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

Daeyun Shin [1], Charless Fowlkes [1], Derek Hoiem [2]
http://www.ics.uci.edu/~daeyuns/pixels-voxels-views

[1] University of California, Irvine
[2] University of Illinois, Urbana-Champaign

The goal of this paper is to compare surface-based and volumetric 3D object shape representations, as well as viewer-centered and object-centered reference frames for single-view 3D shape prediction. Shape is arguably the most important property of objects, providing cues for affordance, function, category, and interaction. This paper examines the problem of predicting the 3D object shape from a single image (Fig. 1). The availability of large 3D object model datasets [1] and flexible deep network learning methods has made this an increasingly active area of research. Recent methods predict complete 3D shape using voxel [2, 5] or octree [12] volumetric representations, multiple depth map surfaces [11], point cloud [3], or a set of cuboid part primitives [14]. However, there is not yet a systematic evaluation of important design choices such as the choice of shape representation and coordinate frame.

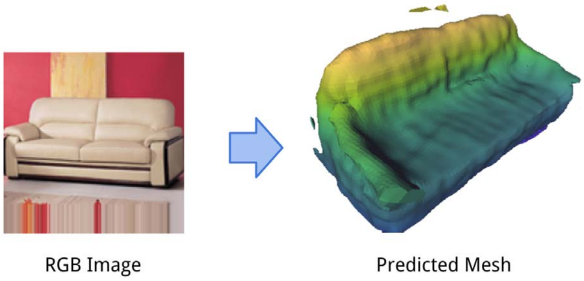

Figure 1: We investigate the problem of predicting the 3D shape of an object from a single depth or RGB image (illustrated above). In particular, we examine the impacts of coordinate frames (viewer-centered vs. object-centered), shape representation (volumetric vs. multi-surface), and familiarity (known instance, novel instance, novel category).

We directly compare the merits of voxels vs. surfaces and viewer-centered vs. object-centered for familiar vs. unfamiliar objects, as predicted from RGB or depth images. Among our findings, we show that surface-based methods outperform voxel representations for objects from novel classes and produce higher resolution outputs. We also find that using viewer-centered coordinates is advantageous for novel objects, while object-centered representations are better for more familiar objects. Interestingly, the coordinate frame significantly affects the shape representation learned, with object-centered placing more importance on implicitly recognizing the object category and viewer-centered producing shape representations with less dependence on category recognition.

In experiments on 2D symbols, Tarr and Pinker [10] found that human perception is largely tied to viewer-centered coordinates; this was confirmed by McMullen and Farah [7] for line drawings, who also found that object-centered coordinates seem to play more of a role for familiar exemplars. Note that in the human vision literature, “viewer-centered” usually means that the object shape is represented as a set of images in the viewer’s coordinate frame, and “object-centered” usually means a volumetric shape is represented in the object’s coordinate frame. In our work, we consider both the shape representation (volumetric or surface) and coordinate frame (viewer or object) as separate design choices. We do not claim our computational approach has any similarity to human visual processing, but it is interesting to see that in our experiments with 3D objects, we also find a preference for object-centered coordinates for familiar exemplars (i.e., novel view of known object) and for viewer-centered coordinates in other cases.

In the commonly used object-centered setting, the shape is predicted in canonical model coordinates specified by the training data. For example, in the ShapeNetCore dataset, the x-axis or ($\phi_{\text{az}} = 0,\ \theta_{\text{el}} = 0$) direction corresponds to the commonly agreed upon front of the object, and the relative transformation parameters from the input view to this coordinate system is unknown. In our viewer-centered approach, we supervise the network to predict a pre-aligned 3D shape in the input image’s reference frame — e.g. so that ($\phi_{\text{az}} = 0,\ \theta_{\text{el}} = 0$) in the output coordinate system always corresponds to the input viewpoint.

Our multi-surface shape prediction system uses an encoder-decoder network to predict a set of silhouettes and depth maps. We compare this with a volumetric prediction network by replacing the decoder with a voxel generator. In all experiments, we train the networks on synthetically generated images.

Reconstructing multi-surface representations: We convert the predicted multiview depth images to a single triangulated mesh using Floating Scale Surface Reconstruction (FSSR) [4], which we found to produce better results than Poisson Reconstruction [6] in our experiments. FSSR is widely used for surface reconstruction from oriented 3D points derived from multiview stereo or depth sensors. Our experiments are unique in that surface reconstruction methods are used to resolve noise in predictions generated by neural networks rather than sensor observations. We have found that 3D surface reconstruction reduces noise and error in surface distance measurements.

3D shape from single depth: We use the SHREC’12 dataset for comparison with the exemplar retrieval approach by Rock *et al.* [8] on predicting novel views, instances, and classes. Novel views require the least generalization (the same shape is seen in training), and novel classes require the most (no instances from the same category seen during training). This dataset has a training set consisting of 22,500 training + 6,000 validation examples and has 600 examples in each of the three test evaluation sets, using the standard splits [8]. The 3D models in the dataset are aligned to each other, so that they can be used for both viewer-centered and object-centered prediction. Results are shown in Tables 1, 3, 4, and 5.

3D shape from real-world RGB images: We also perform novel model experiments on RGB images. We use RenderForCNN’s [9] rendering pipeline and generate 2.4M synthetic training examples using the ShapeNetCore dataset along with target depth and voxel representations. We perform quantitative evaluation of the resulting models on real-world RGB images using the PASCAL 3D+ dataset [13]. We train 3D-R2N2’s network [2] from scratch using the same dataset and compare evaluation results. Results are shown in Tables 2 and in Figure 3.

Evaluation: We evaluate with voxel intersection-over-union and a surface distance metric similar to [8], which tends to correspond better to qualitative judgments of accuracy when there are thin structures. The distance between surfaces is approximated as the mean of *point-to-triangle* distances from i.i.d. sampled points on the ground truth mesh to the closest points on surface of the reconstructed mesh, and vice versa. To evaluate surface distance for voxel-prediction models, we use Marching Cubes to obtain the mesh from the prediction. For multi-surface experiments, we also evaluate using silhouette intersection-over-union and depth error averaged over the predicted views. Sometimes, even when the predictions for individual views are quite accurate, slight inconsistencies or oversmoothing by the final surface estimation can reduce the accuracy of the 3D model.

Multi-surface vs. voxel shape representations: Table 1 compares performance of multi-surface and voxel-based representations for shape prediction. Quantitatively, multi-surface outperforms for novel class and performs similarly for novel view and novel instance. We also find that the 3D shapes produced by the multi-surface model look better qualitatively, as they can encode higher resolution. We observe that it is generally difficult to learn and reconstruct thin structures such as the legs of chairs and tables. In part this is a learning problem, as discussed in Choy *et al.* . [2]. Our qualitative results suggest that silhouettes are generally better for learning and predicting thin object parts than voxels, but the information is often lost during surface reconstruction due to the sparsity

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

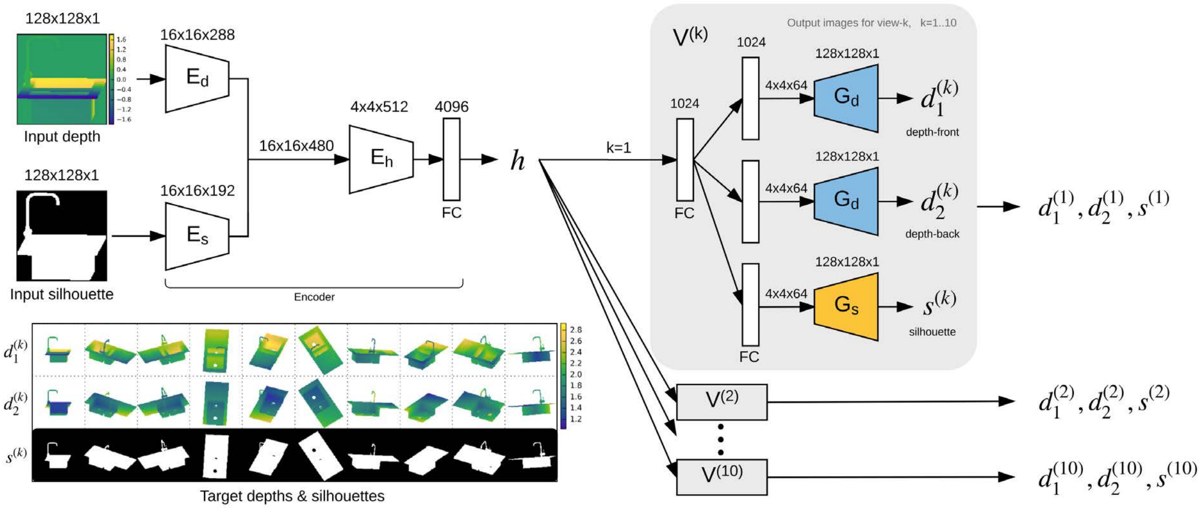

Figure 2: Network architecture: Encoders E_d, E_s, E_h learn view-specific shape features h extracted from the input depth and silhouette. h is used by the 10 output decoder branches $V^{(k)}$, k=1..10 which each synthesize one silhouette and two, front and back, depth images. The branches have independently parameterized fully connected layers, but the up-convolutional decoders G_d, G_s share parameters across all output branches.

Mean	Surface Distance			Voxel IoU		
	NovelClass	NovelModel	NovelView	NovelClass	NovelModel	NovelView
Voxels	0.0950	0.0619	0.0512	0.4569	0.5176	**0.6969**
Multi-surfaces	**0.0759**	0.0622	**0.0494**	0.4914	0.5244	0.6501
Rock *et al.* [8]	0.0827	**0.0604**	0.0639	**0.5320**	**0.5888**	0.6374

Table 1: 3D shape prediction from a single depth image on the SHREC'12 dataset used by [8], comparing results for voxel and multi-surface decoders trained to produce models in a viewer-centered coordinate frame. Rock *et al.* . [8] also predicts in viewer-centered coordinates.

of available data points. We expect that improved depth fusion and mesh reconstruction would likely yield even better results. As shown in Fig. 3, the multi-surface representation can more directly be output as a point cloud by skipping the reconstruction step. This avoids errors that can occur during the surface reconstruction but is more difficult to quantitatively evaluate.

Viewer-centered vs. object-centered coordinates: When comparing performance of predicting in viewer-centered coordinates vs. object-centered coordinates, it is important to remember that only viewer-centered encodes pose and, thus, is more difficult. Sometimes, the 3D shape produced by viewer-centered prediction is very good, but the pose is misaligned, resulting in poor quantitative results for that example. Even so, in Tables 3, 4, and 5, we observe a clear advantage for viewer-centered prediction for novel models and novel classes, while object-centered outperforms for novel views of object instances seen during training. For object-centered prediction, two views of the same object should produce the same 3D shape, which encourages memorizing the observed meshes. Under viewer-centered, the predicted mesh must be oriented according to the input viewpoint, so multiple views of the same object should produce different 3D shapes (which are related by a 3D rotation). This requirement seems to improve the generalization capability of viewer-centered prediction to shapes not seen during training. Qualitative results support our initial hypothesis that object-centered models tend to correspond more directly to category recognition. We see in Figure 3, that the object-centered model often predicts a shape that looks good but is an entirely different object category than the input image. The viewer-centered model tends not to make these kinds of mistakes and, instead, errors tend to be overly simplified shapes or slightly incorrect poses.

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. Technical report, Stanford University, Princeton University, Toyota Technological Institute at Chicago, 2015.

[2] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[3] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 38, 2017.

Category	[2] (OC)	[2] (VC)	Ours (OC)	Ours (VC)
aero	0.359	0.201	**0.362**	0.289
bike	**0.535**	0.106	0.362	0.272
boat	**0.366**	0.236	0.331	0.259
bottle	0.617	0.454	**0.643**	0.576
bus	0.387	0.273	0.497	**0.556**
car	0.462	0.400	0.566	**0.582**
chair	0.325	0.221	**0.362**	0.332
d.table	0.081	0.023	**0.122**	0.118
mbike	0.474	0.167	**0.487**	0.366
sofa	**0.602**	0.447	0.555	0.538
train	**0.340**	0.192	0.301	0.257
tv	0.376	0.164	0.383	**0.397**
mean	0.410	0.240	**0.414**	0.379

Table 2: Per-category voxel IoU on PASCAL 3D+ using our multi-surface network and the voxel-based 3D-R2N2 network [2]. Although the network trained to produce object-centered (OC) models performs slightly better quantitatively (for multi-surface), the viewer-centered (VC) model tends to produce better qualitative results, sometimes with misaligned pose.

	NovelView	NovelModel	NovelClass
View-centered	0.714	**0.570**	**0.517**
Obj-centered	**0.902**	0.474	0.309

Table 3: Voxel IoU of predicted and ground truth values (mean, higher is better), using the voxel network. Trained for 45 epochs with batch size 150, learning rate 0.0001.

	NovelView	NovelModel	NovelClass
View-centered	0.807	**0.706**	**0.670**
Obj-centered	**0.921**	0.586	0.416

Table 4: Silhouette IoU, using the 6-view multi-surface network (mean, higher is better).

	NovelView	NovelModel	NovelClass
View-centered	0.011	**0.016**	**0.0207**
Obj-centered	**0.004**	0.035	0.0503

Table 5: Depth error, using the 6-view multi-surface network (mean, lower is better).

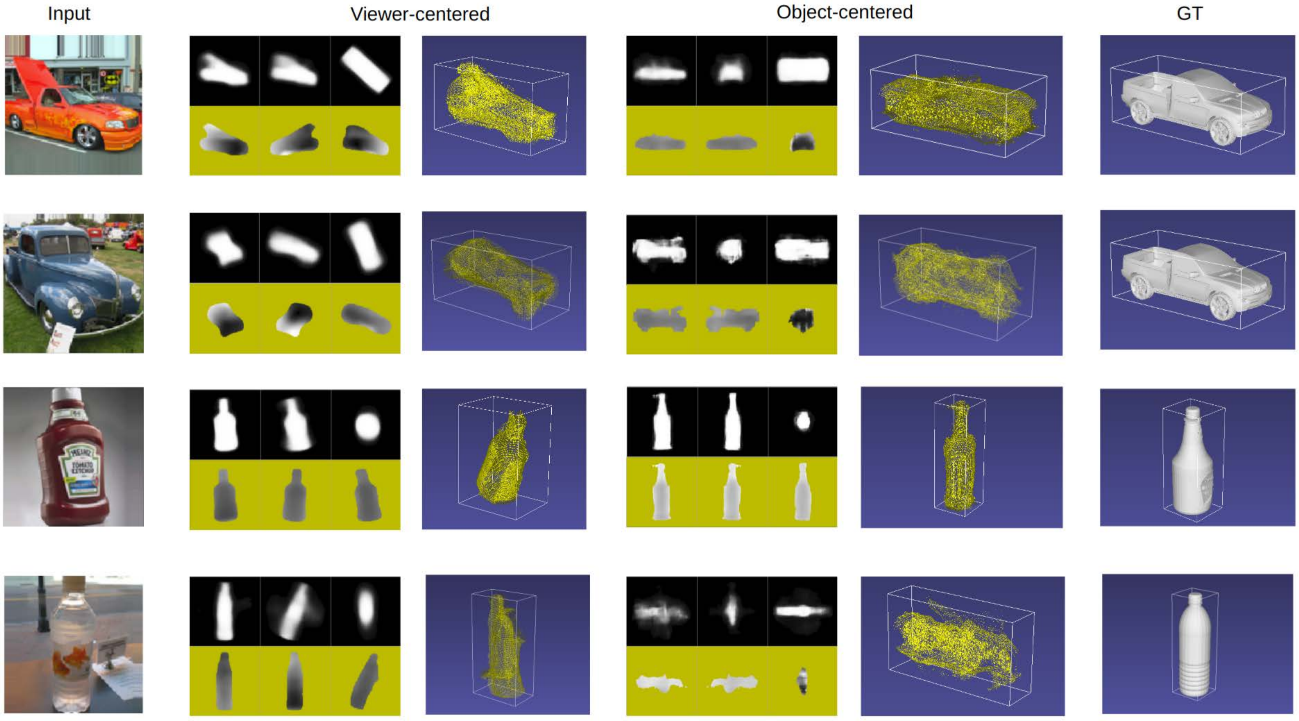

Figure 3: **RGB-based shape prediction examples:** On left, is the input image. We show predicted depth maps and silhouettes from three views and a merged point cloud from all views, produced by the networks trained with object-centered coordinates and with viewer-centered coordinates. Viewer-centered tends to generalize better while object-centered sometimes produces a model that looks good but is from entirely the wrong category. In viewer-centered, the encoder learns to map inputs together if they correspond to similar shapes in similar poses, learning a viewpoint-sensitive representation. In object-centered, the encoder learns to map different views of the same object together, learning a viewpoint-invariant representation.

[4] Simon Fuhrmann and Michael Goesele. Floating scale surface reconstruction. *ACM Transactions on Graphics (TOG)*, 33(4):46, 2014.

[5] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1966–1974. IEEE, 2015.

[6] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG)*, 32(3):29, 2013.

[7] Patricia A. McMullen and Martha J. Farah. Viewer-centered and object-centered representations in the recognition of naturalistic line drawings. *Psychological Science*, 2(4):275–278, 1991.

[8] Jason Rock, Tanmay Gupta, Justin Thorsen, JunYoung Gwak, Daeyun Shin, and Derek Hoiem. Completing 3d object shape from one depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2484–2493, 2015.

[9] Hao Su, Charles R. Qi, Yangyan Li, and Leonidas J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[10] Michael J Tarr and Steven Pinker. When does human object recognition use a viewer-centered reference frame? *Psychological Science*, 1(4):253–256, 1990.

[11] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In *European Conference on Computer Vision*, pages 322–337. Springer, 2016.

[12] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[13] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[14] Chuhang Zou, Ersin Yumer, Jimei Yang, Duygu Ceylan, and Derek Hoiem. 3d-prnn, generating shape primitives with recurrent neural networks. In *ICCV*, 2017.