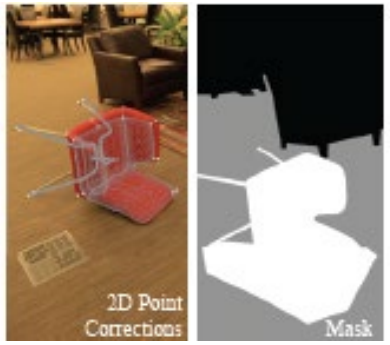
Shape from contour

"Teddy" (SG'99)

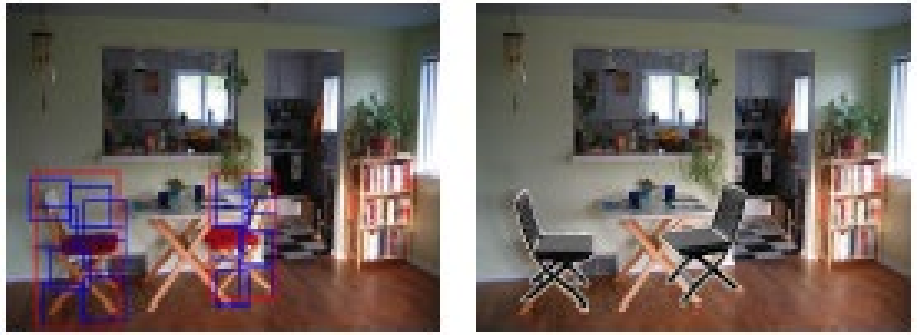
"3-Sweep" (SGA'14)

2.5D from Shading/Contour
(Barron et al. 2014)

Exemplar-based completion



Interactive retrieval/alignment
(Kholgade et al. SG'14)



Automatic Retrieval/Alignment
(Aubrey et al. 2014)

Rock et al. CVPR 2015

Ground Truth

Input Depth Image

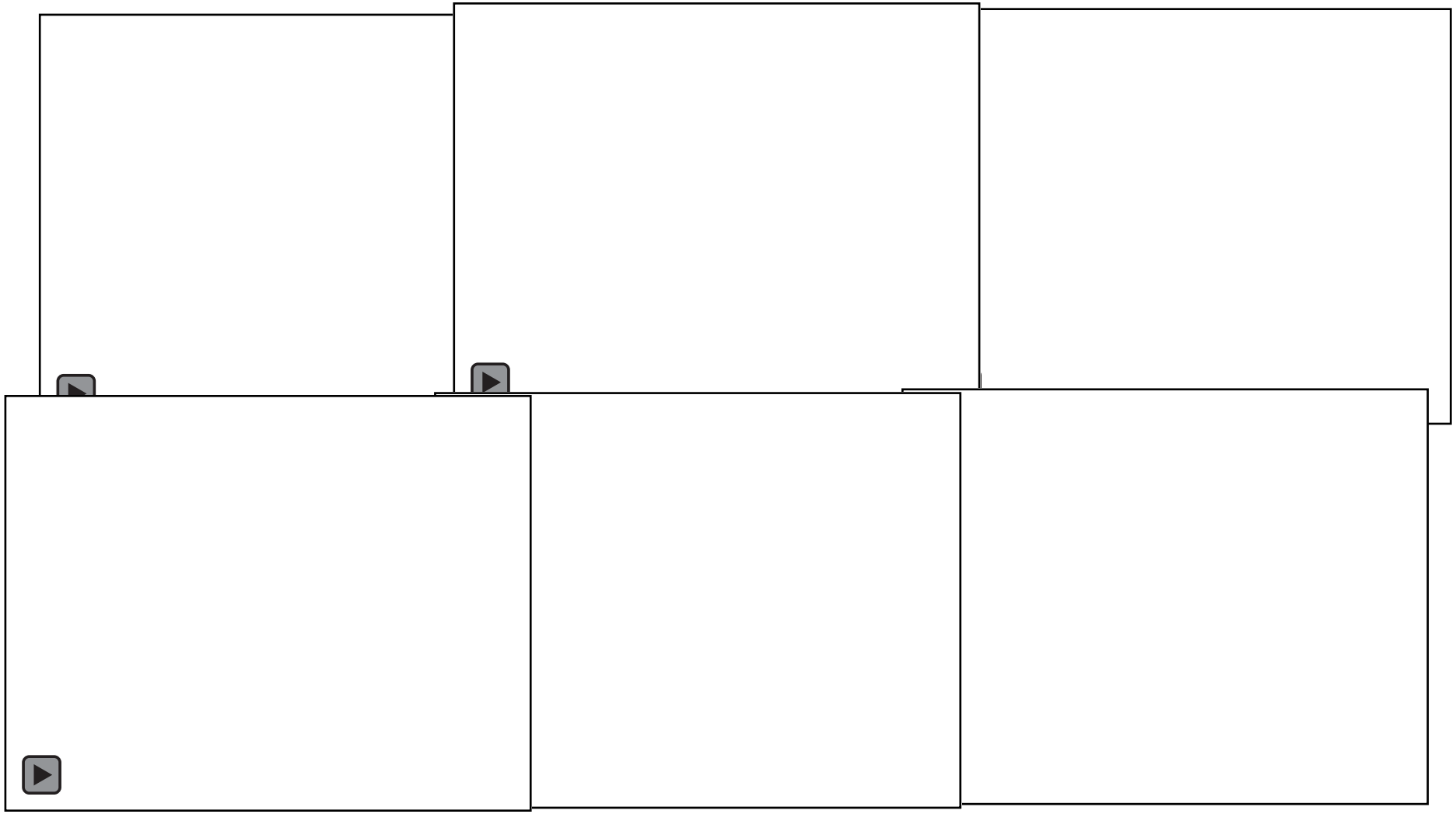
Predicted 3D shape (voxels)

3D exemplars rendered in multiple views

Matched Exemplar Model

Aligned and Deformed Exemplar w/ symmetry constrained thin plate spline deformation

Examples of deformations



Input Depth and Model Matched Model Deformed Model

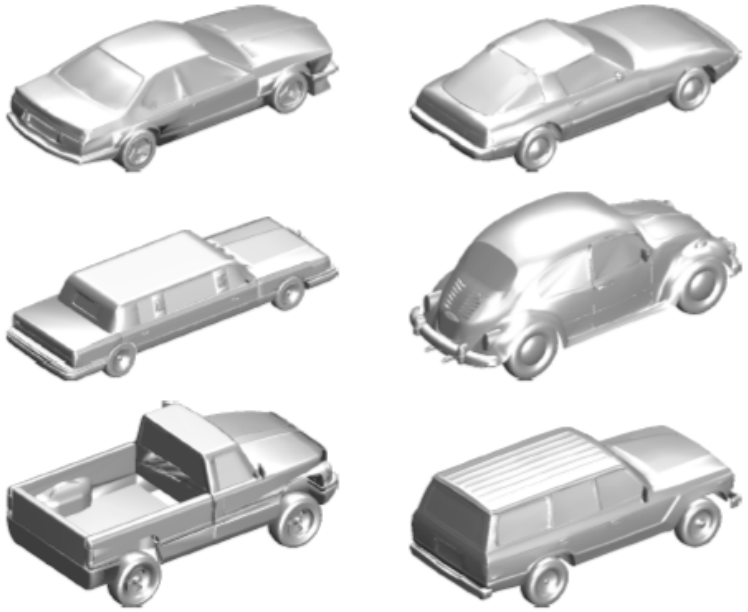
Experiments

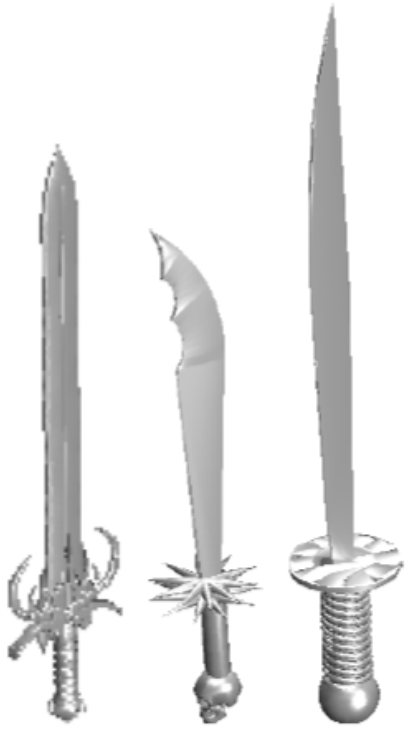
- Three difficulty settings
  - **Novel view**: new view of model that is in exemplar set
  - **Novel model**: new model from a category that is in exemplar set
  - **Novel category**: new model from a category that is not in the exemplar set

- Two measures of reconstruction accuracy
  - Voxel intersection/union
  - Surface-to-surface distance

- Same procedure applied in all cases (system is not told whether examples of the model or category are available)

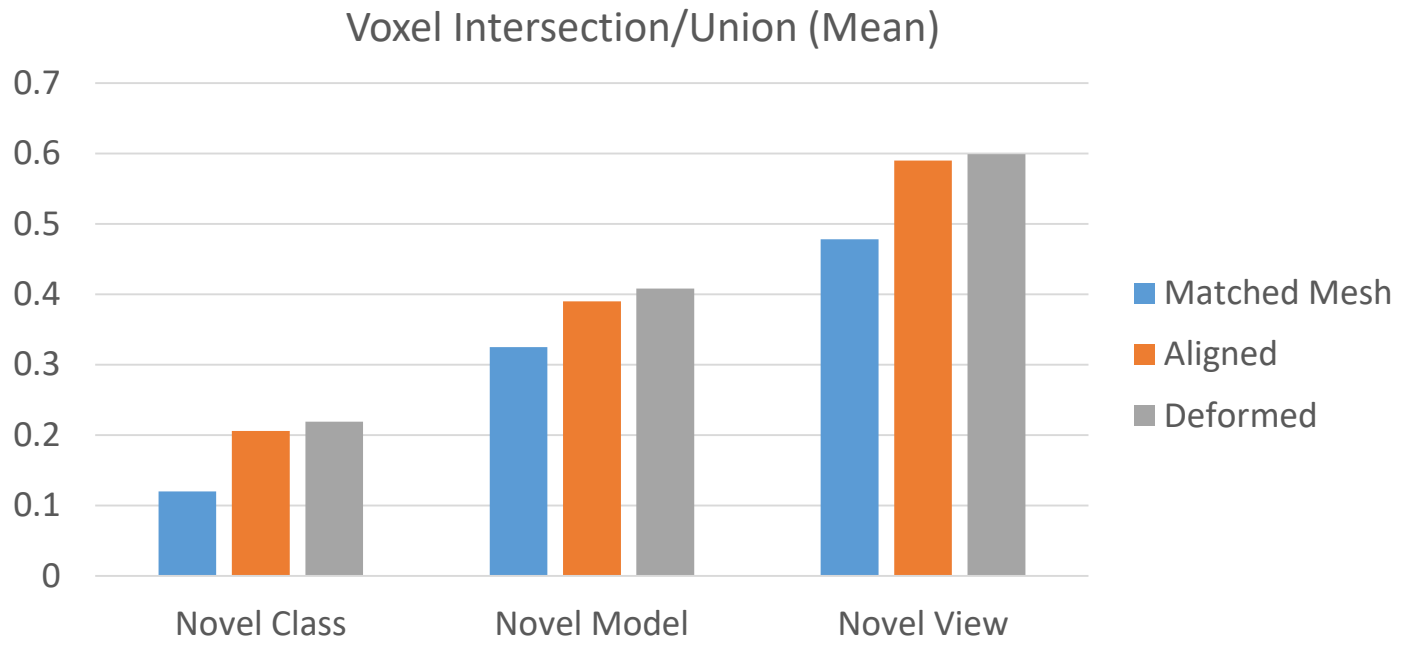
SHREC 2012 Dataset



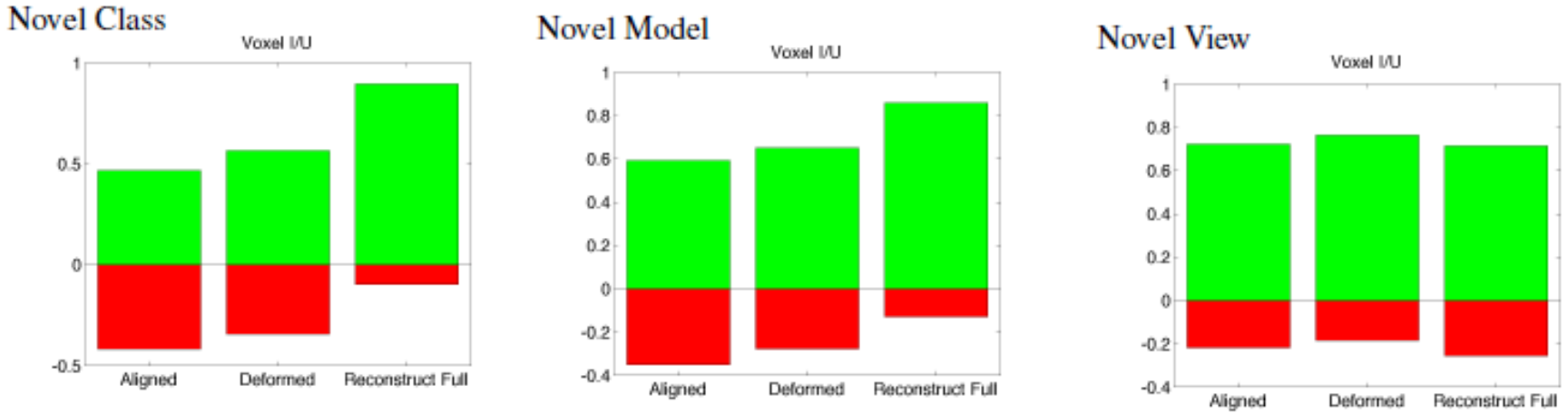


- 60 different classes
  - Instruments
  - Cars
  - Swords
  - Humans
  - Houses
  - …etc
- 20 models per class

Deformation makes exemplars more helpful

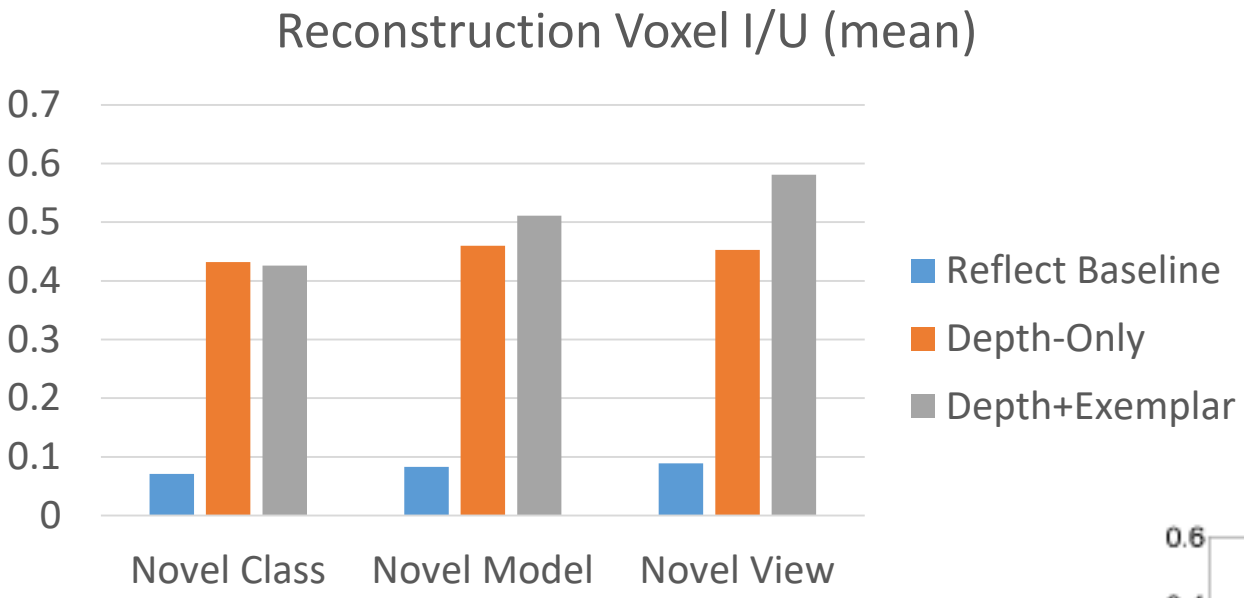


How often aligned/deformed/final model outperforms original match as 3D shape estimate

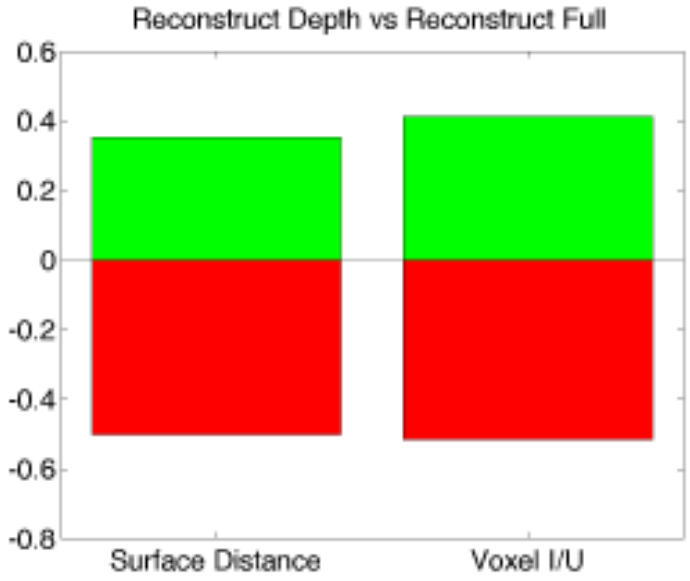


Do retrieved exemplars help reconstruction?

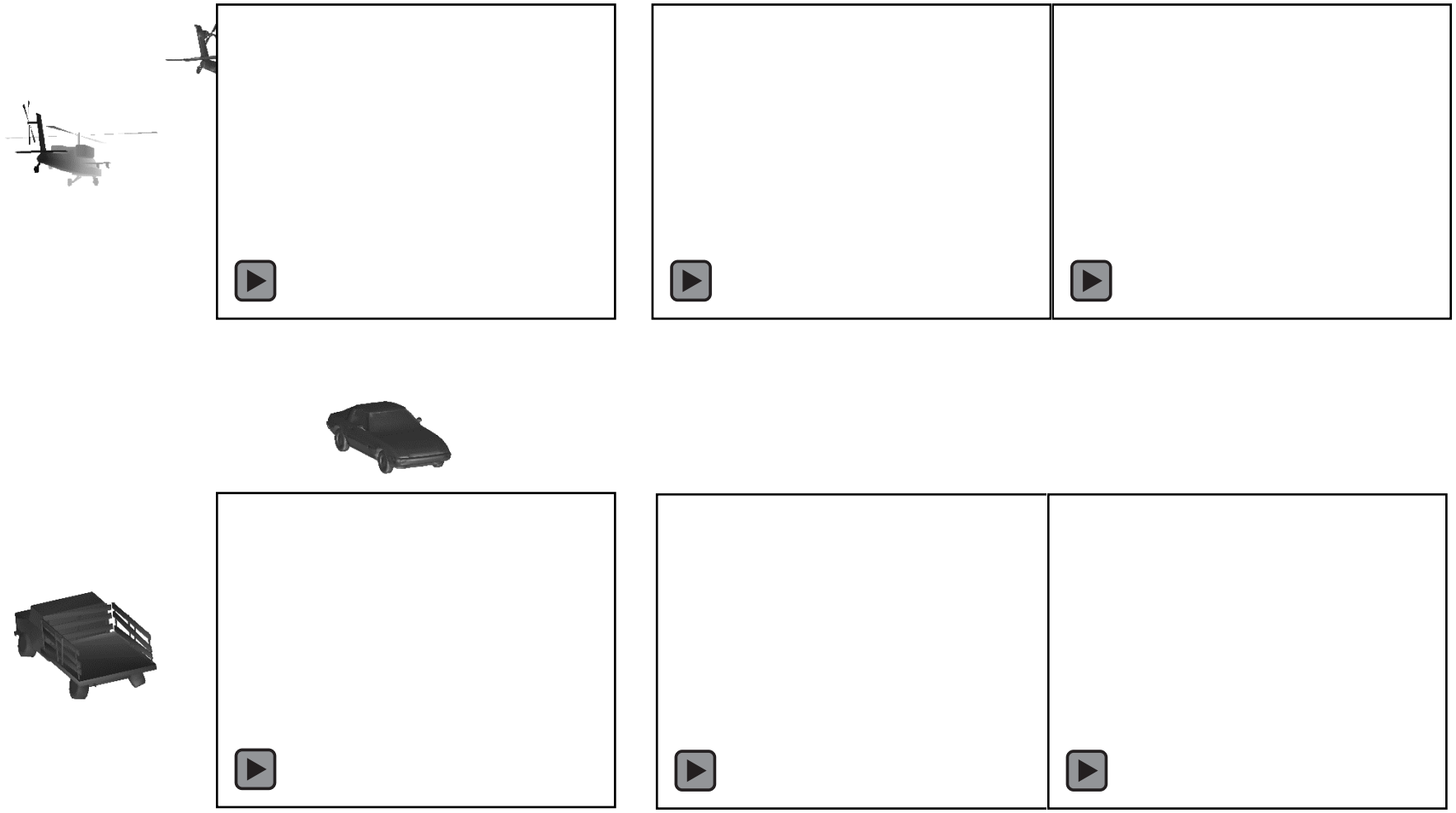
- Usually, but depends on similarity of retrieved model



Fraction of cases that improve for Novel Class



Results: novel model



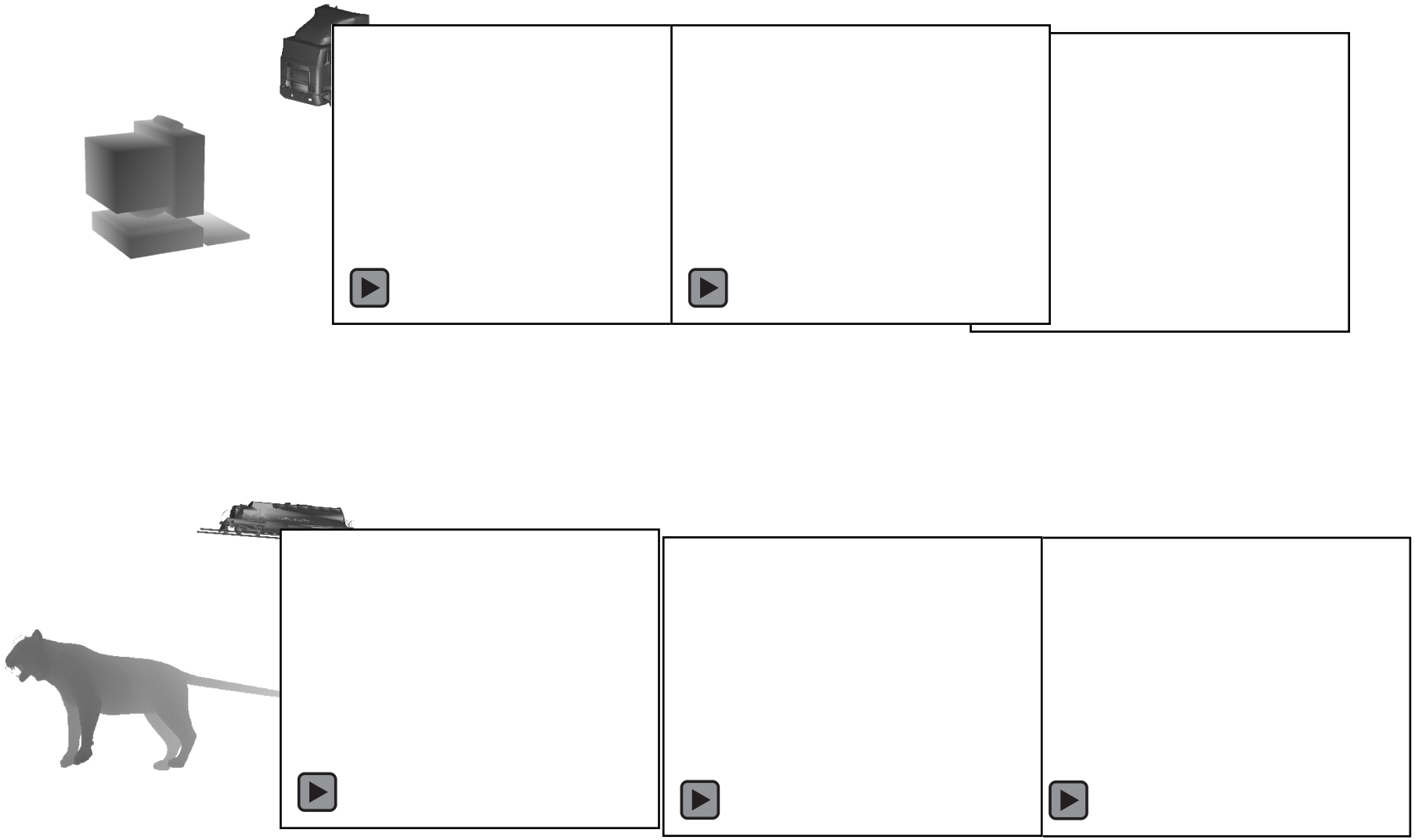
Input

Matched Exemplar

Reconstruction

Ground Truth

Results: novel category



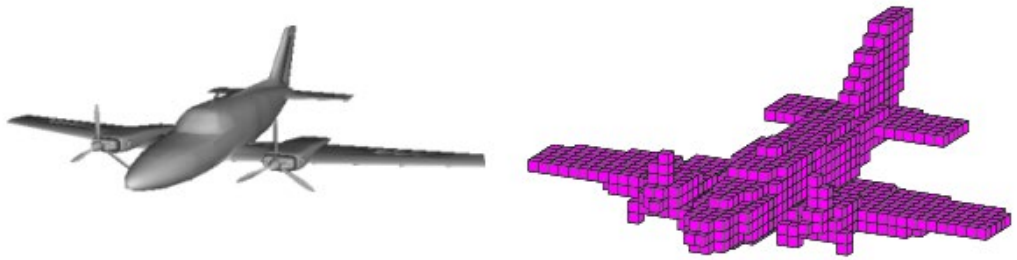
Pixels, voxels, and views: A study of shape representations for single view 3D object shape prediction

(Shin, Fowlkes, Hoiem CVPR 2018)

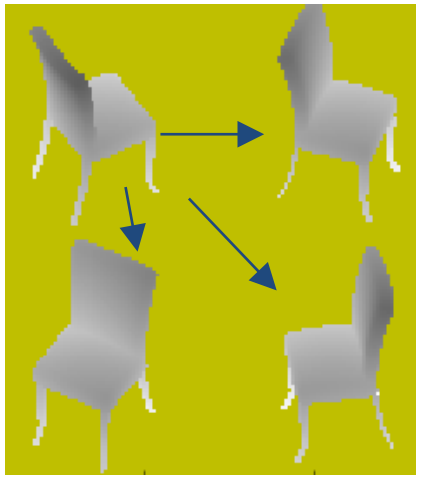
What effect does object representation have on prediction?

- Volumetric vs. Surface based
- Object-centric vs. Viewer -centric

Volumetric vs. Surface-based representations



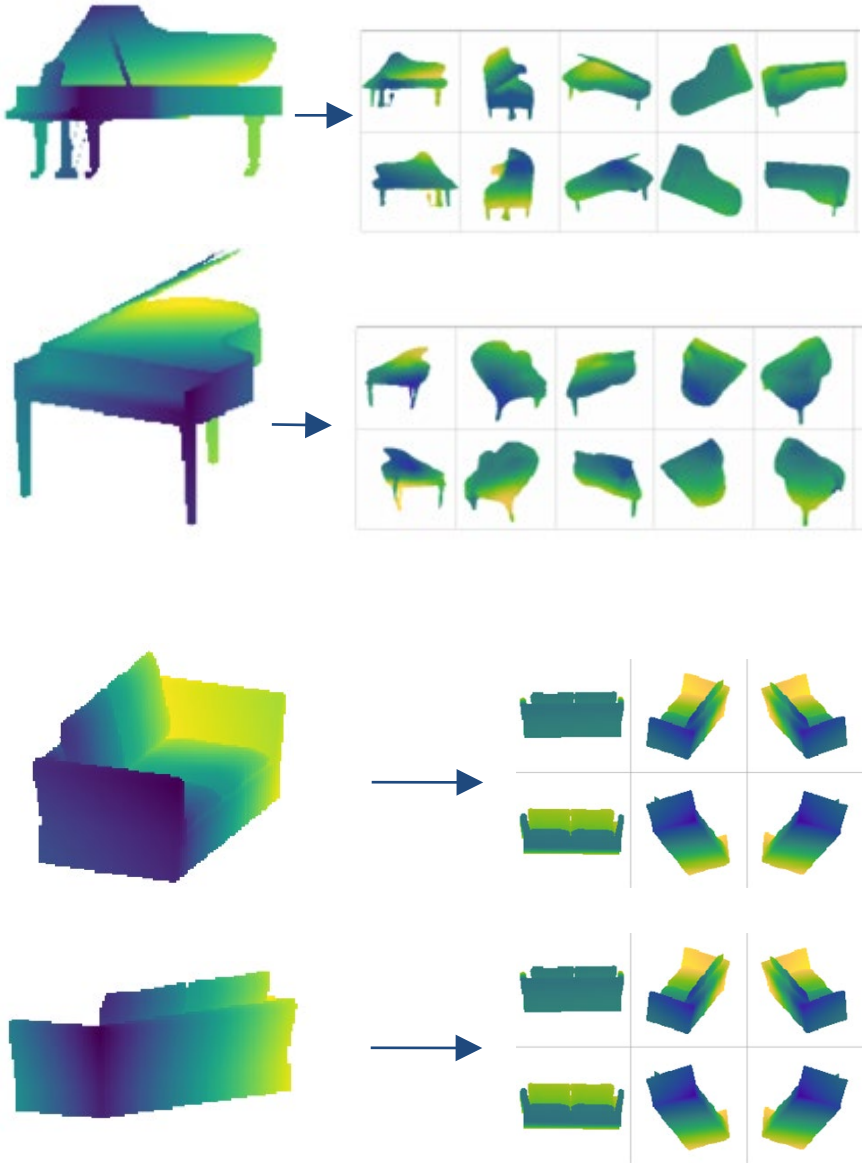
- **3D Voxels**
  - The focus of most previous studies
  - Low resolution
  - Hard to capture compositions, symmetries



Front, back, left, right, etc.

- **2.5D Surfaces**
  - Multiple silhouettes, depth or normal maps
  - Infer volume through multi-view reconstruction software
  - Can adopt existing texture-based image generation techniques

Viewer-centered vs. Object-centered



Viewer-centered

- Object shape/orientation modeled relative to input view
- Input has known viewpoint wrt model
- Good approximation by retrieving a model corresponding to a similar depth map

Object-centered

- Object shape modeled wrt canonical view
- Output models within class will be more similar
- Viewer orientation may need to be inferred
- Good approximation by retrieving a model corresponding to the same class

Multi-view representation for shape completion

- **Object-centered** output coordinates
  - Problem: For shape **completion**, the viewpoint of the input image needs to be guessed separately
  - Problem: Requires 3D model alignment
  - "Interpolation" of the learned outputs happens in object coordinates.
    i. Not good for Novel Class

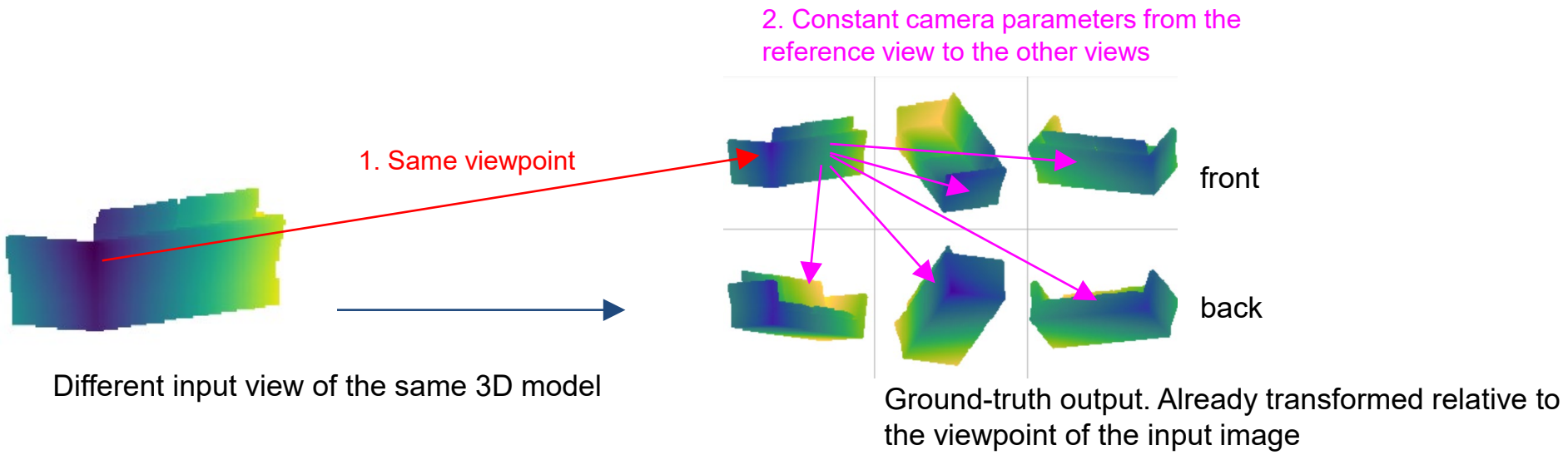
Object in **novel** category

Reference viewpoint of the output?

- Arbitrary and dataset dependent
- Camera transformation to match the input is unknown.
  (needed for shape completion)

Multi-view representation for shape completion

- **Viewer-centered** output coordinates



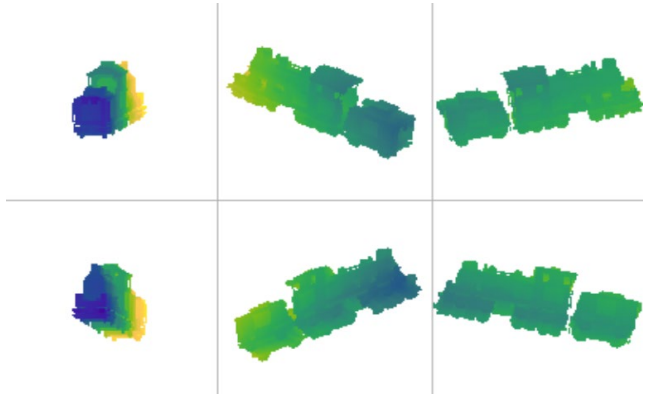
Multi-view representation for shape completion

- **Viewer-centered** output coordinates
  - Shape can be completed without knowing the viewpoint of the input image
  - Does not require 3D model alignment within category or across categories
    - Cross-category alignment is a difficult problem on its own

Object in **novel** category.
Unknown alignment

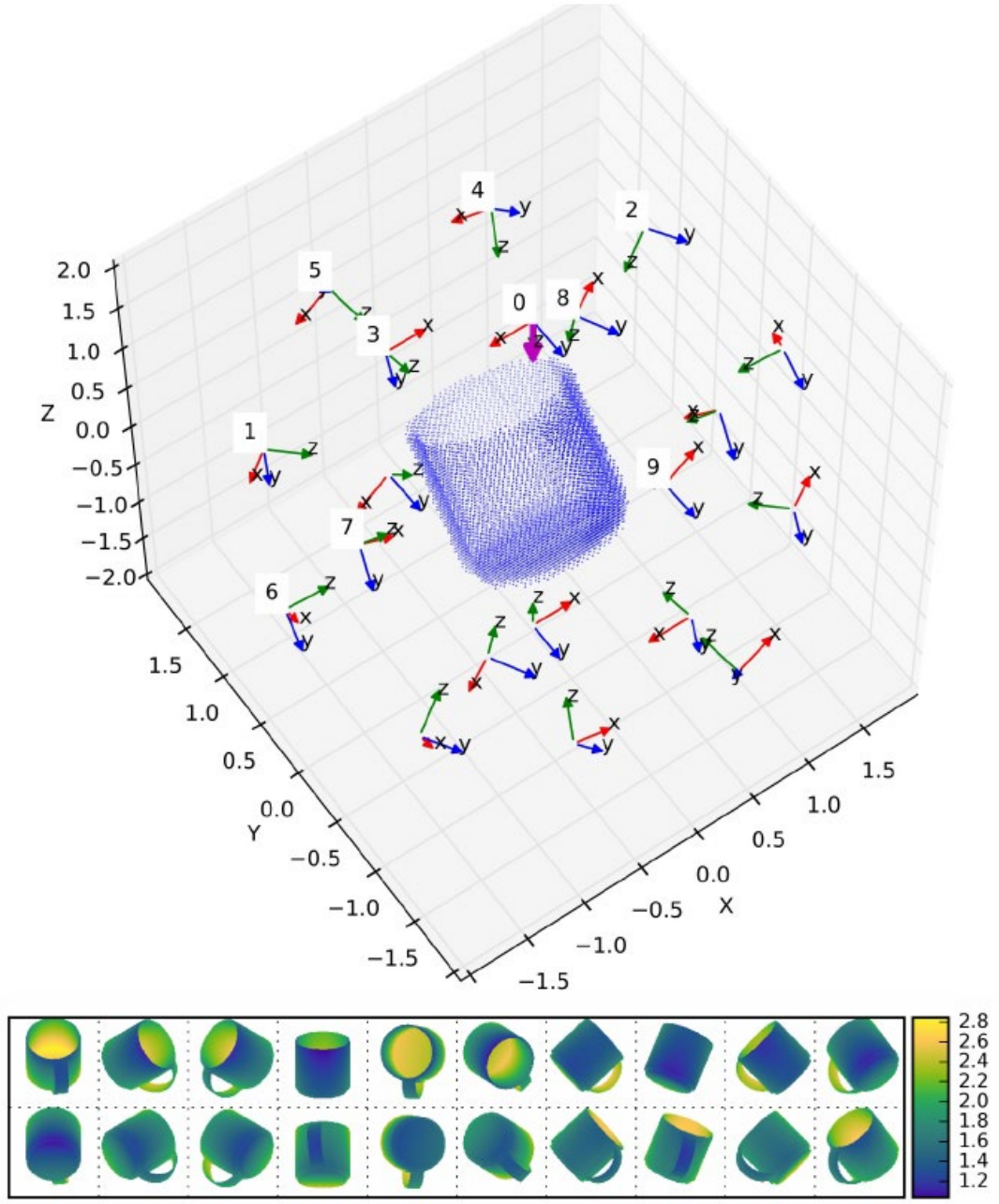
Reference viewpoint of the output for novel category?

- Always the same: relative to the input

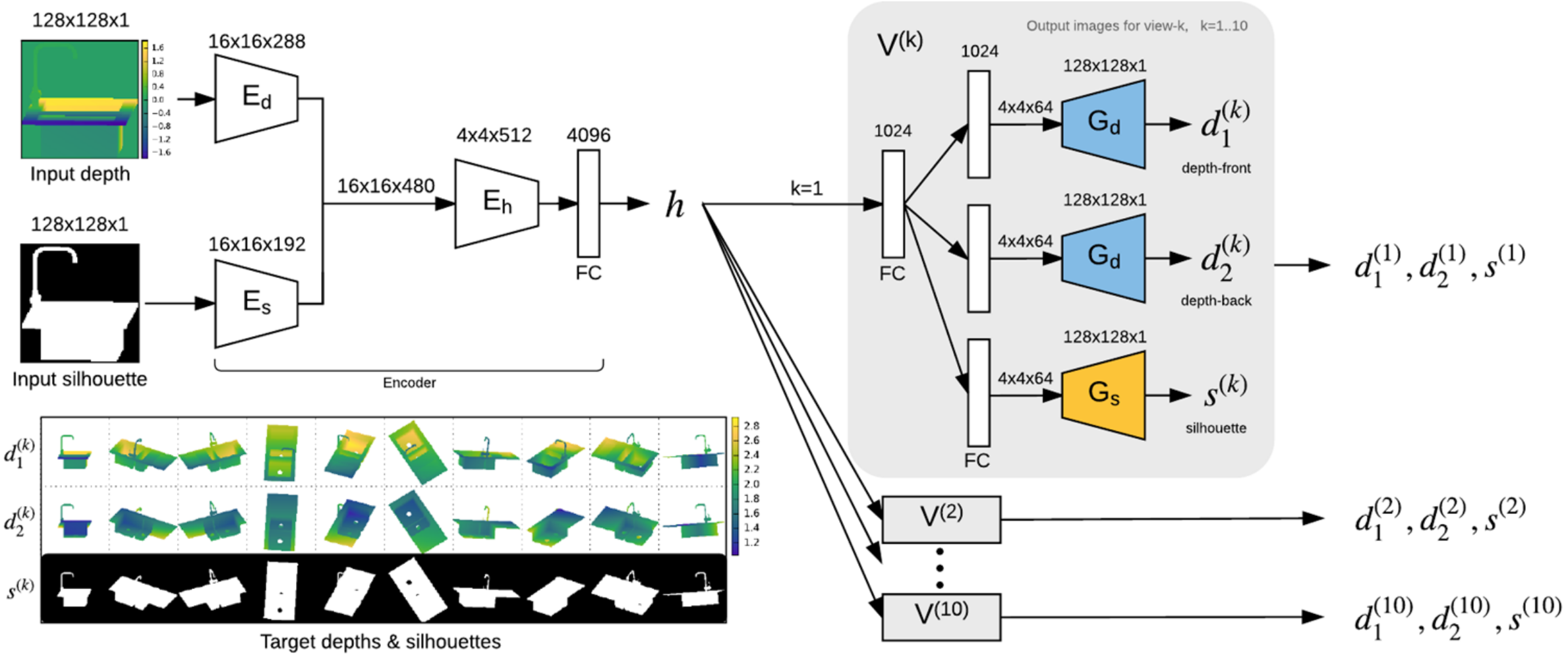


Training Set

Generate 20 relative views (depth + silhouette) of meshes

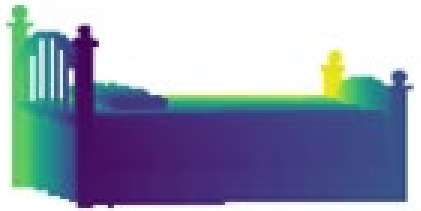


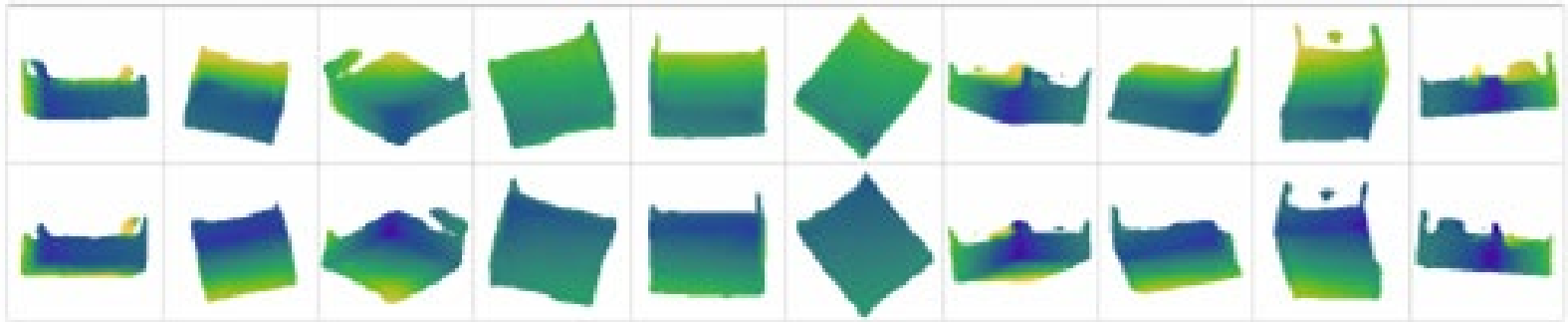
Network architecture



Surface-based 3D Prediction

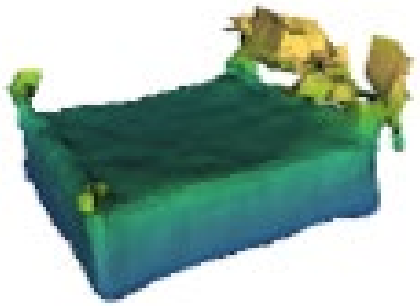
Input





CNN predicts depth and silhouette for each view

Create surface from all 3D points (FSSR: Floating Scale Surface Reconstruction)



FSSR: Fuhrmann and Goesele, SIGGRAPH 2014

Surface-based prediction outperforms, especially for novel class

Distance from Predicted to Ground Truth Surfaces (median over dataset)

	Novel View	Novel Model	Novel Class
Rock et al. (2015)	0.064	**0.060**	0.083
CNN **Voxel**	0.051	0.062	0.095
CNN **2.5D** + fusion (FSSR)	**0.049**	0.062	**0.076**

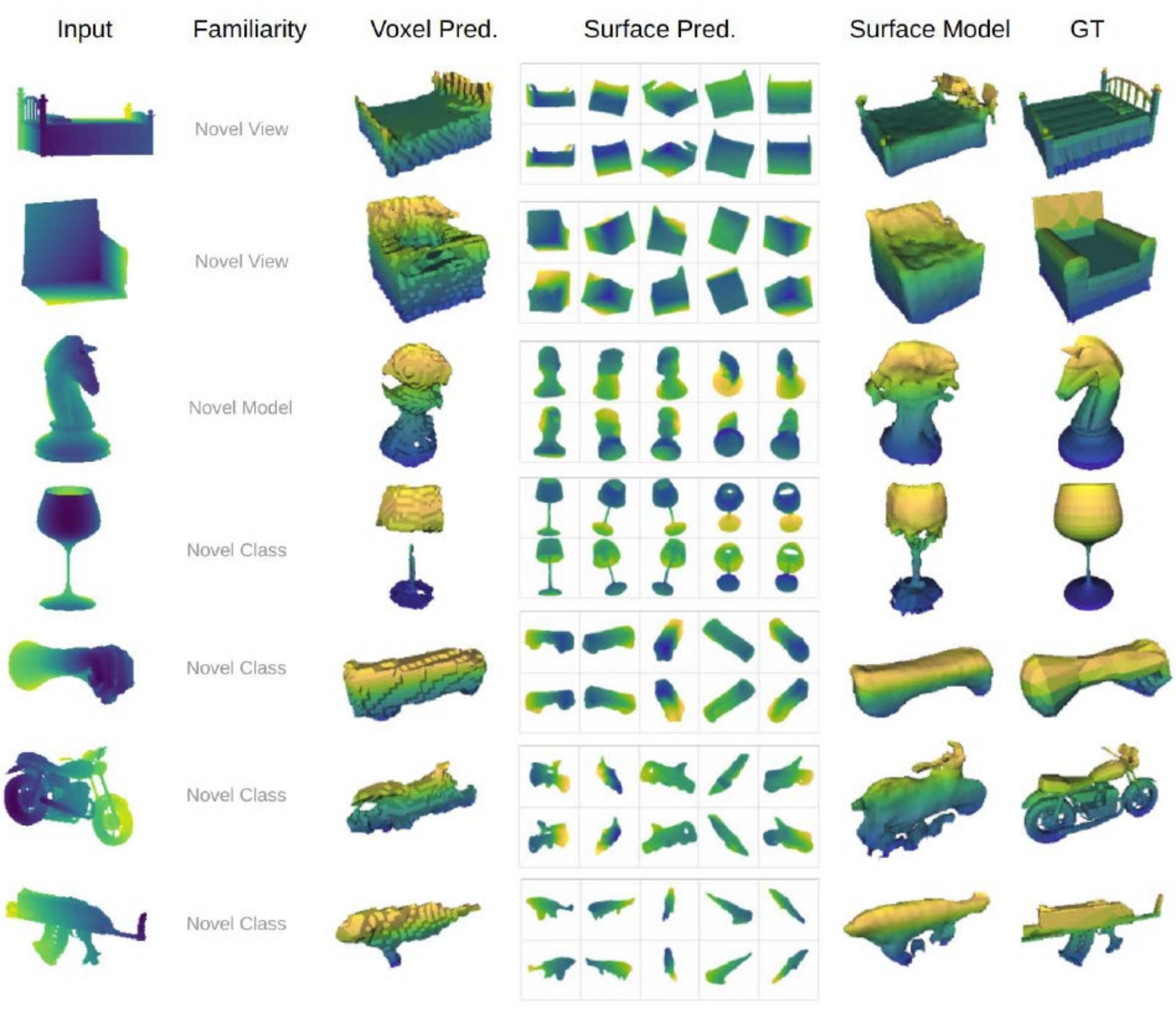
Lower is better!

Viewer-centric model vs. Object-centric model

IoU of Predicted and Ground Truth Values (mean)

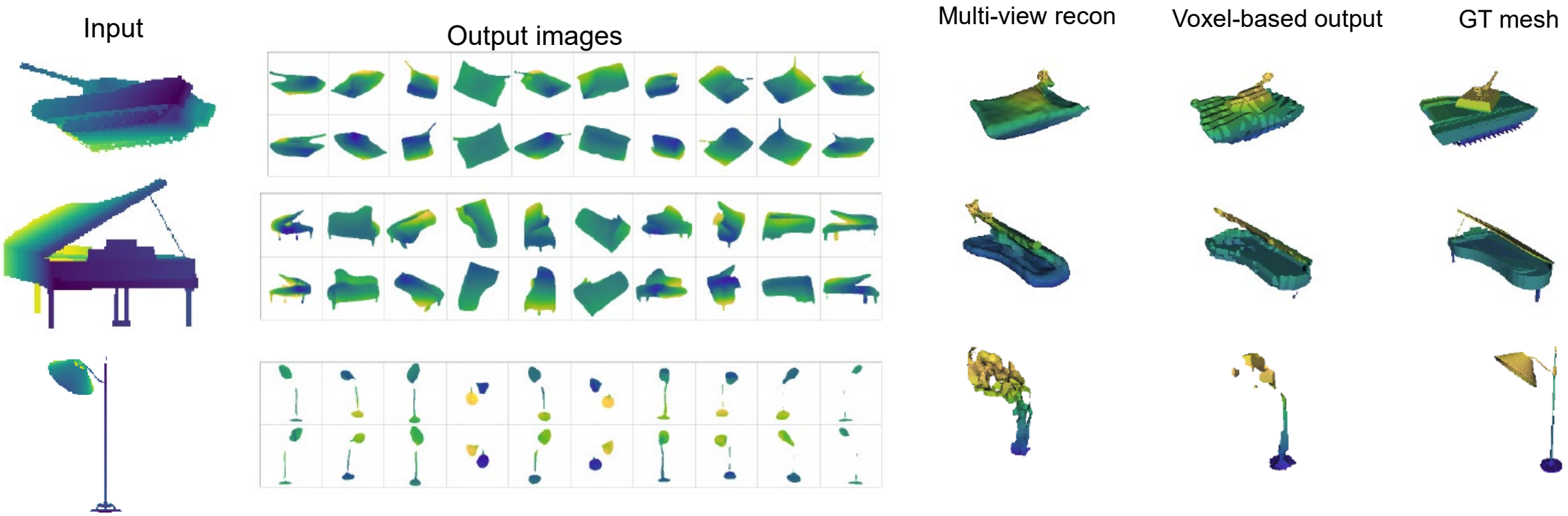
	Novel View	Novel Model	Novel Class
Viewer-Centric	0.714	**0.570**	**0.517**
Object-Centric	**0.902**	0.474	0.309

Higher is better!

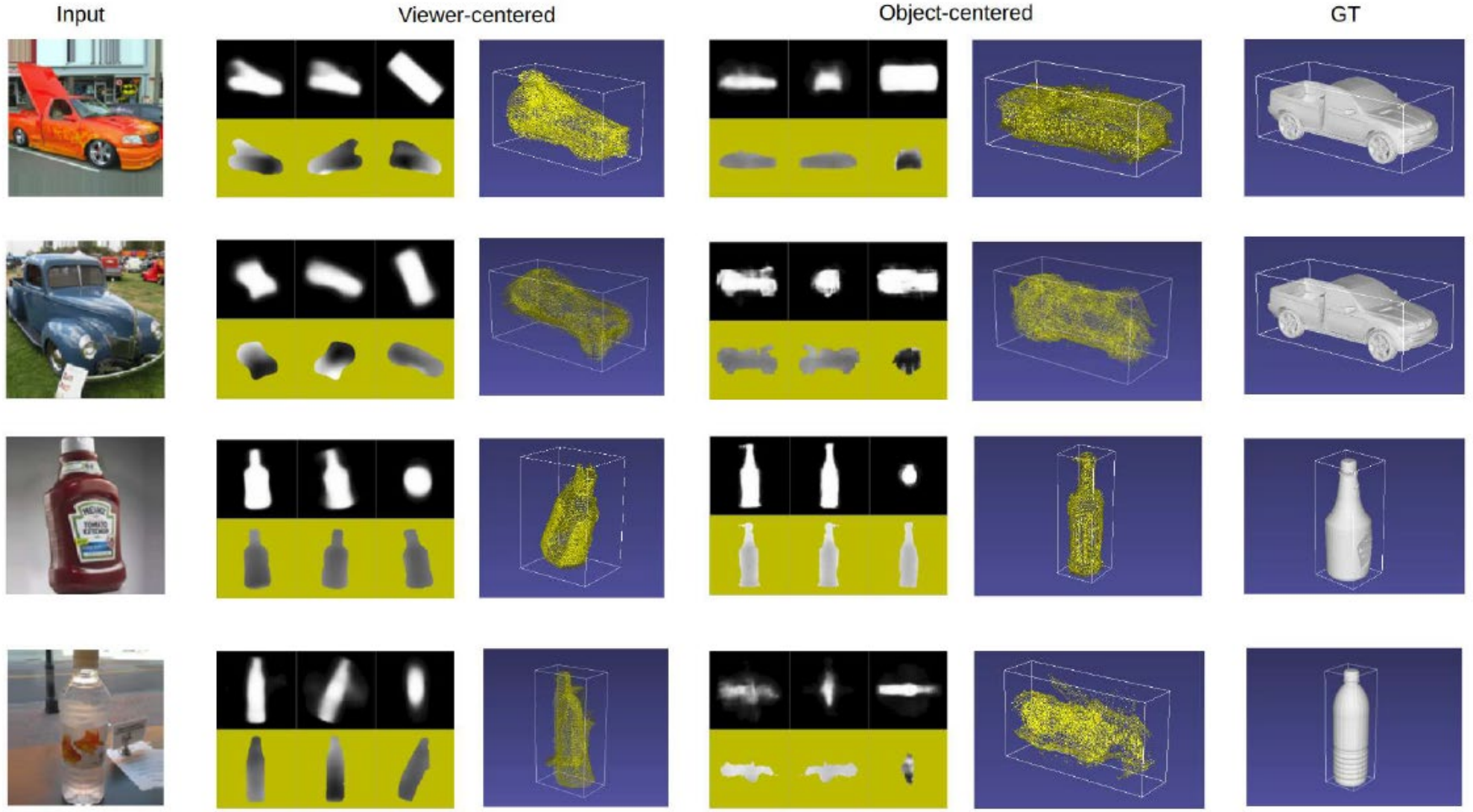

Input
Familiarity
Voxel Pred.
Surface Pred.
Surface Model
GT
Novel View
Novel View
Novel Model
Novel Class
Novel Class
Novel Class
Novel Class

Fusion is a challenge for the surface-based method

- Output depth maps may not be exactly consistent
- Difficulties in reconstructing thin object parts
- Could improve with view alignment



RGB-based prediction

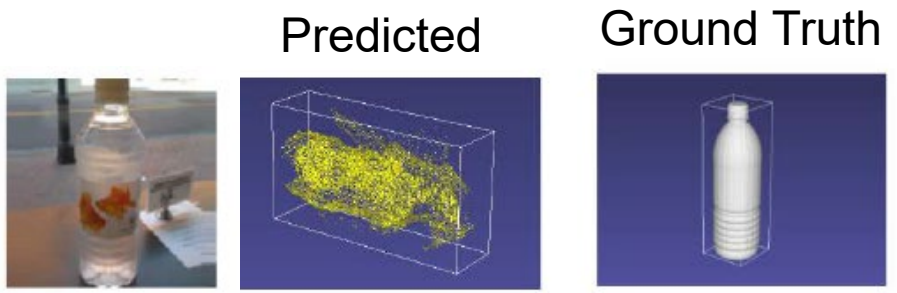


Conclusions so far

- Object-centric shape completion is basically recognition/retrieval

- Viewpoint-centric shape completion forces better generalization and leads to better performance for novel categories

- Predicting in terms of multiview surfaces may be better than voxels



What Do Single-view 3D Reconstruction Networks Learn?

CVPR 2019

Maxim Tatarchenko*[1], Stephan R. Richter*[2], René Ranftl[2], Zhuwen Li[2],
Vladlen Koltun[2], and Thomas Brox[1]

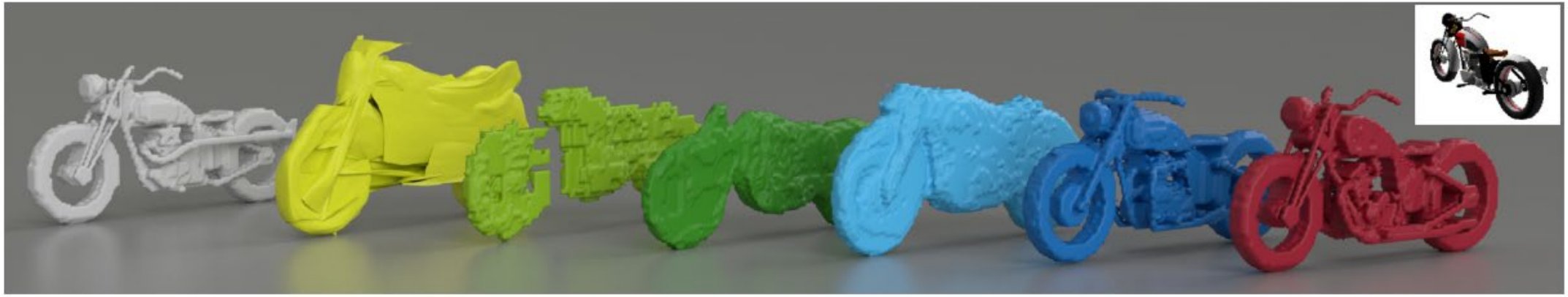
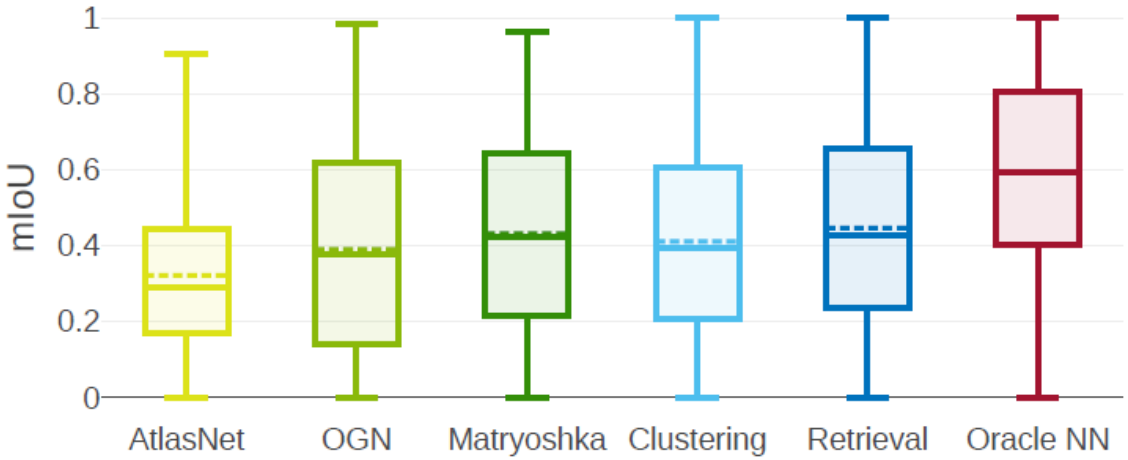
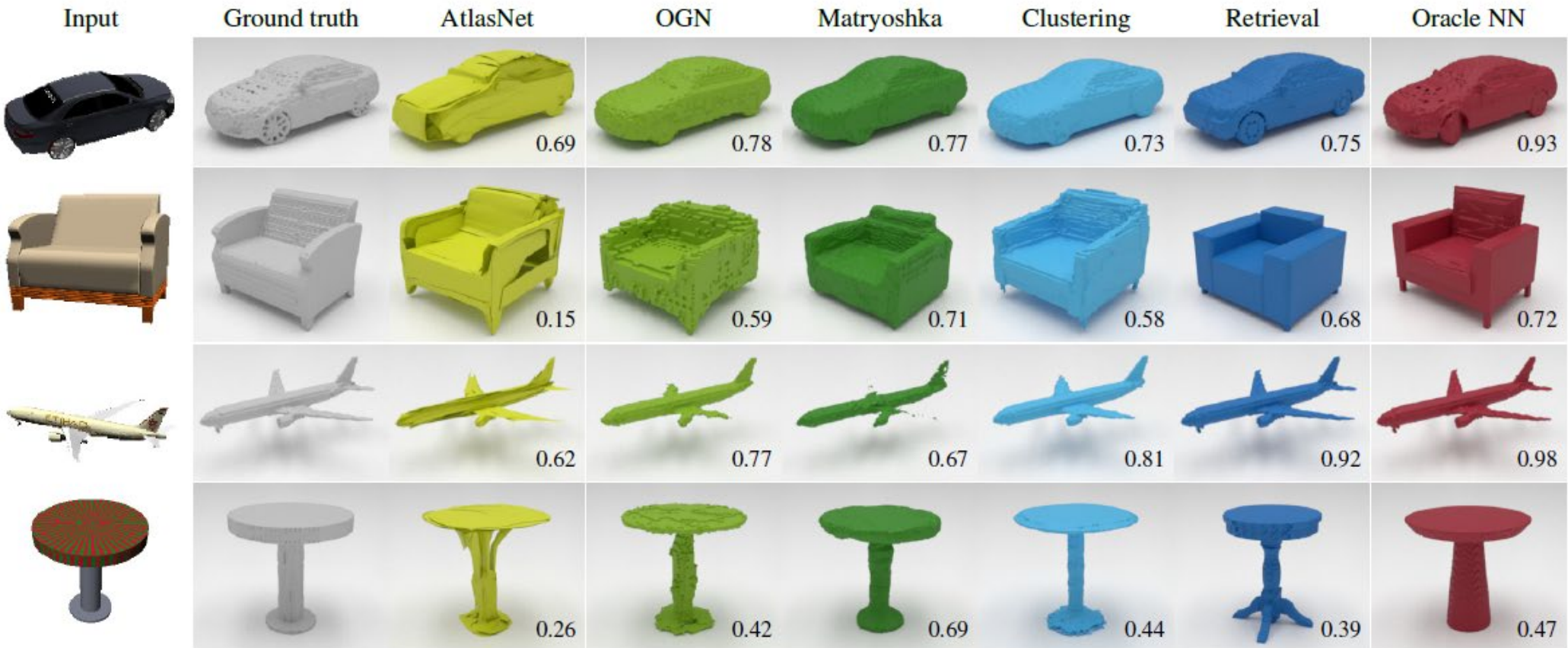


Figure 1. We provide evidence that state-of-the-art single-view 3D reconstruction methods (AtlasNet (light green, 0.38 IoU) [12], OGN (green, 0.46 IoU) [46], Matryoshka Networks (dark green, 0.47 IoU) [37]) do not actually perform reconstruction but image classification. We explicitly design pure recognition baselines (Clustering (light blue, 0.46 IoU) and Retrieval (dark blue, 0.57 IoU)) and show that they produce similar or better results both qualitatively and quantitatively. For reference, we show the ground truth (white) and a nearest neighbor from the training set (red, 0.76 IoU). The inset shows the input image.

Object-centered





Viewer-centered

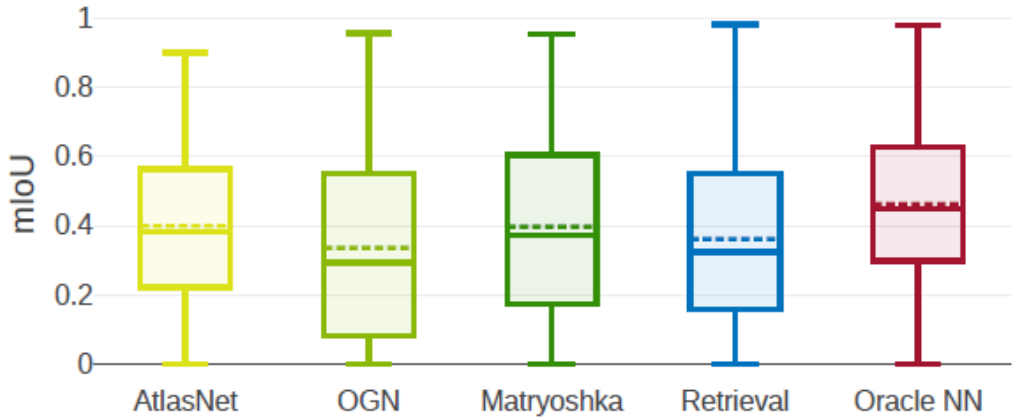


Figure 7. Mean IoU in viewer-centered mode. The retrieval baseline does not perform as well in this mode.

Difficulty measuring shape similarity for evaluation

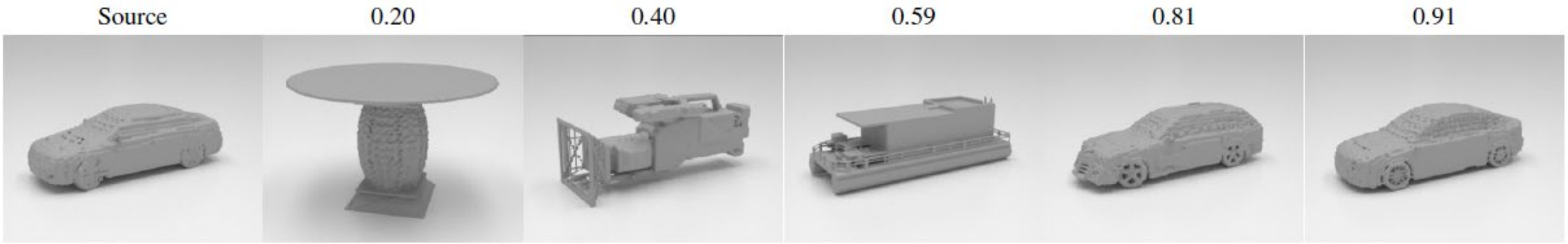


Figure 8. IoU between a source shape and various target shapes. Low to mid-range IoU values are a poor indicator of shape similarity.

Figure 9. The Chamfer distance is sensitive to outliers. Compared to the source, both target shapes exhibit non-matching parts that are equally wrong. While the F@1% is 0.56 for both shapes, the Chamfer distance differs significantly.

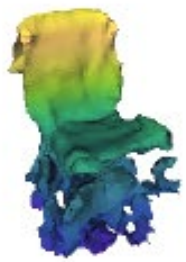
- F-score is proposed (geometric mean of surface precision/recall at some threshold)

Recommendations from Tatarchenko et al.

- Use viewer-centric problem formulation (otherwise, it's just retrieval)

- F1-score is a better metric

Objects are structured.



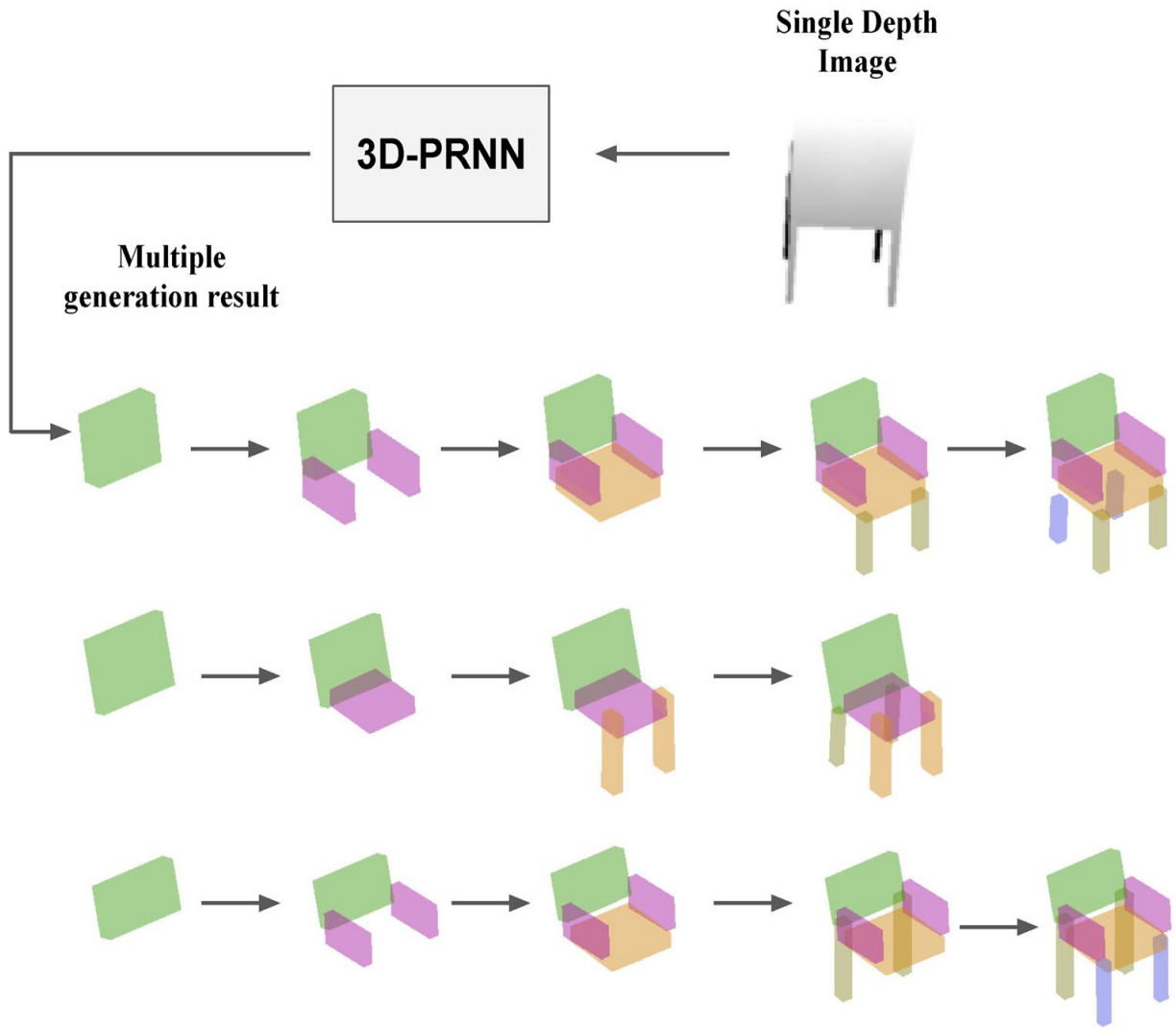
Why aren’t predictions?

3D-PRNN: Generating Shape Primitives with RNNs

Zou et al. ICCV 2017

Output multiple guesses of 3D structure (parts layouts)

- Variable number of parts
- Varying classes



Training annotations automatically generated from meshes

Volume of the primitive

Rotation matrix

Uniformly sampled point from the primitive

Point cloud

Truncation parameter

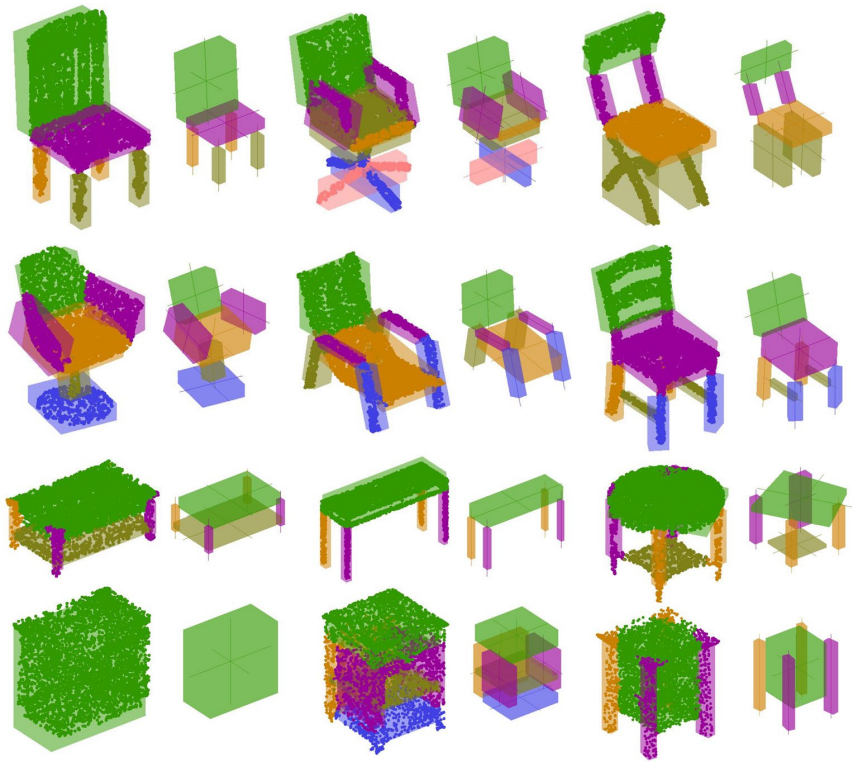
$$E_p = -\sum_{m,n} V_p \min\left(\exp\left(-\frac{\|R(\theta)Sp_m + T - q_n\|^2}{\sigma^2}\right), \xi\right) \quad (1)$$

$$E_w = E_P^+ - \alpha E_P^- \quad (2)$$

Algorithm 1 Primitive fitting

Given shape point clouds Q and empty primitive set X;
$\beta = 0.97|Q|$, t = 0;
while $|Q| < \beta$ or $i <$maxPrimNum **do**
  $E_{best} =$ Inf;
  for $i = 1$:maxRandNum **do**
    $\theta = [0, 0, 0]$, random initialize S, T, $j = 0$;
    while $\delta < 0.01$ or $j <$maxIter **do**
      fix θ, solve $S, T \rightarrow S^*, T^*$ by Eq .2;
      fix S^*, T^*, update $\theta \rightarrow \theta^*$ by Eq .2;
      calculate $E_w(S^*, T^*, \theta^*)$ by Eq .2;
      if $E_w < E_{best}$ **then**
        $E_{best} = E_w$, $x_{best} = [S^*, T^*, \theta^*]$;
      $\delta = \|[S, T, \theta] - [S^*, T^*, \theta^*]\|^2$;
      $S = S^*, T = T_p^*, k = k + 1$;
  $x_t = x_{best}$, add x_t to X, $t = t + 1$;
  Remove fitted points from Q and add to non-occupied space Q^-
return X

Object-centric model



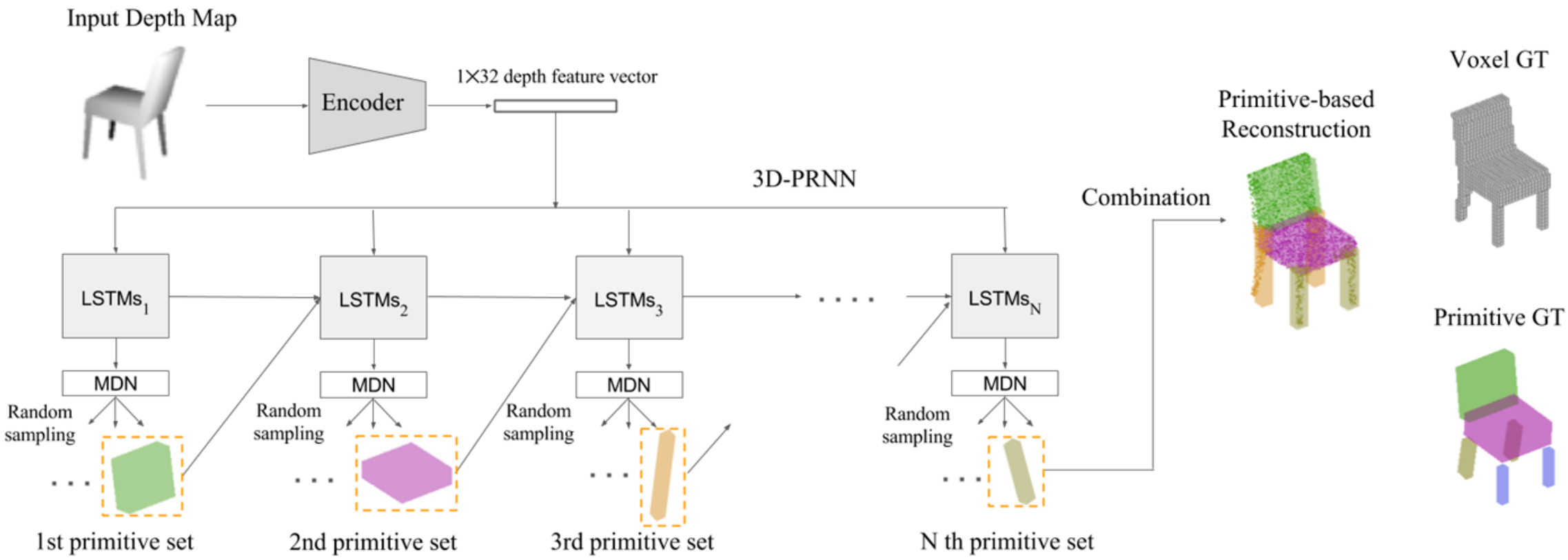
$$x = [s_x, s_y, s_z, t_x, t_y, t_z, \theta_x, \theta_y, \theta_z]$$

Scale of a unit cube along three orthogonal axes (S)

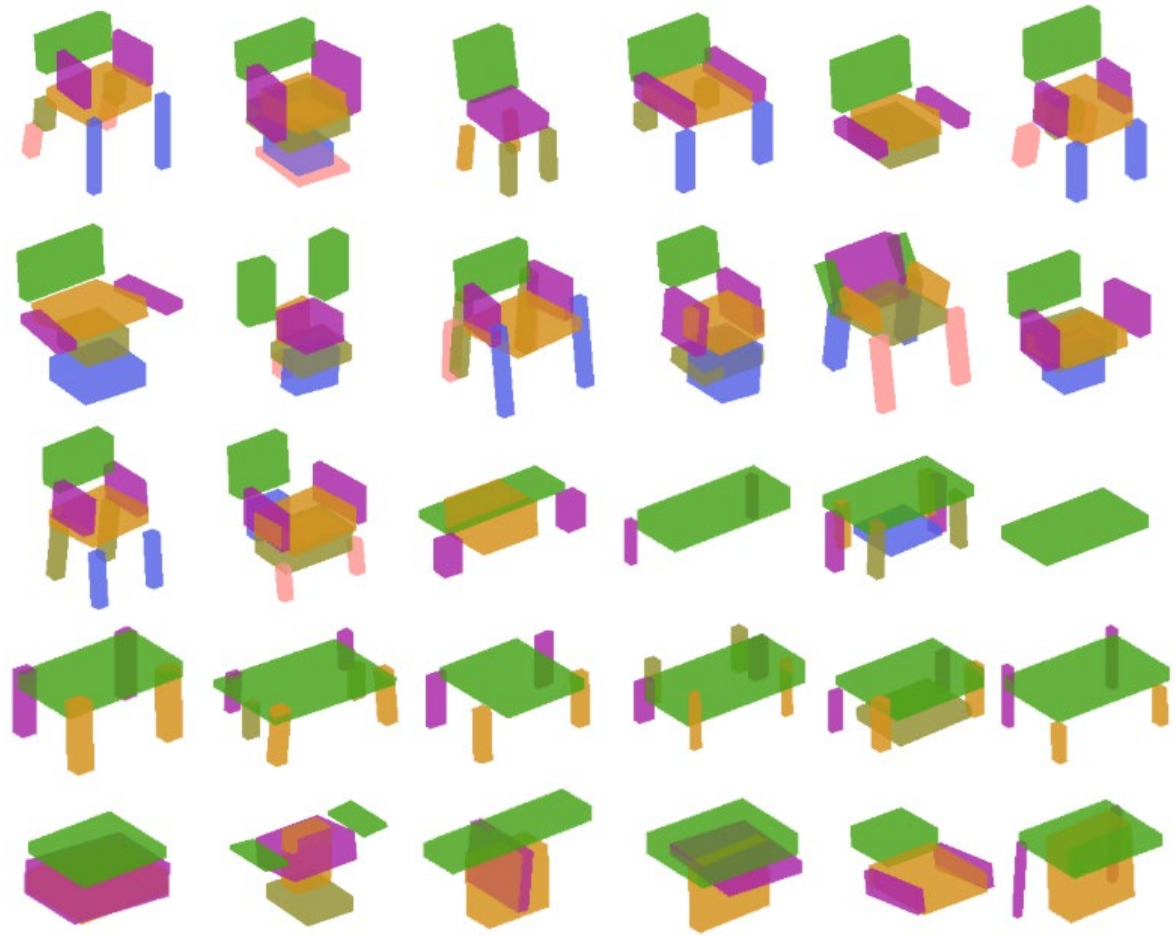
Translation (T)

Rotation (θ)

Network structure

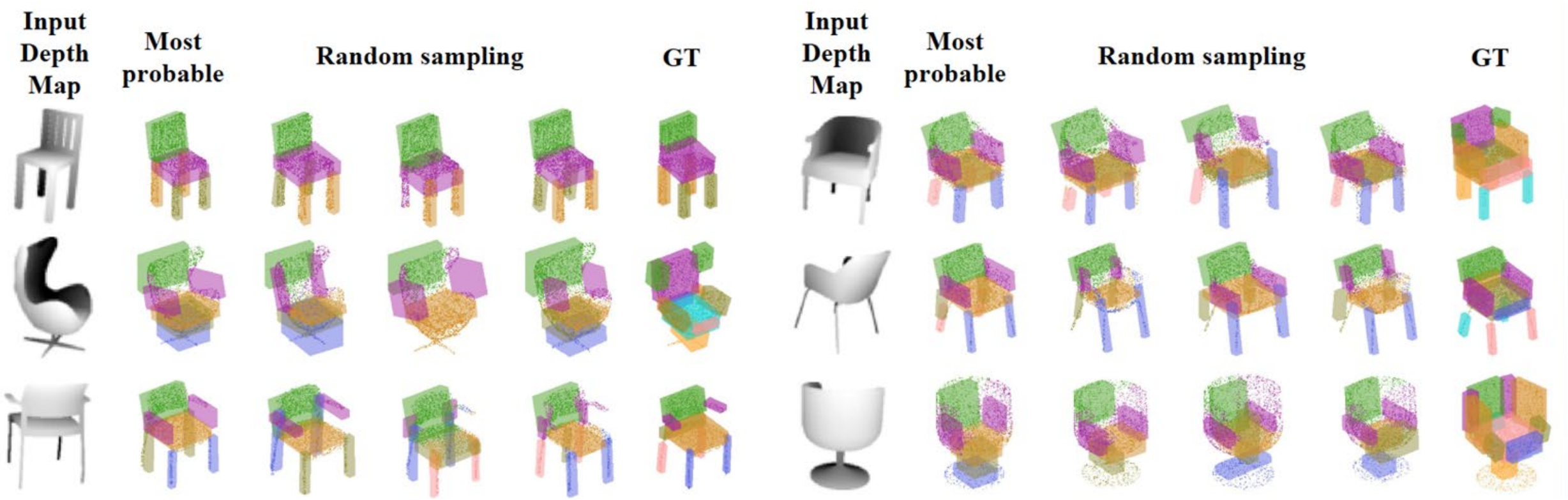


Experiment 1: shape synthesis

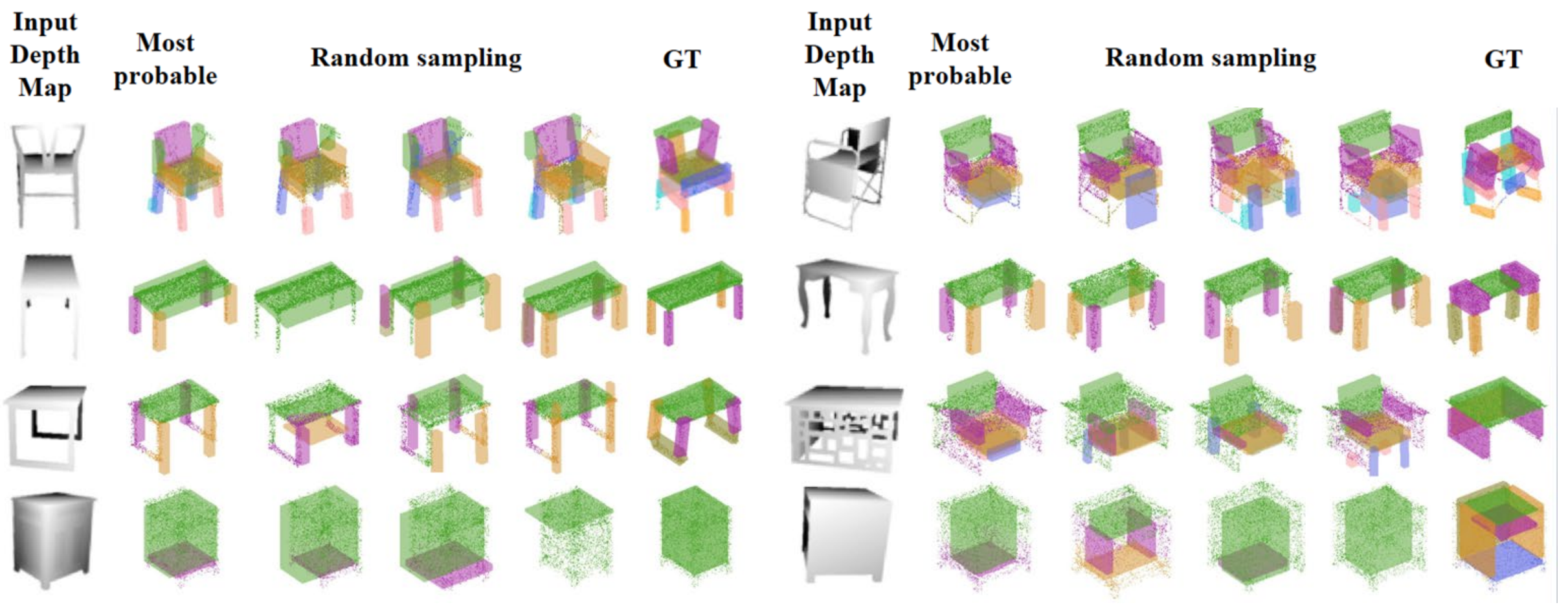


Random synthesis

Experiment 2: Shape reconstruction from single depth view on synthetic data (ModelNet10)



Experiment 2: Shape reconstruction from single depth view on synthetic data (ModelNet10)



Experiment 2: Shape reconstruction from single depth view on synthetic data (ModelNet10)

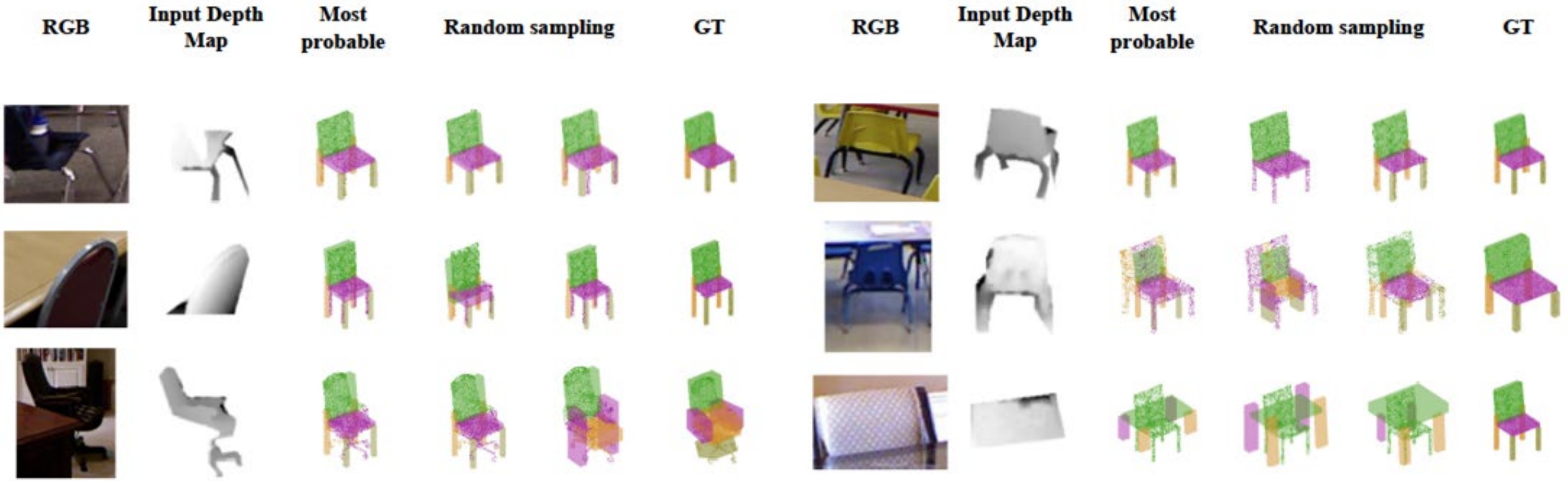
	chair	table	night stand
GT prim	0.473	0.533	0.657
NN Baseline	**0.269**	0.220	0.256
Wu et al. [40] (mean)	0.253	0.250	**0.295**
3D-PRNN	0.245	0.188	0.204
3D-PRNN + rot loss	0.238	0.**263**	0.266

Table 1. Shape IoU evaluation in synthetic depth map in ModelNet. We explore two settings of 3D-PRNN with or without rotation axis constrains, and compare it with ground truth primitive and the nearest neighbor baseline. We also compare to the Wu et al. [40] deep network voxel generation method.

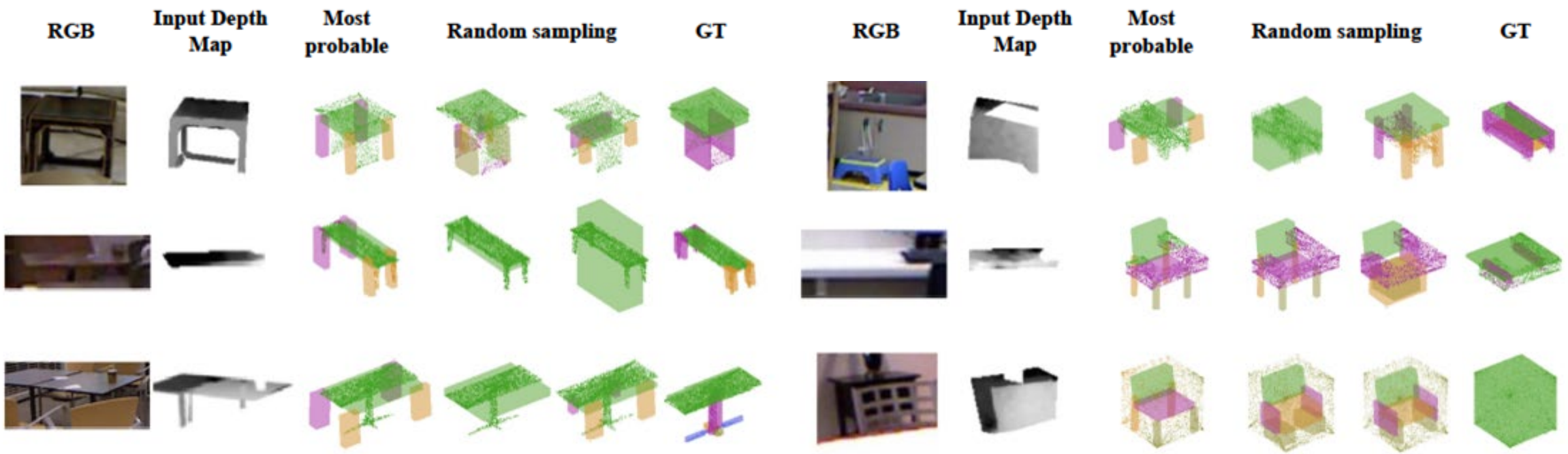
	chair	table	night stand
GT prim	0.049	0.044	0.044
NN baseline	0.075	0.089	0.100
Wu et al. [40] (mean)	**0.045**	**0.035**	**0.057**
3D-PRNN	0.074	0.080	0.104
3D-PRNN + rot loss	0.074	0.078	0.092

Table 2. Surface-to-surface distance evaluation in synthetic depth map in ModelNet. We explore two settings of 3D-PRNN with or without rotation axis constrains, and compare it with ground truth primitive and the nearest neighbor baseline.

Experiment 2: Shape reconstruction from single depth view on real data (NYUd v2)



Experiment 2: Shape reconstruction from single depth view on real data (NYUd v2)



Mini-conclusions

- Can generate part-based models of objects using RNN

Quick survey of additional works worth knowing

AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation

CVPR 2018

Thibault Groueix[1*], Matthew Fisher[2], Vladimir G. Kim[2], Bryan C. Russell[2], Mathieu Aubry[1]
[1]LIGM (UMR 8049), École des Ponts, UPE, [2]Adobe Research

http://imagine.enpc.fr/~groueixt/atlasnet/

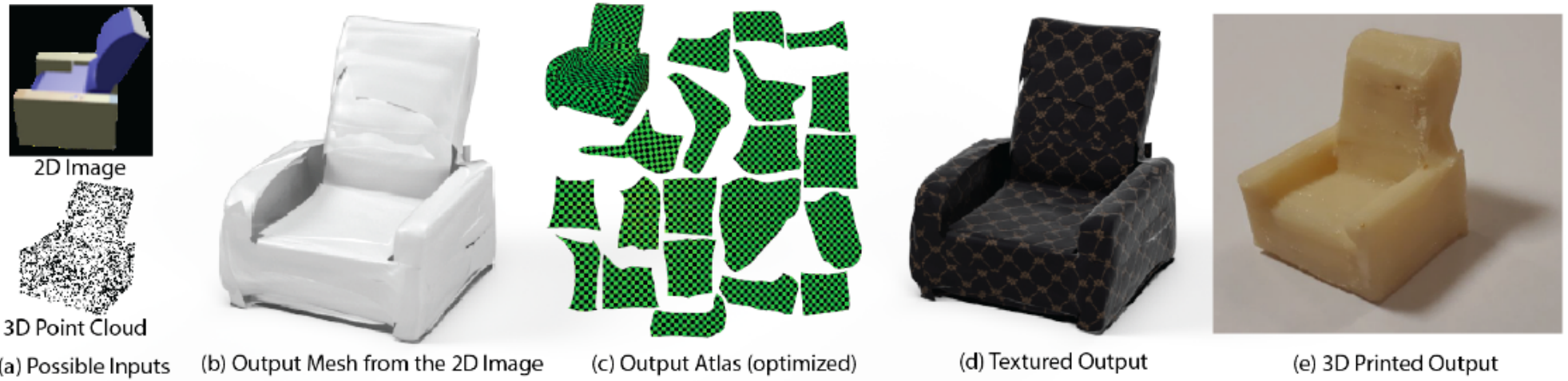
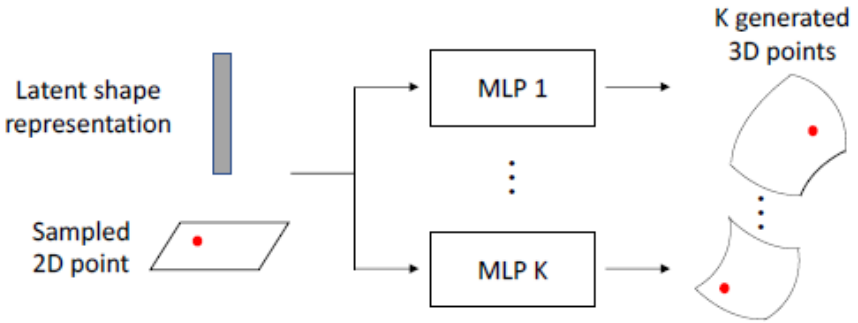


Figure 1. Given input as either a 2D image or a 3D point cloud (a), we automatically generate a corresponding 3D mesh (b) and its atlas parameterization (c). We can use the recovered mesh and atlas to apply texture to the output shape (d) as well as 3D print the results (e).

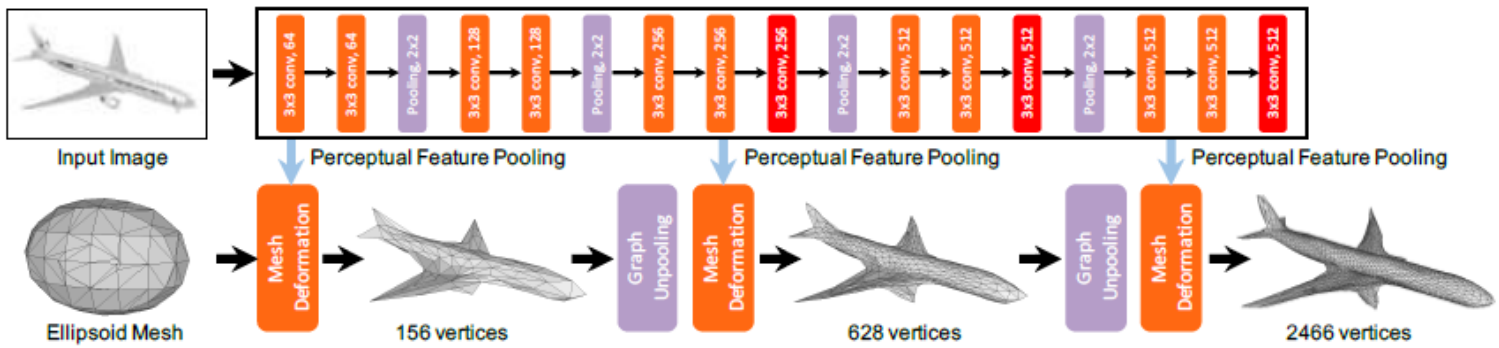
- Generate set of local parametric surfaces that are stitched together



Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images

ECCV 2018

Nanyang Wang[1]*, Yinda Zhang[2]*, Zhuwen Li[3]*,
Yanwei Fu[4], Wei Liu[5], Yu-Gang Jiang[1]†

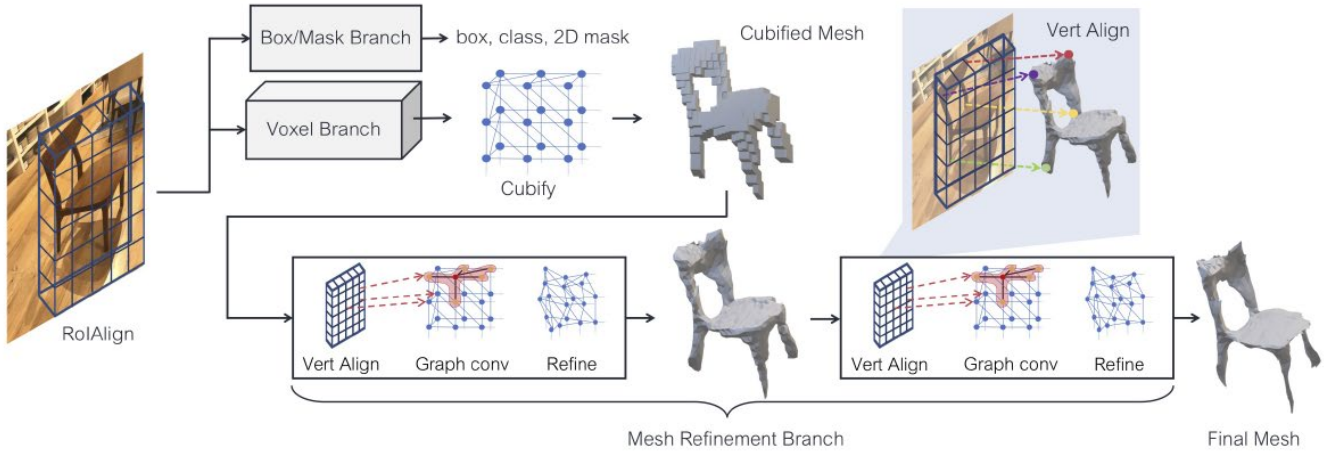


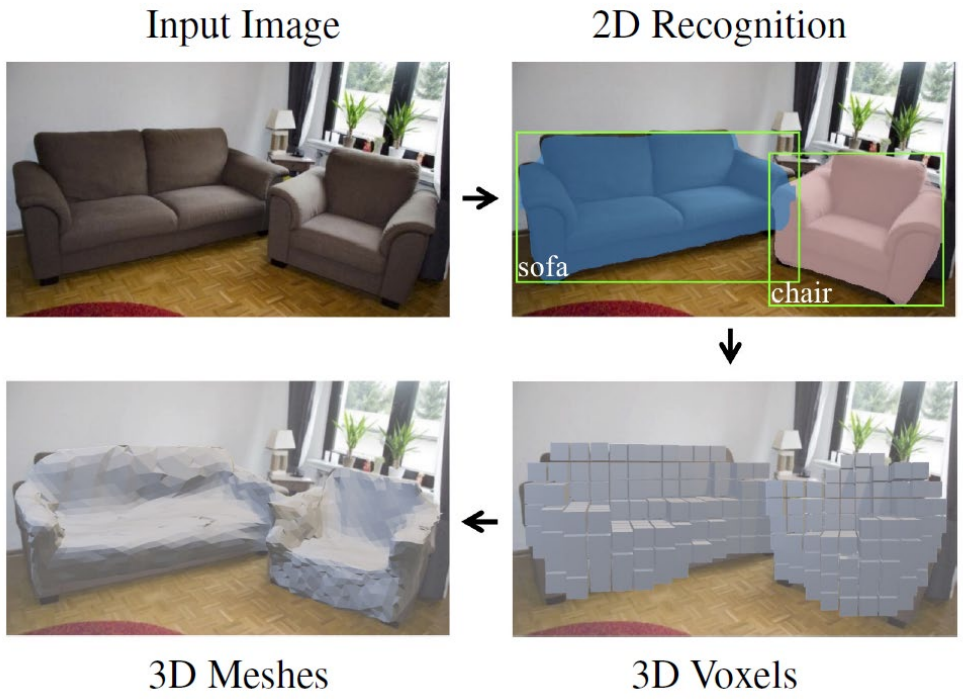
Mesh R-CNN

ICCV 2019

Georgia Gkioxari Jitendra Malik Justin Johnson

Facebook AI Research (FAIR)





Pix2Vox++: Multi-scale Context-aware 3D Object Reconstruction from Single and Multiple Images

Haozhe Xie[1,2] · Hongxun Yao[1] · Shengping Zhang[1,4] · Shangchen Zhou[3] · Wenxiu Sun[2]

2020

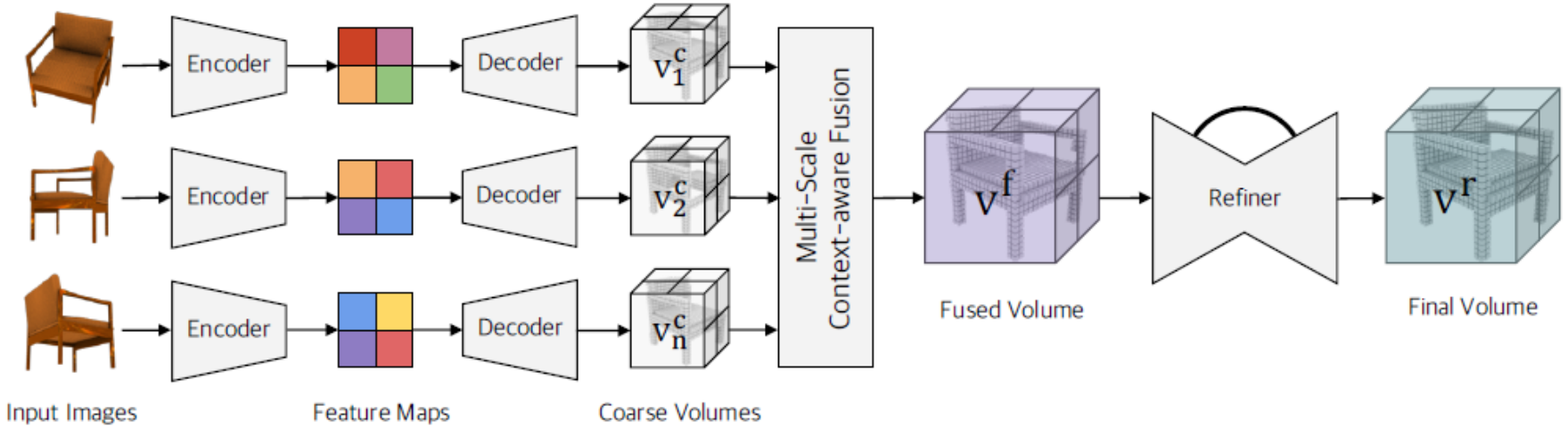
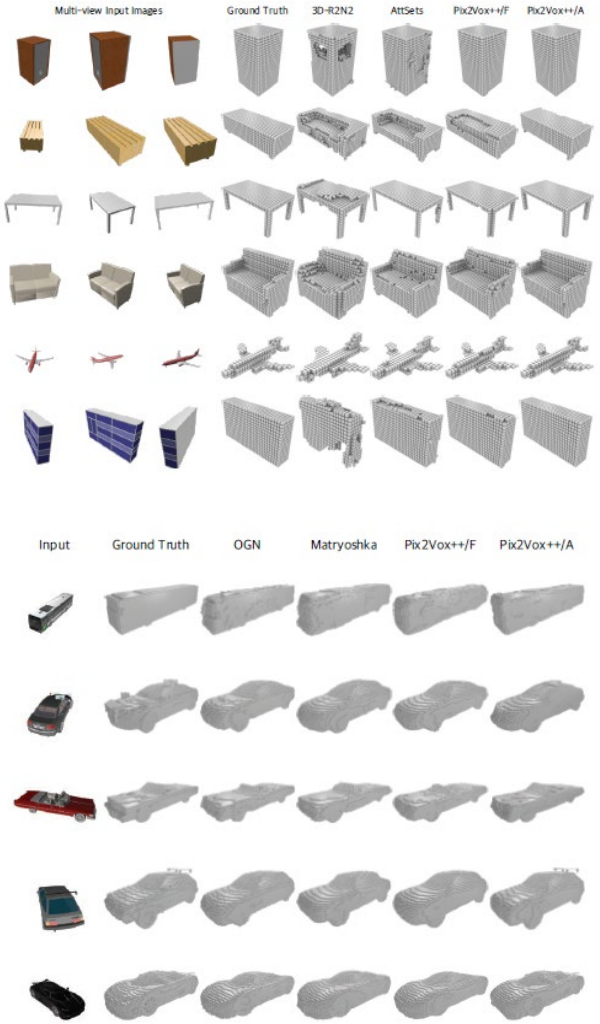


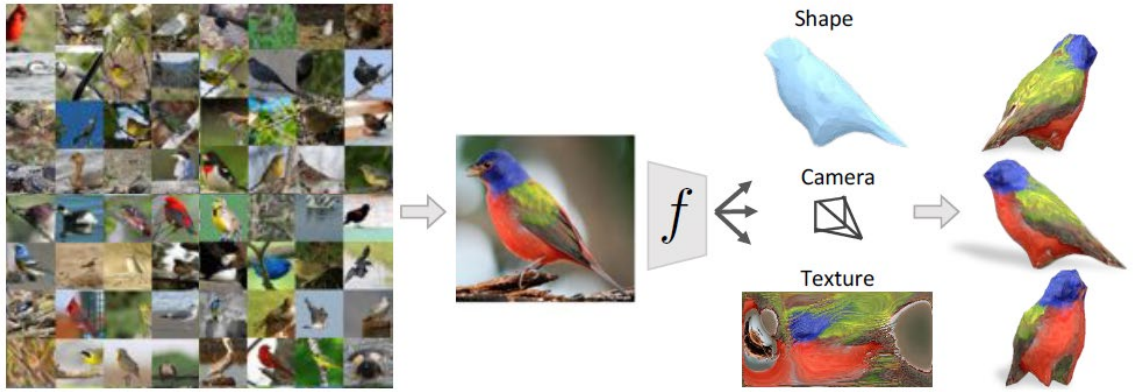
Fig. 1 Overview of the proposed Pix2Vox++. The network recovers the 3D shape of an object from arbitrary (uncalibrated) single or multiple images. The reconstruction result can be refined when more input images are available. Note that the weights

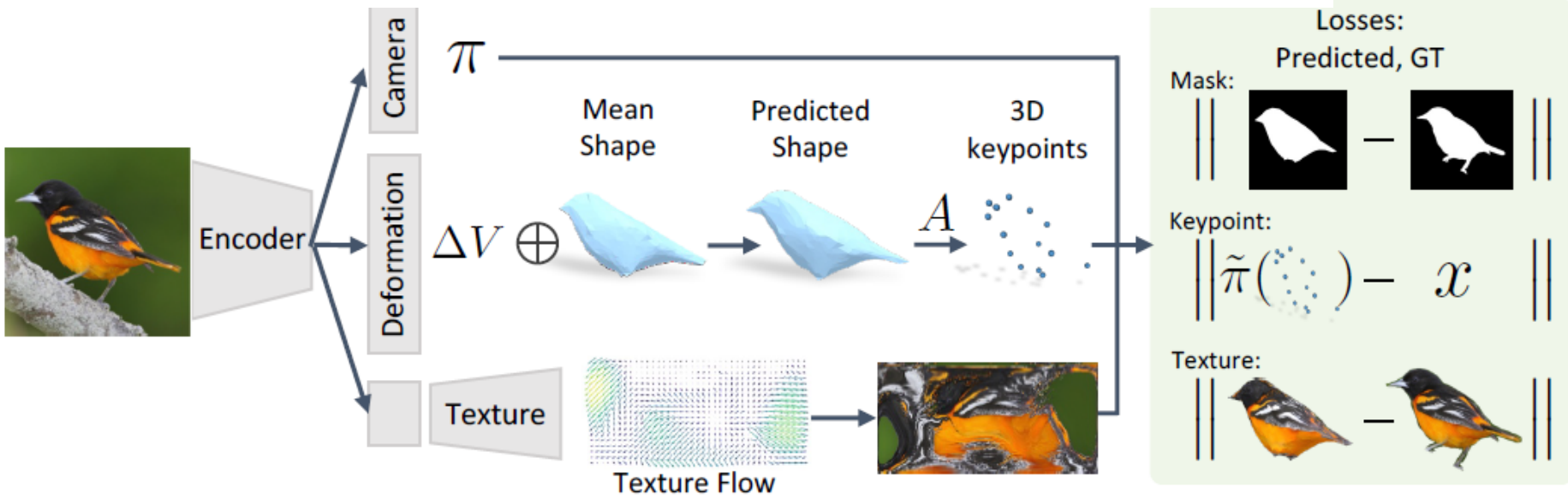


Learning Category-Specific Mesh Reconstruction from Image Collections

ECCV 2018

Angjoo Kanazawa*, Shubham Tulsiani*, Alexei A. Efros, Jitendra Malik





Things to remember

- Two different problem formulations lead to very different challenges
  1. Reconstruct a known object category in a canonical viewpoint
     - Relatively easy to solve via retrieval, so research focuses on learning in a loosely supervised way
  2. Reconstruct any object in the current viewpoint
     - Harder to solve, so not as many people work on it
     - Good solution may still factor shape and pose

- Many shape representations have been tried: voxels, multiview depth, deformed sphere, mesh, multiple local surfaces, primitives