Inverse Graphics GAN: Supplemental Material

A. FID Scores Calculated with Inception Network Trained on ImageNet Classification

In the main paper all FID scores were calculated using an Inception network which was trained to classify gray-scale ShapeNet renders that look similar to the training data used for all of our models. Below we report, for the same set of experiments, FID scores calculated using the traditional ImageNet trained Inception network.

Table 1. Equivalent to Table 1 in the main paper. ImageNet FID scores computed on ShapeNet objects (bathtubs, couches and chairs).

# of Images	500			**One Per Model** ($\approx$ 3000)			**Unlimited**			
Dataset	**Tubs**	**Couches**	**Chairs**	**Tubs**	**Couches**	**Chairs**	**Tubs**	**Couches**	**Chairs**	**LVP**
2D-DCGAN	356.9	324.6	291.9	211.0	156.5	196.8	210.1	117.8	78.8	101.5[1]
Visual Hull	117.3	130.0	153.2	66.5	78.7	47.3	29.7	41.4	**22.0**	36.5
Absorbtion Only	117.6	115.9	**110.7**	74.3	70.3	46.6	37.6	36.6	31.4	41.8
IG-GAN (Ours)	**95.1**	**91.4**	131.4	**41.1**	**36.5**	**32.1**	**23.4**	**18.6**	22.6	**29.2**

Table 2. Equivalent to Table 2 in the main paper. Ablation results without discriminator output matching (DOM) when training on chairs/couches "one per model" datasets. We either fix the pre-trained neural renderer ("Fixed"), or continuing to train it during GAN training ("Retrained"). The generator samples fed to the discriminator are rendered using either OpenGL or the neural renderer. For reference, our model is equivalent to the Retrained OpenGL setup with the addition of the DOM loss and achieves FID scores 32.1/36.5. FID scores calculated using an Inception network trained on ImageNet.

	OpenGL	RenderNet
Retrained	93.6/146.6	58.1/88.2
Fixed	149.3/238.6	61.6/100.0

Table 3. Equivalent to Table 3 in the main paper. Comparisons of neural renderer pre-trainings on different 3D shapes. FIDs are reported for the 'One per model' chairs. FID scores calculated using an Inception network trained on ImageNet.

	Chairs	Random	Tables
Ours	32.7	31.4	32.1
Fixed	43.0	84.4	69.5

B. A note on Emission-Absorption

We chose to compare to the Absorption-only (AO) model from Henzler et al. (2019) and not the Emission-Absorption (EA) model. The EA model was designed to incorporate color information into the differentiable rendering engine. In addition to the occupancy/absorbtion value generated at each voxel, this model also generates one or more emission values at each voxel that can represent either 3-channel color, or a single grey-scale value. The focus of our paper was only on shape, however, leaving color generation for future work. Thus the underlying ShapeNet voxel data used in our experiments does not have any color channel information, and consists of only a single 0-1 occupancy value for each voxel. Therefore, including the additional emission value would only result in providing the EA model additional freedom that it should not use when modeling the data. In the following we show that if we assume that the model generates only a single occupancy channel and the emitted color is fixed globally, then the EA model naturally reduces to the AO model.

[1] A fair comparison to 2D-DCGAN is impossible, as the generator is trained on LVP (Limited View Point) data, but to facilitate easy comparison all FID evaluations are computed with same test data (which includes views from all 360°).

In Henzler et al. (2019) the expression

$$\rho_{EA}(\mathbf{v}) = \frac{\sum_{i=1}^{n_z} v_{a,i} v_{e,i} \prod_{j=1}^{i}(1 - v_{a,j})}{\sum_{i=1}^{n_z} v_{a,i} \prod_{j=1}^{i}(1 - v_{a,j}) + \epsilon} \left[1 - \prod_{j=1}^{n_z}(1 - v_{a,j}) \right]$$

is used for the Emission-Absorption model, where $v_{e,j}$ denotes the emission coefficient, $v_{e,j}$ the absorption, n_z the number of voxels along the chosen dimension and the index j refers to the j-th occupancy of the 3D model along a straight line through the volume. The regularization parameter ϵ is chosen small to numerically stabilize the quotient. In the case of data generated from shape information alone, we can consider that all objects are perfectly white, which would equate to $v_{e,j} = 1$. In this case, the quotient

$$\frac{\sum_{i=1}^{n_z} v_{a,i} v_{e,i} \prod_{j=1}^{i}(1 - v_{a,j})}{\sum_{i=1}^{n_z} v_{a,i} \prod_{j=1}^{i}(1 - v_{a,j})}$$

naturally reduces to one, leaving the expression

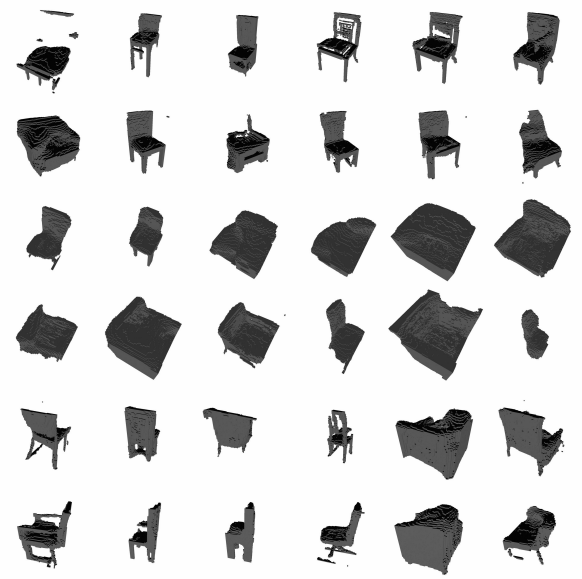
$$\rho_{EA}(\mathbf{v}) = 1 - \prod_{j=1}^{n_z}(1 - v_{a,j}),$$

for the EA model, which is identical to the AO model. Hence, the two imaging models are the same in our setting of where images are obtained from 3D shapes with a single globally emitted color.

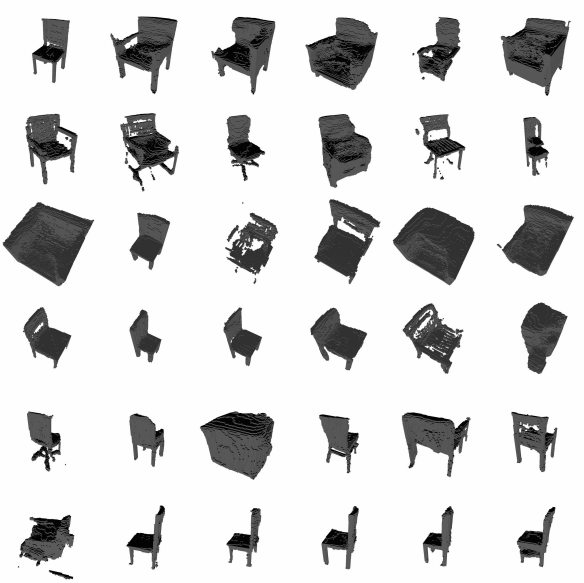
C. Random Samples from Each Model Trained on Each of the Datasets

The rest of the supplemental contains a set of tables where each table contains random samples from one of the models trained on one of the datasets. For each set of random samples we show black and white renders as well as normal map renders. In each figure the view angle is held fixed across all samples in a single row, but each image represents a completely independently sampled underlying 3D model. Note that we cannot show normal map renderers for 2D-DCGAN generations since the 2D-DCGAN only generates images, not 3D models. At the end we also show samples from the dataset rendered in the same way for comparison.

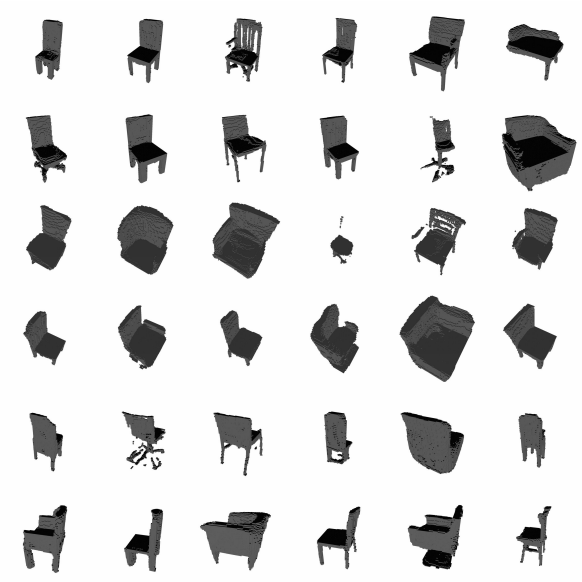
(a) 2D-DCGAN



(b) Absorption Only

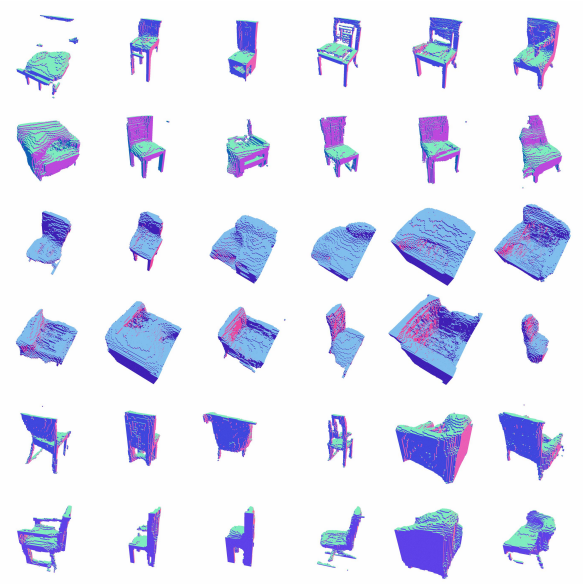


(c) Visual Hull

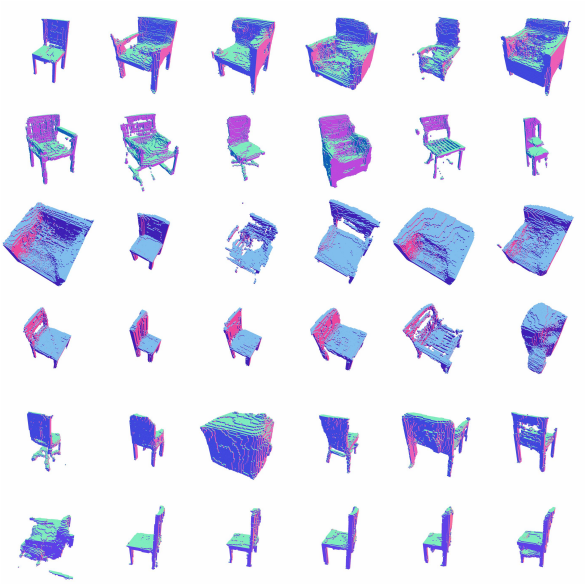


(d) IG-GAN (Ours)

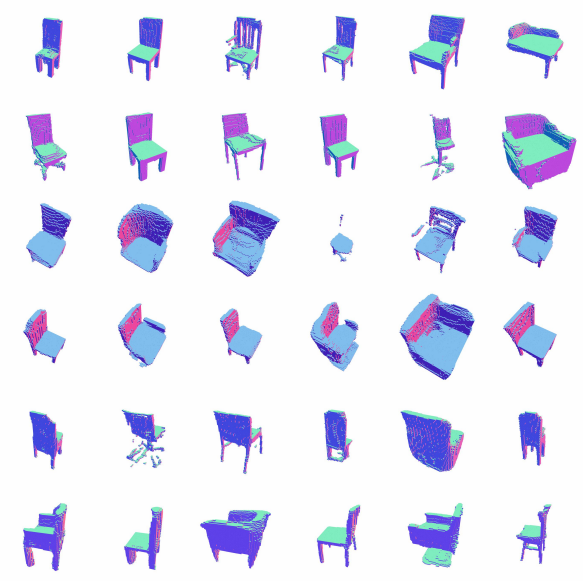
Figure 1. Samples from models trained on the Chairs data in the 'one sample per object' setting (6667 training images).



(a) Absorption Only

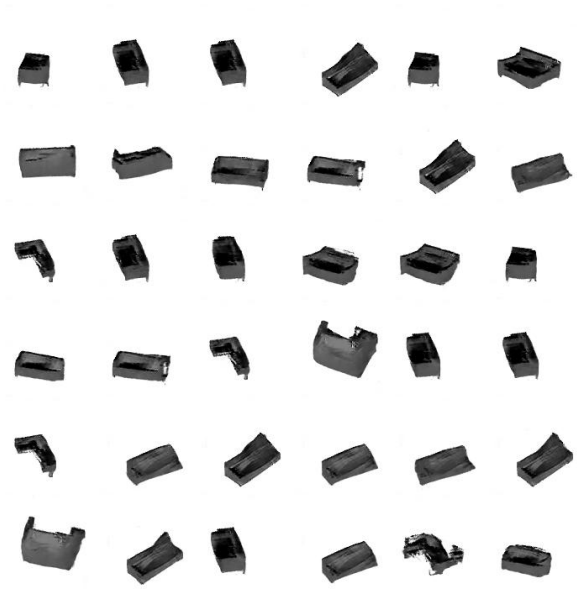


(b) Visual Hull

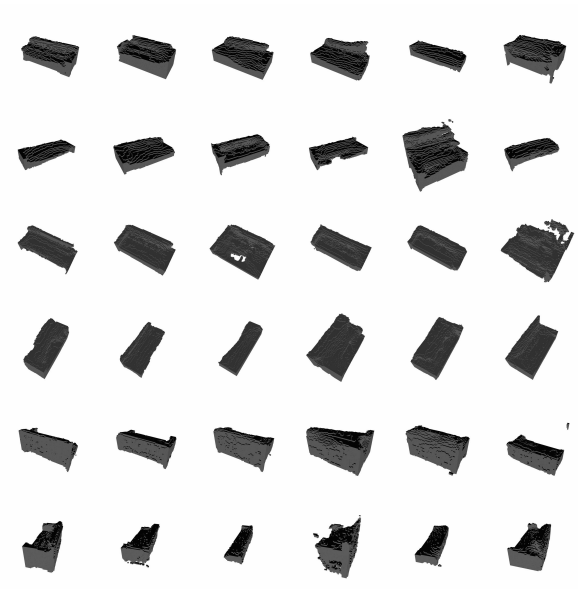


(c) IG-GAN (Ours)

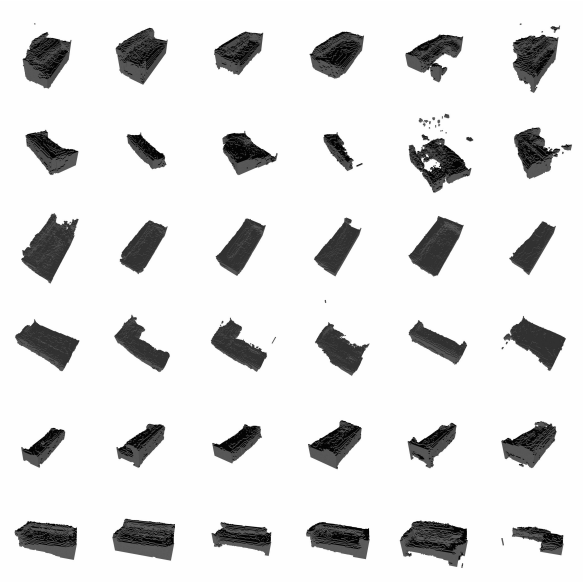
Figure 2. Samples from models trained on the Chairs data in the 'one sample per object' setting (6667 training images), r,endered as normal maps.



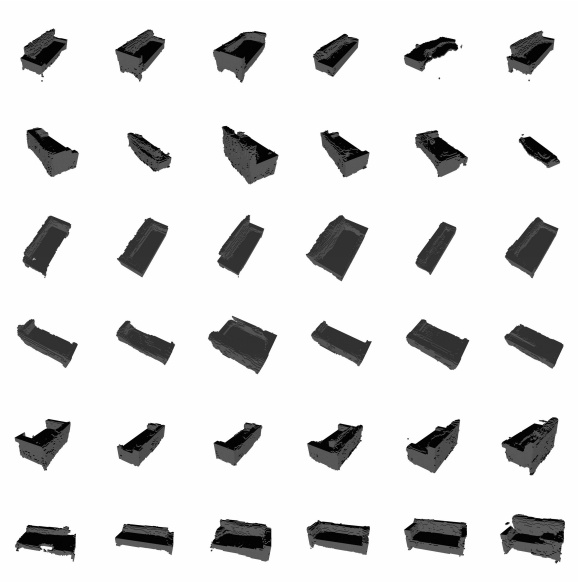
(a) 2D-DCGAN



(b) Absorption Only

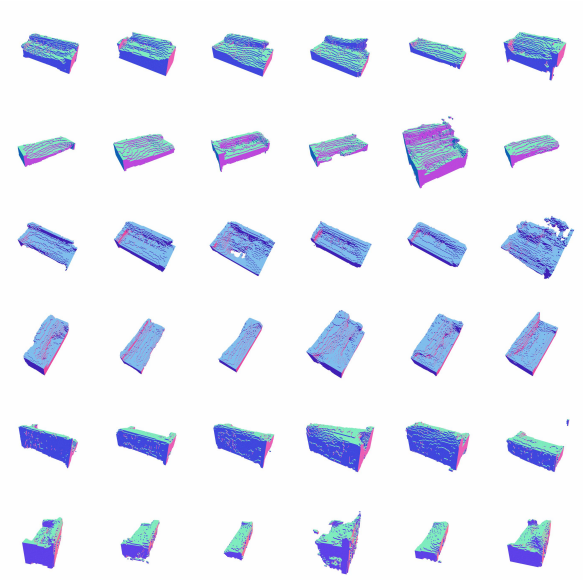


(c) Visual Hull

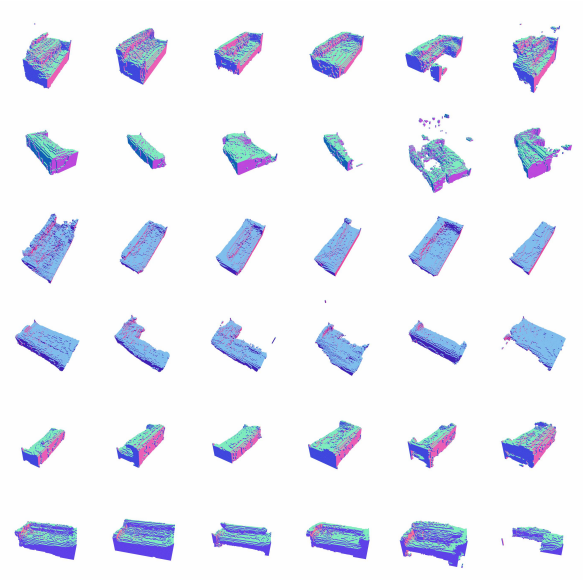


(d) IG-GAN (Ours)

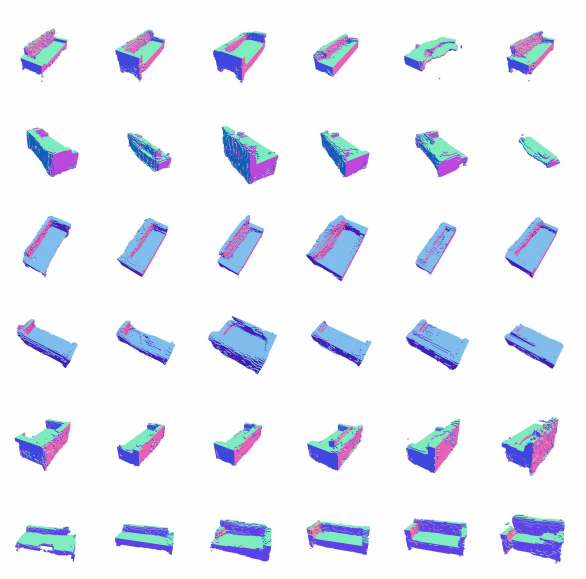
Figure 3. Samples from models trained on the Couches data in the 'one sample per object' setting (3173 training images).



(a) Absorption Only

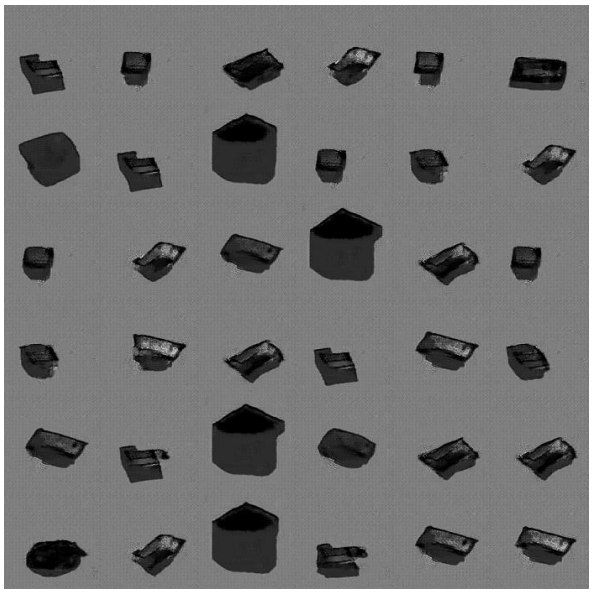


(b) Visual Hull

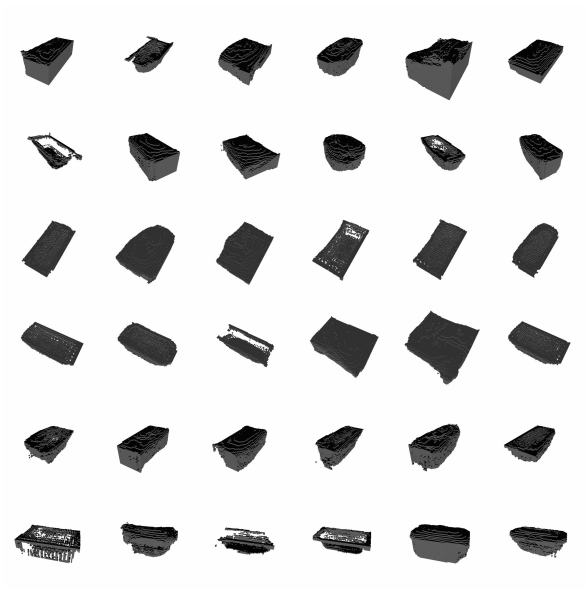


(c) IG-GAN (Ours)

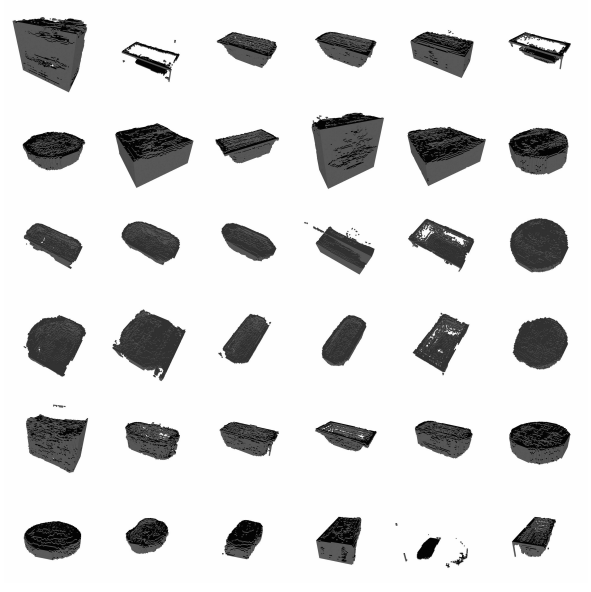
Figure 4. Samples from models trained on the Couches data in the 'one sample per object' setting (3173 training images), rendered as normal maps.



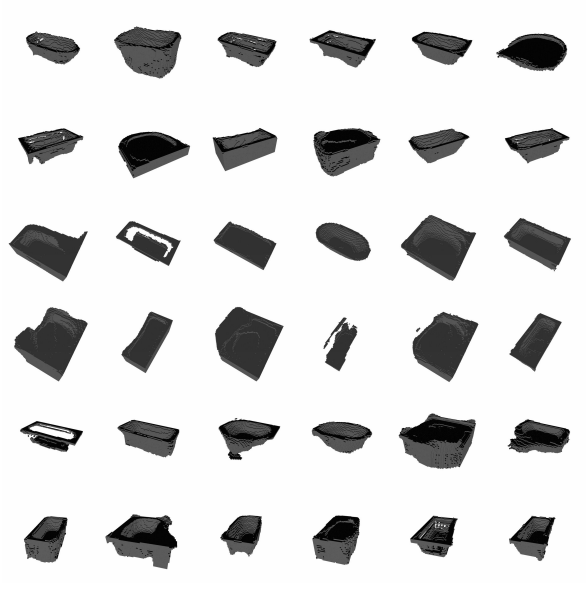
(a) 2D-DCGAN



(b) Absorption Only

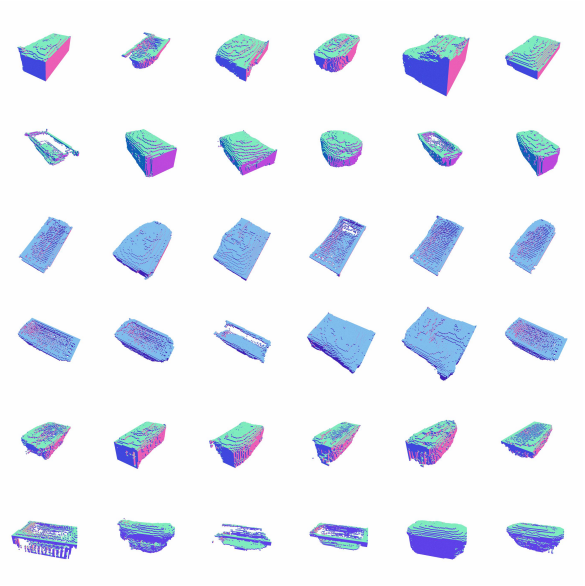


(c) Visual Hull

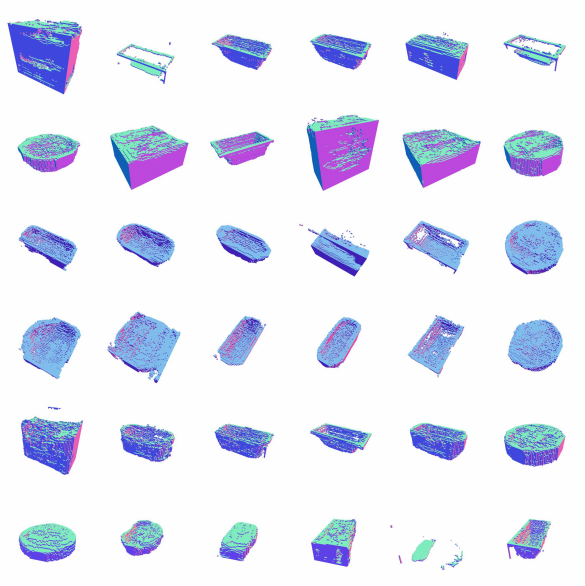


(d) IG-GAN (Ours)

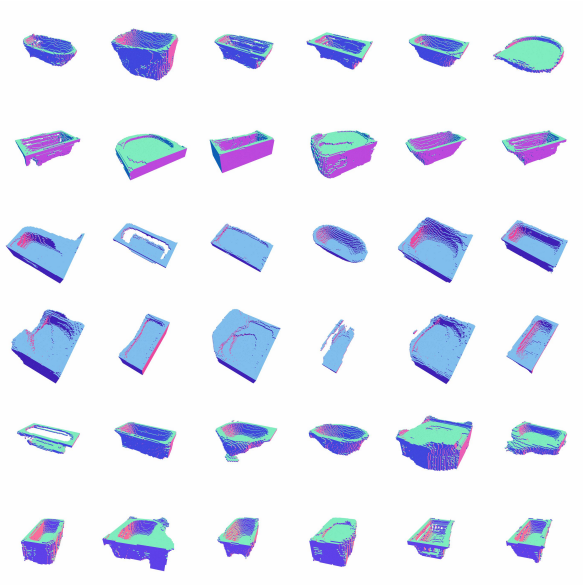
Figure 5. Samples from models trained on the Bathtubs data in the 'four samples per object' setting (3424 training images). With this small dataset, we were unable to get the 2D-DCGAN model to stably train. The resulting mode collapse causes the unusual grey background, as well as the high FID scores seen in Table 1.



(a) Absorption Only



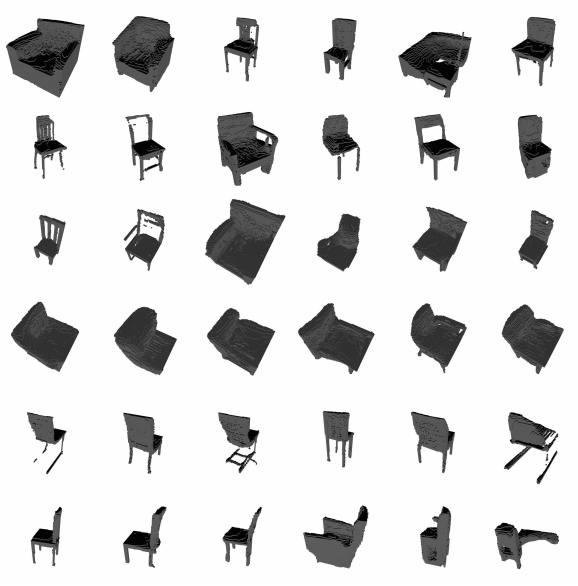
(b) Visual Hull



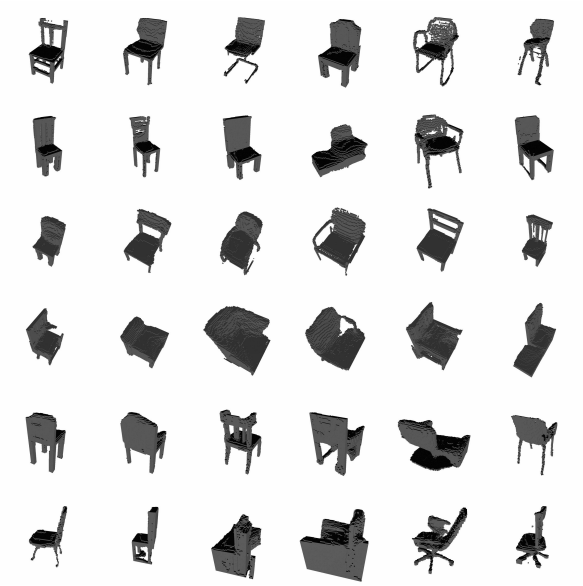
(c) IG-GAN (Ours)

Figure 6. Samples from models trained on the Bathtubs data in the 'four samples per object' setting (3424 training images), rendered as normal maps.

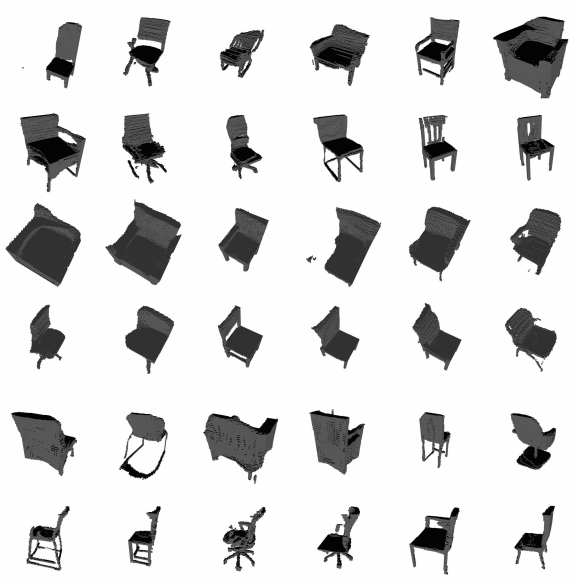
(a) 2D-DCGAN



(b) Absorption Only

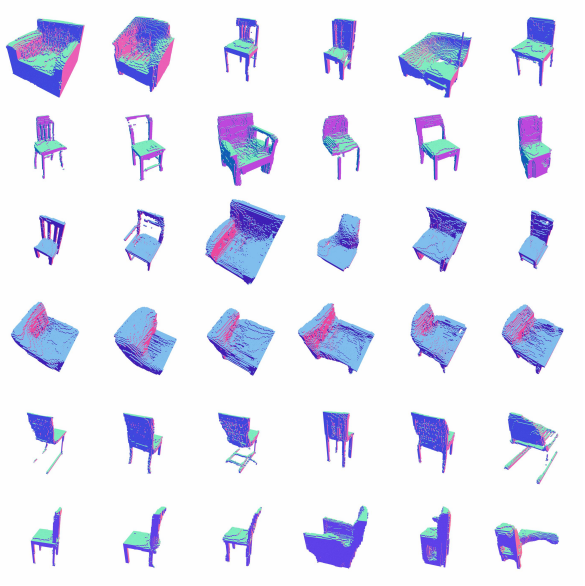


(c) Visual Hull

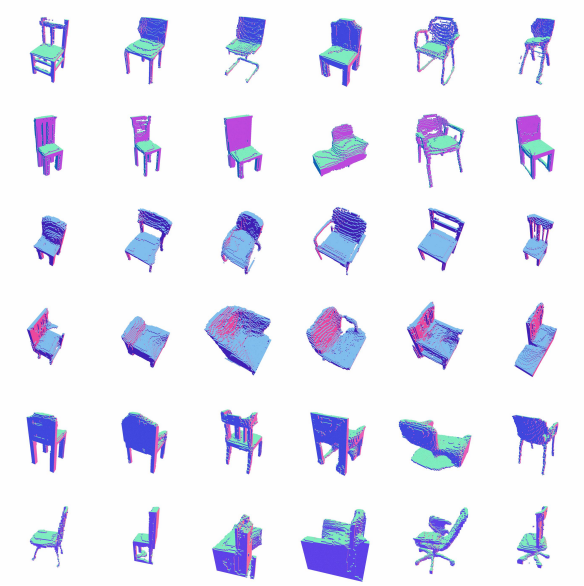


(d) IG-GAN (Ours)

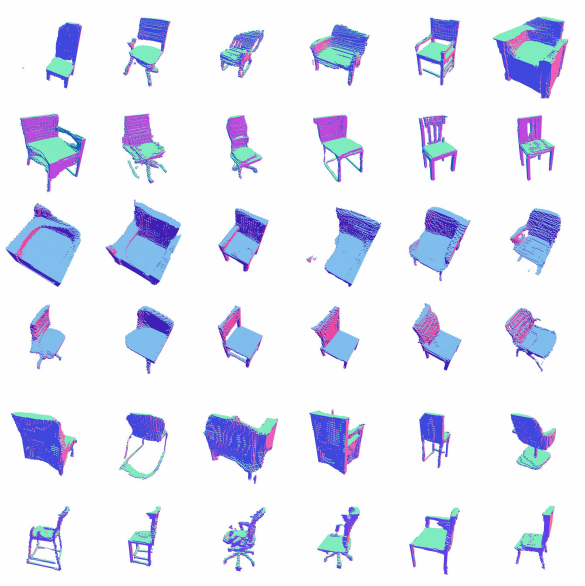
Figure 7. Samples from models trained on the Chairs data in the 'unlimited' setting.



(a) Absorption Only

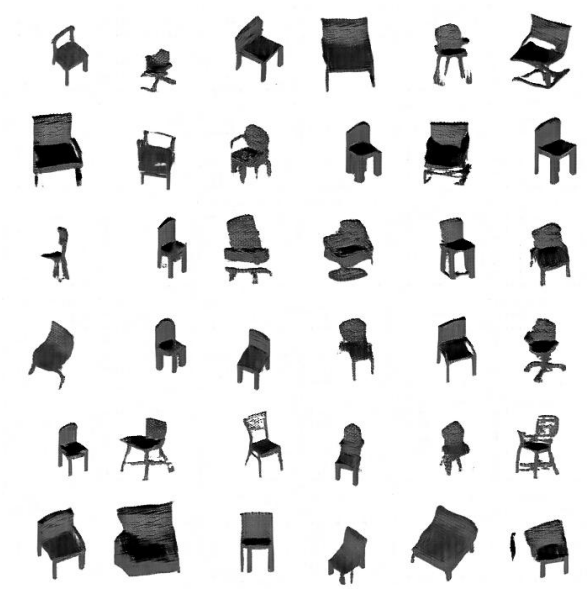


(b) Visual Hull

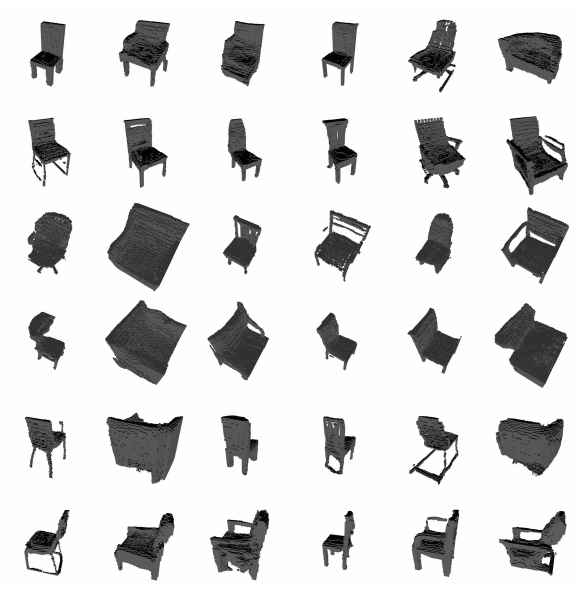


(c) IG-GAN (Ours)

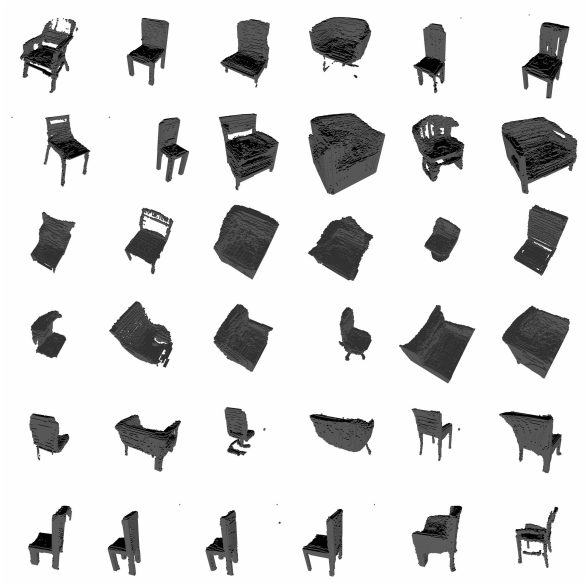
Figure 8. Samples from models trained on the Chairs data in the 'unlimited' setting, rendered as normal maps.



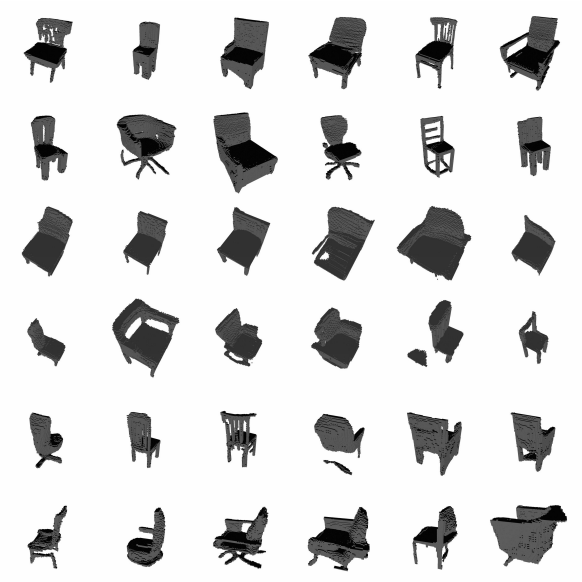
(a) 2D-DCGAN



(b) Absorption Only

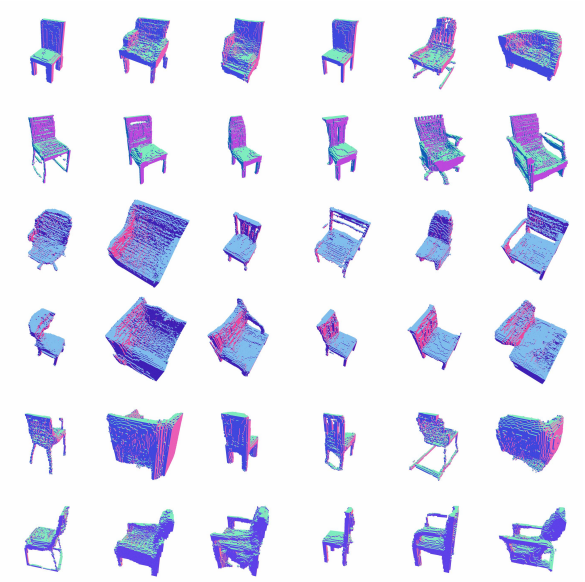


(c) Visual Hull

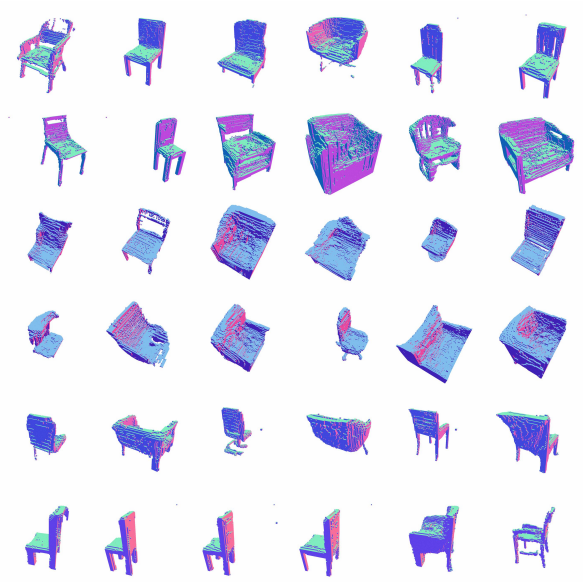


(d) IG-GAN (Ours)

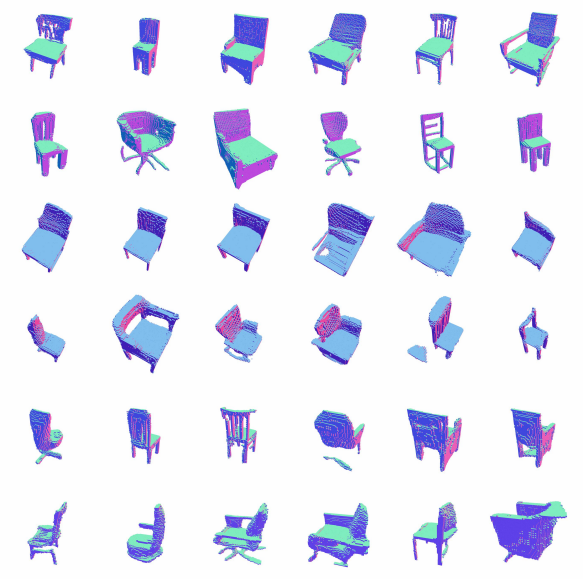
Figure 9. Samples from models trained on the Chairs data in the 'unlimited' setting, using a limited viewpoint distribution.



(a) Absorption Only

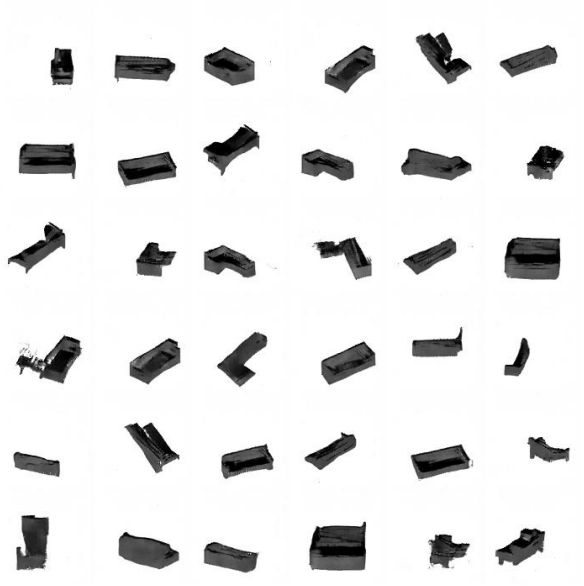


(b) Visual Hull

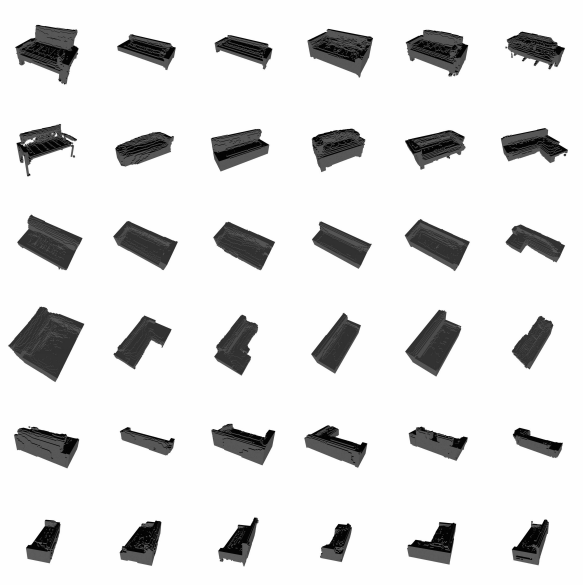


(c) IG-GAN (Ours)

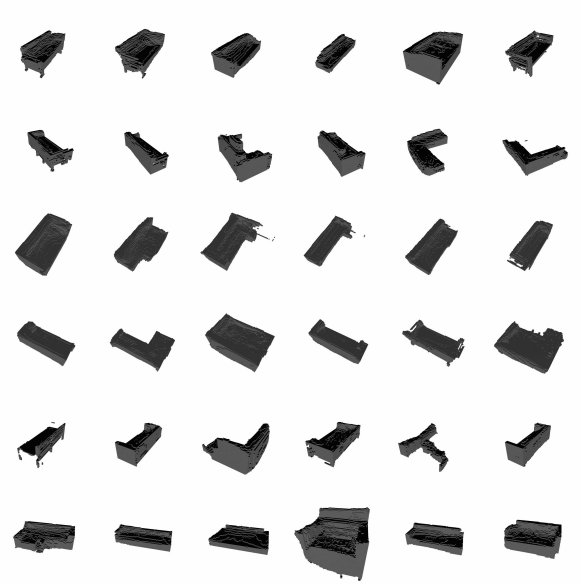
Figure 10. Samples from models trained on the Chairs data in the 'unlimited' setting, using a limited viewpoint distribution, and rendered as normal maps.



(a) 2D-DCGAN



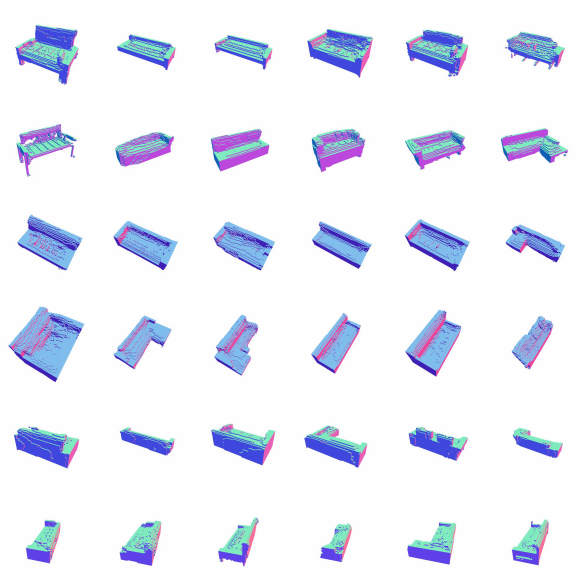
(b) Absorption Only



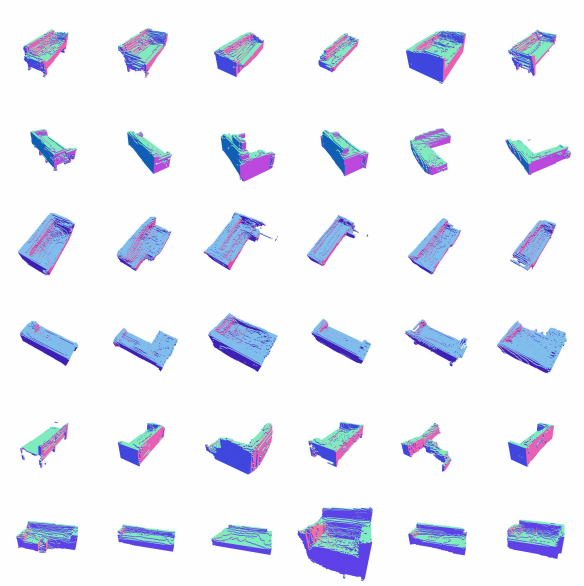
(c) Visual Hull

(d) IG-GAN (Ours)

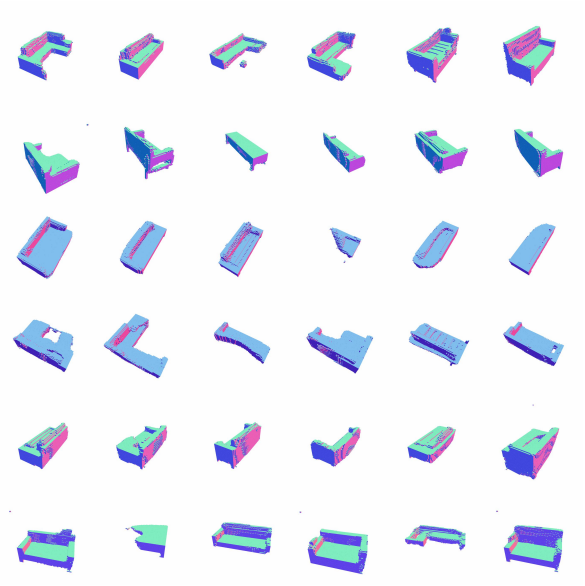
Figure 11. Samples from models trained on the Couches data in the 'unlimited' setting.



(a) Absorption Only

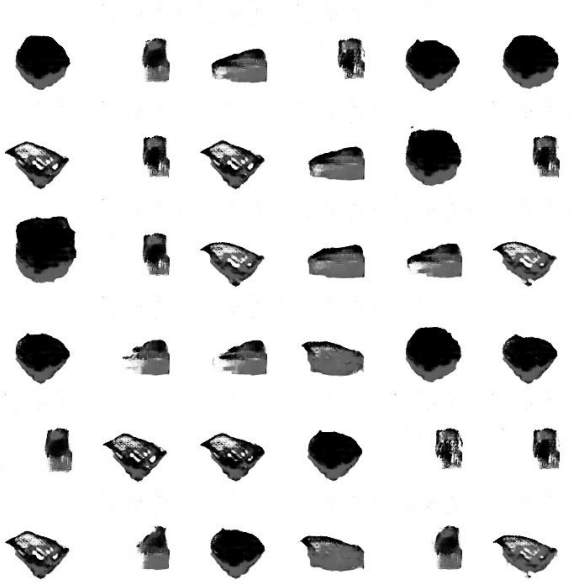


(b) Visual Hull

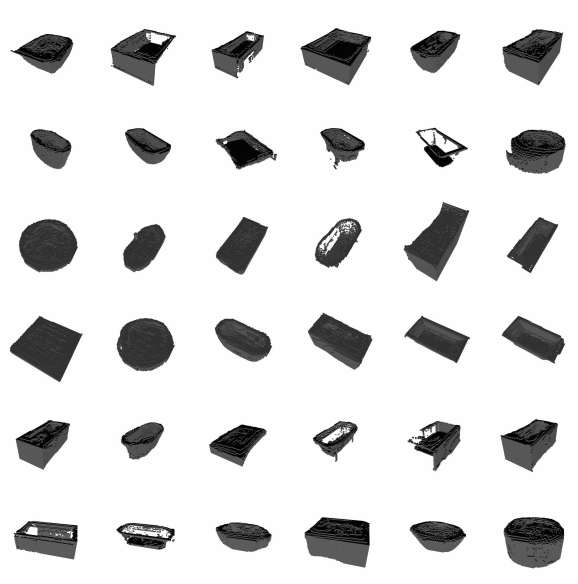


(c) IG-GAN (Ours)

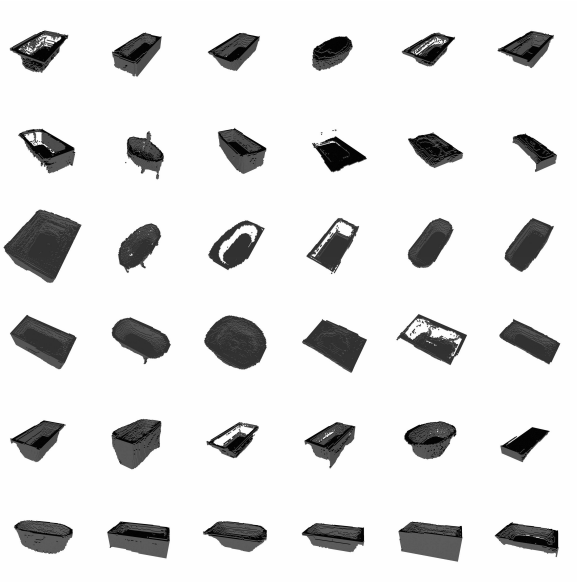
Figure 12. Samples from models trained on the Couches data in the 'unlimited' setting, rendered as normal maps.



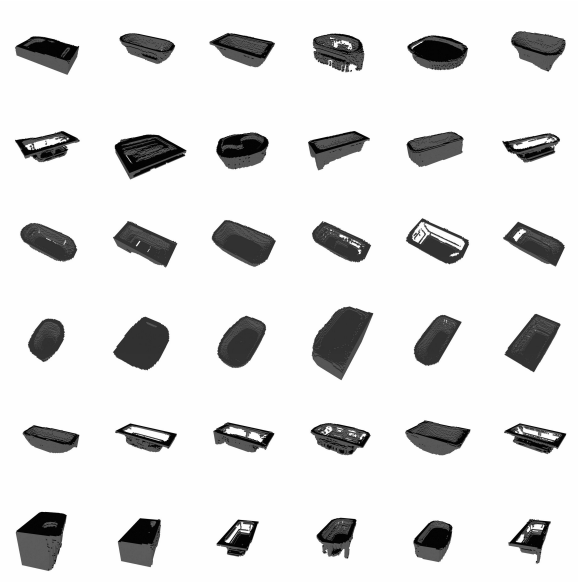
(a) 2D-DCGAN



(b) Absorption Only

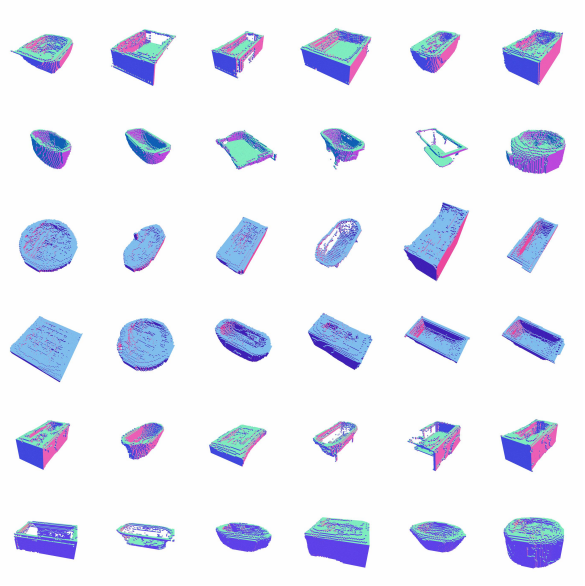


(c) Visual Hull

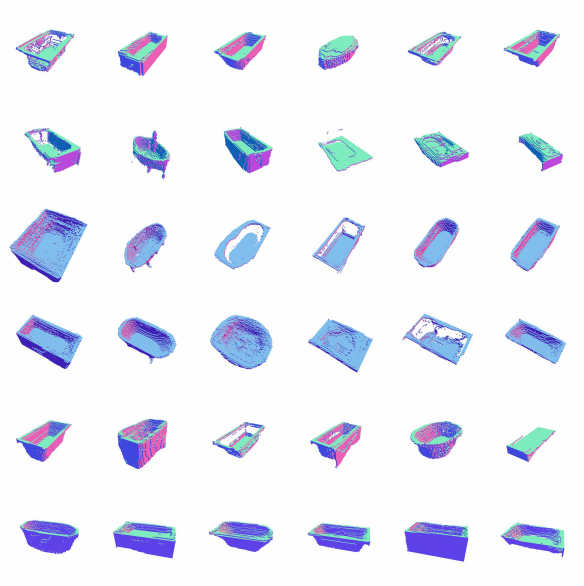


(d) IG-GAN (Ours)

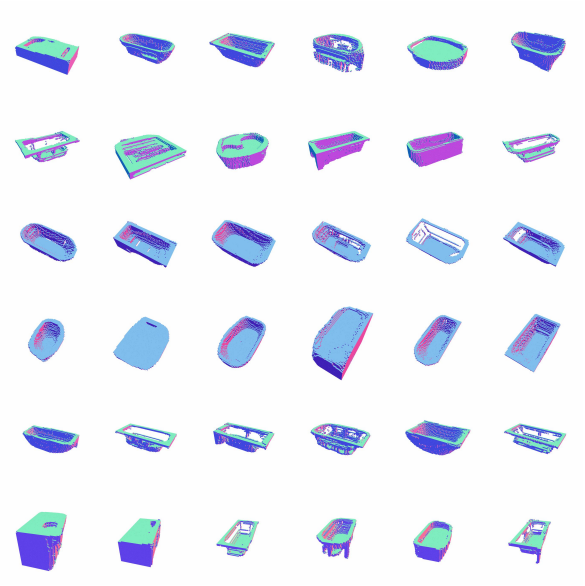
Figure 13. Samples from models trained on the Bathtubs data in the 'unlimited' setting.



(a) Absorption Only

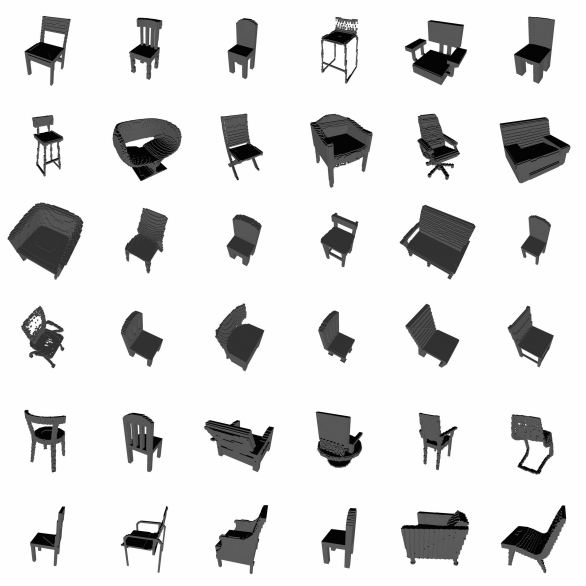


(b) Visual Hull

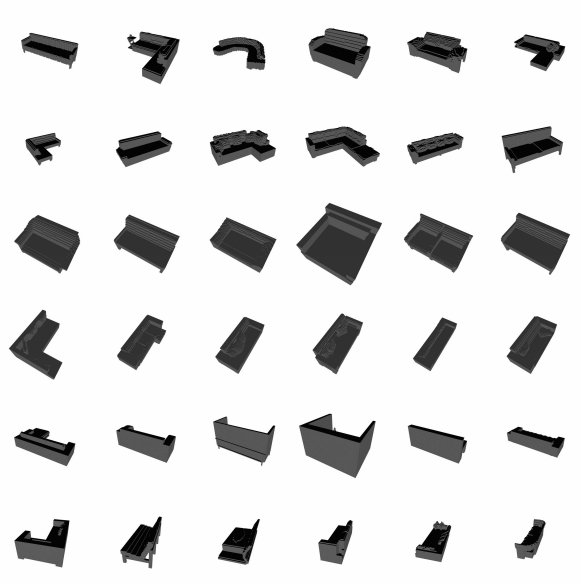


(c) IG-GAN (Ours)

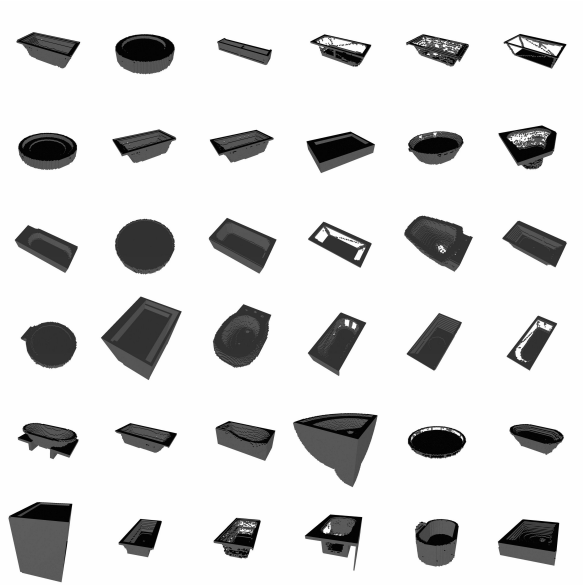
Figure 14. Samples from models trained on the Bathtubs data in the 'unlimited' setting, rendered as normal maps.



(a) Chairs



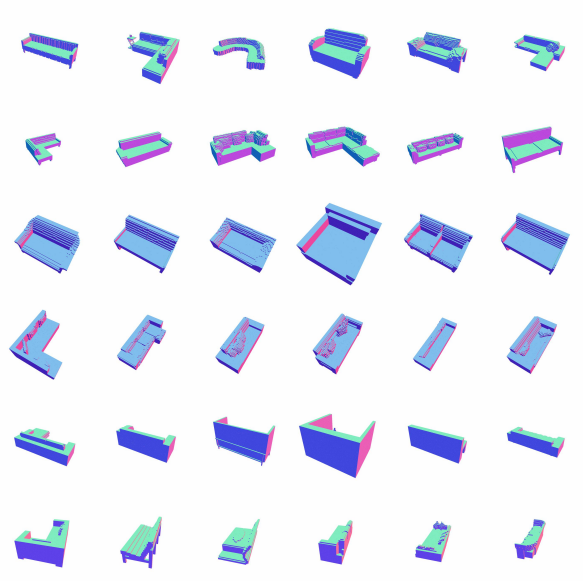
(b) Couches



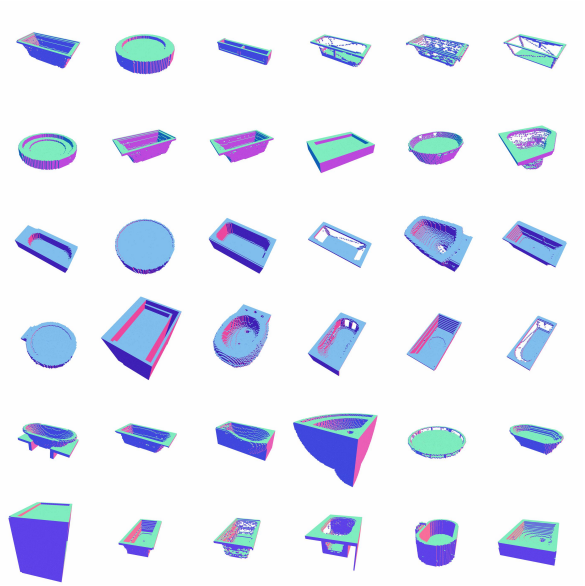
(c) Bathtubs

Figure 15. Samples from the dataset, for reference.

(a) Chairs



(b) Couches



(c) Bathtubs

Figure 16. Samples from the actual data, rendered as normal map for reference.

References

Henzler, P., Mitra, N. J., and Ritschel, T. Escaping plato's cave: 3d shape from adversarial rendering. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.