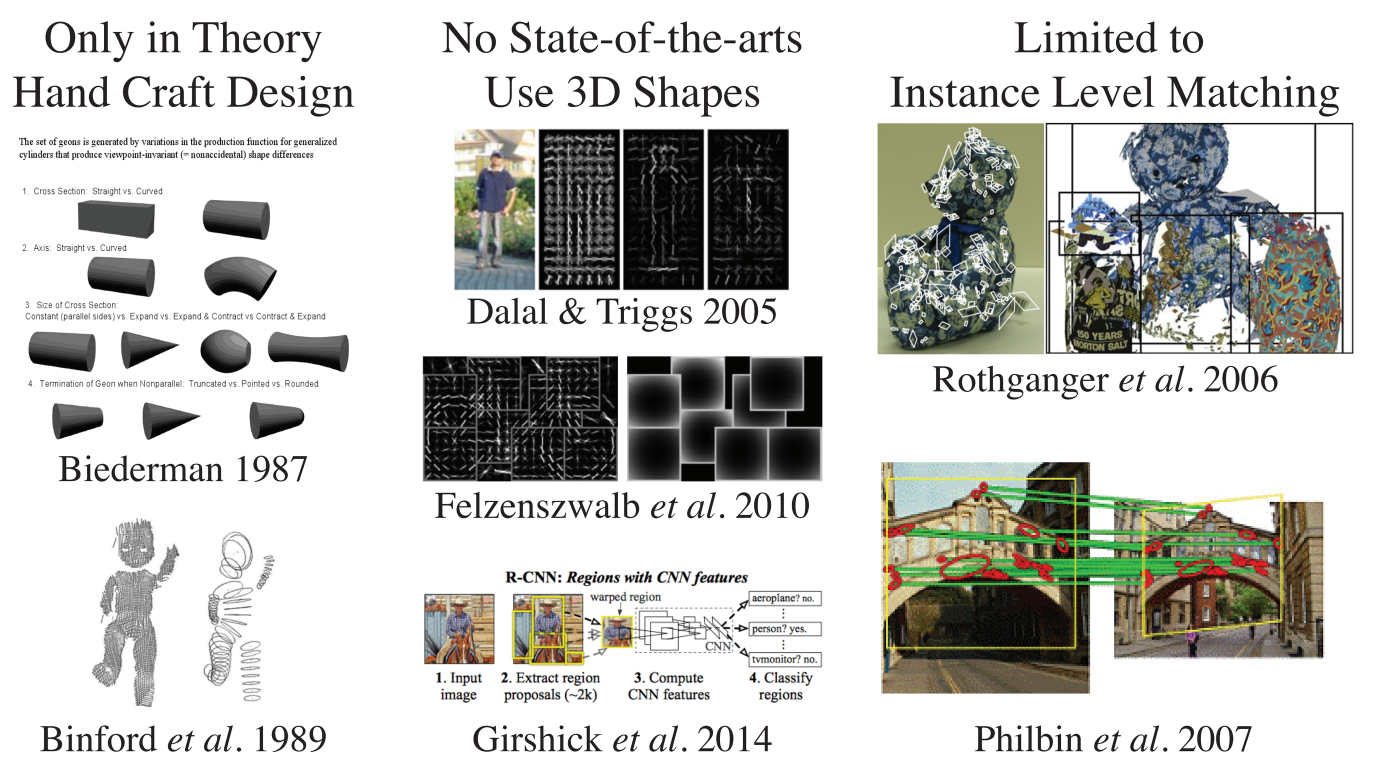
3D ShapeNets: A Deep Representation for Volumetric Shapes

Zhirong Wu Shuran Song Aditya Khosla Fisher Yu Linguang Zhang Xiaoou Tang Jianxiong Xiao

PRINCETON VISION GROUP

This work is funded by intel, Princeton, NVIDIA

Motivation: 3D Shape Representation



Desirable Properties of a Good 3D Shape Representation:

- **Data-driven:** learn from data rather than predefined shape routines.
- **Generic:** any complex shapes rather than simple shape primitives.
- **Compositional:** compose complex shapes by assembling simple ones.
- **Versatile:** applicable to various vision tasks.

3D Deep Learning

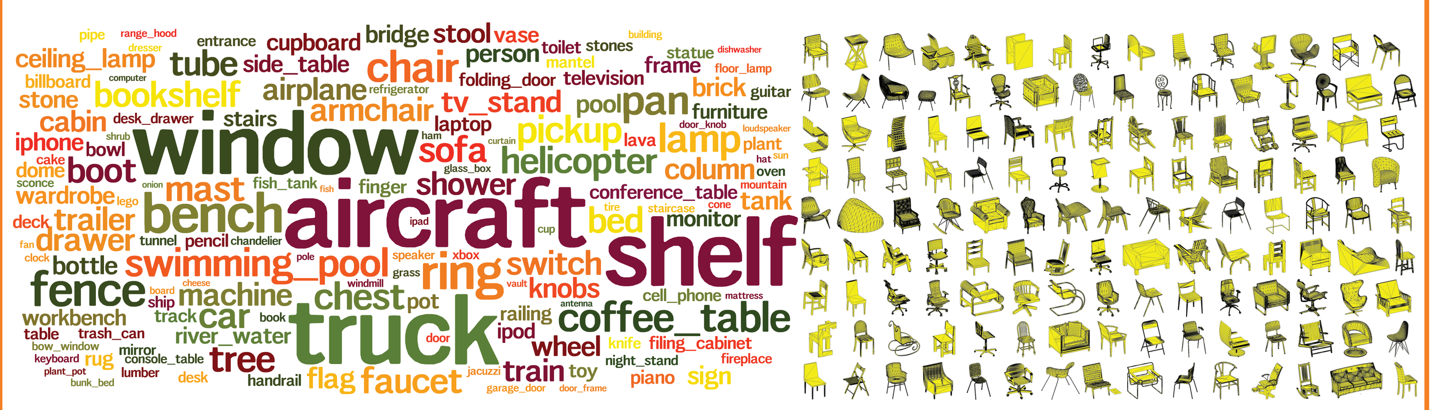
- 3D ShapeNets (a Convolutional Deep Belief Network) learns the joint distribution of generic 3D shapes across object category.
- Meaningful volumetric shapes could be sampled from the model.

Configuration of 3D ShapeNets

Layer	Type	Learning
1-3	convolutional RBM	Contrastive Divergence
4	fully connected RBM	Contrastive Divergence
5	multinomial label + Bernoulli feature forms an associate memory	Fast Persistent Contrastive Divergence

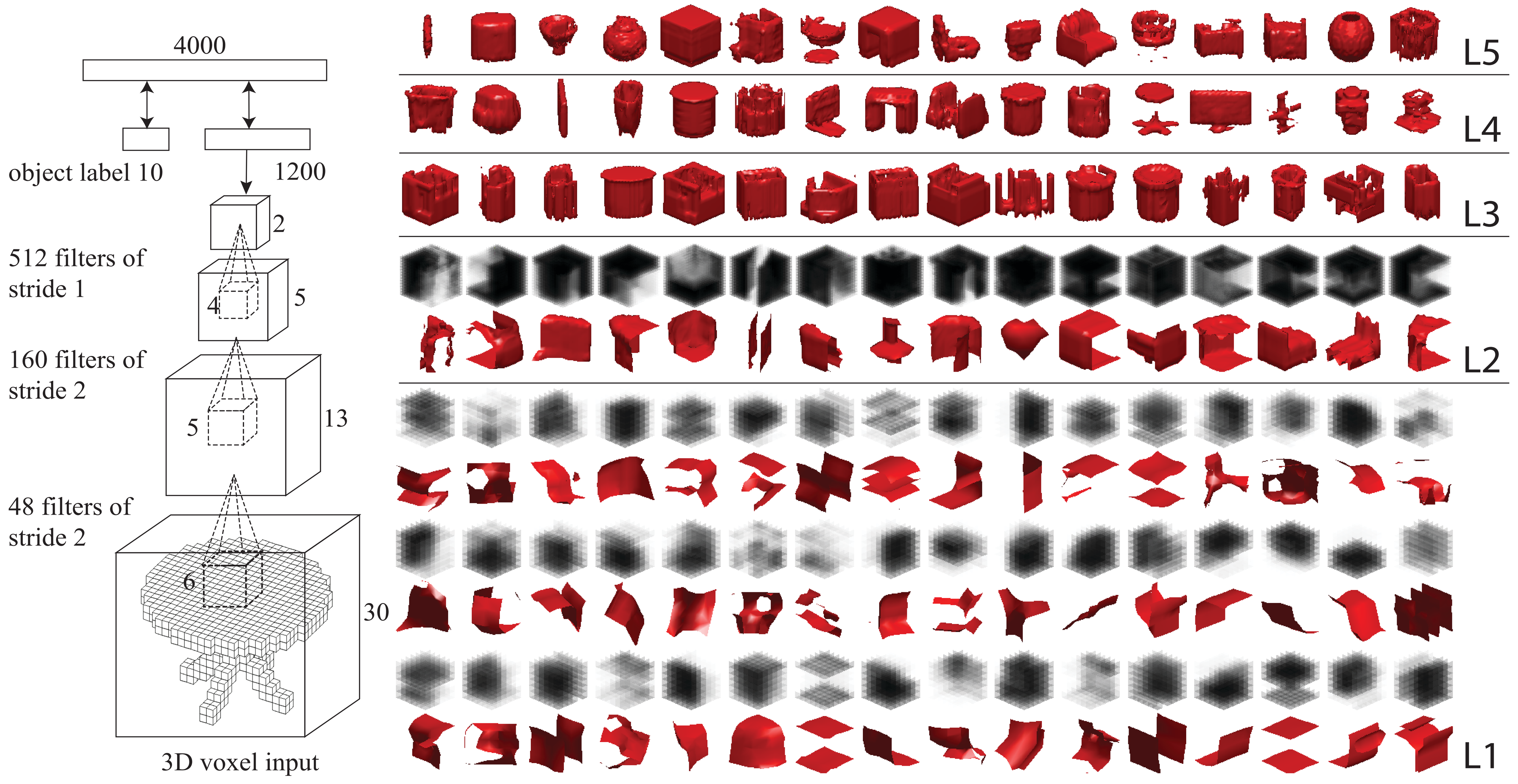
Princeton ModelNet Dataset

A large-scale CAD dataset of more than 150,000 models in 660 categories.



Object Categories Examples of Chairs

Code and Data: http://3DShapeNets.cs.princeton.edu



Recognition and Completion

Inference Process

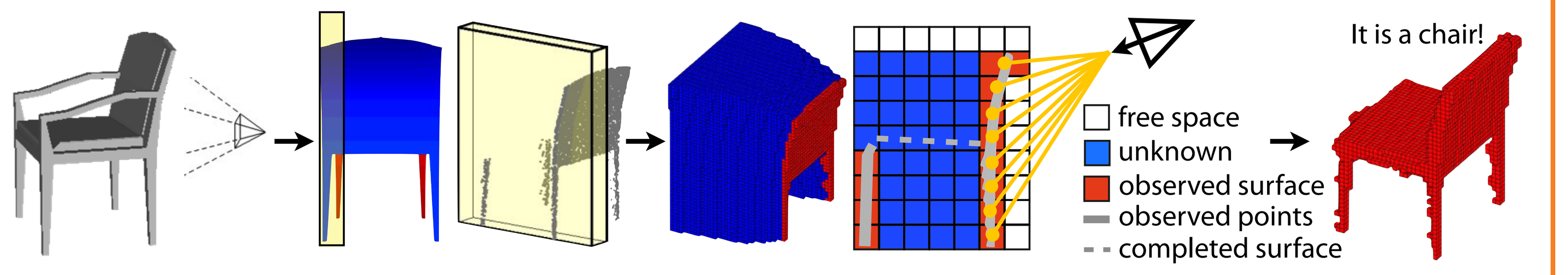
Convert the depth map into volumetric representation

↓

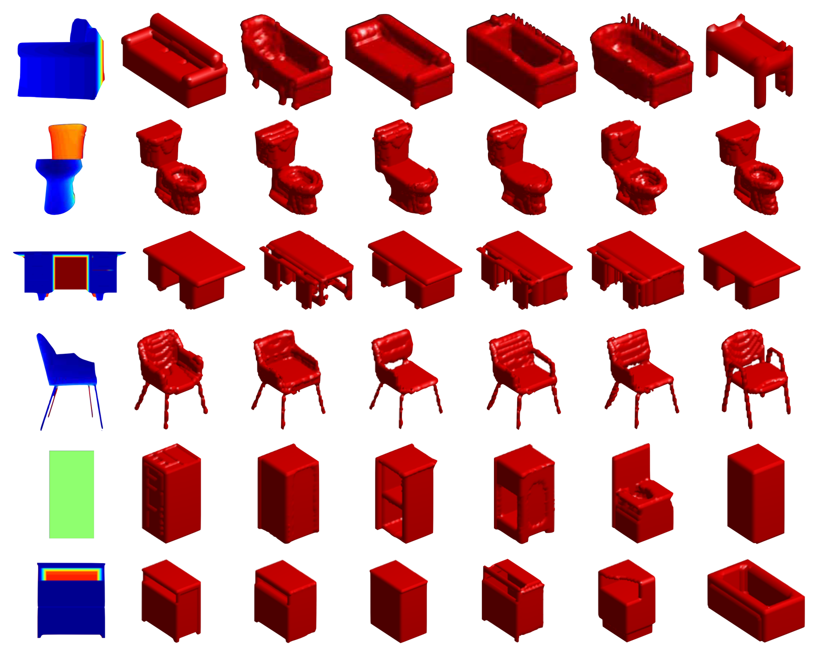
Identify free space, observed surface, unknow space

↓

Infer unknown space and label jointly by gibbs sampling



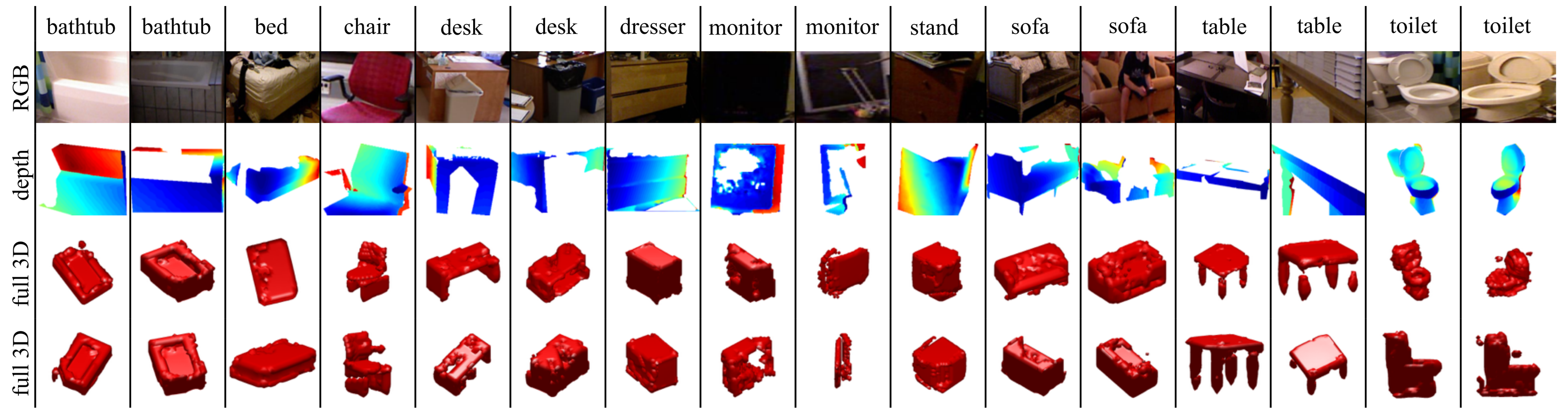
Shape Completion



results on synthetic depth

Recognition and Reconstruction on NYU depth

Training only uses CAD models!



	bathtub	bed	chair	desk	dresser	monitor	nightstand	sofa	table	toilet	all
[29] Depth	0.000	0.729	0.806	0.100	0.466	0.222	0.343	0.481	0.415	0.200	0.376
NN	0.429	0.446	0.395	0.176	0.467	0.333	0.188	0.458	0.455	0.400	0.374
ICP	0.571	0.608	0.194	**0.375**	**0.733**	0.389	0.438	0.349	0.052	**1.000**	0.471
3D ShapeNets	0.142	0.500	0.685	0.100	0.366	**0.500**	**0.719**	0.277	0.377	0.700	0.437
3D ShapeNets fine-tuned	**0.857**	0.703	**0.919**	0.300	0.500	**0.500**	0.625	**0.735**	0.247	0.400	**0.579**
[29] RGB	0.142	**0.743**	0.766	0.150	0.266	0.166	0.218	0.313	0.376	0.200	0.334
[29] RGBD	0.000	**0.743**	0.693	0.175	0.466	0.388	0.468	0.602	**0.441**	0.500	0.448

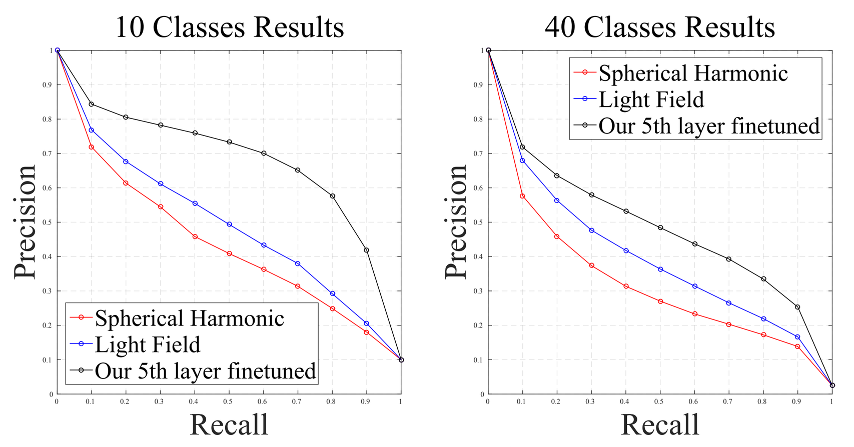
Table 2: Accuracy for View-based 2.5D Recognition on NYU dataset.

3D Feature Extraction

1. Convert 3D ShapeNets to a 3D CNN
2. Feature learning by back-propagation
3. Test for classification and retrieval

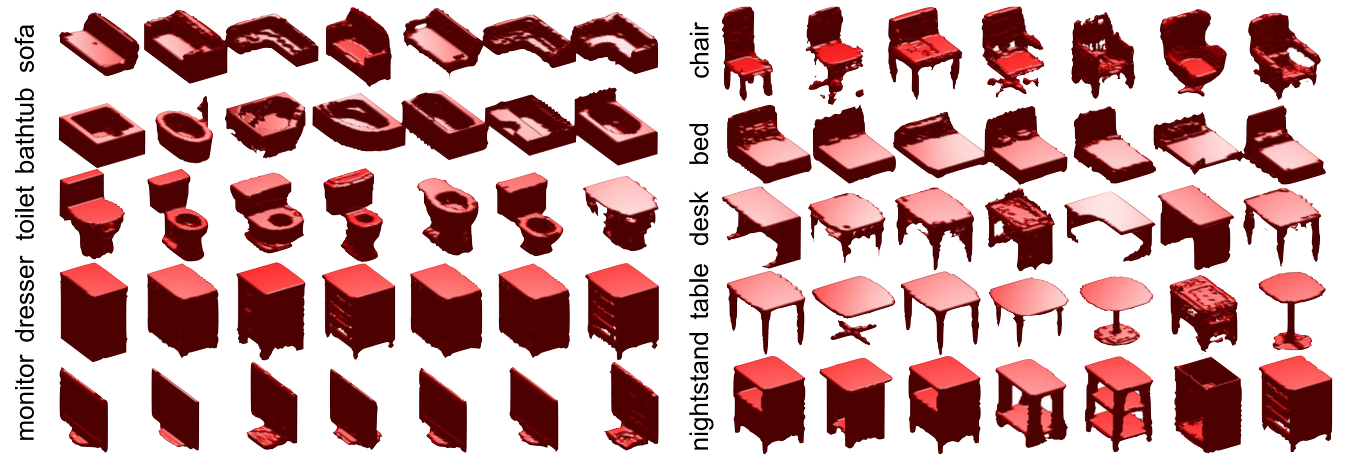
10 classes	SPH [18]	LFD [8]	Ours
classification	79.79 %	79.87 %	**83.54%**
retrieval AUC	45.97%	51.70%	**69.28%**
retrieval MAP	44.05%	49.82%	**68.26%**
40 classes	**SPH [18]**	**LFD [8]**	**Ours**
classification	68.23%	75.47%	**77.32%**
retrieval AUC	34.47%	42.04%	**49.94%**
retrieval MAP	33.26%	40.91%	**49.23%**

Table 1: Shape Classification and Retrieval Results.

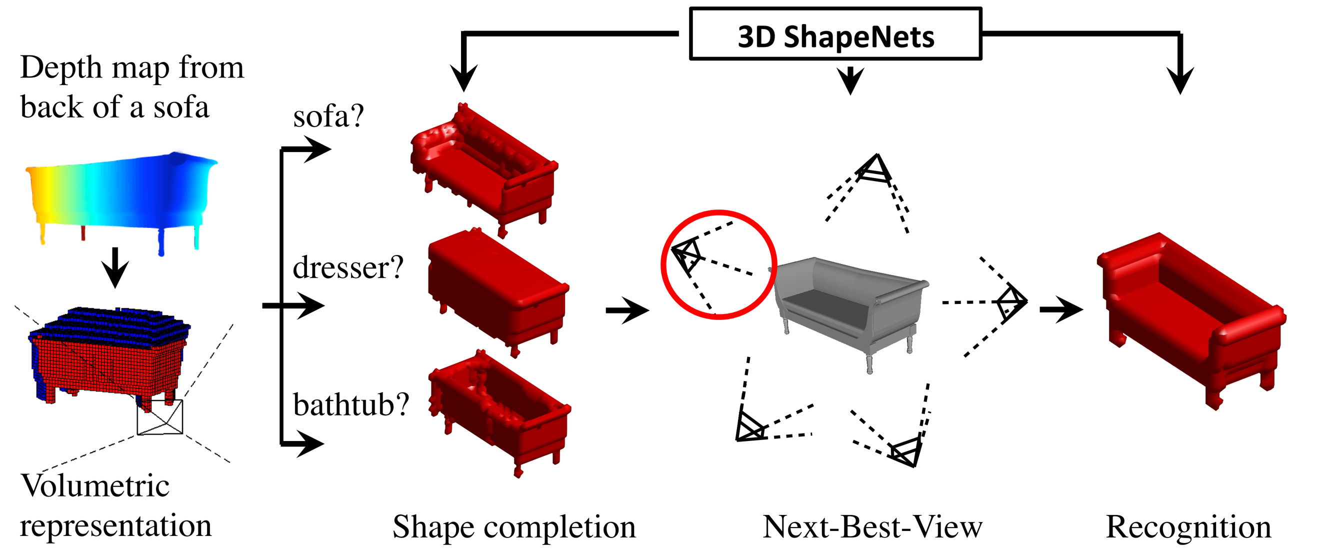


Precision Recall curve

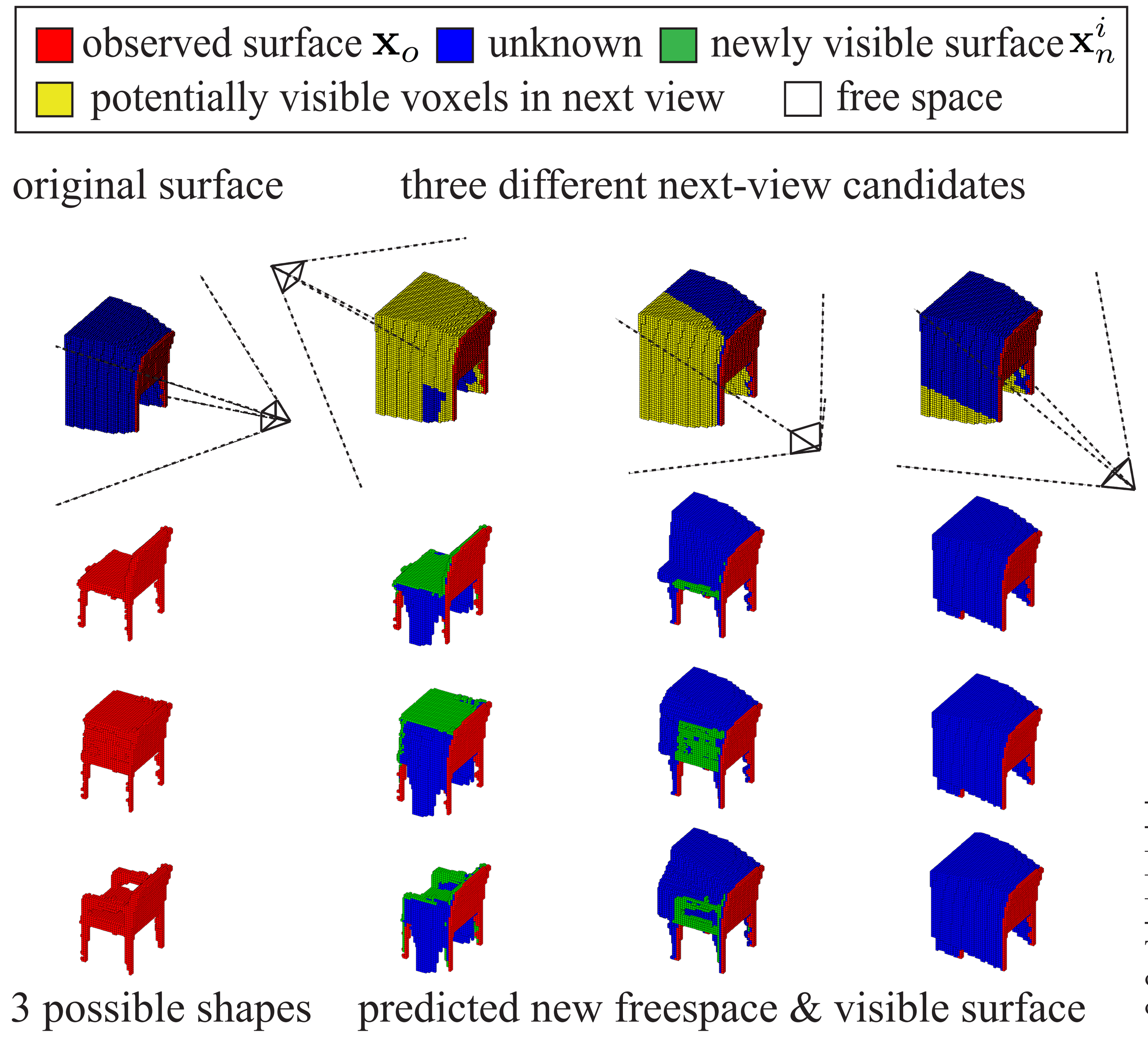
3D Shape Generation



3D ShapeNets for 2.5 Depth Recognition



Deep View Planning



The original entropy of the first view,

$$H = H(p(y|\mathbf{x_o} = \mathbf{x_o}))$$

Given a fixed number of views, the conditional entropy for i^{th} view,

$$H_i = H(p(y|\mathbf{x_n^i}, \mathbf{x_o} = \mathbf{x_o}))$$

$$= \sum_{\mathbf{x_n^i}} p(\mathbf{x_n^i}|\mathbf{x_o} = \mathbf{x_o}) \mathbf{H}(\mathbf{y}|\mathbf{x_n^i}, \mathbf{x_o} = \mathbf{x_o})$$

The reduction of entropy is the mutual information between the new voxel and the label,

$$H - H_i = I(y; \mathbf{x_n^i}|\mathbf{x_o} = \mathbf{x_o}) \geq 0$$

We choose the view with the maximum mutual information

$$V^* = \mathrm{argmax}_{V^i} I(y; \mathbf{x_n^i}|\mathbf{x_o} = \mathbf{x_o})$$

experiements on synthetic depth for two views

	bathtub	bed	chair	desk	dresser	monitor	nightstand	sofa	table	toilet	all
Ours	0.80	**1.00**	**0.85**	0.50	0.45	0.85	**0.75**	**0.85**	0.95	**1.00**	**0.80**
Max Visibility	**0.85**	0.85	**0.85**	0.50	**0.45**	0.85	**0.75**	**0.85**	0.90	0.95	0.78
Furthest Away	0.65	0.85	0.75	**0.55**	0.25	0.85	0.65	0.50	**1.00**	0.85	0.69
Random Selection	0.60	0.80	0.75	0.50	**0.45**	**0.90**	0.70	0.65	0.90	0.90	0.72

Table 3: Comparison of Different Next-Best-View Selections Based on Recognition Accuracy from Two Views. Based on an algorithm's choice, we obtain the actual depth map for the next view and recognize the object using those two views in our 3D ShapeNets representation.