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

1 RGB-based shape prediction

1.1 Synthetic training examples

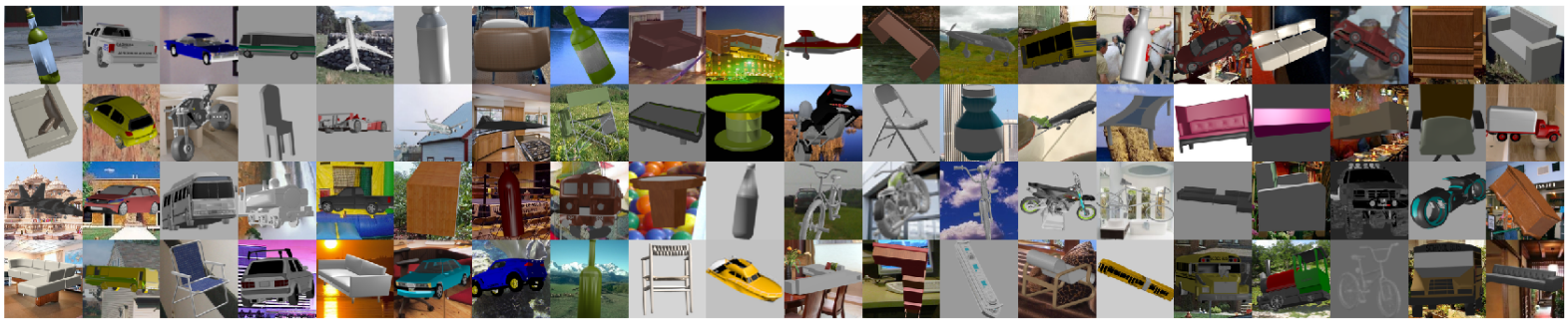


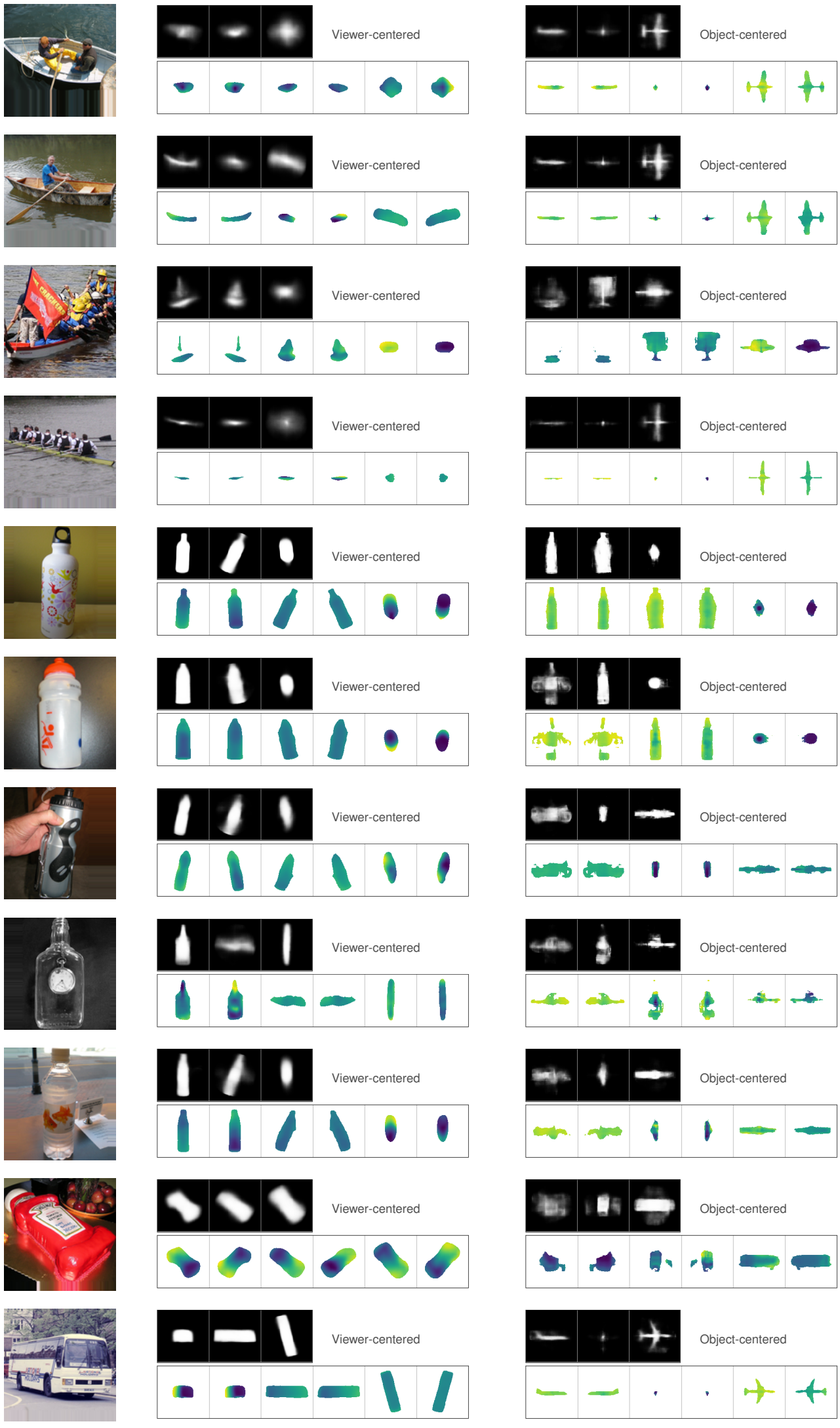
Figure 1: Sample RGB training images generated using RenderForCNN. Our dataset has 2.4M renderings of 34,000 3D CAD models from 12 object categories. In the RGB multi-surface experiment, each example has six output ground truth depth images rendered on the faces of a cube: front, back, left, right, top, bottom.

Figure 2: Viewer-centered (top) and object-centered (bottom) output ground truth depth images for the corresponding input images in Figure 1. Only showing first view in each example.

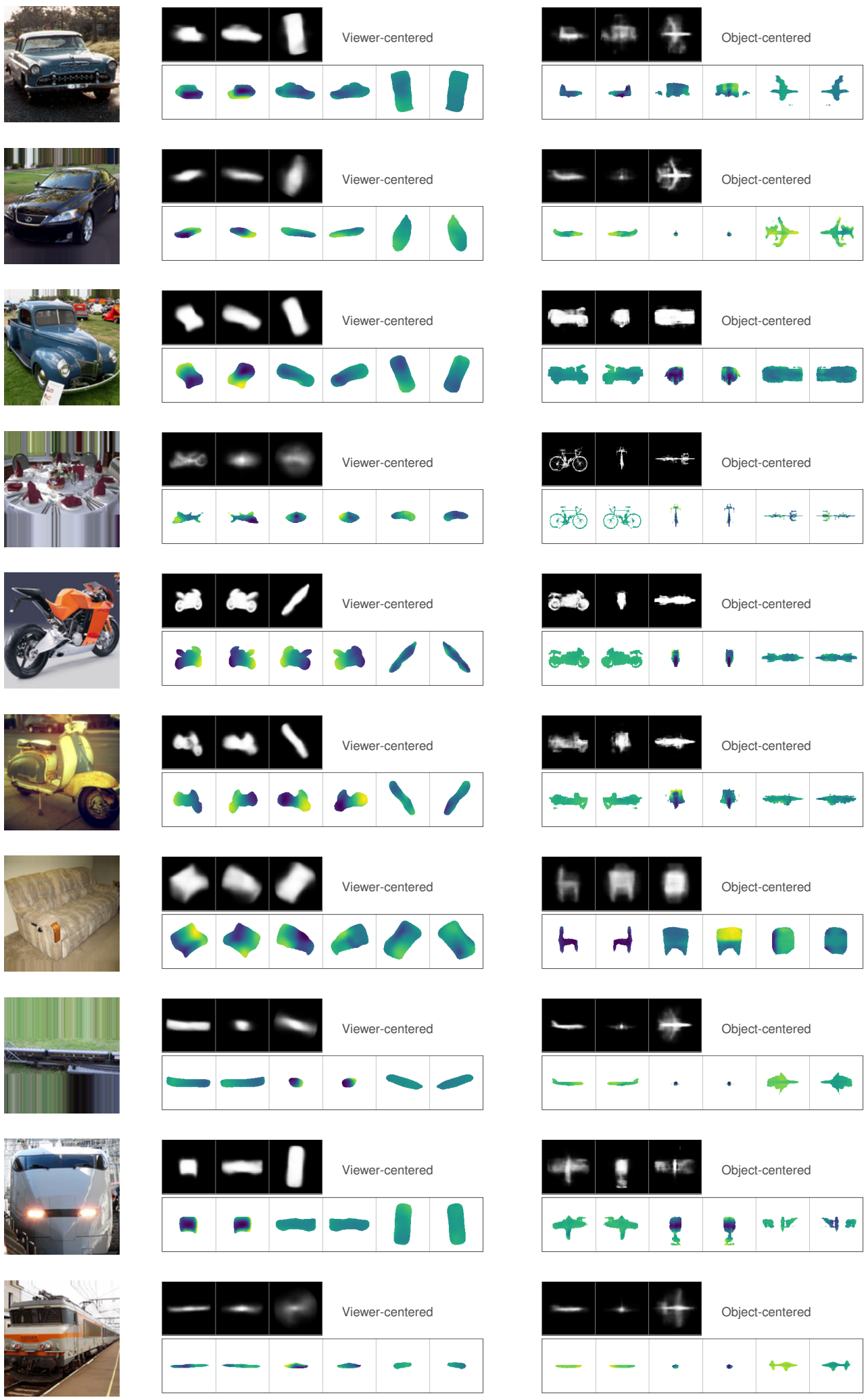
1.2 Viewer-centered and object-centered coordinates

Does object-centered representation make shape prediction more of a categorization problem? Qualitative results support our initial hypothesis that object-centered models tend to correspond more directly to category recognition. Object-centered model often predicts a shape that looks good but is in an entirely different object category than the input image (Section 1.2.1). The viewer-centered model tends not to make these kinds of mistakes and, instead, errors tend to be overly simplified shapes or incorrect poses (Section 1.2.3).

1.2.1 Object-centered failure examples



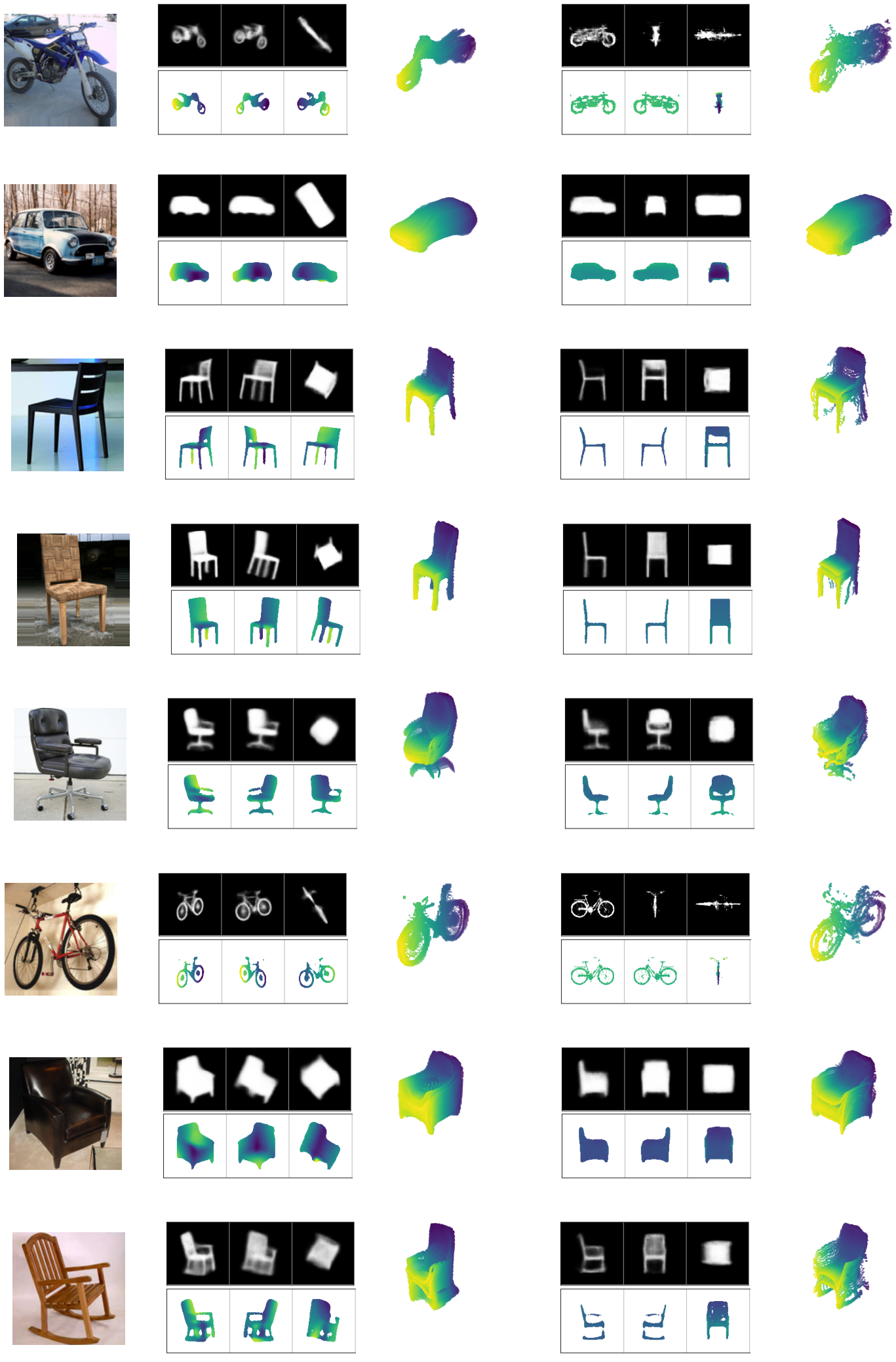
Input **Multi-surface Pred.**



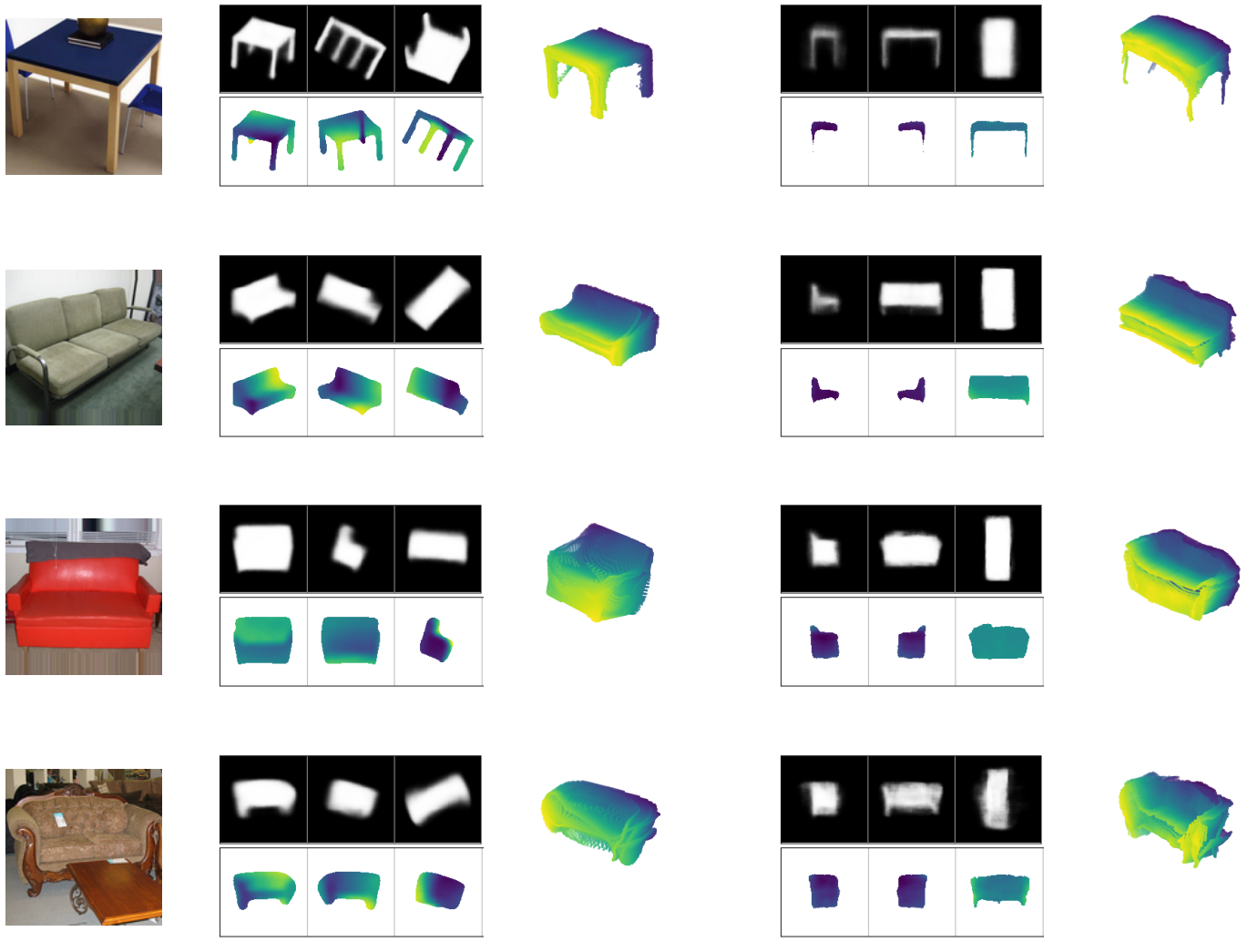
1.2.2 Viewer-centered failure examples

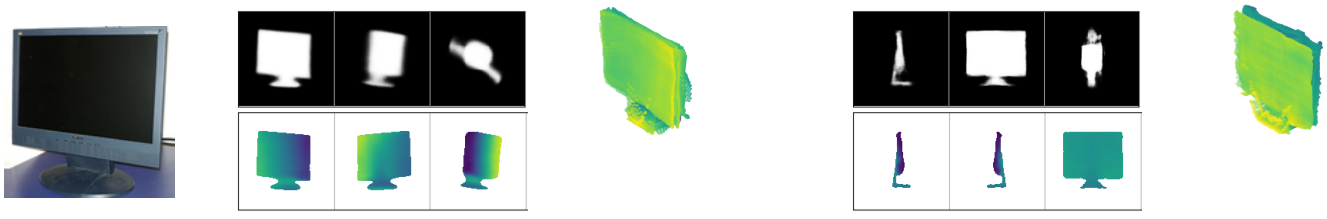
1.2.3 Successful examples

This section shows selected examples that are successful in both coordinate systems.



Input | Viewer-centered Pred. | Object-centered Pred.





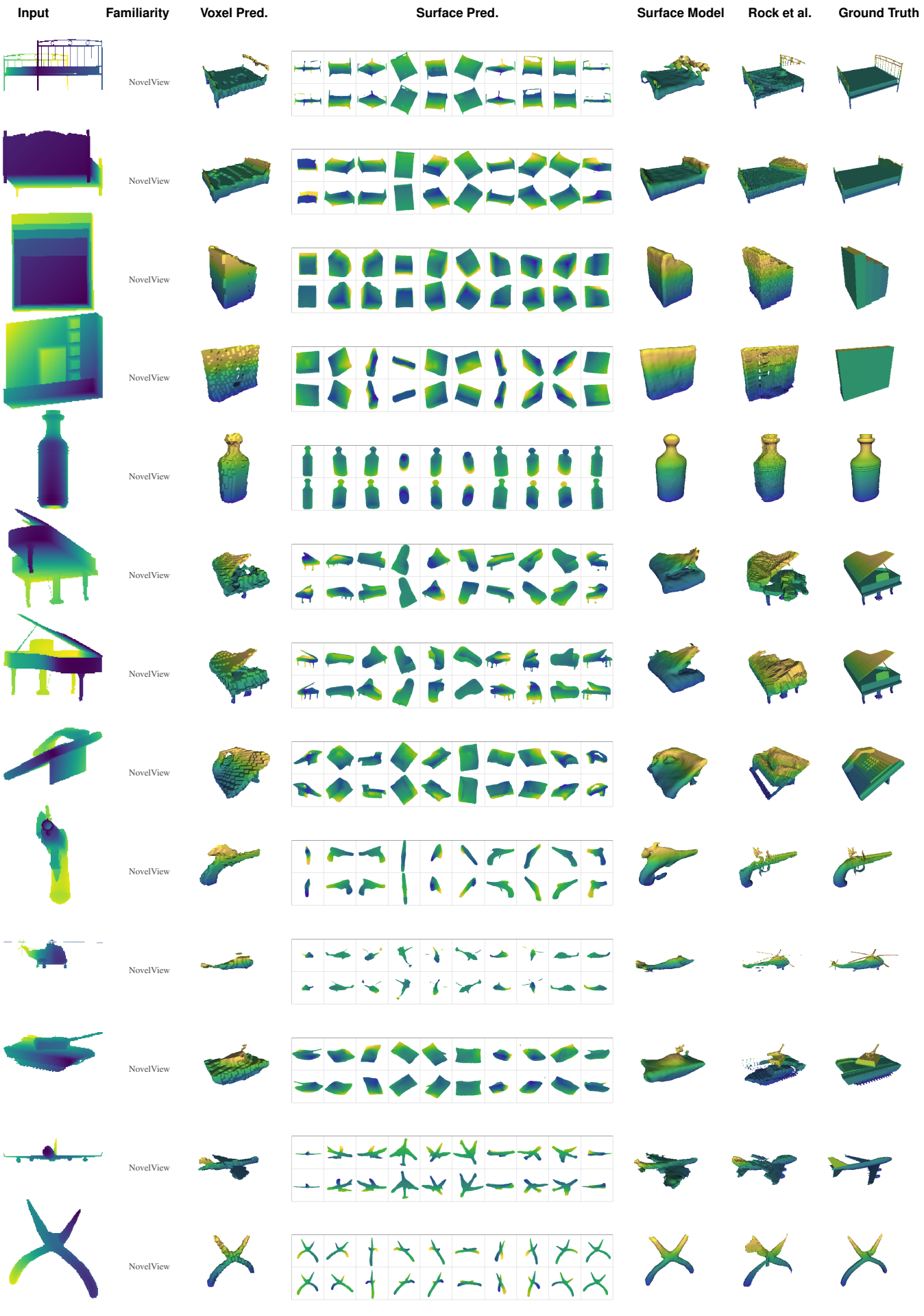
2 Depth-based shape prediction

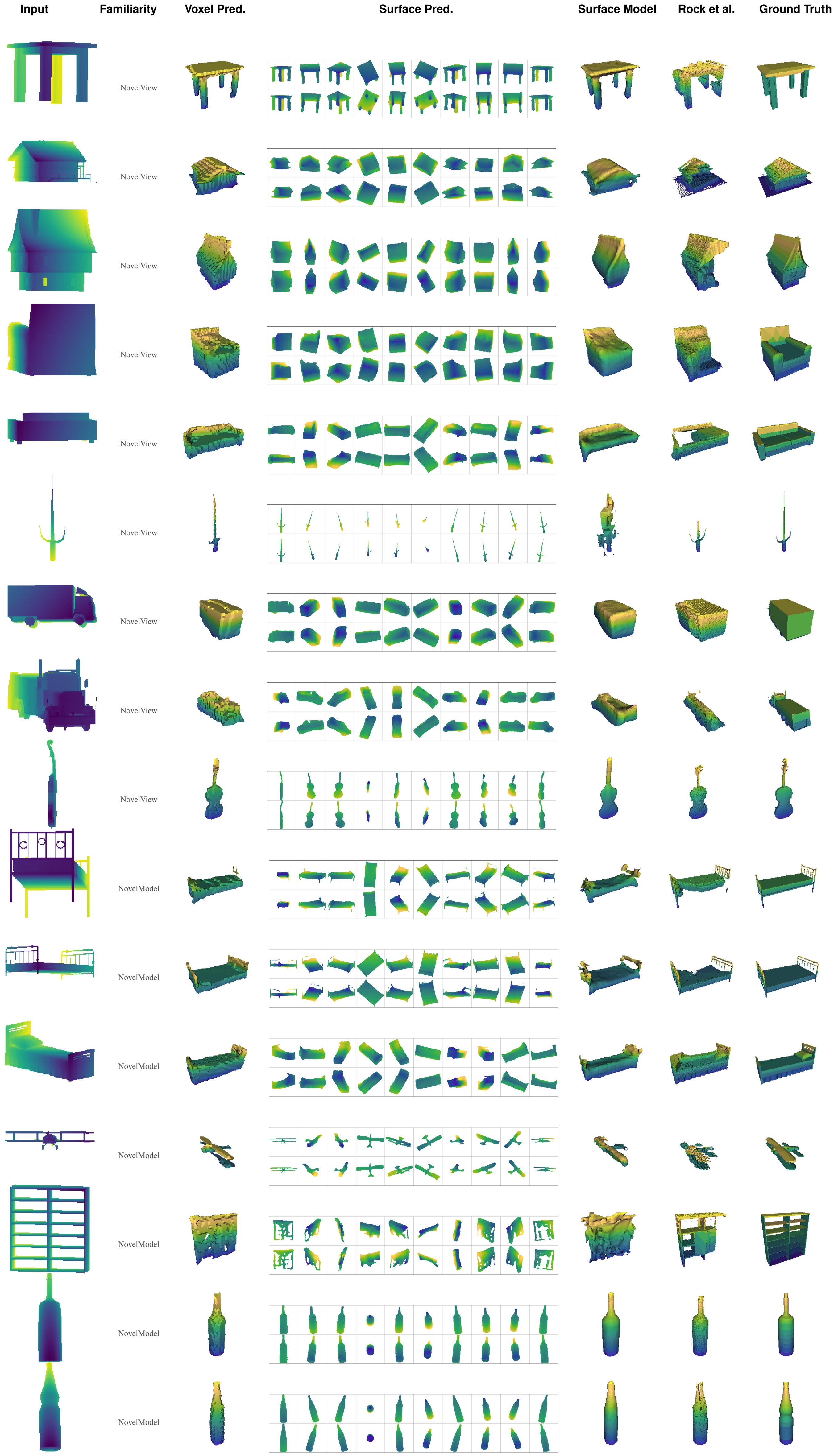
2.1 Multi-surface vs. voxel shape representations

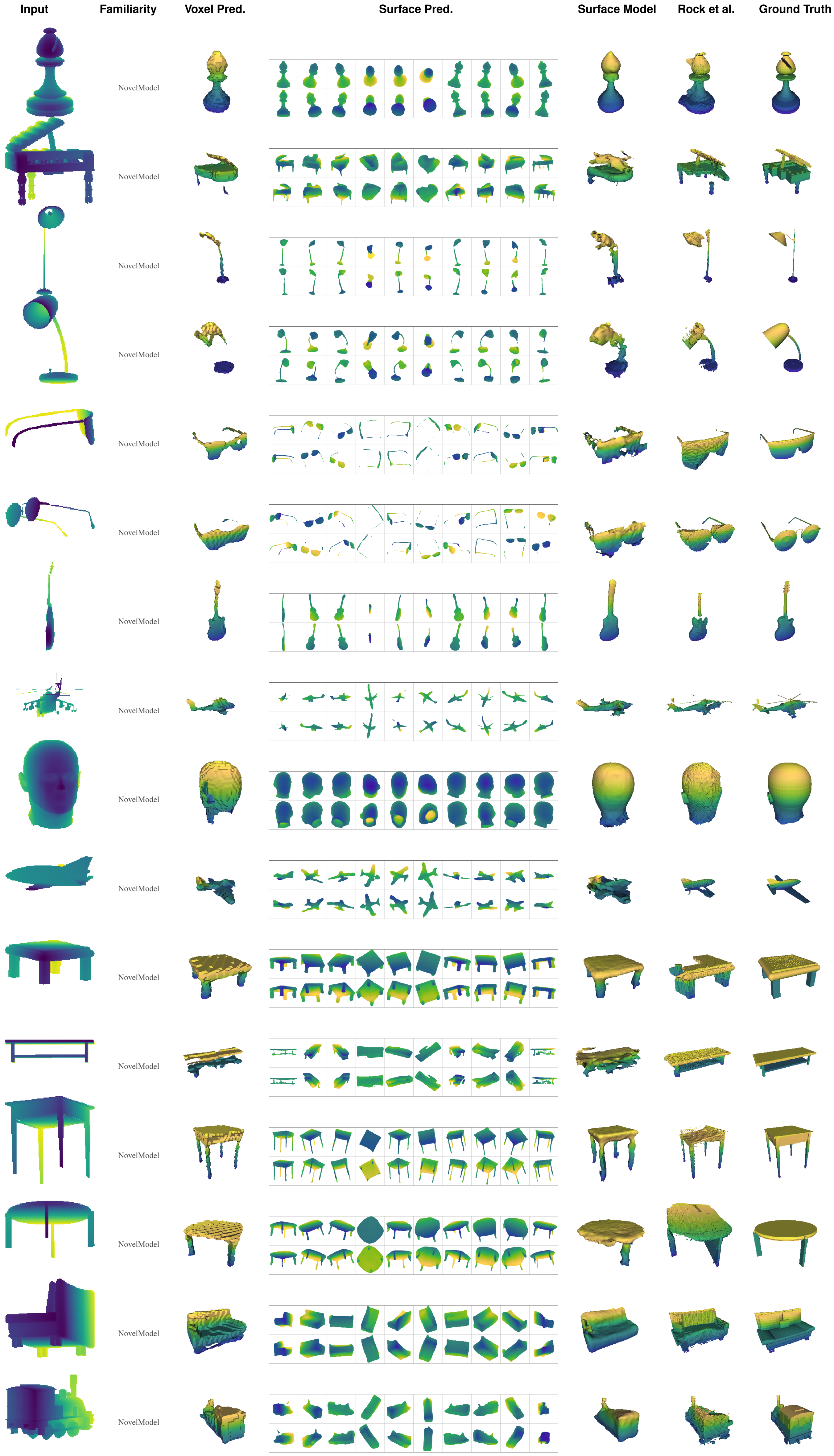
Categories in the SHREC'12 dataset:

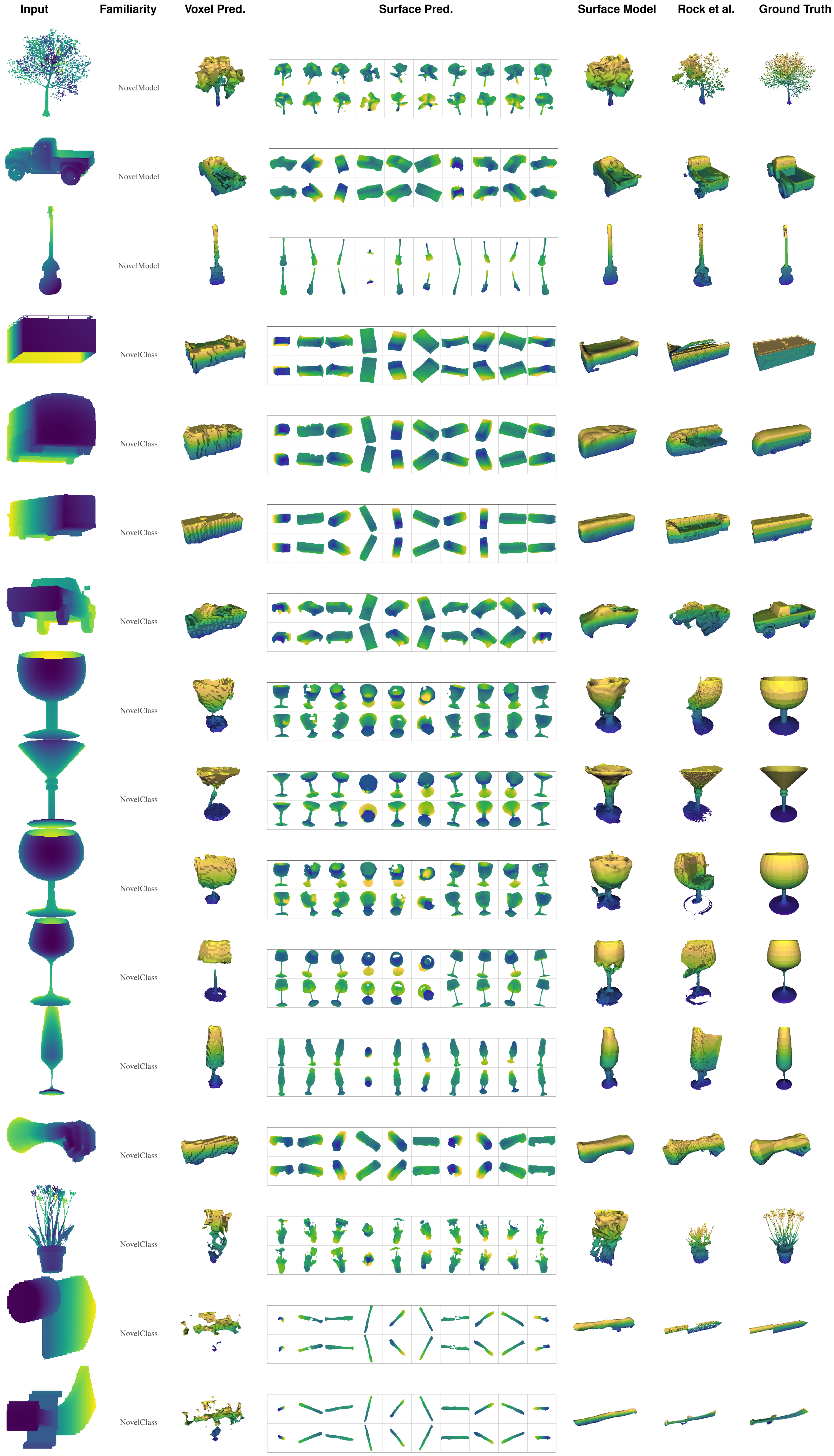
- Train+val: bed, biplane, bookset, bookshelf, cellphone, city, classicpiano, computer, computerkeyboard, desklamp, door, face, glasses, guitar, handgun, helicopter, militaryvehicle, monitor, monoplane, mug, plier, quadruped, rectangletable, roundtable, singlehouse, skyscraper, sofa, spoon, tree, violin, bicycle, biped, fish, floorlamp, flyinginsect, bird, bottle, chess, deskphone, drum, humanhead, sword, train, truck, wheelchair
- Test (NovelClass): apartmenthouse, bus, car, cup, hand, homeplant, knife, motocycle, nonflyinginsect, nonwheelchair, pianoboard, rocket, ship, submachinegun, trucknoncontainer

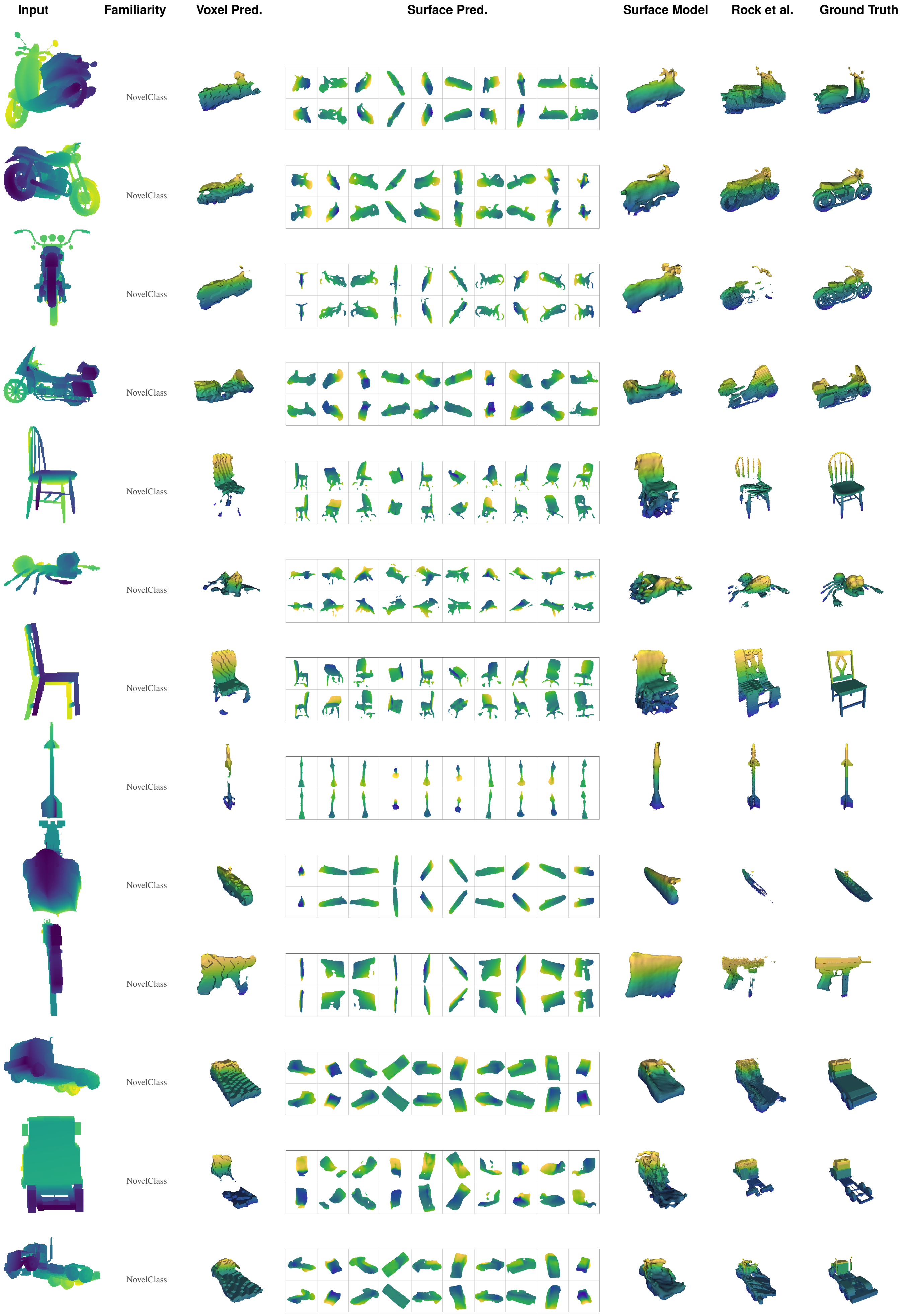
Qualitative reconstruction results:




Input
Familiarity
Voxel Pred.
Surface Pred.
Surface Model
Rock et al.
Ground Truth
NovelView
NovelView
NovelView
NovelView
NovelView
NovelView
NovelView
NovelView
NovelView
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel


Input
Familiarity
Voxel Pred.
Surface Pred.
Surface Model
Rock et al.
Ground Truth
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel
NovelModel


Input
Familiarity
Voxel Pred.
Surface Pred.
Surface Model
Rock et al.
Ground Truth
NovelModel
NovelModel
NovelModel
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass


Input
Familiarity
Voxel Pred.
Surface Pred.
Surface Model
Rock et al.
Ground Truth
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass
NovelClass

3 Network Architecture

layer	output size	kernel	stride	repeats
Input depth	128x128x1			1
conv2d	64x64x48	7	2	1
residual unit	64x64x48	3	1	3
residual unit	32x32x144	3	2	1
residual unit	32x32x144	3	1	3
residual unit	16x16x288	3	2	1

layer	output size	kernel	stride	repeats
Input silhouette	128x128x1			1
conv2d	64x64x32	7	2	1
residual unit	64x64x32	3	1	3
residual unit	32x32x96	3	2	1
residual unit	32x32x96	3	1	3
residual unit	16x16x192	3	2	1

layer	output size	kernel	stride	repeats
residual unit	16x16x480	3	1	10
residual unit	8x8x256	3	2	1
residual unit	8x8x256	3	1	4
residual unit	4x4x512	3	2	1
residual unit	4x4x512	3	1	4
FC $\rightarrow h$	4096			1

Table 1: Top to bottom: Network parameters for encoders E_d, E_s, E_h. ReLU and batch normalization layers are not shown.

layer	output size	kernel	stride	repeats
upconv2d	8x8x256	5	2	1
upconv2d	16x16x128	5	2	1
upconv2d	32x32x64	5	2	1
upconv2d	64x64x32	5	2	1
upconv2d	128x128x1	5	2	1

Table 2: Network parameters for multi-surface decoders G_d, G_s.

layer	output size	kernel	stride	repeats
upconv3d	6x6x6x512	5	2	1
upconv3d	12x12x12x256	5	2	1
upconv3d	12x12x12x128	5	1	1
upconv3d	24x24x24x64	5	2	1
upconv3d	48x48x48x1	5	2	1

Table 3: Network parameters for the volume decoder used in the voxel network.

layer	output size	repeats
$h \rightarrow$ FC	256	1
FC, softmax	40	1

Table 4: Network parameters for the classification layer in the 2.5D shape classification experiment. The classification experiment uses 370K renderings of 3D CAD models from the ModelNet40 dataset and is evaluated using 80/20 train/test model split ratio.

layer	output size	kernel	stride	repeats
Input image	128x128x3			1
conv2d	64x64x64	7	2	1
max pool	32x32x64	3	2	1
residual unit	32x32x256	3	1	4
residual unit	16x16x768	3	2	7
residual unit	8x8x2048	3	2	4
average pool	1x1x2048	8	8	1

Table 5: Network parameters for the 6-view network encoder.

layer	output size	kernel	stride	repeats
upconv2d	4x4x512	4	2	1
upconv2d	8x8x256	4	2	1
upconv2d	16x16x128	4	2	1
upconv2d	32x32x64	4	2	1
upconv2d	64x64x32	4	2	1
upconv2d	128x128x1	4	2	1

Table 6: Network parameters for multi-surface decoders G_d, G_s used in the 6-view network.

layer	output size	kernel	stride	repeats
upconv3d	4x4x4x512	4	2	1
upconv3d	8x8x8x256	4	2	1
upconv3d	16x16x16x128	4	2	1
upconv3d	32x32x32x1	4	2	1

Table 7: Network parameters for the volume decoder used in the voxel network (corresponds to Tables 5 and 6).