Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling

Xingyuan Sun*[1,2] Jiajun Wu*[1] Xiuming Zhang[1] Zhoutong Zhang[1]
Chengkai Zhang[1] Tianfan Xue[3] Joshua B. Tenenbaum[1] William T. Freeman[1,3]

[1]Massachusetts Institute of Technology [2]Shanghai Jiao Tong University [3]Google Research

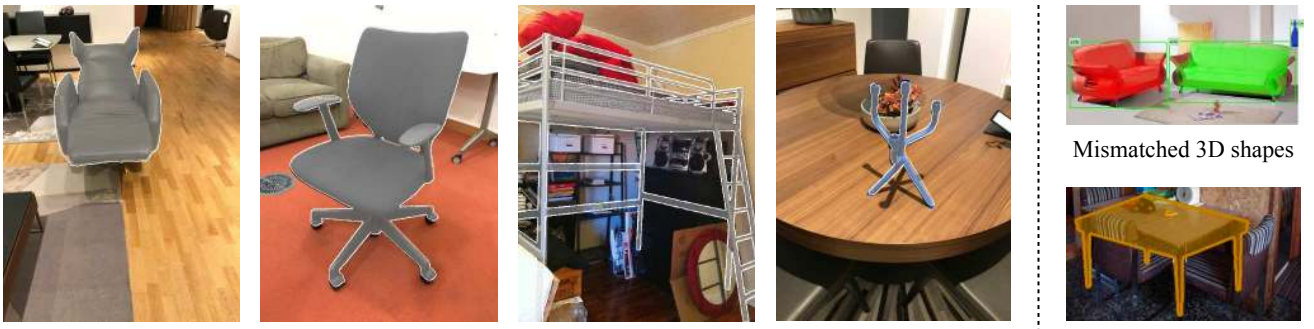

Figure 1: We present Pix3D, a new large-scale dataset of diverse image-shape pairs. Each 3D shape in Pix3D is associated with a rich and diverse set of images, each with an accurate 3D pose annotation to ensure precise 2D-3D alignment. In comparison, existing datasets have limitations: 3D models may not match the objects in images; pose annotations may be imprecise; or the dataset may be relatively small.

Abstract

We study 3D shape modeling from a single image and make contributions to it in three aspects. First, we present Pix3D, a large-scale benchmark of diverse image-shape pairs with pixel-level 2D-3D alignment. Pix3D has wide applications in shape-related tasks including reconstruction, retrieval, viewpoint estimation, etc. *Building such a large-scale dataset, however, is highly challenging; existing datasets either contain only synthetic data, or lack precise alignment between 2D images and 3D shapes, or only have a small number of images. Second, we calibrate the evaluation criteria for 3D shape reconstruction through behavioral studies, and use them to objectively and systematically benchmark cutting-edge reconstruction algorithms on Pix3D. Third, we design a novel model that simultaneously performs 3D reconstruction and pose estimation; our multi-task learning approach achieves state-of-the-art performance on both tasks.*

1. Introduction

The computer vision community has put major efforts in building datasets. In 3D vision, there are rich 3D CAD model repositories like ShapeNet [7] and the Princeton Shape Benchmark [49], large-scale datasets associating images and shapes like Pascal 3D+ [64] and ObjectNet3D [63], and benchmarks with fine-grained pose annotations for shapes in images like IKEA [38]. Why do we need one more?

Looking into Figure 1, we realize existing datasets have limitations for the task of modeling a 3D object from a single image. ShapeNet is a large dataset for 3D models, but does not come with real images; Pascal 3D+ and ObjectNet3D have real images, but the image-shape alignment is rough because the 3D models do not match the objects in images; IKEA has high-quality image-3D alignment, but it only contains 90 3D models and 759 images.

We desire a dataset that has all three merits—a large-scale dataset of real images and ground-truth shapes with precise 2D-3D alignment. Our dataset, named Pix3D, has 395 3D shapes of nine object categories. Each shape associates with a set of real images, capturing the exact object in diverse environments. Further, the 10,069 image-shape pairs have precise 3D annotations, giving pixel-level alignment between shapes and their silhouettes in the images.

Building such a dataset, however, is highly challenging. For each object, it is difficult to simultaneously collect its high-quality geometry and in-the-wild images. We can crawl

∗ indicates equal contributions.

many images of real-world objects, but we do not have access to their shapes; 3D CAD repositories offer object geometry, but do not come with real images. Further, for each image-shape pair, we need a precise pose annotation that aligns the shape with its projection in the image.

We overcome these challenges by constructing Pix3D in three steps. First, we collect a large number of image-shape pairs by crawling the web and performing 3D scans ourselves. Second, we collect 2D keypoint annotations of objects in the images on Amazon Mechanical Turk, with which we optimize for 3D poses that align shapes with image silhouettes. Third, we filter out image-shape pairs with a poor alignment and, at the same time, collect attributes (*i.e.*, truncation, occlusion) for each instance, again by crowdsourcing.

In addition to high-quality data, we need a proper metric to objectively evaluate the reconstruction results. A well-designed metric should reflect the visual appealingness of the reconstructions. In this paper, we calibrate commonly used metrics, including intersection over union, Chamfer distance, and earth mover's distance, on how well they capture human perception of shape similarity. Based on this, we benchmark state-of-the-art algorithms for 3D object modeling on Pix3D to demonstrate their strengths and weaknesses.

With its high-quality alignment, Pix3D is also suitable for object pose estimation and shape retrieval. To demonstrate that, we propose a novel model that performs shape and pose estimation simultaneously. Given a single RGB image, our model first predicts its 2.5D sketches, and then regresses the 3D shape and the camera parameters from the estimated 2.5D sketches. Experiments show that multi-task learning helps to boost the model's performance.

Our contributions are three-fold. First, we build a new dataset for single-image 3D object modeling; Pix3D has a diverse collection of image-shape pairs with precise 2D-3D alignment. Second, we calibrate metrics for 3D shape reconstruction based on their correlations with human perception, and benchmark state-of-the-art algorithms on 3D reconstruction, pose estimation, and shape retrieval. Third, we present a novel model that simultaneously estimates object shape and pose, achieving state-of-the-art performance on both tasks.

2. Related Work

Datasets of 3D shapes and scenes. For decades, researchers have been building datasets of 3D objects, either as a repository of 3D CAD models [4, 5, 49] or as images of 3D shapes with pose annotations [34, 47]. Both directions have witnessed the rapid development of web-scale databases: ShapeNet [7] was proposed as a large repository of more than 50K models covering 55 categories, and Xiang *et al*. built Pascal 3D+ [64] and ObjectNet3D [63], two large-scale datasets with alignment between 2D images and the 3D shape inside. While these datasets have helped to advance the field of 3D shape modeling, they have their respective limitations: datasets like ShapeNet or Elastic2D3D [32] do not have real images, and recent 3D reconstruction challenges using ShapeNet have to be exclusively on synthetic images [67]; Pascal 3D+ and ObjectNet3D have only rough alignment between images and shapes, because objects in the images are matched to a pre-defined set of CAD models, not their actual shapes. This has limited their usage as a benchmark for 3D shape reconstruction [59].

With depth sensors like Kinect [24, 27], the community has built various RGB-D or depth-only datasets of objects and scenes. We refer readers to the review article from Firman [14] for a comprehensive list. Among those, many object datasets are designed for benchmarking robot manipulation [6, 23, 33, 51]. These datasets often contain a relatively small set of hand-held objects in front of clean backgrounds. Tanks and Temples [30] is an exciting new benchmark with 14 scenes, designed for high-quality, large-scale, multi-view 3D reconstruction. In comparison, our dataset, Pix3D, focuses on reconstructing a 3D object from a single image, and contains much more real-world objects and images.

Probably the dataset closest to Pix3D is the large collection of object scans from Choi *et al*. [8], which contains a rich and diverse set of shapes, each with an RGB-D video. Their dataset, however, is not ideal for single-image 3D shape modeling for two reasons. First, the object of interest may be truncated throughout the video; this is especially the case for large objects like sofas. Second, their dataset does not explore the various contexts that an object may appear in, as each shape is only associated with a single scan. In Pix3D, we address both problems by leveraging powerful web search engines and crowdsourcing.

Another closely related benchmark is IKEA [38], which provides accurate alignment between images of IKEA objects and 3D CAD models. This dataset is therefore particularly suitable for fine pose estimation. However, it contains only 759 images and 90 shapes, relatively small for shape modeling*. In contrast, Pix3D contains 10,069 images (13.3x) and 395 shapes (4.4x) of greater variations.

Researchers have also explored constructing scene datasets with 3D annotations. Notable attempts include LabelMe-3D [46], NYU-D [50], SUN RGB-D [53], KITTI [16], and modern large-scale RGB-D scene datasets [10, 40, 54]. These datasets are either synthetic or contain only 3D surfaces of real scenes. Pix3D, in contrast, offers accurate alignment between 3D object shape and 2D images in the wild.

Single-image 3D reconstruction. The problem of recovering object shape from a single image is challenging, as it requires both powerful recognition systems and prior shape knowledge. Using deep convolutional networks, researchers have made significant progress in recent years [9, 17, 21, 29, 41, 43, 56, 59, 60, 62, 66, 52, 61]. While most of these approaches represent objects in voxels, there have also been

*Only 90 of the 219 shapes in the IKEA dataset have associated images.

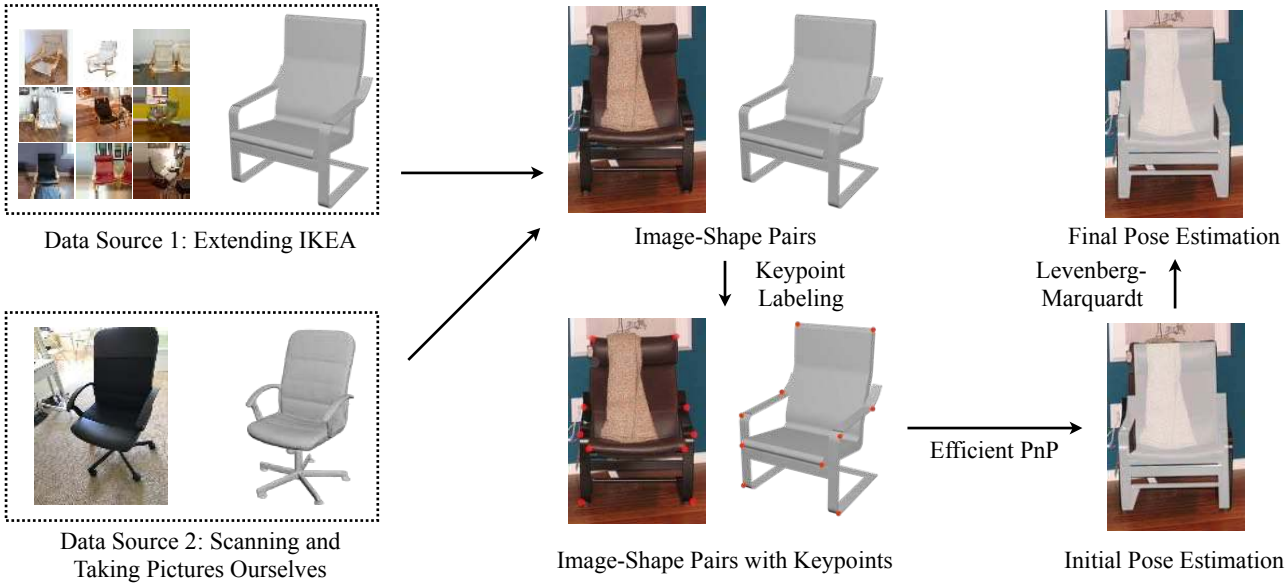

Figure 2: We build the dataset in two steps. First, we collect image-shape pairs by crawling web images of IKEA furniture as well as scanning objects and taking pictures ourselves. Second, we align the shapes with their 2D silhouettes by minimizing the 2D coordinates of the keypoints and their projected positions from 3D, using the Efficient PnP and the Levenberg-Marquardt algorithm.

attempts to reconstruct objects in point clouds [12] or octave trees [44, 57]. In this paper, we demonstrate that our newly proposed Pix3D serves as an ideal benchmark for evaluating these algorithms. We also propose a novel model that jointly estimates an object's shape and its 3D pose.

Shape retrieval. Another related research direction is retrieving similar 3D shapes given a single image, instead of reconstructing the object's actual geometry [1, 15, 19, 48]. Pix3D contains shapes with significant inter-class and intra-class variations, and is therefore suitable for both general-purpose and fine-grained shape retrieval tasks.

3D pose estimation. Many of the aforementioned object datasets include annotations of object poses [34, 38, 47, 63, 64]. Researchers have also proposed numerous methods on 3D pose estimation [13, 42, 55, 58]. In this paper, we show that Pix3D is also a proper benchmark for this task.

3. Building Pix3D

Figure 2 summarizes how we build Pix3D. We collect raw images from web search engines and shapes from 3D repositories; we also take pictures and scan shapes ourselves. Finally, we use labeled keypoints on both 2D images and 3D shapes to align them.

3.1. Collecting Image-Shape Pairs

We obtain raw image-shape pairs in two ways. One is to crawl images of IKEA furniture from the web and align them with CAD models provided in the IKEA dataset [38]. The other is to directly scan 3D shapes and take pictures.

Extending IKEA. The IKEA dataset [38] contains 219 high-quality 3D models of IKEA furniture, but has only 759 images for 90 shapes. Therefore, we choose to keep the 3D shapes from IKEA dataset, but expand the set of 2D images using online image search engines and crowdsourcing.

For each 3D shape, we first search for its corresponding 2D images through Google, Bing, and Baidu, using its IKEA model name as the keyword. We obtain 104,220 images for the 219 shapes. We then use Amazon Mechanical Turk (AMT) to remove irrelevant ones. For each image, we ask three AMT workers to label whether this image matches the 3D shape or not. For images whose three responses differ, we ask three additional workers and decide whether to keep them based on majority voting. We end up with 14,600 images for the 219 IKEA shapes.

3D scan. We scan non-IKEA objects with a Structure Sensor[†] mounted on an iPad. We choose to use the Structure Sensor because its mobility enables us to capture a wide range of shapes.

The iPad RGB camera is synchronized with the depth sensor at 30 Hz, and calibrated by the Scanner App provided by Occipital, Inc.[‡] The resolution of RGB frames is 2592×1936, and the resolution of depth frames is 320×240. For each object, we take a short video and fuse the depth data to get its 3D mesh by using fusion algorithm provided by Occipital, Inc. We also take 10–20 images for each scanned object in front of various backgrounds from different viewpoints, making sure the object is neither cropped nor occluded. In total, we have scanned 209 objects and taken 2,313 images. Combining these with the IKEA shapes and images, we have 418 shapes and 16,913 images altogether.

[†] https://structure.io
[‡] https://occipital.com

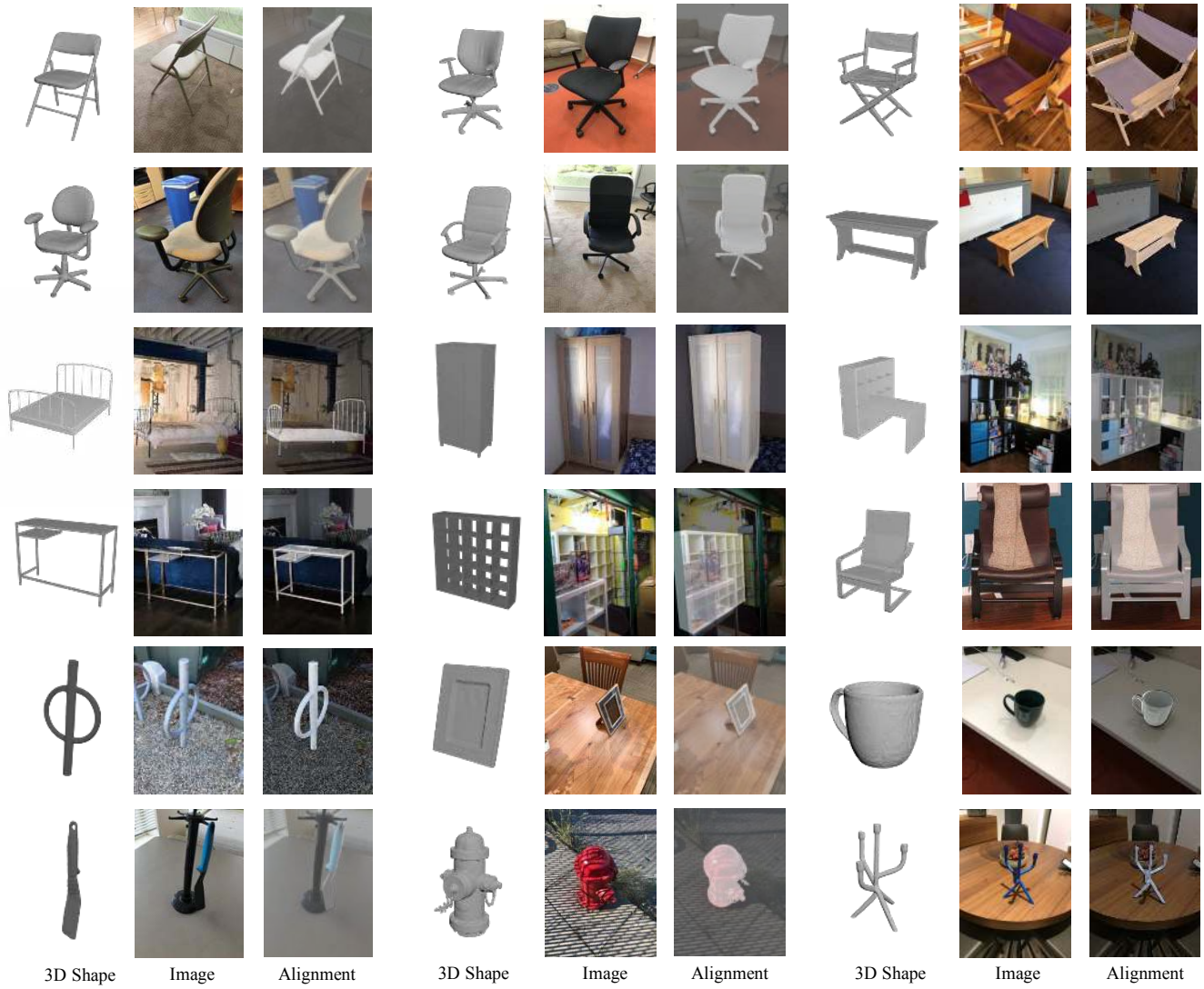

Figure 3: Sample images and shapes in Pix3D. From left to right: 3D shape, image, and alignment. Rows 1–2 show some chairs we scanned, rows 3–4 show a few IKEA objects, and rows 5–6 show some objects of other categories we scanned.

3.2. Image-Shape Alignment

To align a 3D CAD model with its projection in a 2D image, we need to solve for its 3D pose (translation and rotation), and the camera parameters used to capture the image.

We use a keypoint-based method inspired by Lim *et al*. [38]. Denote the keypoints' 2D coordinates as $\boldsymbol{X}_{\text{2D}} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ and their corresponding 3D coordinates as $\boldsymbol{X}_{\text{3D}} = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$. We solve for camera parameters and 3D poses that minimize the reprojection error of the keypoints. Specifically, we want to find the projection matrix $\boldsymbol{P}$ that minimizes

$$\mathcal{L}(\boldsymbol{P}; \boldsymbol{X}_{\text{3D}}, \boldsymbol{X}_{\text{2D}}) = \sum_i \|\text{Proj}_{\boldsymbol{P}}(\mathbf{X}_i) - \mathbf{x}_i\|_2^2, \qquad (1)$$

where $\text{Proj}_{\boldsymbol{P}}(\cdot)$ is the projection function.

Under the central projection assumption (zero-skew, square pixel, and the optical center is at the center of the frame), we have $\boldsymbol{P} = \boldsymbol{K}[\boldsymbol{R}|\boldsymbol{T}]$, where $\boldsymbol{K}$ is the camera intrinsic matrix; $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ and $\boldsymbol{T} \in \mathbb{R}^3$ represent the object's 3D rotation and 3D translation, respectively. We know

$$\boldsymbol{K} = \begin{bmatrix} f & 0 & w/2 \\ 0 & f & h/2 \\ 0 & 0 & 1 \end{bmatrix}, \qquad (2)$$

where f is the focal length, and w and h are the width and height of the image. Therefore, there are altogether seven parameters to be estimated: rotations θ, ϕ, ψ, translations x, y, z, and focal length f (Rotation matrix R is determined by θ, ϕ, and ψ).

To solve Equation 1, we first calculate a rough 3D pose using the Efficient P*n*P algorithm [35] and then refine it using the Levenberg-Marquardt algorithm [36, 39], as shown in Figure 2. Details of each step are described below.

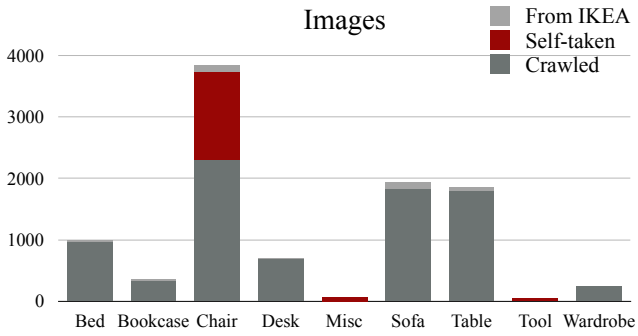

Figure 4: Class distribution of the images

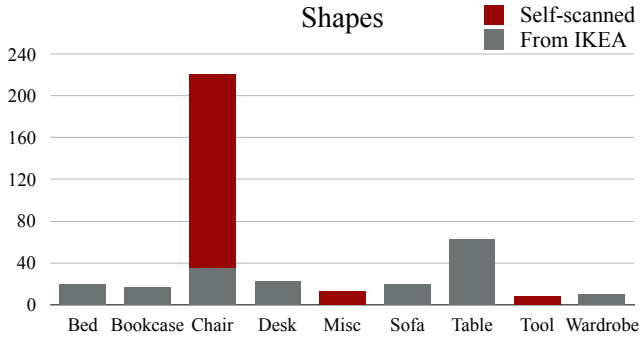

Figure 5: Class distribution of the shapes

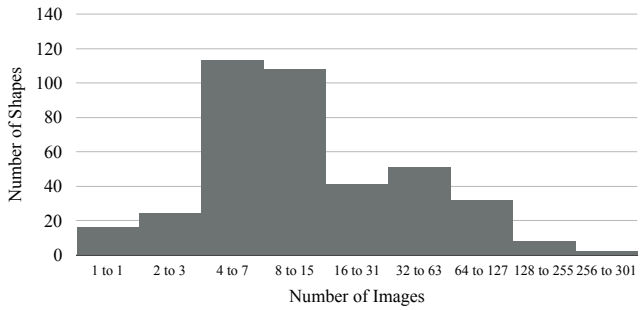

Figure 6: Number of images available for each shape.

Efficient P*n*P. Perspective-n-Point (PnP) is the problem of estimating the pose of a calibrated camera given paired 3D points and 2D projections. The Efficient PnP (EPnP) algorithm solves the problem using virtual control points [36]. Because EPnP does not estimate the focal length, we enumerate the focal length f from 300 to 2,000 with a step size of 10, solve for the 3D pose with each f, and choose the one with the minimum projection error.

The Levenberg-Marquardt algorithm (LMA). We take the output of EPnP with 50 random disturbances as the initial states, and run LMA on each of them. Finally, we choose the solution with the minimum projection error.

Implementation details. For each 3D shape, we manually label its 3D keypoints. The number of keypoints ranges from 8 to 24. For each image, we ask three AMT workers to label if each keypoint is visible on the image, and if so, where it is. We only consider visible keypoints during the optimization.

The 2D keypoint annotations are noisy, which severely hurts the performance of the optimization algorithm. We try two methods to increase its robustness. The first is to use RANSAC. The second is to use only a subset of 2D keypoint annotations. For each image, denote $C = \{c_1, c_2, c_3\}$ as its three sets of human annotations. We then enumerate the seven nonempty subsets $C_k \subseteq C$; for each keypoint, we compute the median of its 2D coordinates in C_k. We apply our optimization algorithm on every subset C_k, and keep the output with the minimum projection error. After that, we let three AMT workers choose, for each image, which of the two methods offers better alignment, or neither performs well. At the same time, we also collect attributes (*i.e.*, truncation, occlusion) for each image. Finally, we fine-tune the annotations ourselves using the GUI offered in ObjectNet3D [63]. Altogether there are 395 3D shapes and 10,069 images. Sample 2D-3D pairs are shown in Figure 3.

4. Exploring Pix3D

We now present some statistics of Pix3D, and contrast it with its predecessors.

Dataset statistics. Figures 4 and 5 show the category distributions of 2D images and 3D shapes in Pix3D; Figure 6 shows the distribution of the number of images each model has. Our dataset covers a large variety of shapes, each of which has a large number of in-the-wild images. Chairs cover the significant part of Pix3D, because they are common, highly diverse, and well-studied by recent literature [11, 59, 20].

Quantitative evaluation. As a quantitative comparison on the quality of Pix3D and other datasets, we randomly select 25 chair and 25 sofa images from PASCAL 3D+ [64], ObjectNet3D [63], IKEA [38], and Pix3D. For each image, we render the projected 2D silhouette of the shape using its pose annotation provided by the dataset. We then manually annotate the ground truth object masks in these images, and calculate Intersection over Union (IoU) between the projections and the ground truth. For each image-shape pair, we also ask 50 AMT workers whether they think the image is picturing the 3D *ground truth* shape provided by the dataset.

From Table 1, we see that Pix3D has much higher IoUs than PASCAL 3D+ and ObjectNet3D, and slightly higher IoUs compared with the IKEA dataset. Humans also feel IKEA and Pix3D have matched images and shapes, but not PASCAL 3D+ or ObjectNet3D. In addition, we observe that many CAD models in the IKEA dataset are of an incorrect scale, making it challenging to align the shapes with images. For example, there are only 15 unoccluded and untruncated images of sofas in IKEA, while Pix3D has 1,092.

5. Metrics

Designing a good evaluation metric is important to encourage researchers to design algorithms that reconstruct high-quality 3D geometry, rather than low-quality 3D reconstruction that overfits to a certain metric.

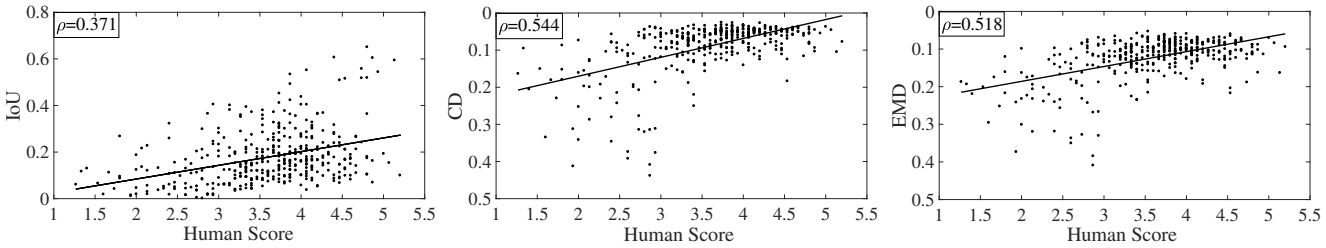

Figure 7: Scatter plots between humans' ratings of reconstructed shapes and their IoU, CD, and EMD. The three metrics have a Pearson's coefficient of 0.371, 0.544, and 0.518, respectively.

	Chairs		Sofas	
	IoU	Match?	IoU	Match?
PASCAL 3D+ [64]	0.514	0.00	0.813	0.00
ObjectNet3D [63]	0.570	0.16	0.773	0.08
IKEA [38]	0.748	**1.00**	0.918	**1.00**
Pix3D (ours)	**0.835**	**1.00**	**0.926**	**1.00**

Table 1: We compute the Intersection over Union (IoU) between manually annotated 2D masks and the projections of 3D shapes. We also ask humans to judge whether the object in the images matches the provided shape.

Many 3D reconstruction papers use Intersection over Union (IoU) to evaluate the similarity between ground truth and reconstructed 3D voxels, which may significantly deviate from human perception. In contrast, metrics like shortest distance and geodesic distance are more commonly used than IoU for matching meshes in graphics [31, 25]. Here, we conduct behavioral studies to calibrate IoU, Chamfer distance (CD) [2], and Earth Mover's distance (EMD) [45] on how well they reflect human perception.

5.1. Definitions

The definition of IoU is straightforward. For Chamfer distance (CD) and Earth Mover's distance (EMD), we first convert voxels to point clouds, and then compute CD and EMD between pairs of point clouds.

Voxels to a point cloud. We first extract the isosurface of each predicted voxel using the Lewiner marching cubes [37] algorithm. In practice, we use 0.1 as a universal surface value for extraction. We then uniformly sample points on the surface meshes and create the densely sampled point clouds. Finally, we randomly sample 1,024 points from each point cloud and normalize them into a unit cube for distance calculation.

Chamfer distance (CD). The Chamfer distance (CD) between $S_1, S_2 \subseteq \mathbb{R}^3$ is defined as

$$\text{CD}(S_1, S_2) = \frac{1}{|S_1|} \sum_{x \in S_1} \min_{y \in S_2} \|x-y\|_2 + \frac{1}{|S_2|} \sum_{y \in S_2} \min_{x \in S_1} \|x-y\|_2. \quad (3)$$

For each point in each cloud, CD finds the nearest point in the other point set, and sums the distances up. CD has been used in recent shape retrieval challenges [67].

	IoU	EMD	CD	Human
IoU	1	0.55	0.60	0.32
EMD	0.55	1	0.78	0.43
CD	0.60	0.78	1	0.49
Human	0.32	0.43	0.49	1

Table 2: Spearman's rank correlation coefficients between different metrics. IoU, EMD, and CD have a correlation coefficient of 0.32, 0.43, and 0.49 with human judgments, respectively.

Earth Mover's distance (EMD). We follow the definition of EMD in Fan *et al.* [12]. The Earth Mover's distance (EMD) between $S_1, S_2 \subseteq \mathbb{R}^3$ (of equal size, *i.e.*, $|S_1| = |S_2|$) is

$$\text{EMD}(S_1, S_2) = \frac{1}{|S_1|} \min_{\phi: S_1 \to S_2} \sum_{x \in S_1} ||x - \phi(x)||_2, \quad (4)$$

where $\phi : S_1 \to S_2$ is a bijection. We divide EMD by the size of the point cloud for normalization. In practice, calculating the exact EMD value is computationally expensive; we instead use a $(1+\epsilon)$ approximation algorithm [3].

5.2. Experiments

We then conduct two user studies to compare these metrics and benchmark how they capture human perception.

Which one looks better? We run three shape reconstructions algorithms (3D-R2N2 [9], DRC [59], and 3D-VAE-GAN [62]) on 200 randomly selected images of chairs. We then, for each image and every pair of its three constructions, ask three AMT workers to choose the one that looks closer to the object in the image. We also compute how each pair of objects rank in each metric. Finally, we calculate the Spearman's rank correlation coefficients between different metrics (*i.e.*, IoU, EMD, CD, and human perception). Table 2 suggests that EMD and CD correlate better with human ratings.

How good is it? We randomly select 400 images, and show each of them to 15 AMT workers, together with the voxel prediction by DRC [59] and the ground truth shape. We then ask them to rate the reconstruction, on a scale of 1 to 7, based on how similar it is to the ground truth. The scatter plot in Figure 7 suggests that CD and EMD have higher Pearson's coefficients with human responses.

Figure 8: Our 3D reconstructions of chairs on Pix3D. For each of the images, we show two views of the predicted voxels.

	IoU	EMD	CD
3D-R2N2 [9]	0.136	0.211	0.239
PSGN [12]	N/A	0.216	0.200
3D-VAE-GAN [62]	0.171	0.176	0.182
DRC [59]	0.265	0.144	0.160
MarrNet* [60]	0.231	0.136	0.144
AtlasNet [18]	N/A	0.128	0.125
Ours (w/o Pose)	0.267	0.124	0.124
Ours (w/ Pose)	**0.282**	**0.118**	**0.119**

Table 3: Results of 3D shape reconstruction. Our model gets the highest IoU, EMD, and CD. We also compare our full model with a variant that does not have the view estimator. Results show that multi-task learning helps boost its performance. As MarrNet and PSGN predict viewer-centered shapes, while the other methods are object-centered, we rotate their reconstructions into the canonical view using ground truth pose annotations before evaluation.

6. Approach

Pix3D serves as a benchmark for shape modeling tasks including reconstruction, retrieval, and pose estimation. Here, we design a new model that simultaneously performs shape reconstruction and pose estimation, and evaluate it on Pix3D.

Our model is an extension of MarrNet [60], both of which use 2.5D sketches (the object's depth, surface normals, and silhouette) as an intermediate representation. It contains four modules: (1) a 2.5D sketch estimator that predicts the depth, surface normals, and silhouette of the object; (2) a 2.5D sketch encoder that encodes the 2.5D sketches into a low-dimensional latent vector; (3) a 3D shape decoder and (4) a view estimator that decodes a latent vector into a 3D shape and camera parameters, respectively. Different from MarrNet [60], our model has an additional branch for pose estimation. We briefly describe them below, and please refer to the supplementary material for more details.

2.5D sketch estimator. The first module takes an RGB image as input and predicts the object's 2.5D sketches (its depth, surface normals, and silhouette). We use an encoder-decoder network. The encoder is based on a ResNet-18 [22] and turns a 256×256 image into 384 feature maps of size 16×16; the decoder has three branches for depth, surface normals, and silhouette, respectively, each consisting of four sets of 5×5 transposed convolutional, batch normalization and ReLU layers, followed by one 5×5 convolutional layer. All output sketches are of size 256×256.

2.5D sketch encoder. We use a modified ResNet-18 [22] that takes a four-channel image (three for surface normals and one for depth). Each channel is masked by the predicted silhouette. A final linear layer outputs a 200-D latent vector.

3D shape decoder. Our 3D shape decoder has five sets of 4×4×4 transposed convolutional, batch-norm, and ReLU layers, followed by a 4×4×4 transposed convolutional layer. It outputs a voxelized shape of size 128×128×128 in the object's canonical view.

View estimator. The view estimator contains three sets of linear, batch normalization, and ReLU layers, followed by two parallel linear and softmax layers that predict the shape's azimuth and elevation, respectively. Here, we treat pose estimation as a classification problem, where the 360-degree azimuth angle is divided into 24 bins and the 180-degree elevation angle is divided into 12 bins.

Training paradigm. For training, we use Mitsuba [26] to render each chair in ShapeNet [7] from 20 random views using three types of backgrounds: 1/3 on a white background, 1/3 on high-dynamic-range backgrounds with illumination channels, and 1/3 on backgrounds randomly sampled from the SUN database [65]. We augment our training data by random color and light jittering.

We first train the 2.5D sketch estimator. We then train the 2.5D sketch encoder and the 3D shape decoder (and the view estimator if we're predicting the pose) jointly. We finally concatenate them for prediction.

7. Experiments

We now evaluate our model and state-of-the-art algorithms on single-image 3D shape reconstruction, retrieval, and pose estimation, all using Pix3D. For all experiments, we use the 2,894 untruncated and unoccluded chair images.

3D shape reconstruction. We compare our model, with and without the pose estimation branch, with the state-of-the-art systems, including 3D-VAE-GAN [62], 3D-R2N2 [9],

Figure 9: Our shape retrievals of chairs on Pix3D. Here we show top-8 retrieval results from our proposed method (both *with pose estimation* version and *without pose estimation* version). The *with pose estimation* version tends to retrieval images with a similar pose.

Figure 10: Our pose estimations of chairs on Pix3D. For each of the images, our method only predicts azimuth and elevation.

	R@1	R@2	R@4	R@8	R@16	R@32
3D-VAE-GAN [62]	0.02	0.03	0.07	0.12	0.21	0.34
MarrNet [60]	0.42	0.51	0.57	0.64	0.71	0.78
Ours (w/ Pose)	0.42	0.48	0.55	0.63	0.70	0.76
Ours (w/o Pose)	**0.53**	**0.62**	**0.71**	**0.78**	**0.85**	**0.90**

Table 4: Results for image-based shape retrieval, where R@K stands for Recall@K. Our model (without the pose estimation module) achieves the highest numbers. Our model (with the pose estimation module) does not perform as well, because it sometimes retrieves images of objects with the same pose, but not exactly the same shape.

DRC [59], and MarrNet [60]. We use pre-trained models offered by the authors and we crop the input images as required by each algorithm. The results are shown in Table 3 and Figure 8. Our model outperforms the state-of-the-arts in all metrics. Our full model gets better results compared with the variant without the view estimator, suggesting multi-task learning helps to boost its performance. Also note the discrepancy among metrics: MarrNet has a lower IoU than DRC, but according to EMD and CD, it performs better.

Image-based, fine-grained shape retrieval. For shape retrieval, we compare our model with 3D-VAE-GAN [62] and MarrNet [60]. We use the latent vector from each algorithm as its embedding of the input image, and use L2 distance for image retrieval. For each test image, we retrieve its K nearest neighbors from the test set, and use Recall@K [28] to compute how many retrieved images are actually depicting the same shape. Here we do not consider images whose shape is not captured by any other images in the test set. The results are shown in Table 4 and Figure 9. Our model (without the pose estimation module) achieves the highest numbers; our model (with the pose estimation module) does not perform as well, because it sometimes retrieves images of objects with the same pose, but not exactly the same shape.

	Azimuth				Elevation		
# of views	4	8	12	24	4	6	12
Render for CNN	0.71	0.63	0.56	0.40	0.57	0.56	0.37
Ours	**0.76**	**0.73**	**0.61**	**0.49**	**0.87**	**0.70**	**0.61**

Table 5: Results of 3D pose estimation. Our model outperforms Render for CNN [55] in both azimuth and elevation.

3D pose estimation. We compare our method with Render for CNN [55]. We calculate the classification accuracy for both azimuth and elevation, where the azimuth is divided into 24 bins and the elevation into 12 bins. Table 5 suggests that our model outperforms Render for CNN in pose estimation. Qualitative results are included in Figure 10.

8. Conclusion

We have presented Pix3D, a large-scale dataset of well-aligned 2D images and 3D shapes. We have also explored how three commonly used metrics correspond to human perception through two behavioral studies and proposed a new model that simultaneously performs shape reconstruction and pose estimation. Experiments showed that our model achieved state-of-the-art performance on 3D reconstruction, shape retrieval, and pose estimation. We hope our paper will inspire future research in single-image 3D shape modeling.

Acknowledgements. This work is supported by NSF #1212849 and #1447476, ONR MURI N00014-16-1-2007, the Center for Brain, Minds and Machines (NSF STC award CCF-1231216), the Toyota Research Institute, and Shell Research. J. Wu is supported by a Facebook fellowship.

References

[1] M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014. 3

[2] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *IJCAI*, 1977. 6

[3] D. P. Bertsekas. A distributed asynchronous relaxation algorithm for the assignment problem. In *CDC*, 1985. 6

[4] F. Bogo, J. Romero, M. Loper, and M. J. Black. Faust: Dataset and evaluation for 3d mesh registration. In *CVPR*, 2014. 2

[5] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008. 2

[6] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE RAM*, 22(3):36–52, 2015. 2

[7] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv:1512.03012*, 2015. 1, 2, 7

[8] S. Choi, Q.-Y. Zhou, S. Miller, and V. Koltun. A large dataset of object scans. *arXiv:1602.02481*, 2016. 2

[9] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016. 2, 6, 7

[10] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2

[11] A. Dosovitskiy, J. Springenberg, M. Tatarchenko, and T. Brox. Learning to generate chairs, tables and cars with convolutional networks. *IEEE TPAMI*, 39(4):692–705, 2017. 5

[12] H. Fan, H. Su, and L. Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017. 3, 6, 7

[13] S. Fidler, S. J. Dickinson, and R. Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In *NIPS*, 2012. 3

[14] M. Firman. Rgbd datasets: Past, present and future. In *CVPR Workshop*, 2016. 2

[15] M. Fisher and P. Hanrahan. Context-based search for 3d models. *ACM TOG*, 29(6):182, 2010. 3

[16] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2

[17] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 2

[18] T. Groueix, M. Fisher, V. G. Kim, B. Russell, and M. Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *CVPR*, 2018. 7

[19] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik. Aligning 3d models to rgb-d images of cluttered scenes. In *CVPR*, 2015. 3

[20] J. Gwak, C. B. Choy, M. Chandraker, A. Garg, and S. Savarese. Weakly supervised 3d reconstruction with adversarial constraint. In *3DV*, 2017. 5

[21] C. Häne, S. Tulsiani, and J. Malik. Hierarchical surface prediction for 3d object reconstruction. In *3DV*, 2017. 2

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2015. 7

[23] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *WACV*, 2017. 2

[24] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. A. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. J. Davison, and A. W. Fitzgibbon. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In J. S. Pierce, M. Agrawala, and S. R. Klemmer, editors, *UIST*, 2011. 2

[25] V. Jain and H. Zhang. Robust 3d shape correspondence in the spectral domain. In *Shape Modeling and Applications*, 2006. 6

[26] W. Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org. 7

[27] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3-d object dataset: Putting the kinect to work. In *ICCV Workshop*, 2011. 2

[28] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE TPAMI*, 33(1):117–128, 2011. 8

[29] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Category-specific object reconstruction from a single image. In *CVPR*, 2015. 2

[30] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM TOG*, 36(4):78, 2017. 2

[31] V. Kreavoy, D. Julius, and A. Sheffer. Model composition from interchangeable components. In *PG*, 2007. 6

[32] Z. Lahner, E. Rodola, F. R. Schmidt, M. M. Bronstein, and D. Cremers. Efficient globally optimal 2d-to-3d deformable shape matching. In *CVPR*, 2016. 2

[33] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *ICRA*, 2011. 2

[34] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *CVPR*, 2003. 2, 3

[35] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o (n) solution to the pnp problem. *IJCV*, 81(2):155–166, 2009. 4

[36] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944. 4, 5

[37] T. Lewiner, H. Lopes, A. W. Vieira, and G. Tavares. Efficient implementation of marching cubes' cases with topological guarantees. *Journal of Graphics Tools*, 8(2):1–15, 2003. 6

[38] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing ikea objects: Fine pose estimation. In *ICCV*, 2013. 1, 2, 3, 4, 5, 6

[39] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963. 4

[40] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *ICCV*, 2017. 2

[41] D. Novotny, D. Larlus, and A. Vedaldi. Learning 3d object categories by looking around them. In *ICCV*, 2017. 2

[42] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *CVPR*, 2009. 3

[43] D. J. Rezende, S. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess. Unsupervised learning of 3d structure from images. In *NIPS*, 2016. 2

[44] G. Riegler, A. O. Ulusoys, and A. Geiger. Octnet: Learning deep 3d representations at high resolutions. In *CVPR*, 2016. 3

[45] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000. 6

[46] B. C. Russell and A. Torralba. Building a database of 3d scenes from user annotations. In *CVPR*, 2009. 2

[47] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *ICCV*, 2007. 2, 3

[48] M. Savva, F. Yu, H. Su, M. Aono, B. Chen, D. Cohen-Or, W. Deng, H. Su, S. Bai, X. Bai, et al. Shrec17 track: large-scale 3d shape retrieval from shapenet core55. In *Eurographics Workshop on 3D Object Retrieval*, 2016. 3

[49] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The princeton shape benchmark. In *Shape Modeling Applications*, 2004. 1, 2

[50] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[51] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel. Bigbird: A large-scale 3d database of object instances. In *ICRA*, 2014. 2

[52] A. A. Soltani, H. Huang, J. Wu, T. D. Kulkarni, and J. B. Tenenbaum. Synthesizing 3d shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In *CVPR*, 2017. 2

[53] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 2

[54] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017. 2

[55] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *ICCV*, 2015. 3, 8

[56] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3d models from single images with a convolutional network. In *ECCV*, 2016. 2

[57] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *ICCV*, 2017. 3

[58] S. Tulsiani and J. Malik. Viewpoints and keypoints. In *CVPR*, 2015. 3

[59] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017. 2, 5, 6, 7, 8

[60] J. Wu, Y. Wang, T. Xue, X. Sun, W. T. Freeman, and J. B. Tenenbaum. MarrNet: 3D Shape Reconstruction via 2.5D Sketches. In *NIPS*, 2017. 2, 7, 8

[61] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. In *ECCV*, 2016. 2

[62] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. In *NIPS*, 2016. 2, 6, 7, 8

[63] Y. Xiang, W. Kim, W. Chen, J. Ji, C. Choy, H. Su, R. Mottaghi, L. Guibas, and S. Savarese. Objectnet3d: A large scale database for 3d object recognition. In *ECCV*, 2016. 1, 2, 3, 5, 6

[64] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014. 1, 2, 3, 5, 6

[65] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 7

[66] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *NIPS*, 2016. 2

[67] L. Yi, H. Su, L. Shao, M. Savva, H. Huang, Y. Zhou, B. Graham, M. Engelcke, R. Klokov, V. Lempitsky, et al. Large-scale 3d shape reconstruction and segmentation from shapenet core55. *arXiv:1710.06104*, 2017. 2, 6