Silhouette Guided Point Cloud Reconstruction beyond Occlusion

Chuhang Zou Derek Hoiem
University of Illinois at Urbana-Champaign
{czou4, dhoiem}@illinois.edu

Abstract

One major challenge in 3D reconstruction is to infer the complete shape geometry from partial foreground occlusions. In this paper, we propose a method to reconstruct the complete 3D shape of an object from a single RGB image, with robustness to occlusion. Given the image and a silhouette of the visible region, our approach completes the silhouette of the occluded region and then generates a point cloud. We show improvements for reconstruction of non-occluded and partially occluded objects by providing the predicted complete silhouette as guidance. We also improve state-of-the-art for 3D shape prediction with a 2D reprojection loss from multiple synthetic views and a surface-based smoothing and refinement step. Experiments demonstrate the efficacy of our approach both quantitatively and qualitatively on synthetic and real scene datasets.

1. Introduction

3D reconstruction from 2D images has many applications in robotics and augmented reality. One major challenge is to infer the complete shape of a partially occluded object. Occlusion frequently occurs in natural scenes: *e.g.* we often see an image of a sofa occluded by a table in front and a dining table partially occluded by a vase on top. Even multi-view approaches [34, 12, 19] may fail to recover complete shape, since occlusions may block most views of the object. Single-view learning-based methods [13, 6, 44] have approached seeing beyond occlusion as a 2D semantic segmentation completion task, but complete 3D shape recovery adds the challenges of predicting 3D shape from a 2D image and being robust to the unknown existence and extent of an occluding region.

In this paper, our goal is to reconstruct a complete 3D shape from a single RGB image, in a way that is robust to occlusions. We follow a data-driven approach, using convolution neural networks (CNNs) to encode shape-relevant features and decode them into an object point cloud. To simplify the shape prediction, we split the task into: (1) determining the visible region of the object; (2) predicting a completed silhouette (filling in any occluded regions); and (3) predicting the object 3D point cloud based on the silhouette and RGB image (Fig. 1). We reconstruct the object in a viewer-centered manner, inferring both object shape and pose. We show that, provided with ground truth silhouettes, shape prediction achieves nearly the same performance for occluded objects as non-occluded objects. We obtain the visible portion of the silhouette using Mask-RCNN [16] and then predict the completed silhouette using an auto-encoder. Using the predicted silhouette as part of shape prediction also yields large improvements for both occluded and non-occluded objects, indicating that providing an explicit foreground/background separation for the object in RGB images is helpful [1].

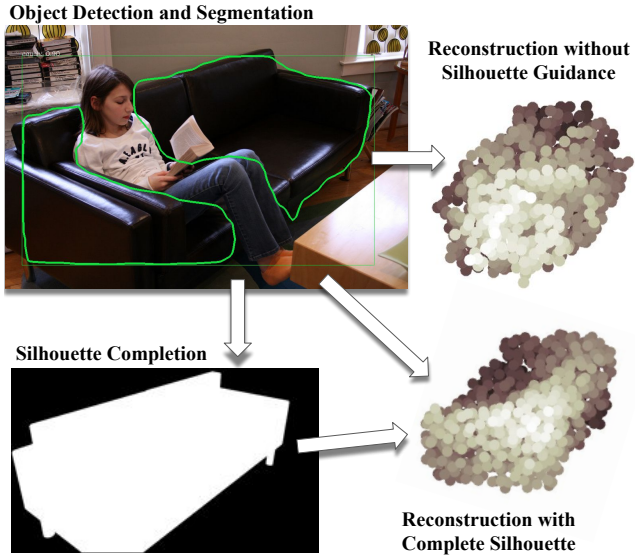

Figure 1. **Illustration.** The person sitting on the sofa blocks much of the sofa from view, causing errors in existing shape prediction methods. We propose improvements to shape prediction, including the prediction and completion of the object silhouette as an intermediate step, and demonstrate more accurate reconstruction of both occluded and non-occluded objects. Best viewed in color.

[1] Code and data is available at https://github.com/zouchuhang/Silhouette-Guided-3D

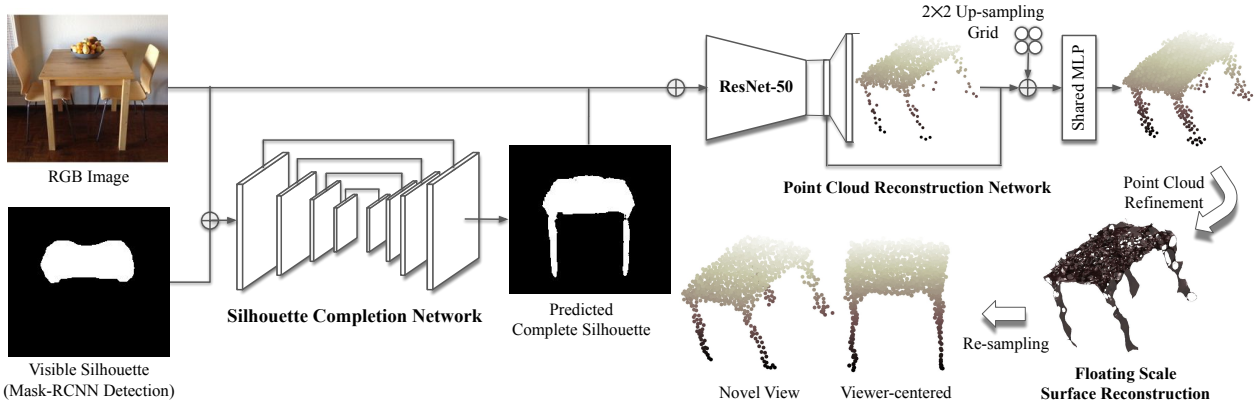

Figure 2. **Approach**. Our approach is composed of three steps. In the first step, the silhouette completion network takes an RGB image and the visible silhouette as input, and predict the complete silhouette of the object beyond foreground occlusions. In the second step, given the RGB image and the predicted complete silhouette, the reconstruction network predicts point clouds in viewer-centered coordinates. Finally, we perform a post refinement step to produce smooth and uniformly distributed point clouds. Best viewed in color.

Our reconstruction represents a 3D shape as a set of point clouds, which is flexible and easy to transform. Our method follows an encoder-decoder strategy, and we demonstrate performance gains using a 2D reprojection loss from multiple synthetic views and a surface-based post refinement step, achieving state-of-the-art. Our silhouette guidance approach is related to shape from silhouette [21, 3, 24], but our silhouette guidance is part of learning approach rather than explicit constraint.

Our contributions:

- We improve the state-of-the-art for 3D point clouds reconstruction from a single RGB image. We show performance gains by using a 2D reprojection loss on multiple synthetic views and a surface-based refinement step.
- We demonstrate that completing the visible silhouette leads to better object shape completion. We propose a silhouette completion network that achieves the state-of-the-art. We show improvements for reconstruction of non-occluded and partially occluded objects.

2. Related Work

Single image 3D shape reconstruction is an active topic of research. Approaches use RGB images [36, 39, 37, 11, 31], depth images [42, 46, 40, 30] or both [14, 9, 15]. Approaches include exemplar based shape retrieval and alignment [2, 1, 14, 17], deformations from meshes [20, 25, 36], or a direct prediction via convolution neural networks [39, 41, 38]. Qi *et al*. [28] propose a novel deep net architecture suitable for consuming unordered point sets in 3D; Fan *et al*. [7] propose to generate point clouds from a single RGB image using generative models. More recent approaches improve point set reconstruction performance by learning representative latent features [27] or by imposing constraints of geometric appearance in multiple views [18].

Most of these approaches are applied to non-occluded objects with clean backgrounds and no occlusions, which may prevent their application to natural images. Sun *et al*. [32] conduct experiments on real images from Pix3D, a large-scale dataset with aligned ground-truth 3D shapes, but do not consider the problem of occlusion. We are concerned with predicting shape of objects in natural scenes, which may be partly occluded. Our approach improves the state-of-the-art for object point set generation, and is extended to reconstruct beyond occlusion with the guidance of completed silhouettes. Our silhouettes guidance is closely related to the human depth estimation by Rematas *et al*. [29]. However, Rematas *et al*. use the visible silhouette (semantic segmentation) rather than a complete silhouette, making it hard to predict overlapped (occluded) regions. Differently, our approach conditions on predicted silhouette to resolve occlusion ambiguity, and is able to predict complete 3D shape rather than 2.5D depth points.

Seeing beyond occlusion. Occlusions have long been an obstacle in multi-view reconstruction. Solutions have been proposed to recover portions of surfaces from single views, *e.g*. with synthetic apertures [35, 8], or to otherwise improve robustness of matching and completion functions from multiple views [34, 12, 19]. Other work decompose a scene into layered depth maps from RGBD [26] images or video [45] and then seek to complete the occluded portions of the maps. But errors in layered segmentation can severely degrade the recovery of the occluded region. Learning-based approaches [13, 6, 44] have posed recovery from occlusion as a 2D semantic segmentation completion task. Ehsani *et al*. [6] propose to complete the silhouette and texture of an occluded object. Our silhouette completion network is most similar to Ehsani *et al*., but we ease

the task by predicting the complete silhouette rather than the full texture. We demonstrate better performance with our up-sampling based convolution decoder instead of fully connected layers used in Ehsani *et al*. Moreover, We go further to try to predict the complete 3D shape of the occluded object.

3. Point Clouds from a Single RGB Image

Direct point set prediction from a single image is challenging due to the unknown camera viewpoint or object pose and large intraclass variations in shape. This requires a careful design choice on the network architecture. We aim to have an encoder that can capture object pose and shape from a single image, and a decoder that is flexible in producing unordered, dense point clouds.

In this section, we introduce our point prediction network architecture, the training scheme including a 2D reprojection loss on multiple synthetic views to improve performance (Sec. 3.1). We then introduce a post-refinement step via surface-based fitting (Sec. 3.2) to produce smooth and uniformly distributed point sets.

3.1. Point Cloud Reconstruction Network

Our network architecture is illustrated in Fig. 2. The network predicts 3D point clouds in viewer-centered coordinates. The encoder is based on ResNet-50 [16] to better capture object shape and pose feature. The decoder follows a coarse-to-fine multi-stage generation scheme in order to efficiently predict dense points with limited memory. Our decoder follows the design of PCN [43]. The coarse predictor predicts $N = 1024$ sparse points. The refinement branch produces $4N$ finer points, by learning a 2×2 up-sampling surface grid centered on each coarse point via local folding operation. We experimented with a higher up-sampling rate (*e.g*. 9, 16) as PCN but observed repetitive patterns across all surface patches, missing local shape details. Note that our network is able to generate a denser prediction with another up-sampling branch on top, but the current structure best balances accuracy and training/inference speed. Our reconstruction network does not require features from partial points like PCN, and produces an on-par performance with PSG [7], a state-of-the-art method in point set generation (see experiments in Sec. 5.3), even without the refinement step we will introduce in Sec. 3.2.

Loss function. We consider the training loss in 3D space using the bidirectional Chamfer distance. Given predicted point clouds $\hat{p} \in \hat{P}$ and the ground truth $p \in P$, we have:

$$\begin{aligned} L_{rec} &= d_{Chamfer}(P, \hat{P}) \\ &= \frac{1}{|P|} \sum_{p \in P} \min_{\hat{p} \in \hat{P}} \|p - \hat{p}\|_2^2 + \frac{1}{|\hat{P}|} \sum_{\hat{p} \in \hat{P}} \min_{p \in P} \|p - \hat{p}\|_2^2 \end{aligned} \tag{1}$$

To further boost the performance, we propose a 2D reprojection loss on point sets as follows:

$$\begin{aligned} L_{proj} &= d(Proj(P), Proj(\hat{P})) \\ &= \frac{1}{|P|} \sum_{p \in P} \min_{\hat{p} \in \hat{P}} \|K[R\ t]p - K[R\ t]\hat{p}\|_2^2 \end{aligned} \tag{2}$$

Where $Proj(\cdot)$ is a 2D projection operation from 3D space, with 3D rotation R and translation t in world coordinates and a known camera intrinsic K. Since our reconstruction is viewer-centered, we can simply set $R = I, t = 0$ assuming projections on the image plane. Our 2D reprojection loss is an unidirectional Chamfer distance; we only penalize the average distance from each projected ground truth point to the nearest projection of predicted point cloud. This is because the Chamfer distance on another direction tends to be redundant. When the predicted point is projected inside the ground truth 2D segmentation, the distance to the nearest projected ground truth points tends to zero, resulting in a small gradient and having less effect for learning better 3D point clouds. Although we project points instead of surfaces or voxel occupancy, producing non-continuous 2D segments, our 2D reprojection loss is computational efficient and shows promising improvements in experiment. Moreover, fitting a surface for post-refinement (Sec. 3.2) to these points is effective.

We can extend Eq. 2 to project 3D points onto multiple orthographic synthetic views: *e.g*. projecting to $x - y$ plane, $y - z$ plane (image plane) or $x - z$ plane in world coordinates (detailed illustration in Appx. **??**). In this case we can simply change the rotation matrix R based on each view. Our 2D reprojection loss does not require additional rendered 2D silhouette ground truth of known view points, which makes the training possible on the dataset where the 3D ground truth is available.

When being projected on the $y - z$ plane (image plane), the ground truth is a subset of the ground truth silhouette. We thus use 2D points sampled from the ground truth segmentation mask S instead of projecting ground truth 3D points P:

$$L_{silhouette} = \frac{1}{|S|} \sum_{s \in S} \min_{\hat{p} \in \hat{P}} \|s - \pi(K[R\ t]\hat{p})\|_2^2 \tag{3}$$

where π is the 3D projection wrap function.

The **overall loss function** is shown below:

$$L = w_{rec} L_{rec} + w_{silhouette} L_{silhouette} + w_{proj} L_{proj} \tag{4}$$

which is the weighted summation over the 3D Chamfer loss and the 2D reprojection losses. Here $L_{silhouette}$ is the 2D reprojection loss on the image plane, and L_{proj} is the projection on $x-y, x-z$ planes. Note that different from PCN, our network only penalizes on the finest output, which helps ease training and shows no performance degrades.

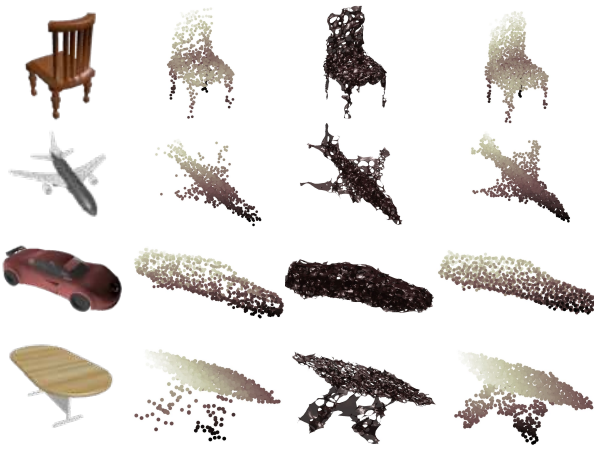

Figure 3. Surface-based point clouds refinement. We show from left to right: input RGB image, network prediction, FSSR surface fitting and smoothing and the re-sampled point clouds from the fitted surfaces. Each sample consists the same number of points and we visualize each predicted shape in a novel view for better illustration. Our refinement step is able to produce smooth and uniformly distributed point sets. Best viewed in color.

Implementation details. Our network gets as input a 224×224 image with pixel values normalized to $[0, 1]$. The bottleneck feature size is 1024. The coarse decoder consists two fc layers, with feature size of 1024 and 3072 and ReLU in between. We set the surface grid for point up-sampling to be zero-centered with a side length of 0.1. We use the ResNet encoder pre-trained from ImageNet and apply a stage-wise training scheme for faster convergence and easier training: first train to predict coarse point cloud, fix the trained layers, then train the up-sampling header, and finally train the whole network end-to-end. We use ADAM [23] to update network parameters with a learning rate of $1e^{-4}$ and $\epsilon = 1e^{-6}$ and batch size 32. We set $w_{rec} = 1$, $w_{silhouette} = 1e^{-9}$ and $w_{proj} = 1e^{-10}$ in Eq. 4 based on grid search in the validation set.

Data augmentation. We augment the training samples by gamma correction with γ between 0.5-2. We re-scale image intensity with a minimum intensity ranges between 0-127 and a fixed maximum intensity of 255. We add color jittering to each RGB channel independently by multiplying a factor ranges in 0.8-1.2. Each augmentation parameter is uniformly and randomly sampled from the defined range.

3.2. Surface-based Point Clouds Refinement

One important 3D shape property is the smooth and continuous shape surfaces, especially for thin structures like chair legs and light stands. To impose this property, we perform a post-refinement step (Fig. 3), fitting surfaces from dense points, smoothing surfaces and uniformly re-sampling points from the surfaces again. Our surface fitting method is based on Floating Scale Surface Reconstruction (FSSR) [10], which is the state-of-the-art for surface-based reconstruction from dense point clouds. We set the parameter of point-wise normal by plane fitting on 6 near-east neighbors, and set the per-point scale as the average distance to the two closest points. A mesh cleaning step goes after FSSR to remove small and redundant patches. We also experimented with Poisson surface reconstruction [22], but FSSR produces a better surface in our case.

Smoothing. Given the fitted surfaces, we perform smoothing by implicit integration method [5]. We use curvature flow as the smoothing operator for 5 iterations. We then uniformly sample point clouds based on Poisson disc sampling to obtain our final output.

Our surface fitting, smoothing, and re-sampling enables production of evenly distributed points that can model thin structures and smooth surfaces, but the predicted shape may not be closed, preventing volumetric-based evaluation. Also, small and disconnected sections of points that model details may be lost, and points that are incorrectly connected to the mesh surfaces may increase errors in some area. Overall, though, our proposed post-processing refinement step improves the performance (large gain in earth-mover's distance with a small cost to Chamfer distance).

4. Reconstruction of Occluded Objects

So far, our proposed point cloud reconstruction approach does not consider foreground occlusions: such as a table in front of a sofa, or a person or pillows on the sofa. Standard approaches do not handle occlusions well since the model does not know whether or where an object is occluded. Starting with an RGB image and an initial silhouette of the visible region, which can be acquired by recent approaches such as Mask-RCNN [16], we propose a 2D silhouette completion approach to generate the complete silhouette of the object. We show that prediction based on the true completed silhouette greatly improves shape prediction and brings performance on occluded objects close to non-occluded; using predicted silhouettes also improves performance, with completed silhouettes outperforming predicted silhouettes of the visible portion.

4.1. Silhouette Completion Network

We assume a detected object, its RGB image crop I and the segmentation of the visible region S_v. Our silhouette completion network (Fig. 2) predicts the complete 2D silhouette S_f of the object based on S_v. The network follows an encoder-decoder strategy, gets as input the concatenation of I and S_v, and predicts S_f with the same resolution as S_v. The encoder is a modified ResNet-50 and the decoder consists 5 up-sampling layers, producing a single channel silhouette. Intuitively, when occlusion occurs, we want the network to complete silhouette, and when no occlusion oc-

curs, we want the network to predict the original segmentation. We add skip connections to obtain this property. We concatenate the feature after the $i-th$ conv layer of the encoder to the input feature layer of the $(6-i)-th$ decoder layer. A full skip connection helps ease training and produces the best results.

Implementation details. For each input image I and visible silhouette S_v, we resize them to 224×224 with preserved aspect ratio and white pixel padded. The value of I is re-scaled to $[0, 1]$ and S_v is a binary map with 1 indicating the object. For the encoder, we remove the top fc layer and the average pooling layer of a pre-trained ResNet on ImageNet and obtain a bottleneck feature of 7×7. The decoder applies nearest neighbor up-sampling with a scale factor of 2, followed by a convolution layer (kernel size 3×3, stride 1) and ReLU. The decoder feature sizes are 2048, 1024, 512, 256, 64 and 1 respectively, with a Sigmoid operation on top. We train the network using binary cross entropy loss between the prediction and the ground truth complete silhouette. We use ADAM to update network parameters with a learning rate of $1e^{-4}$ and $\epsilon = 1e^{-6}$ and batch size 32. Our final prediction is a binary mask obtained by a threshold of 0.5. To account for the truncation of the full object due to the unknown extent of occluded region, we expand the bounding box around each object by 0.3 on each side.

Data augmentation. For each training sample, we perform random cropping on the input image. We crop with a uniformly and randomly sampled ratio ranges between $[0.2, 0.4]$ on each side of the input image. Other augmentations include left-right flipping with 50% probability and random rotation uniformly sampled within ± 5 degrees on the image plane. We also perform image gamma correction, intensity changes and color jittering as in Sec. 3.1.

4.2. Silhouette Guided 3D Reconstruction

Given the predicted complete silhouette S_f, we modify our point cloud reconstruction network to be robust to occlusion. We concatenate our predicted complete silhouette S_f as an additional input channel to the input RGB image I, which can effectively guide reconstruction for both partially occluded and non-occluded objects. We show in experiments the importance of using silhouette compared to the approach with no silhouette guidance at all.

Synthetic occlusion dataset. Since there is no large-scale 3D dataset of rendered 2D object image with occlusion and the complete silhouette ground truth available, we propose to generate a synthetic occlusion dataset. Instead of off-line rendering samples which is time consuming and has limited variety of occlusion, we propose a "cut-and-paste" strategy to create random foreground occlusion. Starting with a set of pre-rendered 2D images without occlusion and with known ground truth silhouettes, for each input image I, we randomly select another image I' from the same split (train/val/test) of the dataset as I, cutting out the object segment O' from I', pasting and overlaying O' on the input image I. To be more specific, we paste on the location uniformly sampled from $[(h_0-h', w_0-w'), (h_1+h', w_1+w')]$, where $[(h_0, w_0), (h_1, w_1)]$ denotes the top left and bottom right position of the bounding box around the object segment O in I. h', w' are the height and width of the pasted segment O' which is considered as foreground occlusion. To ease training, we exclude input samples with pasted occlusion covering over 50% of the complete object segment. We perform "cut-and-paste" with 50% probability in training and further add randomly sampled real-scene background with 50% probability. We use the same data augmentation as in Sec. 3.1 to train the network. We penalize the network on the ground truth complete 3D point clouds and the 2D reprojection loss assuming full shape 2D reprojection. We use the ground truth visible and complete silhouettes to train the network.

5. Experiments

5.1. Setup

We verify the following three aspects of our proposed framework: (1) the performance of our reconstruction network compared with the state-of-the-art, the positive impact of surface refinement and reprojection loss (Sec. 5.3); (2) the performance of our silhouette completion network compared with the state-of-the-art (Sec. 5.4); (3) the impact of silhouette guidance with robustness to occlusion (Sec. 5.5). Please find more results in the appendix.

We use ShapeNet to evaluate the overall reconstruction approach. However, since ShapeNet features already-segmented objects, it is not suitable for evaluating silhouette guidance. For that, we use the Pix3D dataset [32], which contains real images of occluded objects with aligned 3D shape ground truth. We consider the following two standard metrics for 3D point cloud reconstruction:

1. Chamfer distance (CD), defined in Eq. 1.
2. Earth mover's distance (EMD), defined as follows:

$$d_{EMD}(P, \hat{P}) = \min_{\phi: P \to \hat{P}} \frac{1}{|P|} \sum_{p \in P} \|p - \phi(p)\| \quad (5)$$

where $\phi : P \to \hat{P}$ denotes a bijection that minimizes the average distance between corresponding points. Since ϕ is expensive to compute, we follow the approximation solution as in Fan *et al.* [7].

For silhouette completion, we evaluate our approach on the synthetic DYCE dataset [6] and compare with the state-of-the-art. DYCE contains segmentation mask for both visible and occluded region in photo-realistic indoor scenes. We also report performance on the Pix3D real dataset since we perform silhouette guided reconstruction on Pix3D. We

Figure 4. Qualitative results of point cloud reconstruction on ShapeNet dataset. Each from left to right: input RGB image, reconstruction in viewer-centered coordinates and a novel view of the predicted 3D shape. Our approach is able to reconstruct thin structures and smooth surfaces. Best viewed in color.

use the evaluation metric of 2D IoU between the predicted and the ground truth complete silhouette. For detailed comparison, we consider the 2D IoU of the visible potion, the invisible portion and the complete silhouette.

To verify the impact of silhouette guidance, we compare with the following three baselines:

1. Without silhouette guidance (ours w/o seg);
2. Guided by visible silhouette (ours w/ vis seg). We use the predicted visible silhouette (semantic segmentation by Mask-RCNN) to guide reconstruction;
3. Guided by ground truth complete silhouette (ours w/ gt full seg). This shows the upper bound performance of our approach.

5.2. Implementation Details

We implement our network using PyTorch, train and test on a single NVIDIA Titan X GPU with 12GB memory. A single forward pass of the network takes 15.2 ms. The point cloud refinement step is C++ based and takes around 1s on a Linux machine with Intel Xeon 3.5G Hz in CPU mode.

To train our reconstruction approach on ShapeNet, we follow the train-test split defined by Choy *et al*. [4]. The dataset consists 13 objects classes. Each object has 24 randomly selected views rendered in 2D. We randomly select 10% of the shapes from each class of the train set to form the validation set. The viewer-centered point clouds ground truth is generated by Wang *et al*. [36]. Since ShapeNet objects are non-occluded, we train our approach without silhouette guidance.

To train the network with silhouette guidance, we construct a synthetic occlusion dataset (Sec. 4.2) based on ShapeNet and add real background from LSUN dataset. We fine-tune the network we previously trained on ShapeNet on our generated synthetic occlusion data. We train the baseline network having predicted visible silhouette as input with the ground truth visible silhouette generated from ShapeNet. We train the baseline network having no silhouette guidance with RGB images as input only. All configurations use the same training settings. It's worth noting that we do not train reconstructions on Pix3D in any of our experiments to avoid the classification/retrieval problem as discussed by Tatarchenko *et al*. [33].

For silhouette completion, since Pix3D has fewer images, we first train our approach on the synthetic DYCE dataset. We use DYCE's official train-val split and use ground truth visible 2D silhouette as input. On Pix3D, we use Mask-RCNN to detect and segment the visible silhouette of the object in each image, then perform completion upon the segmentation. We obtain valid detections with correctly detected object class and a 2D IoU > 0.5 compared to the ground truth bounding box. We use 5-fold cross validation to fine-tune the silhouette completion network pretrained on DYCE. We further split out 10% val data from each train split in each fold to tune network parameters. Note that images in Pix3D with the same 3D ground truth model are in the same split of either train, val or test.

Category	CD			EMD		
	PSG	Pixel2Mesh	Ours	PSG	Pixel2Mesh	Ours
plane	0.430	0.477	**0.386**	**0.396**	0.579	0.527
bench	0.629	0.624	**0.436**	1.113	0.965	**0.815**
cabinet	0.439	0.381	**0.373**	2.986	2.563	**2.147**
car	0.333	**0.268**	0.308	1.747	**1.297**	1.306
chair	0.645	0.610	**0.606**	1.946	1.399	**1.257**
monitor	0.722	0.755	**0.501**	1.891	1.536	**1.314**
lamp	1.193	1.295	**0.969**	1.222	1.314	**1.007**
speaker	0.756	0.739	**0.632**	3.490	2.951	**2.441**
firearm	**0.423**	0.453	0.463	**0.397**	0.667	0.572
couch	0.549	0.490	**0.439**	2.207	1.642	**1.536**
table	0.517	**0.498**	0.589	2.121	1.480	**1.340**
cellphone	0.438	0.421	**0.332**	1.019	0.724	**0.674**
watercraft	0.633	0.670	**0.478**	0.945	0.814	**0.730**
mean	0.593	0.591	**0.501**	1.653	1.380	**1.205**

Table 1. Viewer-centered single image shape reconstruction performance compared with the state-of-the-art on ShapeNet. We report both the Chamfer distance (CD, left) and the Earth mover's distance (EMD, right).

5.3. Single Image 3D Reconstruction without Occlusion

We show in Fig. 4 sample qualitative results of our reconstructions on ShapeNet. We report in Tab. 1 our quantitative comparison with the state-of-the-art viewer-centered reconstruction approaches. PSG [7] generates point clouds and Pixel2Mesh [36] produces meshes. We sample the same 2466 points as PSG and Pixel2Mesh for fair comparison. Our approach outperforms the two methods on both metrics.

Method	CD	EMD
Ours w/o surface refine & w/o 2D proj loss	0.398	1.784
Ours w/o surface refine	**0.389**	1.660
Ours w/o 2D proj loss	0.502	1.220
Ours Full	0.501	**1.205**

Table 2. Ablation study. We evaluate our performance on ShapeNet based on CD (left) and EMD (right) with different configurations. Our full approach seeks for a balanced performance on both two metrics.

Method	CD	EMD
3D-LMNet [27]	**5.40**	7.00
Ours	5.54	**5.93**

Table 3. Comparison with the state-of-the-art object-centered point cloud reconstruction approach on ShapeNet, reported in both CD (left) and EMD (right). Although our method is trained on a harder viewer-centered prediction, our method achieves a much better EMD and a slightly worse CD compared to 3D-LMNet.

Method	Full	Visible	Occluded
SeGAN [6]	76.4	63.9	27.6
Ours (ResNet-18)	82.8	82.9	33.9
Ours (ResNet-50)	**84.3**	**83.4**	**36.2**

Table 4. Silhouette completion performance on DYCE dataset. We report the 2D IoU of visible, occluded and complete silhouette. "Ours (ResNet-18)" use the same encoder as SeGAN.

Ablation study. We show in Tab. 2 our performance with different configurations: without both surface-based refinement step and 2D reprojection loss, without refinement step only, without 2D reprojection loss only and our full approach. We observe the improvement with 2D reprojection loss on both CD and EMD. The improvement with 2D loss is less significant when we have the surface-based refinement step, showing the refinement step mitigates some problems with point cloud quality that are otherwise remedied by training with the reprojection loss. The refinement step increases Chamfer distance, mainly due to the smoothed out small and sparse point pieces that model the thin and complex shape structure like railings or handles (Fig. 4, 1st row, 2nd column), and the enhanced error point predictions by connecting sparse point sets (Fig. 4, 3rd row, 2nd column). It's also worth noting that, even without the post-refinement step, our approach achieves a better CD than PSG and a slightly worse EMD, proving the superiority of our network architecture.

Comparison with object-centered approach. Tab. 3 shows our comparison with the state-of-the-art object-centered point cloud reconstruction approach: 3D-LMNet [27] on ShapeNet. We follow the evaluation procedure as 3D-LMNet, sample 1024 points and re-scale our prediction (and ground truth) to be zero-centered and unit length of 1, then perform ICP to fit to the ground truth. Although our approach targets the more difficult task of joint shape prediction and view point estimation, we achieve a much better EMD and only a slightly worse CD. Note that the reported CD and EMD are of different scales compared to that in Tab. 1, this is because 3D-LMNet takes a squared value when computing CD and the ground truth is re-sized.

Method	Training data	Occluded			Non-occluded		
		sofa	chair	table	sofa	chair	table
Mask-RCNN	real	84.34	59.05	60.14	91.99	69.96	60.88
Ours	syn	87.58	**59.61**	58.12	92.02	69.88	**64.94**
Ours	syn+real	**88.56**	59.25	**68.83**	**92.19**	**72.01**	56.90

Table 5. Silhouette completion performance on Pix3D dataset. We report the 2D IoU between the predicted and the ground truth complete silhouette for occluded and non-occluded objects.

Method	CD			EMD		
	sofa	chair	table	sofa	chair	table
Ours w/o seg	15.54	17.96	24.35	16.63	16.51	22.56
Ours w/ pred vis seg	9.15	13.20	17.96	9.29	13.38	17.81
Ours w/ pred full seg	**8.70**	**13.14**	**16.50**	**8.81**	**13.04**	**16.36**
Ours w/ gt full seg	8.27	10.16	11.36	8.11	10.47	11.44

Table 6. Quantitative results for reconstructing *occluded* objects in the Pix3D dataset. We report both CD (left) and EMD (right).

Method	CD			EMD		
	sofa	chair	table	sofa	chair	table
Ours w/o seg	12.62	16.00	20.65	13.18	15.44	19.92
Ours w/ pred vis seg	8.75	11.34	15.55	8.66	11.84	15.75
Ours w/ pred full seg	**8.42**	**10.82**	**13.65**	**8.40**	**11.13**	**13.85**
Ours w/ gt full seg	8.24	9.21	10.50	8.18	9.66	11.44

Table 7. Quantitative results for reconstructing *non-occluded* objects in the Pix3D dataset. We report CD (left) and EMD (right).

5.4. Silhouette Completion

Tab. 4 shows the comparison with the state-of-the-art silhouette completion approach SeGAN [6] on DYCE test set. SeGAN uses ResNet-18 encoder, our approach with ResNet-18 outperforms SeGAN, due to our better upsampling based decoder that predicts better region beyond occlusion, and the skip connection architecture that preserves the visible region. We use ResNet-50 encoder which yields the best performance to complete silhouettes for the downstream reconstruction task.

Tab. 5 reports our silhouette completion performance on Pix3D. Our approach achieves a better performance than the Mask-RCNN baseline (visible region). We show better performance by fine-tuning our completion approach on Pix3D ("syn+real" v.s. "syn" for training data).

5.5. Robustness to Occlusion

In Fig. 5 we show our qualitative performance on Pix3D for the three object classes that co-occur in ShapeNet. Tab. 6 and Tab. 7 present our quantitative performance of reconstructing occluded objects and non-occluded objects respectively. For evaluation, we use the ground truth point cloud provided by Mandikal *et al*. [27] and sample the same number of points from our approach for evaluation. Since Pix3D evaluates object-centered reconstruction, we rotate each

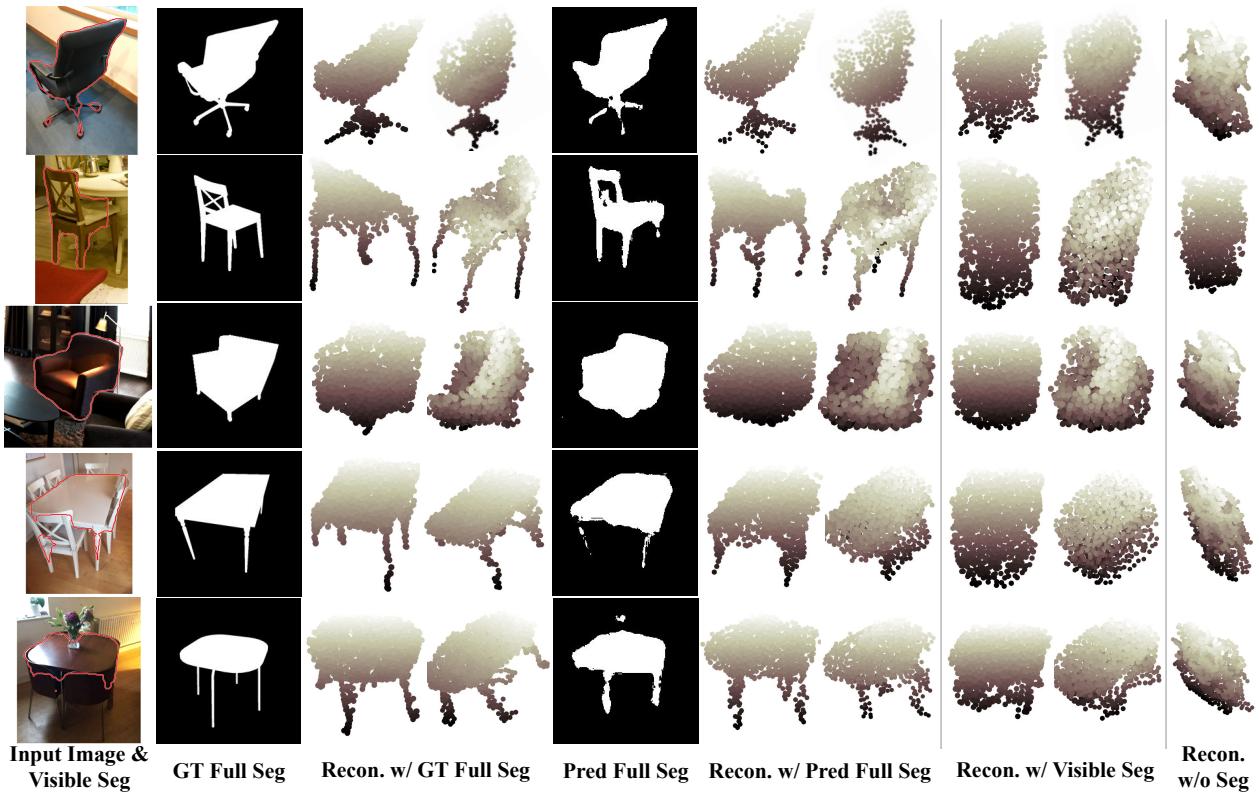

Figure 5. Qualitative results on Pix3D dataset. We show in each row from left to right: input RGB image with predicted visible silhouette obtained by Mask-RCNN (outlined in red), ground truth complete silhouette, reconstruction guided by ground truth complete silhouette in two views (viewer-centered and a novel view), our predicted complete silhouette, reconstruction guided by predicted complete silhouette, reconstruction guided by visible silhouette and reconstruction without silhouette guidance. The first row shows a non-occluded object, and other rows show occluded objects. Best viewed in color.

ground truth shape to have the viewer-centered orientation w.r.t. camera. We then re-scale both ground truth and our prediction to be zero-centered and unit-length, and perform ICP with translation only. We follow the officially provided evaluation metrics of Chamfer distance and Earth mover's distance by Sun *et al*. [32]. We compare the performance of our silhouette guided reconstruction "Ours w/ pred full seg" with three baselines: without silhouette guidance, guided with predicted visible silhouette from Mask-RCNN, and guided with ground truth complete silhouette. Compared to the qualitative results on ShapeNet (Fig. 4), we see the challenges of 3D reconstruction given real images due to occlusion and complex background. Using ground truth complete silhouette can make the network be robust to occlusion. Without silhouette guidance, the prediction is difficult because the network does not know whether or where an object is occluded and what to reconstruct; and the network faces the challenges of synthetic to real, since our reconstruction network is only trained on synthetic dataset. Our proposed silhouette guidance is able to bridges the gap between synthetic and real, referring to the large performance boosts from "Ours w/o seg" to other rows. With the guidance of predicted complete silhouettes, our approach is able to narrow down the performance gap between occluded and non-occluded objects, and outperforms the approach with predicted visible silhouettes.

6. Conclusion

We propose a method to reconstruct the complete 3D shape of an object from a single RGB image, with robustness to occlusion. Our point cloud reconstruction approach achieves the state-of-the-art with the major improvement by the surface-based refinement step. We show that, when provided with input ground truth silhouettes, the shape prediction performance is nearly as good for occluded as for non-occluded objects. Using the predicted silhouette also yields large improvements for both occluded and non-occluded objects, indicating that providing an explicit foreground/background separation for the object in RGB images is helpful.

Acknowledgements

This research is supported in part by NSF award 14-21521 and ONR MURI grant N00014-16-1-2007.

References

[1] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3762–3769, 2014. 2

[2] A. Bansal, B. Russell, and A. Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5965–5974, 2016. 2

[3] J. T. Barron and J. Malik. Color constancy, intrinsic images, and shape estimation. In *European Conference on Computer Vision*, pages 57–70. Springer, 2012. 2

[4] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 6

[5] M. Desbrun, M. Meyer, P. Schröder, and A. H. Barr. Implicit fairing of irregular meshes using diffusion and curvature flow. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 317–324. Citeseer, 1999. 4

[6] K. Ehsani, R. Mottaghi, and A. Farhadi. Segan: Segmenting and generating the invisible. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6144–6153, 2018. 1, 2, 5, 7

[7] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 2, 3, 5, 6

[8] P. Favaro and S. Soatto. Seeing beyond occlusions (and other marvels of a finite lens aperture). In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–II. IEEE, 2003. 2

[9] S. Fowler, H. Kim, and A. Hilton. Towards complete scene reconstruction from single-view depth and human motion. In *Proceedings of the 28th British Machine Vision Conference (BMVC 2017)*, 2017. 2

[10] S. Fuhrmann and M. Goesele. Floating scale surface reconstruction. *ACM Transactions on Graphics (ToG)*, 33(4):46, 2014. 4

[11] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 2

[12] L. Guan, J.-S. Franco, and M. Pollefeys. Multi-view occlusion reasoning for probabilistic silhouette-based dynamic scene reconstruction. *International journal of computer vision*, 90(3):283–303, 2010. 1, 2

[13] R. Guo and D. Hoiem. Labeling complete surfaces in scene understanding. *International Journal of Computer Vision*, 112(2):172–187, 2015. 1, 2

[14] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik. Aligning 3d models to rgb-d images of cluttered scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4731–4740, 2015. 2

[15] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik. Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision*, 112(2):133–149, 2015. 2

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3, 4

[17] H. Izadinia, Q. Shan, and S. M. Seitz. Im2cad. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5134–5143, 2017. 2

[18] L. Jiang, S. Shi, X. Qi, and J. Jia. Gal: Geometric adversarial loss for single-view 3d-object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 802–816, 2018. 2

[19] S. B. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multi-view stereo. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001. 1, 2

[20] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Category-specific object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1966–1974, 2015. 2

[21] K. Karsch, Z. Liao, J. Rock, J. T. Barron, and D. Hoiem. Boundary cues for 3d object shape recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2163–2170, 2013. 2

[22] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006. 4

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[24] J. J. Koenderink. What does the occluding contour tell us about solid shape? *Perception*, 13(3):321–330, 1984. 2

[25] C. Kong, C.-H. Lin, and S. Lucey. Using locally corresponding cad models for dense 3d reconstructions from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4857–4865, 2017. 2

[26] C. Liu, P. Kohli, and Y. Furukawa. Layered scene decomposition via the occlusion-crf. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 165–173, 2016. 2

[27] P. Mandikal, K. L. Navaneet, M. Agarwal, and R. V. Babu. 3D-LMNet: Latent embedding matching for accurate and diverse 3d point cloud reconstruction from a single image. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 2, 7

[28] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 2

[29] K. Rematas, I. Kemelmacher-Shlizerman, B. Curless, and S. Seitz. Soccer on your tabletop. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 4738–4747, 2018. 2

[30] J. Rock, T. Gupta, J. Thorsen, J. Gwak, D. Shin, and D. 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. 2

[31] D. Shin, C. C. Fowlkes, and D. Hoiem. Pixels, voxels, and views: A study of shape representations for single view 3d object shape prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3069, 2018. 2

[32] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2974–2983, 2018. 2, 5, 8

[33] M. Tatarchenko, S. R. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2019. 6

[34] A. O. Ulusoy, M. J. Black, and A. Geiger. Semantic multi-view stereo: Jointly estimating objects and voxels. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4531–4540. IEEE, 2017. 1, 2

[35] V. Vaish, M. Levoy, R. Szeliski, C. L. Zitnick, and S. B. Kang. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2331–2338. IEEE, 2006. 2

[36] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018. 2, 6

[37] P.-S. Wang, C.-Y. Sun, Y. Liu, and X. Tong. Adaptive o-cnn: a patch-based deep representation of 3d shapes. In *SIGGRAPH Asia 2018 Technical Papers*, page 217. ACM, 2018. 2

[38] J. Wu, Y. Wang, T. Xue, X. Sun, B. Freeman, and J. Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. In *Advances in neural information processing systems*, pages 540–550, 2017. 2

[39] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. In *European Conference on Computer Vision*, pages 365–382. Springer, 2016. 2

[40] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 2

[41] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, pages 1696–1704, 2016. 2

[42] B. Yang, S. Rosa, A. Markham, N. Trigoni, and H. Wen. Dense 3d object reconstruction from a single depth view. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 2

[43] W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert. Pcn: Point completion network. In *2018 International Conference on 3D Vision (3DV)*, pages 728–737. IEEE, 2018. 3

[44] Y. Zhu, Y. Tian, D. Metaxas, and P. Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1464–1472, 2017. 1, 2

[45] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *ACM transactions on graphics (TOG)*, volume 23, pages 600–608. ACM, 2004. 2

[46] C. Zou, E. Yumer, J. Yang, D. Ceylan, and D. Hoiem. 3d-prnn: Generating shape primitives with recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 900–909, 2017. 2