Learning Shape Priors for Single-View 3D Completion and Reconstruction

Jiajun Wu*[1], Chengkai Zhang*[1], Xiuming Zhang[1], Zhoutong Zhang[1], William T. Freeman[1,2], and Joshua B. Tenenbaum[1]

[1] MIT CSAIL, Cambridge MA 02139, USA
[2] Google Research, Cambridge MA 02139, USA

Abstract. The problem of single-view 3D shape completion or reconstruction is challenging, because among the many possible shapes that explain an observation, most are implausible and do not correspond to natural objects. Recent research in the field has tackled this problem by exploiting the expressiveness of deep convolutional networks. In fact, there is another level of ambiguity that is often overlooked: among plausible shapes, there are still multiple shapes that fit the 2D image equally well; *i.e.*, the ground truth shape is non-deterministic given a single-view input. Existing fully supervised approaches fail to address this issue, and often produce blurry mean shapes with smooth surfaces but no fine details. In this paper, we propose *ShapeHD*, pushing the limit of single-view shape completion and reconstruction by integrating deep generative models with adversarially learned shape priors. The learned priors serve as a regularizer, penalizing the model only if its output is unrealistic, not if it deviates from the ground truth. Our design thus overcomes both levels of ambiguity aforementioned. Experiments demonstrate that ShapeHD outperforms state of the art by a large margin in both shape completion and shape reconstruction on multiple real datasets.

Keywords: Shape priors · Shape completion · 3D reconstruction

1 Introduction

Let's start with a game: each of the two instances in Figure 1 shows a depth or color image and two different 3D shape interpretations. Which one looks better?

We asked this question to 100 people on Amazon Mechanical Turk. 59% of them preferred interpretation A of the airplane, and 35% preferred interpretation A of the car. These numbers suggest that people's opinions diverge on these two cases, indicating that these reconstructions are close in quality, and their perceptual differences are relatively minor.

Actually, for each instance, one of the reconstructions is the output of the model introduced in this paper, and the other is the ground truth shape. Answers are available in the footnote.

In this paper, we aim to push the limits of 3D shape completion from a single depth image, and of 3D shape reconstruction from a single color image. Recently,

* J. Wu and C. Zhang contributed equally to this work.

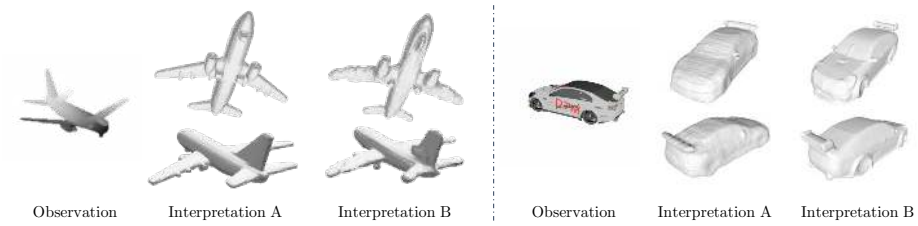


Fig. 1. Our model completes or reconstructs the object's full 3D shape with fine details from a single depth or RGB image. In this figure, we show two examples, each consisting of an input image, two views of its ground truth shape, and two views of our results. Our reconstructions are of high quality with fine details, and are preferred by humans 41% and 35% of the time in behavioral studies, respectively. Our model takes a single feed-forward pass without any post-processing during testing, and is thus highly efficient (< 100 ms) and practically useful. Answers are available in the footnote.

researchers have made impressive progress on the these tasks [7, 52, 8], making use of gigantic 3D datasets [5, 60, 59]. Many of these methods tackle the ill-posed nature of the problem by using deep convolutional networks to regress possible 3D shapes. Leveraging the power of deep generative models, their systems learn to avoid producing implausible shapes (Figure 2b).

However, from Figure 2c we realize that there is still ambiguity that a supervisedly trained network fails to model. From just a single view, there exist multiple natural shapes that explain the observation equally well. In other words, there is no deterministic ground truth for each observation. Through pure supervised learning, the network tends to generate mean shapes that minimize its penalty precisely due to this ambiguity.

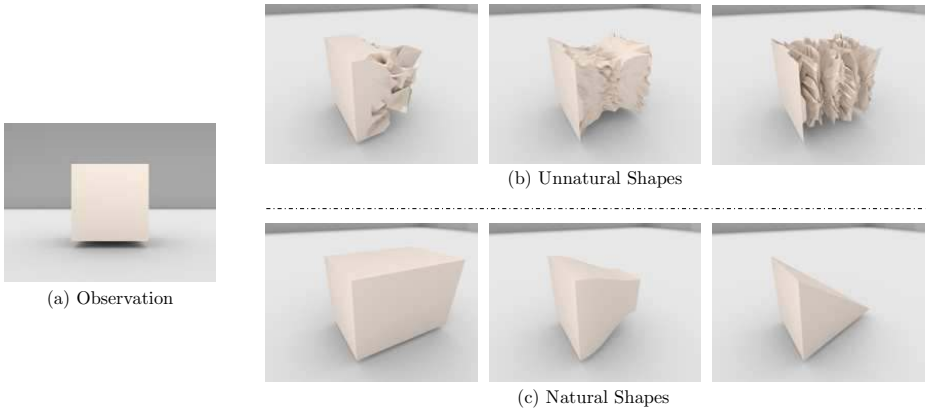
To tackle this, we propose ShapeHD, which completes or reconstructs a 3D shape by combining deep volumetric convolutional networks with adversarially learned shape priors. The learned shape priors penalize the model only if the generated shape is unrealistic, not if it deviates from the ground truth. This overcomes the difficulty discussed above. Our model characterizes this naturalness loss through adversarial learning, a research topic that has received immense attention in recent years and is still rapidly growing [14, 37, 57].

Experiments on multiple synthetic and real datasets suggest that ShapeHD performs well on single-view 3D shape completion and reconstruction, achieving better results than state-of-the-art systems. Further analyses reveal that the network learns to attend to meaningful object parts, and the naturalness module indeed helps to characterize shape details over time.

2 Related Work

3D shape completion. Shape completion is an essential task in geometry processing and has wide applications. Traditional methods have attempted to complete shapes with local surface primitives, or to formulate it as an optimization problem [35, 44], *e.g.*, Poisson surface reconstruction solves an indicator function on a voxel grid via the Poisson equation [29, 28]. Recently, there have also been

** Our reconstructions: B, A



(a) Observation

(b) Unnatural Shapes

(c) Natural Shapes

Fig. 2. Two levels of ambiguity in single-view 3D shape perception. For each 2D observation (a), there exist many possible 3D shapes that explain this observation equally well (b, c), but only a small fraction of them correspond to real, daily shapes (c). Methods that exploit deep networks for recognition reduce, to a certain extent, ambiguity on this level. By using an adversarially learned naturalness model, our ShapeHD aims to model ambiguity on the next level: even among the realistic shapes, there are still multiple shapes explaining the observation well (c).

a growing number of papers on exploiting shape structures and regularities [34, 51], and papers on leveraging strong database priors [46, 32, 4]. These methods, however, often require the database to contain exact parts of the shape, and thus have limited generalization power.

With the advances in large-scale shape repositories like ShapeNet [5], researchers began to develop fully data-driven methods, some building upon deep convolutional networks. To name a few, Voxlets [12] employs random forests for predicting unknown voxel neighborhoods. 3D ShapeNets [58] uses a deep belief network to obtain a generative model for a given shape database, and Nguyen *et al*. [50] extend the method for mesh repairing.

Probably the most related paper to ours is the 3D-EPN from Dai *et al*. [8]. 3D-EPN achieves impressive results on 3D shape completion from partial depth scans by levering 3D convolutional networks and nonparametric patch-based shape synthesis methods. Our model has advantages over 3D-EPN in two aspects. First, with naturalness losses, ShapeHD can choose among multiple hypotheses that explain the observation, therefore reconstructing a high-quality 3D shape with fine details; in contrast, the output from 3D-EPN without nonparametric shape synthesis is often blurry. Second, our completion takes a single feed-forward pass without any post-processing, and is thus much faster (<100ms) than 3D-EPN.

Single-image 3D reconstruction. The problem of recovering the object shape from a single image is challenging, as it requires both powerful recognition systems and prior shape knowledge. As an early attempt, Huang *et al*. [21] propose to borrows shape parts from existing CAD models. With the development of large-scale shape repositories like ShapeNet [5] and methods like deep convolutional networks, researchers have built more scalable and efficient models in recent years [7, 13, 18, 27, 36, 38, 48, 52, 56, 57, 62]. While most of these approaches encode

objects in voxels from vision, there have also been attempts to reconstruct objects in point clouds [11, 15] or octave trees [40, 49, 39], or using tactile signals [53].

A related direction is to estimate 2.5D sketches (*e.g.*, depth and surface normal maps) from an RGB image. In the past, researchers have explored recovering 2.5D sketches from shading, texture, or color images [2, 3, 20, 47, 55, 63]. With the development of depth sensors [23] and larger-scale RGB-D datasets [33, 42, 43], there have also been papers on estimating depth [6, 10], surface normals [1, 54], and other intrinsic images [25, 41] with deep networks. Inspired by MarrNet [56], we reconstructs 3D shapes via modeling 2.5D sketches, but incorporating a naturalness loss for much higher quality.

Perceptual losses and adversarial learning. Researchers recently proposed to evaluate the quality of 2D images using perceptual losses [26, 9]. The idea has been applied to many image tasks like style transfer and super-resolution [26, 31]. Furthermore, the idea has been extended to learn a perceptual loss function with generative adversarial nets (GAN) [14]. GANs incorporate an adversarial discriminator into the procedure of generative modeling, and achieve impressive performance on tasks like image synthesis [37]. Isola *et al.* [22] and Zhu *et al.* [65] use GANs for image translation with and without supervision, respectively.

In 3D vision, Wu *et al.* [57] extends GANs for 3D shape synthesis. However, their model for shape reconstruction (3D-VAE-GAN) often produces a noisy, incomplete shape given an RGB image. This is because training GANs jointly with recognition networks could be highly unstable. Many other researchers have also noticed this issue: although adversarial modeling of 3D shape space may resolve the ambiguity discussed earlier, its training could be challenging [8]. Addressing this, when Gwak *et al.* [17] explored adversarial nets for single-image 3D reconstruction and chose to use GANs to model 2D projections instead of 3D shapes. This weakly supervised setting, however, hampers their reconstructions. In this paper, we develop our naturalness loss by adversarial modeling of the 3D shape space, outperforming the state-of-the-art significantly.

3 Approach

Our model consists of three components: a 2.5D sketch estimator and a 3D shape estimator that predicts a 3D shape from an RGB image via 2.5D sketches (Figure 3-I,II, inspired by MarrNet [56]), and a deep naturalness model that penalizes the shape estimator if the predicted shape is unnatural (Figure 3-III). Models trained with a supervised reconstruction loss alone often generate blurry mean shapes. Our learned naturalness model helps to avoid this issue.

2.5D sketch estimation network. Our 2.5D sketch estimator has an encoder-decoder structure that predicts the object's depth, surface normals, and silhouette from an RGB image (Figure 3-I). We use a ResNet-18 [19] to encode a 256×256 image into 512 feature maps of size 8×8. The decoder consists of four transposed convolutional layers with a kernel size of 5×5 and a stride and padding of 2. The predicted depth and surface normal images are then masked by the predicted silhouette and used as the input to our shape completion network.

3D shape completion network. Our 3D estimator (Figure 3-II) is an encoder-decoder network that predicts a 3D shape in the canonical view from 2.5D sketches.

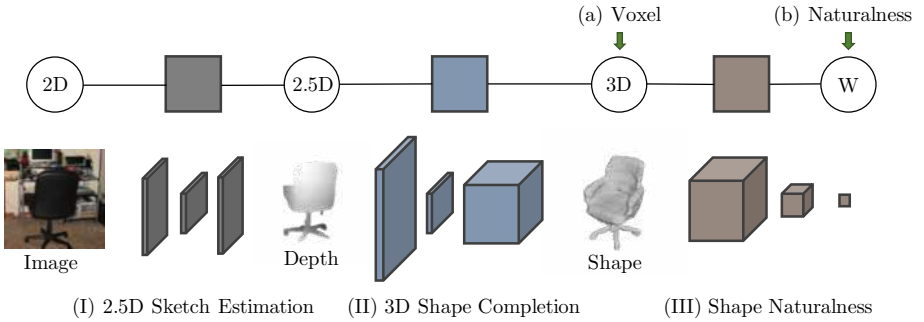


Fig. 3. For single-view shape reconstruction, ShapeHD contains three components: (I) a 2.5D sketch estimator that predicts depth, surface normal and silhouette images from a single image; (II) a 3D shape completion module that regresses 3D shapes from silhouette-masked depth and surface normal images; (III) an adversarially pretrained convolutional net that serves as the naturalness loss function. While fine-tuning the 3D shape completion net, we use two losses: a supervised loss on the output shape, and a naturalness loss offered by the pretrained discriminator.

The encoder is adapted from ResNet-18 [19] to encode a four-channel 256×256 image (one for depth, three for surface normals) into a 200-D latent vector. The vector then goes through a decoder of five transposed convolutional and ReLU layers to generate a 128×128×128 voxelized shape. Binary cross-entropy losses between predicted and target voxels are used as the supervised loss L_{voxel}.

3.1 Shape Naturalness Network

Due to the inherent uncertainty of single-view 3D shape reconstruction, shape completion networks with only a supervised loss usually predict unrealistic mean shapes. By doing so, they minimize the loss when there exist multiple possible ground truth shapes. We instead introduce an adversarially trained deep naturalness regularizer that penalizes the network for such unrealistic shapes.

We pre-train a 3D generative adversarial network [14] to determine whether a shape is realistic. Its generator synthesizes a 3D shape from a randomly sampled vector, and its discriminator distinguishes generated shapes from real ones. Therefore, the discriminator has the ability to model the real shape distribution and can be used as a naturalness loss for the shape completion network. The generator is not involved in our later training process. Following 3D-GAN [57], we use 5 transposed convolutional layers with batch normalization and ReLU for the generator, and 5 convolutional layers with leaky ReLU for the discriminator.

Due to the high dimensionality of 3D shapes (128×128×128), training a GAN becomes highly unstable. To deal with this issue, we follow Gulrajani *et al.* [16] and use the Wasserstein GAN loss with a gradient penalty to train our adversarial generative network. Specifically,

$$L_{\text{WGAN}} = \mathop{\mathbb{E}}_{\tilde{x}\sim P_g}[D(\tilde{x})] - \mathop{\mathbb{E}}_{x\sim P_r}[D(x)] + \lambda \mathop{\mathbb{E}}_{\hat{x}\sim P_x}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2], \qquad (1)$$

where D is the discriminator, P_g and P_r are distributions of generated shapes and real shapes, respectively. The last term is the gradient penalty from Gulrajani *et al.* [16]. During training, the discriminator attempts to minimize the overall loss

L_{WGAN} while the generator attempts to maximize the loss via the first term in Equation 1, so we can define our naturalness loss as $L_{\text{natural}} = -\mathop{\mathbb{E}}_{\tilde{x}\sim P_c}[D(\tilde{x})]$, where P_c are the reconstructed shapes from our completion network.

3.2 Training Paradigm

We train our network in two stages. We first pre-train the three components of our model separately. The shape completion network is then fine-tuned with both voxel loss and naturalness losses.

Our 2.5D sketch estimation network and 3D completion network are trained with images rendered with ShapeNet [5] objects (see Sections 4.1 and 5 for details). We train the 2.5D sketch estimator using a L2 loss and SGD with a learning rate of 0.001 for 120 epochs. We only use the supervised loss L_{voxel} for training the 3D estimator at this stage, again with SGD, a learning rate of 0.1, and a momentum of 0.9 for 80 epochs. The naturalness network is trained in an adversarial manner, where we use Adam [30] with a learning rate of 0.001 and a batch size of 4 for 80 epochs. We set $\lambda = 10$ as suggested in Gulrajani *et al.* [16].

We then fine-tune our completion network with both voxel loss and naturalness losses as $L = L_{\text{voxel}} + \alpha L_{\text{natural}}$. We compare the scale of gradients from the losses and train our completion network with $\alpha = 2.75 \times 10^{-11}$ using SGD for 80 epochs. Our model is robust to these parameters; they are only for ensuring gradients of various losses are of the same magnitude.

An alternative is to jointly train the naturalness module with the completion network from scratch using both losses. It seems tempting, but in practice we find that Wasserstein GANs have large losses and gradients, resulting in unstable outputs. We therefore choose to use our pre-training and fine-tuning setup.

4 Single-View Shape Completion

For 3D shape completion from a single depth image, we only use the last two modules of the model: the 3D shape estimator and deep naturalness network.

4.1 Setup

Data. We render each of the ShapeNet Core55 [5] objects from the aeroplane, car and chair categories in 20 random, fully unconstrained views. For each view, we randomly set the azimuth and elevation angles of the camera, but the camera up vector is fixed to be the world $+y$ axis, and the camera always looks at the object center. The focal length is fixed at 50mm with a 35mm film. We use Mitsuba [24], a physically-based graphics engine, for all our renderings. We used 90% of the data for training and 10% for testing.

We render the ground-truth depth image of each object in all 20 views. Depth values are measured from the camera center (*i.e.*, ray depth), rather than from the image plane. To approximate depth scanner data, we also generate the accompanying ground-truth surface normal images from the raw depth data, as surface normal maps are the common by-products of depth scanning. All our rendered surface normal vectors are defined in the camera space.

Baselines. We compare with the state of the art: 3D-EPN [8]. To ensure a fair comparison, we convert depth maps to partial surfaces registered in a

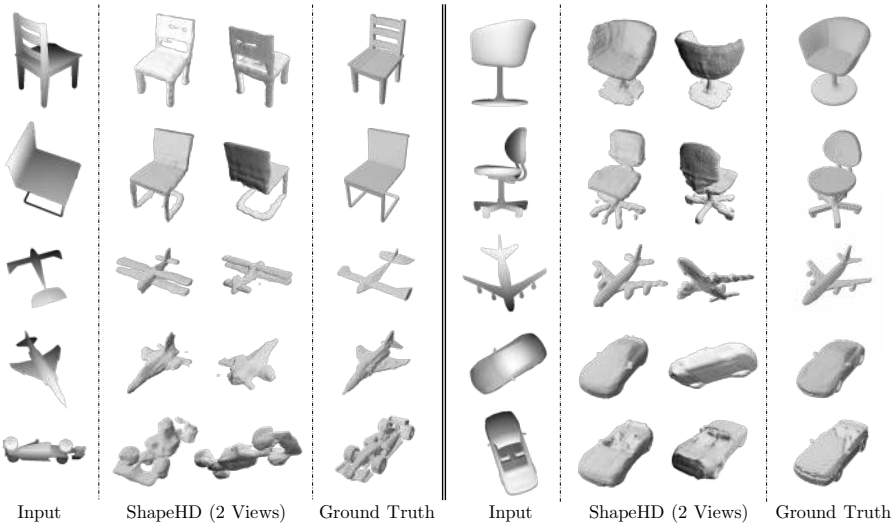


Fig. 4. Results on 3D shape completion from single-view depth. From left to right: input depth maps, shapes reconstructed by ShapeHD in the canonical view and a novel view, and ground truth shapes in the canonical view. Assisted by the adversarially learned naturalness losses, ShapeHD recovers highly accurate 3D shapes with fine details. Sometimes the reconstructed shape deviates from the ground truth, but can be viewed as another plausible explanation of the input (*e.g.*, the airplane on the left, third row).

canonical global coordinate defined by ShapeNet Core55 [5], which is required by 3D-EPN. While the original 3D-EPN paper generates their partial observations by rendering and fusing multi-view depth maps, our method takes a single-view depth map as input and is solving a more challenging problem.

Metrics. We use two standard metrics for quantitative comparisons: Intersection over Union (IoU) and Chamfer Distance (CD). In particular, Chamfer distance can be applied to various shape representations including voxels (by sampling points on the isosurface) and point clouds.

4.2 Results on ShapeNet

Qualitative results. In Figure 4, we show 3D shapes predicted by ShapeHD from single-view depth images. While common encoder-decoder structure usually generates mean shapes with few details, our ShapeHD predicts shapes with large variance and fine details. In addition, even when there is strong occlusion in the depth image, our model can predict a high-quality, plausible 3D shape that looks good perceptually, and infer parts not present in the input images.

Ablation. When using naturalness loss, the network is penalized for generating mean shapes that are unreasonable but minimize the supervised loss. In Figure 5, we show reconstructed shapes from our ShapeHD with and without naturalness loss (*i.e.* before fine-tuning with L_{natural}), together with ground truth shapes and shapes predicted by 3D-EPN [8]. Our results contain finer details compared with those from 3D-EPN. Also, the performance of ShapeHD improves greatly with the naturalness loss, which predicts more reasonable and complete shapes.

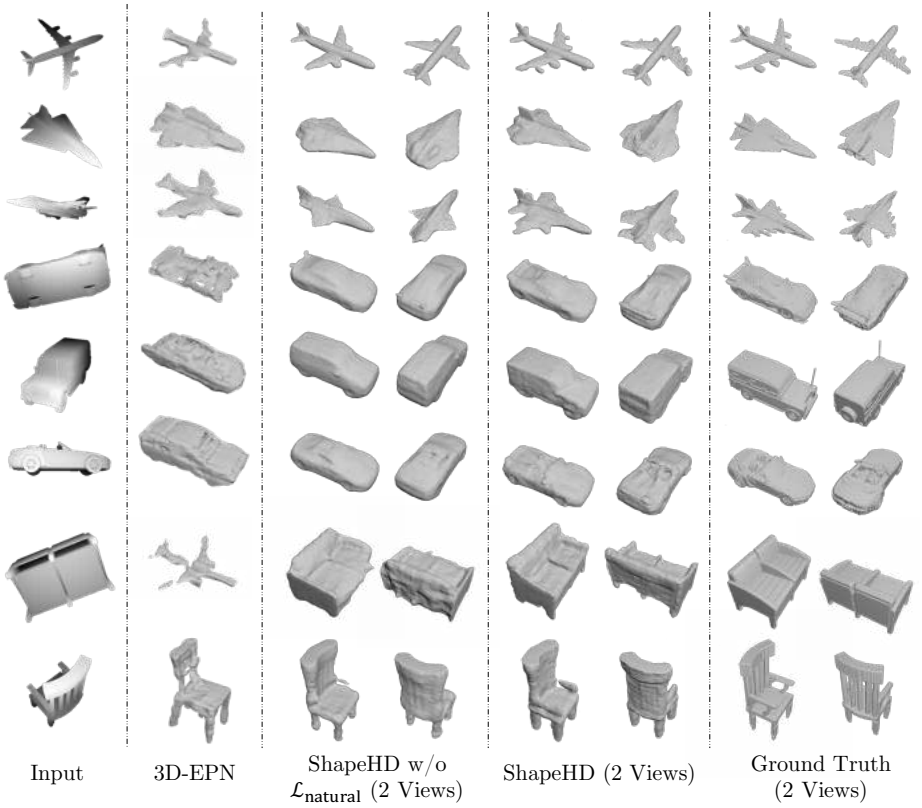


Fig. 5. Our results on 3D shape completion, compared with the state of the art, 3D-EPN [8], and our model but without naturalness losses. Our results contain more details than 3D-EPN. We observe that the adversarially trained naturalness losses help fix errors, add details (*e.g.*, the plane wings in row 3, car seats in row 6, and chair arms in row 8), and smooth planar surfaces (*e.g.*, the sofa back in row 7).

Methods	IoU				CD			
	chair	car	plane	avg	chair	car	plane	avg
3D-EPN [8]	.147	.274	.155	.181	.227	.200	.125	.192
ShapeHD w/o L_{natural}	.466	**.698**	**.488**	**.529**	.112	.083	.071	.093
ShapeHD	**.488**	**.698**	.452	**.529**	**.096**	**.078**	**.068**	**.084**

Table 1. Average IoU scores (32^3) and CDs for 3D shape completion on ShapeNet [5]. Our model outperforms the state of the art by a large margin. The learned naturalness losses consistently improve the CDs between our results and ground truth.

Quantitative results. We present quantitative results in Table 1. Our ShapeHD outperforms the state of the art by a margin in all metrics. Our method outputs shapes at the resolution of 128^3, while shapes produced by 3D-EPN are of resolution 32^3. Therefore, for a fair comparison, we downsample our predicted shapes to 32^3 and report results of both methods in that resolution. The original 3D-EPN paper suggests a post-processing step that retrieves similar patches

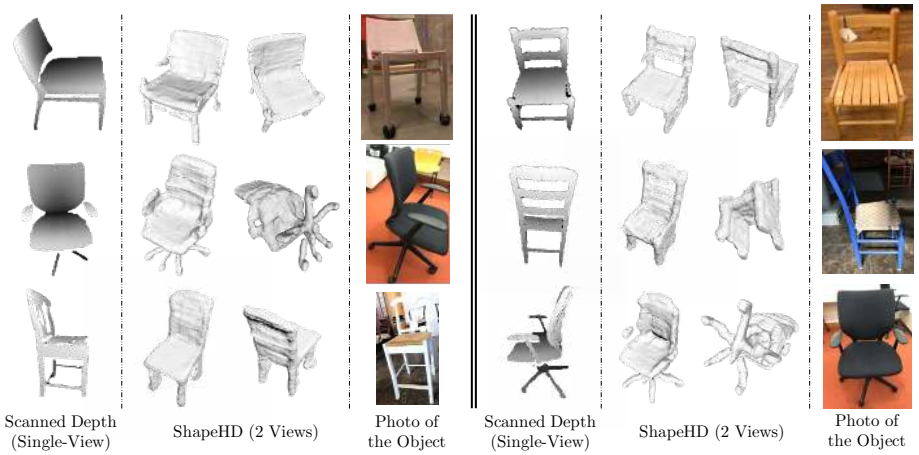


Fig. 6. Results of 3D shape completion on depth data from a physical scanner. Our model is able to reconstruct the shape well from just a single view. From left to right: input depth, two views of our result, and a color image of the object.

from a shape database for results of a higher resolution. Practically, we find this steps takes 18 hours for a single image. We therefore report results without post-processing for both methods.

Table 1 also suggests the naturalness loss improve the completion results, achieving comparable IoU scores and better (lower) CDs. CD has been reported to be better at capturing human perception of shape quality [45].

4.3 Results on Real Depth Scans

We now show results of ShapeHD on real depth scans. We capture six depth maps of different chairs using a Structure sensor (`http://structure.io`) and use the captured depth maps to evaluate our model. All the corresponding normal maps used as inputs are estimated from depth measurements. Figure 6 shows that ShapeHD completes 3D shapes well given a single-view depth map. Our ShapeHD is more flexible than 3D-EPN, as we do not need any camera intrinsics or extrinsics to register depth maps. In our case, none of these parameters are known and thus 3D-EPN cannot be applied.

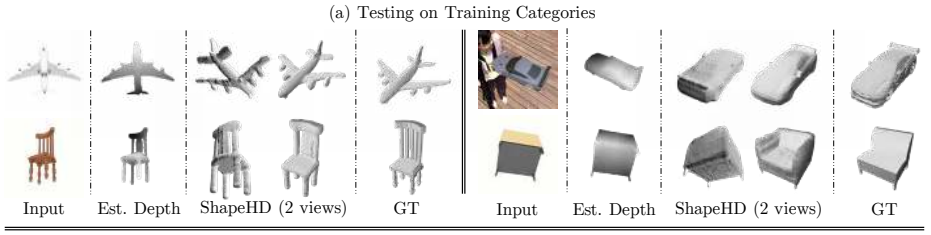
5 3D Shape Reconstruction

We now evaluate ShapeHD on 3D shape reconstruction from a single color image.

RGB image preparation. For the task of single-image 3D reconstruction, we need to render RGB images that correspond to the depth images for training. We follow the same camera setup specified earlier. Additionally, to boost the realism of the rendered RGB images, we put three different types of backgrounds behind the object during rendering. One third of the images are rendered in a clean white background; one third are rendered in high-dynamic-range backgrounds with illumination channels that produce realistic lighting. We render the remaining one third images with backgrounds randomly sampled from the SUN database [61].

Methods	bench	boat	cabin	car	chair	disp	lamp	phone	plane	rifle	sofa	speak	table	avg
DRC (3D) [52]	.122	.131	.127	.077	.128	.128	.168	.102	.166	.107	.106	**.138**	.138	.126
AtlasNet [15]*	.123	.130	.169	.107	.141	.162	.171	.138	.105	.096	.131	.172	.161	.139
ShapeHD (ours)	**.121**	**.103**	**.126**	**.066**	**.125**	**.124**	**.157**	**.084**	**.073**	**.053**	**.102**	.141	**.124**	**.108**

Fig. 7. Qualitative results and CDs for 3D shape reconstruction on ShapeNet [5]. Our rendering of ShapeNet is more challenging than that from Choy *et al.* [7]; as such, the numbers of the other methods may differ from those in the original paper. All methods are trained with full 3D supervision on our rendering of the largest 13 ShapeNet categories. *DRC and ShapeHD take a single image as input, while AltasNet requires ground truth object silhouettes as additional input.



Methods	bench	boat	cabin	disp	lamp	phone	rifle	sofa	speak	table	avg
DRC (3D) [52]	.175	.161	.189	.278	**.225**	.268	.153	.149	.203	.221	.202
AtlasNet [15]*	**.155**	**.114**	.202	**.244**	.261	.263	**.121**	**.126**	.206	.262	**.195**
ShapeHD (ours)	.166	.129	**.182**	.252	.235	**.229**	.232	.133	**.193**	**.199**	**.195**

Fig. 8. Qualitative results and CDs for 3D shape reconstruction on novel categories from ShapeNet [5]. All methods are trained with full 3D supervision on our rendering of ShapeNet cars, chairs, and planes, and tested on the next 10 largest categories. *DRC and ShapeHD take a single image as input, while AltasNet requires ground truth object silhouettes as additional input.

Baselines. We compare our ShapeHD with the state-of-the-art in 3D shape reconstruction, including 3D-R2N2 [7], point set generation network (PSGN) [11], differentiable ray consistency (DRC) [52], octree generating network (OGN) [49], and AtlasNet [15]. 3D-R2N2, DRC, OGN, and our ShapeHD take a single image as input, while PSGN and AltasNet require object silhouettes as additional input.

Results on synthetic data. We first evaluate on renderings of ShapeNet objects [5]. We present reconstructed 3D shapes and quantitative results in Figures 7. All these models are trained on our rendering of the largest 13 ShapeNet categories (those have at least 1,000 models) with ground truth 3D shapes as

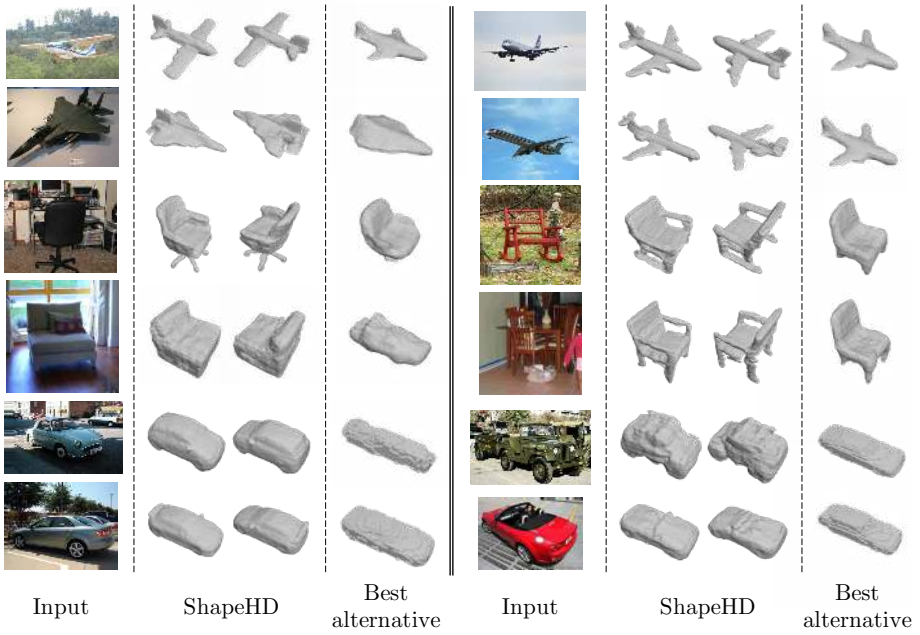
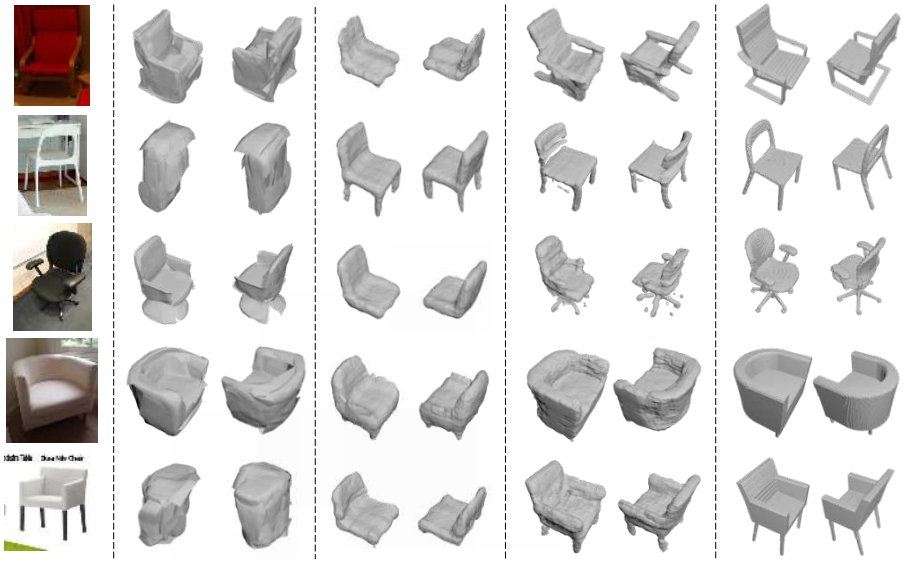


Fig. 9. Single-view 3D shape reconstruction on PASCAL 3D+ [60]. From left to right: input, two views of reconstructions from ShapeHD, and reconstructions by the best alternative in Table 2. Assisted by the learned naturalness losses, ShapeHD recovers accurate 3D shapes with fine details.

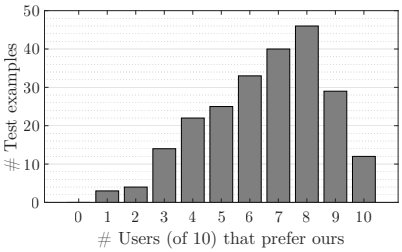


(a) Input (b) AtlasNet (c) DRC (3D) (d) ShapeHD (e) GT

Fig. 10. Single-view 3D reconstruction on Pix3D [45]. For each input image, we show reconstructions by AtlasNet, DRC, our ShapeHD, and ground truth. Our ShapeHD reconstructs complete 3D shapes with fine details that resemble the ground truth.

Methods	CD			
	chair	car	plane	avg
3D-R2N2 [7]	0.238	0.305	0.305	0.284
DRC (3D) [52]	0.158	0.099	0.112	0.122
OGN [49]	-	**0.087**	-	-
ShapeHD (ours)	**0.137**	0.129	**0.094**	**0.119**

(a) CDs on PASCAL 3D+ [60]



(b) Human Study results

Table 2. Results for 3D shape reconstruction on PASCAL 3D+ [60]. (a) We compare our ShapeHD with 3D-R2N2, DRC, and OGN. PSGN and AtlasNet are not evaluated, because they require object masks as additional input, but PASCAL 3D+ has only inaccurate masks. (b) In the behavioral study, most users prefer our constructions on most images. Overall, our reconstructions are preferred 64.5% of the time to OGN's.

supervision. In general, our ShapeHD is able to predict 3D shapes that closely resemble the ground truth shapes, giving fine details that make the reconstructed shapes more realistic. It also performs better quantitatively.

Generalization on novel categories. An important aspect of evaluating shape reconstruction methods is on how well they generalize. Here we train our model and baselines on the largest three ShapeNet classes (cars, chairs, and planes), again with ground truth shapes as supervision, and test them on the next largest ten. Figure 8 shows our ShapeHD performs better than DRC (3D) and is comparable to AtlasNet; however, note that AtlasNet requires ground truth silhouettes as additional input, while ShapeHD works on raw images.

Results on real data. We then evaluate on two real datasets, PASCAL 3D+ [60] and Pix3D [45]. Here, we train our model on synthetic ShapeNet renderings and use the pre-trained models released by the authors as baselines. All methods take ground truth 3D shapes as supervision during training. As shown in Figures 9 and 10, ShapeHD works well, inferring a reasonable shape even in the presence of strong self-occlusions. In particular, in Figure 9, we compare our reconstructions with the best-performing alternatives (DRC on chairs and airplanes, and AtlasNet on cars). In addition to preserving details, our model captures the shape variations of the objects, while the competitors produce similar reconstructions across instances.

Quantitatively, Tables 2 and 3 suggest that ShapeHD performs significantly better than the other methods in almost all metrics. The only exception is the CD on PASCAL 3D+ cars, where OGN performs the best. However, as PASCAL 3D+ only has around 10 CAD models for each object category as ground truth 3D shapes, the ground truth labels and the scores can be inaccurate, failing to reflect human perception [52].

We therefore conduct an additional user study, where we show an input image and its two reconstructions (from ShapeHD and from OGN, each in two views) to users on Amazon Mechanical Turk, and ask them to choose the shape that looks closer to the object in the image. For each image, we collect 10 responses from "Masters" (workers who have demonstrated excellence across a wide range of HITs).

	3D-R2N2 [7]	DRC (3D) [52]	PSGN [11]*	AtlasNet [15]*	ShapeHD
IoU (32^3)	0.136	0.265	-	-	**0.284**
IoU (128^3)	0.089	0.185	-	-	**0.205**
CD	0.239	0.160	0.199	0.126	**0.123**

Table 3. 3D shape reconstruction results on Pix3D [45]. All methods were trained with full 3D supervision on rendered images of ShapeNet objects. *3D-R2N2, DRC, and ShapeHD take a single image as input, while PSGN and AtlasNet require the ground truth mask as input. Also, PSGN and AtlasNet generate surface point clouds without guaranteeing watertight meshes and therefore cannot be evaluated in IoU.

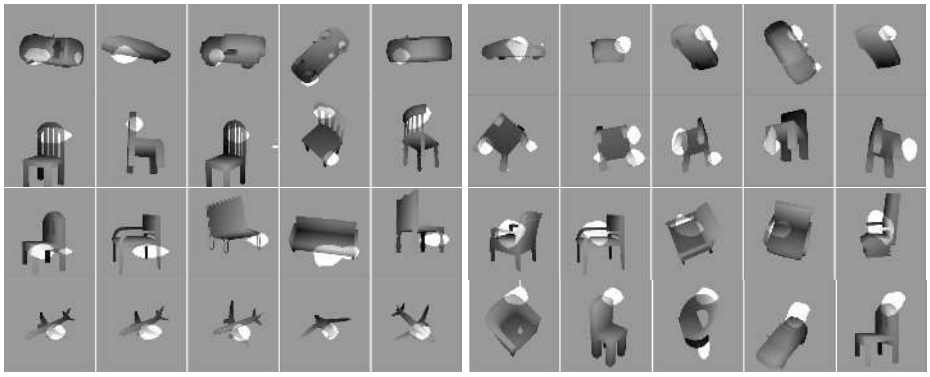


Fig. 11. Visualizations on how ShapeHD attends to details in depth maps. Row 1: car wheel detectors. Row 2: chair back and leg detectors. The left responds to the strided pattern in particular. Row 3: chair arm and leg detectors. Row 4: airplane engine and curved surface detectors. The right responds to a specific pattern across classes.

Table 2b suggests that on most images, most users prefer our reconstruction to OGN's. In general, our reconstructions are preferred 64.5% of the time.

6 Analyses

We want to understand what the network has learned. In this section, we present a few analyses to visualize what the network is learning, analyze the effect of the naturalness loss function over time, and discuss common failure modes.

Network visualization. As the network successfully reconstructs object shape and parts, it is natural to ask if it learns object or part detectors implicitly. To this end, we visualize the top activating regions across all validation images for units in the last convolutional layer of the encoder in our 3D completion network, using the method proposed by Zhou *et al.* [64]. As shown in Figure 11, the network indeed learns a diverse and rich set of object and part detectors. There are detectors that attend to car wheels, chair backs, chair arms, chair legs, and airplane engines. Also note that many detectors respond to certain patterns (*e.g.*, strided) in particular, which is probably contributing to the fine details in the reconstruction. Additionally, there are units that respond to generic shape patterns across categories, like the curve detector in the bottom right.

Training with naturalness loss over time. We study the effect of the naturalness loss over time. In Figure 12, we plot the loss of the completion

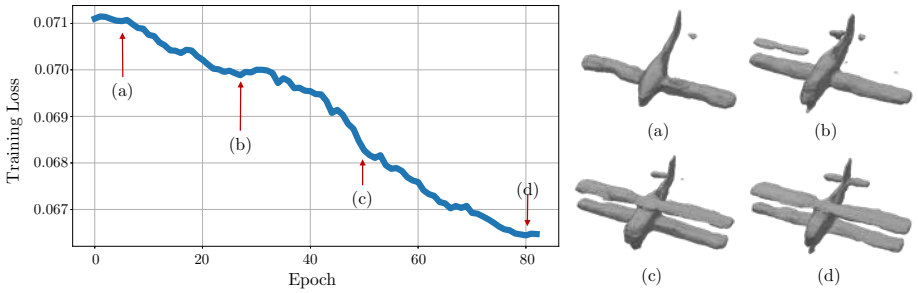


Fig. 12. Visualizations on how ShapeHD evolves over time with naturalness losses: the predicted shape becomes increasingly realistic as details are being added.

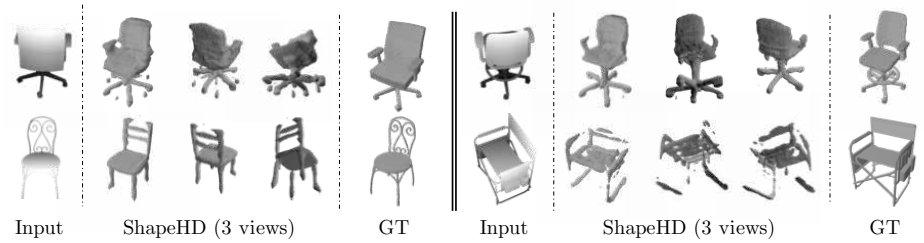


Fig. 13. Common failure modes of our system. Top left: the model sometimes gets confused by deformable object parts (*e.g.*, wheels). Top right: the model might miss uncommon object parts (the ring above the wheels). Bottom row: the model has difficulty in recovering very thin structure, and may generate other structure patterns instead.

network with respect to fine-tuning epochs. We realize the voxel loss goes down slowly but consistently. If we visualize the reconstructed examples at different timestamps, we clearly see details are being added to the shapes. These fine details occupy a small region in the voxel grid, and thus training with supervised loss alone is unlikely to recover them. In contrast, with adversarially training perceptual losses, our model recovers details successfully.

Failure cases. We present failure cases in Figure 13. We observe our model has these common failing modes: it sometimes gets confused by deformable object parts (*e.g.*, wheels on the top left); it may miss uncommon object parts (top right, the ring above the wheels); it has difficulty in recovering very thin structure (bottom right), and may generate other patterns instead (bottom left). While the voxel representation makes it possible to incorporate the naturalness loss, intuitively, it also encourages the network to focus on thicker shape parts, as they carry more weights in the loss function.

7 Conclusion

We have proposed to use learned shape priors to overcome the 2D-3D ambiguity and to learn from the multiple hypotheses that explain a single-view observation. Our ShapeHD achieves state-of-the-art results on 3D shape completion and reconstruction. We hope our results will inspire further research in 3D shape modeling, in particular on explaining the ambiguity behind partial observations.

Acknowledgements: This work is supported by NSF #1231216, ONR MURI N00014-16-1-2007, Toyota Research Institute, Shell Research, and Facebook.

References

1. Bansal, A., Russell, B.: Marr revisited: 2d-3d alignment via surface normal prediction. In: CVPR (2016)
2. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE TPAMI **37**(8), 1670–1687 (2015)
3. Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. ACM TOG **33**(4), 159 (2014)
4. Brock, A., Lim, T., Ritchie, J.M., Weston, N.: Generative and discriminative voxel modeling with convolutional neural networks. In: NIPS Workshop (2016)
5. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv:1512.03012 (2015)
6. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: NIPS (2016)
7. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: ECCV (2016)
8. Dai, A., Qi, C.R., Nießner, M.: Shape completion using 3d-encoder-predictor cnns and shape synthesis. In: CVPR (2017)
9. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: NIPS (2016)
10. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV (2015)
11. Fan, H., Su, H., Guibas, L.: A point set generation network for 3d object reconstruction from a single image. In: CVPR (2017)
12. Firman, M., Aodha, O.M., Julier, S., Brostow, G.J.: Structured Completion of Unobserved Voxels from a Single Depth Image. In: CVPR (2016)
13. Girdhar, R., Fouhey, D.F., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: ECCV (2016)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
15. Groueix, T., Fisher, M., Kim, V.G., Russel, B.C., Aubry, M.: Atlasnet: A papier-mch approach to learning 3d surface generation. In: CVPR (2018)
16. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans. In: NIPS (2017)
17. Gwak, J., Choy, C.B., Chandraker, M., Garg, A., Savarese, S.: Weakly supervised 3d reconstruction with adversarial constraint. In: 3DV (2017)
18. Häne, C., Tulsiani, S., Malik, J.: Hierarchical surface prediction for 3d object reconstruction. In: 3DV (2017)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2015)
20. Horn, B.K., Brooks, M.J.: Shape from shading. MIT press (1989)
21. Huang, Q., Wang, H., Koltun, V.: Single-view reconstruction via joint analysis of image and shape collections. ACM TOG **34**(4), 87 (2015)
22. Isola, P., Zoran, D., Krishnan, D., Adelson, E.H.: Learning visual groups from co-occurrences in space and time. In: ICLR Workshop (2016)
23. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R.A., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A.J., Fitzgibbon, A.W.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: UIST (2011)
24. Jakob, W.: Mitsuba renderer (2010), http://www.mitsuba-renderer.org

25. Janner, M., Wu, J., Kulkarni, T., Yildirim, I., Tenenbaum, J.B.: Self-Supervised Intrinsic Image Decomposition. In: NIPS (2017)
26. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
27. Kar, A., Tulsiani, S., Carreira, J., Malik, J.: Category-specific object reconstruction from a single image. In: CVPR (2015)
28. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: SGP. SGP '06 (2006)
29. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. ACM TOG **32**(3), 29 (2013)
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
31. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv:1609.04802 (2016)
32. Li, Y., Dai, A., Guibas, L., Nießner, M.: Database-assisted object retrieval for real-time 3d reconstruction. CGF **34**(2), 435–446 (2015)
33. McCormac, J., Handa, A., Leutenegger, S., Davison, A.J.: Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In: ICCV (2017)
34. Mitra, N.J., Guibas, L.J., Pauly, M.: Partial and approximate symmetry detection for 3d geometry. ACM TOG **25**(3), 560–568 (2006)
35. Nealen, A., Igarashi, T., Sorkine, O., Alexa, M.: Laplacian mesh optimization. In: Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia. pp. 381–389. ACM (2006)
36. Novotny, D., Larlus, D., Vedaldi, A.: Learning 3d object categories by looking around them. In: ICCV (2017)
37. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR (2016)
38. Rezende, D.J., Eslami, S., Mohamed, S., Battaglia, P., Jaderberg, M., Heess, N.: Unsupervised learning of 3d structure from images. In: NIPS (2016)
39. Riegler, G., Ulusoy, A.O., Bischof, H., Geiger, A.: Octnetfusion: Learning depth fusion from data. In: 3DV (2017)
40. Riegler, G., Ulusoys, A.O., Geiger, A.: Octnet: Learning deep 3d representations at high resolutions. In: CVPR (2017)
41. Shi, J., Dong, Y., Su, H., Yu, S.X.: Learning non-lambertian object intrinsics across shapenet categories. In: CVPR (2017)
42. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
43. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: CVPR (2017)
44. Sorkine, O., Cohen-Or, D.: Least-squares meshes. In: Shape Modeling Applications (2004)
45. Sun, X., Wu, J., Zhang, X., Zhang, Z., Xue, T., Tenenbaum, J.B., Freeman, W.T.: Pix3d: Dataset and methods for single-image 3d shape modeling. In: CVPR (2018)
46. Sung, M., Kim, V.G., Angst, R., Guibas, L.: Data-driven structural priors for shape completion. ACM TOG **34**(6), 175 (2015)
47. Tappen, M.F., Freeman, W.T., Adelson, E.H.: Recovering intrinsic images from a single image. In: NIPS (2003)
48. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Multi-view 3d models from single images with a convolutional network. In: ECCV (2016)

49. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In: ICCV (2017)
50. Thanh Nguyen, D., Hua, B.S., Tran, K., Pham, Q.H., Yeung, S.K.: A field model for repairing 3d shapes. In: CVPR (2016)
51. Thrun, S., Wegbreit, B.: Shape from symmetry. In: ICCV (2005)
52. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: CVPR (2017)
53. Wang, S., Wu, J., Sun, X., Yuan, W., Freeman, W.T., Tenenbaum, J.B., Adelson, E.H.: 3d shape perception from monocular vision, touch, and shape priors. In: IROS (2018)
54. Wang, X., Fouhey, D., Gupta, A.: Designing deep networks for surface normal estimation. In: CVPR (2015)
55. Weiss, Y.: Deriving intrinsic images from image sequences. In: ICCV (2001)
56. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, W.T., Tenenbaum, J.B.: MarrNet: 3D Shape Reconstruction via 2.5D Sketches. In: NIPS (2017)
57. Wu, J., Zhang, C., Xue, T., Freeman, W.T., Tenenbaum, J.B.: Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. In: NIPS (2016)
58. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: CVPR (2015)
59. Xiang, Y., Kim, W., Chen, W., Ji, J., Choy, C., Su, H., Mottaghi, R., Guibas, L., Savarese, S.: Objectnet3d: A large scale database for 3d object recognition. In: ECCV (2016)
60. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3d object detection in the wild. In: WACV (2014)
61. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR (2010)
62. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In: NIPS (2016)
63. Zhang, R., Tsai, P.S., Cryer, J.E., Shah, M.: Shape-from-shading: a survey. IEEE TPAMI **21**(8), 690–706 (1999)
64. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene cnns. In: ICLR (2014)
65. Zhu, J.Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: ECCV (2016)