3D ShapeNets: A Deep Representation for Volumetric Shapes

Zhirong Wu[1,2], Shuran Song[1], Aditya Khosla[3], Fisher Yu[1], Linguang Zhang[1], Xiaoou Tang[2], Jianxiong Xiao[1].
[1]Princeton University. [2]Chinese University of Hong Kong. [3]Massachusetts Institute of Technology.

Since the establishment of computer vision as a field five decades ago, 3D geometric shape has been considered to be one of the most important cues in object recognition. Even though there are many theories about 3D representation [1, 3], the success of 3D-based methods has largely been limited to instance recognition, using model-based keypoint matching [4]. For object category recognition, 3D shape is not used in any state-of-the-art recognition methods, mostly due to the lack of a strong generic representation for 3D geometric shapes. Furthermore, the recent availability of inexpensive 2.5D depth sensors, such as the Microsoft Kinect, has led to a renewed interest in 2.5D object recognition from depth maps. As a result, it is becoming increasingly important to have a strong 3D shape model in modern computer vision systems.

In this paper, we study generic shape representation for both object category recognition and shape completion. While there is some significant progress on shape synthesis [2] and recovery [6], they are mostly limited to part-based assembly and heavily relies on expensive part annotation. Instead of hand-coding shapes by parts, we desire a data-driven way to learn the complicate shape distributions from raw 3D data across object categories and poses, and automatically discover hierarchical compositional part representation. This allows us to infer the full 3D volume from a depth map without the knowledge of object category and pose a priori. We are also able to compute the potential information gain for recognition with regard to some occluded voxels. This would allow an active recognition system [5] to choose an optimal subsequent view for observation, when the category recognition from the first view is not sufficiently confident.

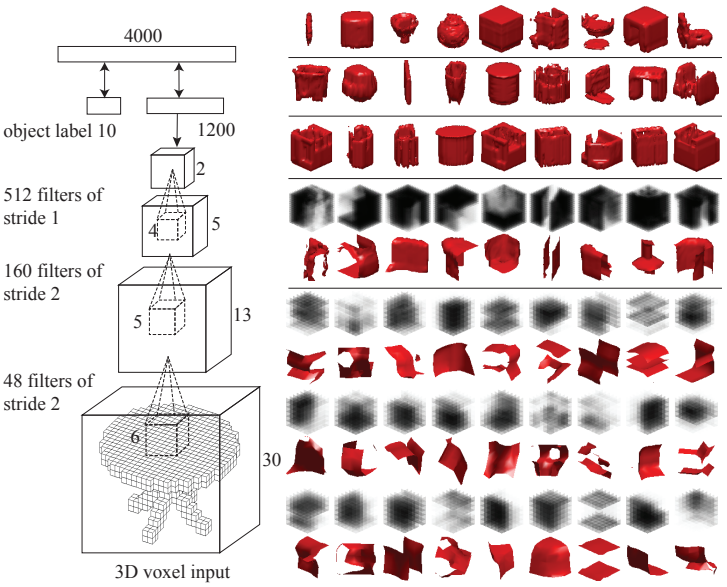
To study 3D shape representation, we propose to represent a geometric 3D shape as a probability distribution of binary variables on a 3D voxel grid. Each 3D mesh is represented as a binary tensor: 1 indicates the voxel is inside the mesh surface, and 0 indicates the voxel is outside the mesh (i.e., it is empty space). We design a Convolutional Deep Belief Network (CDBN) to learn this complex probabilistic distribution. The network is composed of a set of convolution layers and fully-connected layers. We do not use pooling layers as we find it hurts shape completion.

The energy, E, of a convolutional layer in our model can be computed as:

$$E(\mathbf{v},\mathbf{h}) = -\sum_f \sum_j \left(h_j^f \left(W^f * v \right)_j + c^f h_j^f \right) - \sum_l b_l v_l \qquad (1)$$

where v_l denotes each visible unit, h_j^f denotes each hidden unit in a feature channel f, and W^f denotes the convolutional filter. The "$*$" sign represents the convolution operation. In this energy definition, each visible unit v_l is associated with a unique bias term b_l to facilitate reconstruction, and all hidden units $\{h_j^f\}$ in the same convolution channel share the same bias term c^f.

After training the CDBN, the model learns the joint distribution $p(\mathbf{x}, y)$ of voxel data $\mathbf{x}$ and object category label $y \in \{1, \cdots, K\}$. Although the model is trained on complete 3D shapes, it is able to recognize objects in single-view 2.5D depth maps (e.g., from RGB-D sensors). We first convert the 2.5D depth map into a volumetric representation where we categorize each voxel as free space, surface or occluded, depending on whether it is in front of, on, or behind the visible surface (i.e., the depth value) from the depth map. The free space and surface voxels are considered to be observed, and the occluded voxels are regarded as missing data. The test data is represented by $\mathbf{x} = (\mathbf{x}_o, \mathbf{x}_u)$, where $\mathbf{x}_o$ refers to the observed free space and surface voxels, while $\mathbf{x}_u$ refers to the unknown voxels. Recognizing the object category involves estimating $p(y|\mathbf{x}_o)$. This posterior distribution is approximated by Gibbs sampling as follows. We initialize $\mathbf{x}_u$ to random values and propagate data bottom up to sample a label y from $p(y|\mathbf{x}_o, \mathbf{x}_u)$, then we use the sample y and propagate the data down to sample for unknown voxels $\mathbf{x}_u$. 50 iterations of this up-down sampling should suffice to get a shape completion $\mathbf{x}$, and its corresponding label y. The above procedure runs in parallel for a large number of particles resulting in a variety of completion results corresponding to potentially different classes.



http://3DShapeNets.cs.princeton.edu

Figure 1: Architecture and filter visualizations of 3D ShapeNets.

Training a 3D shape model that captures intra-class variance requires a large collection of 3D shapes. Previous CAD datasets (e.g., [7]) are limited both in the variety of categories and the number of examples per category. Therefore, we construct ModelNet, a new large scale 3D CAD model dataset to train our data-hungry deep learning model. Our new dataset is 22 times larger than previous ones, containing 151,128 3D CAD models belonging to 660 unique object categories.

From the experimental results, our model significantly outperforms existing approaches on 3D mesh classification, mesh retrieval, as well as depth map object recognition. It is also a promising approach for next-best-view planning. Source code and data are available at our project website.

[1] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 1987.

[2] Evangelos Kalogerakis, Siddhartha Chaudhuri, Daphne Koller, and Vladlen Koltun. A probabilistic model for component-based shape synthesis. *ACM Transactions on Graphics (TOG)*, 2012.

[3] Joseph L Mundy. Object recognition in the geometric era: A retrospective. In *Toward category-level object recognition*. 2006.

[4] Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *IJCV*, 2006.

[5] William Scott, Gerhard Roth, and Jean-François Rivest. View planning for automated 3d object reconstruction inspection. *ACM Computing Surveys*, 2003.

[6] Chao-Hui Shen, Hongbo Fu, Kang Chen, and Shi-Min Hu. Structure recovery by part assembly. *ACM Transactions on Graphics (TOG)*, 2012.

[7] Philip Shilane, Patrick Min, Michael Kazhdan, and Thomas Funkhouser. The princeton shape benchmark. In *Shape Modeling Applications*, 2004.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.