Discovering 3D Parts from Image Collections

Chun-Han Yao[1] Wei-Chih Hung[2] Varun Jampani[3] Ming-Hsuan Yang[134]
[1]UC Merced [2]Waymo [3]Google [4]Yonsei University

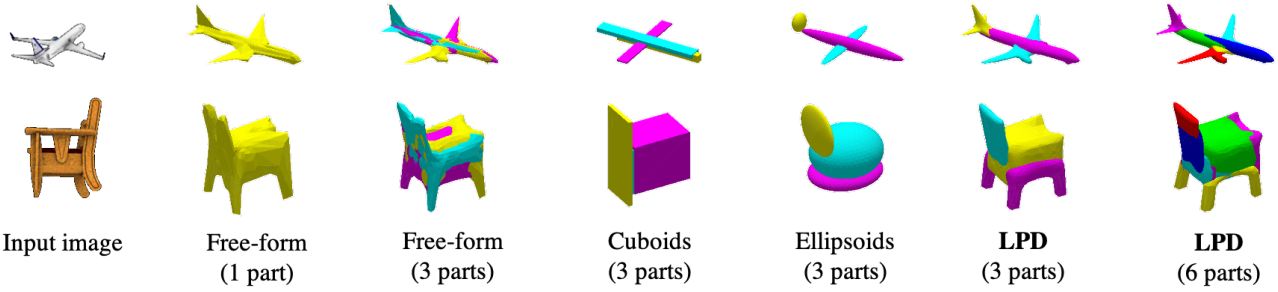

Figure 1: **Discovering 3D parts form single-view image collections.** Our method (LPD) enables self-supervised 3D part discovery while learning to reconstruct object shapes from single-view images. Compared to other methods using different part constraints, LPD discovers more faithful and consistent parts, which improve the reconstruction quality and allow part reasoning/manipulation.

Abstract

Reasoning 3D shapes from 2D images is an essential yet challenging task, especially when only single-view images are at our disposal. While an object can have a complicated shape, individual parts are usually close to geometric primitives and thus are easier to model. Furthermore, parts provide a mid-level representation that is robust to appearance variations across objects in a particular category. In this work, we tackle the problem of 3D part discovery from only 2D image collections. Instead of relying on manually annotated parts for supervision, we propose a self-supervised approach, latent part discovery (LPD). Our key insight is to learn a novel part shape prior that allows each part to fit an object shape faithfully while constrained to have simple geometry. Extensive experiments on the synthetic ShapeNet, PartNet, and real-world Pascal 3D+ datasets show that our method discovers consistent object parts and achieves favorable reconstruction accuracy compared to the existing methods with the same level of supervision. Our project page with code is at https://chhankyao.github.io/lpd/.

1. Introduction

Recognizing and reasoning about objects surrounding us is essential for many computer vision systems. While deep learning models have been shown to perform well at recognizing [27, 46, 17] and localizing [13, 44, 35] objects in a 2D image, reasoning 3D attributes of objects from a single image remains a challenging task. Single-view 3D reasoning is fundamentally ill-posed due to several factors that cause ambiguous object appearance in 2D images, *e.g.*, camera pose, self-occlusions, lighting, and material properties. Although objects in general have complicated shapes, they can usually be decomposed into parts that have simpler geometry and are relatively easy to model. Furthermore, most object instances of a particular category share similar part configurations, *e.g.*, the wings, body, and tail of airplanes. In this work, we propose to tackle the problem by discovering faithful and consistent 3D parts from 2D image collections. Compared to existing single-view 3D reconstruction approaches that directly predict an object shape, we aim to learn rich and dense part configurations which form an entire object when combined.

Although several recent methods [48, 37, 3, 10, 33, 36, 41, 24] leverage part-based representations for 3D object reasoning, they rely on either 3D object shapes or explicit part annotations as supervision. Moreover, the learned parts only serve as additional information and are not exploited to improve 3D reconstruction. Considering that collecting ground-truth 3D shapes and the corresponding part labels is labor intensive, we follow a practical scenario where only single-view images, 2D object silhouettes, and camera viewpoints are available for model training. In contrast to existing techniques, our method automatically discovers 3D parts from image collections in a self-supervised manner.

A common practice to represent 3D parts is to use geometric primitives such as ellipsoids or cuboids [48]. They

provide a strong regularization of part shapes but are usually too coarse to faithfully represent object parts. As an alternative, several approaches adopt meshes [23, 50, 14, 22, 34, 18, 15, 32] or point-cloud [8, 20, 28, 40, 6] representations. Although these representations are more expressive and can faithfully describe part shapes, they lack part shape regularization that is particularly needed in a weakly-supervised or unsupervised setting. The key insight of this work is to represent 3D parts with deep latent embeddings. Specifically, we propose to learn a prior distribution of part shapes with a variational auto-encoder (VAE) [26] that encodes part shapes as latent embeddings. We call this network *Part-VAE* and pre-train it with a set of geometric primitives like cones, cylinders, cuboids, and ellipsoids. We then learn a reconstruction network that takes an input image and predicts part embeddings to obtain a 3D mesh by passing through the decoder of Part-VAE. To further improve the quality of part discovery and reconstruction, we propose a novel part adversarial loss which involves re-assembling parts from different objects in the same category. We name the proposed method *latent part discovery (LPD)*. Figure 1 shows several reconstruction results with and without part prior, which demonstrate that LPD can discover consistent parts and produce faithful reconstruction to the input image.

We evaluate LPD on the synthetic ShapeNet [1], PartNet [39] and the real-world Pascal 3D+ [51] datasets. Both quantitative and qualitative results demonstrate that our approach achieves favorable performance against the state-of-the-art methods using the same level of supervision. In addition to part discovery, our part representation enables object manipulation like selective part swapping, interpolation, and random shape generation from the latent space. In this work, we make the following contributions:

- We propose a part-based single-view 3D reasoning network which can automatically discover object parts. To the best of our knowledge, this is the first work that discovers 3D parts in a self-supervised manner without using any 3D shape or multi-view supervision.

- We develop Part-VAE to learn a latent prior over part shapes. We show that training with geometric primitives can learn useful part embeddings, allowing each part to faithfully represent object shape while constrained to have simple geometry.

- We conduct extensive experiments on both synthetic and natural images. Qualitatively, our method produces more faithful and consistent object parts compared to other part-based methods. Quantitatively, the discovered parts improve whole-object reconstruction and achieve favorable accuracy against the state-of-the-art techniques. In addition, our Part-VAE allows us to manipulate object parts for various applications.

2. Related Work

3D Reconstruction. While 3D representation has been widely studied for decades, the best and unified way to represent general objects remains unclear. Voxel grids [4, 49, 47, 31], point clouds [8, 20, 28, 40, 6], and meshes [23, 50, 14, 22, 34, 2, 18, 15, 32] are commonly used to represent object shapes. Several recent methods [12, 38, 52, 16, 5, 11] explore the possibilities to represent 3D shapes in a functional space. While fine-grained voxels, point clouds, and local functions can represent complicated shapes, the flexibility of representation demands strong 3D or multi-view supervision for training. Meshes, on the other hand, are constrained to form a water-tight surface and are convenient to render 2D images. By deforming from a simple template mesh like sphere or cuboid, it is easier to apply shape regularization and thus can be applied to reconstruction scenarios with weaker supervision. One can learn single-view mesh reconstruction from multi-view or single-view images using naive 2D projection or differentiable rendering [23, 34, 2]. For instance, Henderson *et al*. [18] generate background images and object rendering to learn textured mesh reconstructions from natural images. Kato *et al*. [22] propose view prior learning (VPL) to improve the reconstructed shapes from unseen views. Although they are effective for compact and deformable objects, a single mesh cannot represent complicated shapes with holes or disconnected parts. In this work, we exploit multi-mesh part representation which allows disconnected parts in an object while each part can be well-regularized.

Part Discovery. Parts provide a mid-level representation that is robust to appearance variations across objects in the same category. Hung *et al*. [19] learn 2D co-part segmentation on image collections with self-supervision. Lathuilière *et al*. [29] exploit motion cues in videos for part discovery. In the 3D domain, Tulsiani *et al*. [48] use volumetric cuboids as part abstractions to learn 3D reconstruction. Li *et al*. [33] assume known part shapes and learn to assemble them given an input image. Using a point-cloud representation, Mandikal *et al*. [37] predict part-segmented 3D reconstructions from a single image and Luo *et al*. [36] learn to form object parts by clustering 3D points. Paschalidou *et al*. [41] propose hierarchical part decomposition (HPD) by constraining 3D points with super-quadratic functions [42]. These methods require 3D ground-truth shape of a whole object or its parts as supervision. Furthermore, the part shapes in [48, 41] are limited by the expressiveness of their representations. Li *et al*. [32] leverage 2D semantic parts to improve single-view 3D reconstruction, which, however, does not produce individual part shapes. To the best of our knowledge, we propose the first 3D part reconstruction method without any part annotations or ground-truth 3D shapes for training, and our latent part representation enables each part to fit a given object shape faithfully.

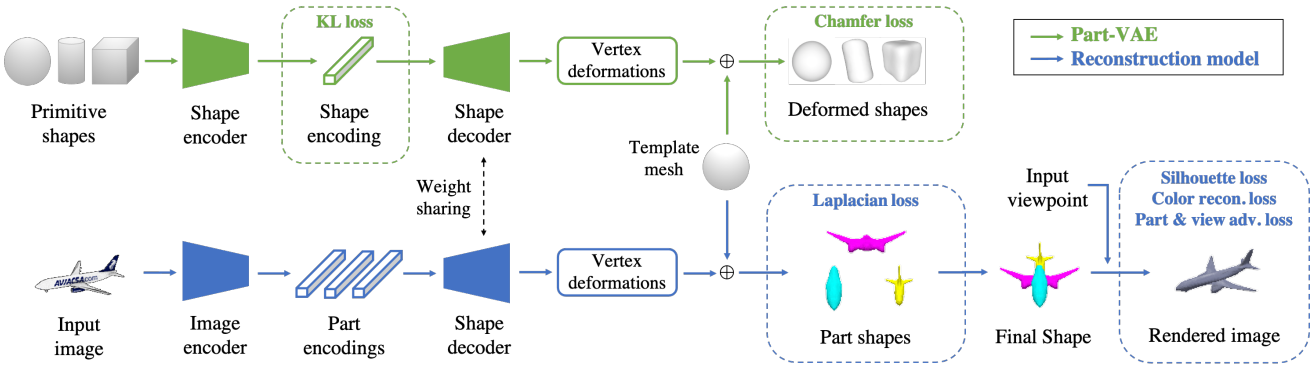

Figure 2: **Approach overview.** (top) Our Part-VAE is trained with geometric primitives. (bottom) Our reconstruction model shares the shape decoder with Part-VAE and predicts object parts. We then composite the reconstructed parts to form a 3D object. The prediction of part centroids and surface texture are detailed in the supplemental material.

3. Approach

Reasoning 3D objects from single-view images is inherently ill-posed since the reconstructed object can overfit to a given view and be highly deformed in the unseen parts. To address this, we propose LPD to represent an object with multiple latent parts. Our intuition is that a complicated object shape can be expressed by assembling simple and regularized parts. Consistent with some recent approaches [22, 18], we propose a method under a weakly-supervised setting where only a single-view image, 2D object silhouette, and its camera viewpoint are available for each object. That is, we do not assume any 3D shape or multi-view images to supervise part discovery or reconstruction. In order to automatically discover underlying parts while reconstructing an object, we propose to learn part embeddings with a variational auto-encoder [26] (VAE) named *Part-VAE*. We train a reconstruction model that predicts part embeddings which are then decoded into part meshes to compose the whole object. Figure 2 illustrates the proposed method with two main modules: Part-VAE and reconstruction network.

3.1. Learning Part Prior with Part-VAE

We propose Part-VAE to learn a latent shape prior for object parts. The proposed method constrains parts with primitive shapes while allowing the flexibility to fit real-world object parts. In addition, it enables smooth part-interpolation and novel shape generation by random sampling in the latent space. Figure 2 (top) illustrates the training process of the Part-VAE with geometric primitives. We first collect a set of primitive shapes such as ellipsoids, cylinders, cones, and cuboids, which are centered at origin but with random scaling and rotation. The Part-VAE network consists of a shape encoder and a shape decoder. The encoder transforms each given primitive shape to a low-dimensional shape encoding, and the decoder reconstructs the input shape by predicting the vertex deformations of a spherical template mesh. To supervise the Part-VAE pre-training, we calculate the Chamfer distance between the input and output vertices as the loss function since the point sets are unordered and not densely corresponding. Given the vertices of an input shape Q and its reconstruction P, the Chamfer loss $\mathcal{L}_c$ can be expressed as:

$$\mathcal{L}_c(P,Q) = \frac{1}{\|P\|}\sum_{p\in P}\min_{q\in Q}\|p-q\|_2^2 + \frac{1}{\|Q\|}\sum_{q\in Q}\min_{p\in P}\|p-q\|_2^2. \quad (1)$$

To encourage the continuity of latent shape distribution, we adopt a standard KL divergence loss $\mathcal{L}_{kl}$ calculated between the shape embeddings and a standard normal distribution $\mathcal{N} \sim (0,1)$. The overall training loss of Part-VAE is: $\mathcal{L}_{vae} = \mathcal{L}_c + \lambda_{kl}\,\mathcal{L}_{kl}$, where λ_{kl} is a weight parameter.

3.2. Part Discovery by Learning to Reconstruct

Figure 2 (bottom) illustrates our reconstruction model. Instead of directly predicting an object mesh, we learn an image encoder that takes an input image and predicts the 3D part centroid, latent shape encodings, and surface texture for each part. The part encodings are then passed through the shape decoder of Part-VAE to generate part meshes. To compose an entire object, we simply shift the mesh vertices using the predicted part centroids and concatenate the vertices and surfaces of each part. In the remainder of the paper, we denote the reconstruction model as $R(\cdot)$, which takes an image as input and outputs a part-composited mesh. The rendering function from viewpoint v is denoted as $G(\cdot, v)$, which produces a rendered image of the input mesh. Note that each training image I includes a silhouette channel I_s and RGB color channels I_c. Likewise, the rendering function G can be separated into G_s and G_c for silhouette projection and color rendering, respectively.

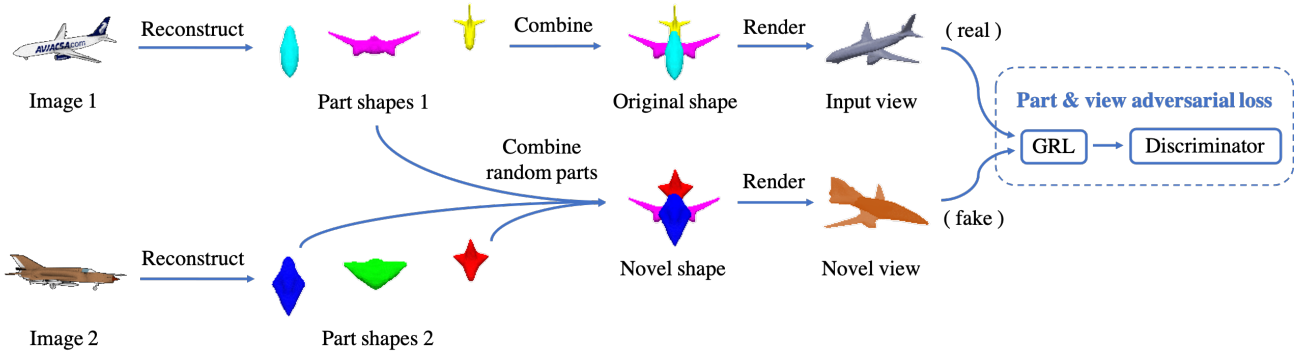

Figure 3: **Part and view adversarial learning.** Given two images with different objects, we randomly combine their reconstructed parts into a novel shape. The novel shape is then rendered from a novel viewpoint, which we treat as a 'fake' sample. We train a discriminator to distinguish the fake and real rendered images. By using a gradient reversal layer (GRL), the reconstruction model learns to produce parts that can compose realistic novel shapes.

Shape Reconstruction Losses. To supervise shape reconstruction, we enforce the 2D projection of a reconstructed shape to be close to the ground-truth silhouette. In particular, we render the predicted meshes with input viewpoints using a differentiable renderer [34], and calculate the intersection-over-union (IoU) ratio between the rendered and ground-truth silhouettes. Then the silhouette loss $\mathcal{L}_{sil}$ is computed as:

$$\mathcal{L}_{sil}(I, v) = \frac{\|I_s \odot G_s(R(I), v)\|_1}{\|I_s + G_s(R(I), v)\|_1 - \|I_s \odot G_s(R(I), v)\|_1}, \tag{2}$$

where $\odot$ denotes element-wise multiplication. We further apply Laplacian regularization $\mathcal{L}_{lap}$ on the reconstructed mesh vertices:

$$\mathcal{L}_{lap}(P) = \frac{1}{\|P\|} \sum_{p \in P} \left\| p - \frac{1}{\|\mathcal{N}(p)\|} \sum_{q \in \mathcal{N}(p)} q \right\|_2^2, \tag{3}$$

where $\mathcal{N}(p)$ denotes the neighboring vertices of vertex p. It aims to smooth the mesh surfaces by pulling each vertex towards the center of its neighboring pixels. Note that this regularization is applied on each part mesh individually so discontinuity between part surfaces is allowed.

Color Reconstruction Loss. We further exploit the color information in input images by generating textured reconstructions. Given the part encodings, our model predicts texture flow to map the input image to a UV texture image. We then color the mesh surfaces by sampling from the texture image using a pre-defined UV mapping function. The texture flow is predicted per object part so each part has more coherent texture. We denote the overall color rendering process as G_c and show more details in the supplemental material. The color reconstruction loss $\mathcal{L}_{cr}$ is defined on the semantic features of input and rendered images:

$$\mathcal{L}_{cr}(I, v) = \|F(I_c) - F(G_c(R(I), v))\|_2^2, \tag{4}$$

where F is the feature extractor of a fixed classification network. We use AlexNet [27] pre-trained on the ImageNet dataset [27] and extract the output of multiple convolutional layers as F. It encourages the rendered image to be perceptually similar to the input at different levels.

3.3. Part and View Adversarial Learning

Unlike the multi-view or 3D-supervised settings, single-view training requires stronger regularization to produce realistic 3D shapes and to discover meaningful parts. Based on the intuition that object parts are interchangeable and they should look realistic from various viewpoints, we extend the view adversarial learning in VPL [22] with part adversarial learning. As illustrated in Figure 3, we assemble a novel shape by randomly combining parts from different objects of the same class in a training batch, then render the novel shape from a novel viewpoint. To make the novel shape realistic, we treat the rendered images of novel shapes as fake examples and those of original shapes as real ones. A discriminator is then trained to classify each rendered image as real or fake. We train the discriminator with a binary cross-entropy between the positive and negative samples:

$$\begin{aligned}\mathcal{L}_{adv}(I, I', v) = &- \log(D(G(R(I), v))) \\ &- \log(1 - D(G(R'(I, I'), v'))),\end{aligned} \tag{5}$$

where I' is a random image different from the input I, $R'(\cdot, \cdot)$ is the reconstruction model with random part selection from two input images, v' is a random novel view, and D is the discriminator. To apply adversarial training, we add a gradient reversal layer (GRL) [9] before the discriminator. As a result, the reconstruction model is trained to fool the discriminator by generating novel objects with realistic shapes. Considering that different object classes may have different view and shape prior, we condition the discriminator with input class labels during training. Note that the

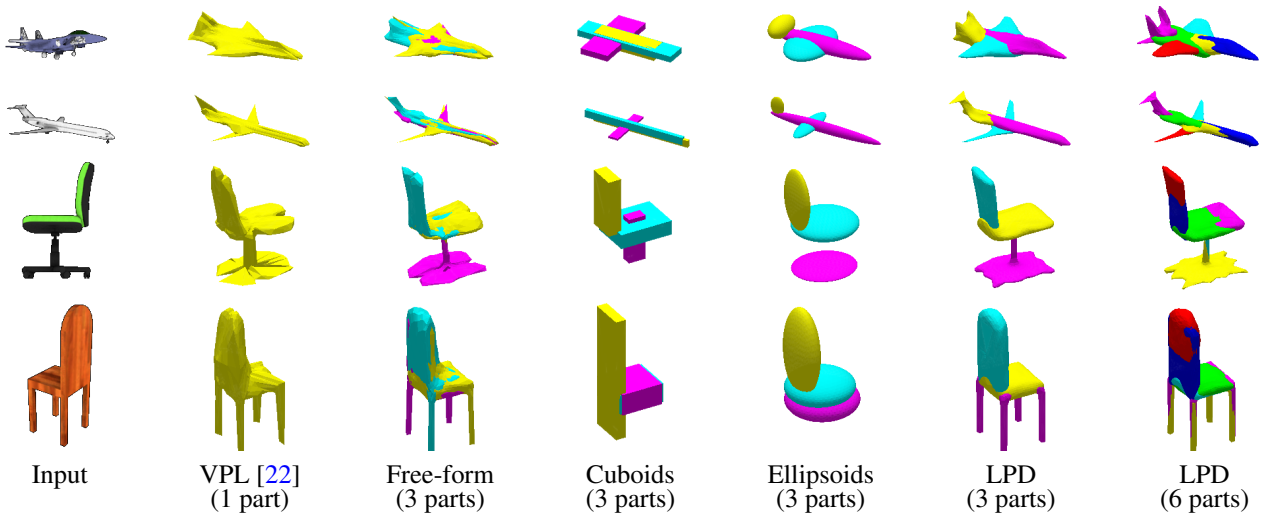

Figure 4: **Qualitative results on the ShapeNet dataset [1].** LPD models (ours) adopt the proposed Part-VAE and adversarial learning. The 3-part free-form model reconstructs a whole object with three fully-deformable meshes without any part prior. Compared to the baselines, our approach can produce more faithful and consistent parts from diverse objects.

class labels are not required during inference. That is, for given data, we train a single part-discovery/reconstruction model that operates across different object categories. This part-based adversarial learning approach is proposed as a semantic constraint to make the global part arrangements more feasible and realistic.

3.4. Model Training and Inference

We first pre-train the Part-VAE with primitive shapes to minimize the loss $\mathcal{L}_{vae}$. Next, the Part-VAE and reconstruction network are jointly trained using image collections together with primitive shapes. The overall objective function of reconstruction network is given by:

$$\mathcal{L}_{sil} + \lambda_{lap}\,\mathcal{L}_{lap} + \lambda_{cr}\,\mathcal{L}_{cr} - \lambda_{adv}\,\mathcal{L}_{adv}, \qquad (6)$$

where $(\lambda_{lap},\ \lambda_{cr},\ \lambda_{adv})$ are weight parameters. The discriminator is trained to minimize $\mathcal{L}_{adv}$, whose gradients are reversed and back-propagated to the reconstruction network to perform adversarial learning. Note that we still fine-tune the Part-VAE with primitive shapes in this stage so that the shape decoder is regularized while adapting to various part shapes in the training images. The Part-VAE, reconstruction network, and discriminator are parametrized as deep neural networks, and the weights are optimized by mini-batch gradient descent. In the model inference phase, we discard the Part-VAE encoder and discriminator. An input image is simply passed through the image encoder and Part-VAE decoder to reconstruct the 3D parts. We represent each object part by a deformable mesh with $N_v = 642$ vertices and $N_f = 1280$ faces. The size of texture image is 64×64. We implement the proposed method in PyTorch [43] framework and use Adam optimizer [25] for training. The hyper-parameters are tuned on a validation set.

4. Experiments and Analysis

Metrics and Baselines. Evaluating self-supervised part discovery can be ambiguous and subjective as the discovered parts need not correspond to human-annotated ones. Due to the lack of standard metrics or benchmark for 3D part discovery, we qualitatively compare the discovered parts with other methods. As a reference, we also quantitatively evaluate the reconstruction accuracy at both object level and part level. We convert each predicted mesh into a volume of 32^3 voxels and calculate the intersection-over-union (IoU) ratio between the voxelized object and the ground-truth voxels. We report results of our model with $k = 3$ parts and latent part embedding dimension of $d = 64$ if not specified otherwise. Since our work is the first to discover 3D parts using single-view supervision, we mainly compare LPD against three part-based baselines: cuboids, ellipsoids, and free-form meshes. We implement the cuboid and ellipsoid models by reconstructing each object part with a scalable cuboid/ellipsoid. The free-form model adopts fully-deformable meshes without any part shape prior. For object-level reconstruction, we evaluate our method against SoftRas [34] and VPL [22] as they adopt a similar training setting with single-view images and known viewpoints. While there are several other methods for single-view 3D reconstruction, we omit the comparison with them since whole-object reconstruction is not the major focus of this work. We experiment on the synthetic ShapeNet [1], PartNet [39], and real-world Pascal 3D+ [51] datasets. We present the main findings here and additional results in the supplemental material.

Table 1: **Ablative evaluations on the ShapeNet dataset [1].** The base model reconstructs an object shape with 3 meshes, each is fully-deformable as in SoftRas [34] (PP: part prior, VA: view adversarial learning, PA: part adversarial learning, CR: color reconstruction).

PP	VA	PA	CR	Airplane	Bench	Dresser	Car	Chair	Display	Lamp	Speaker	Rifle	Sofa	Table	Phone	Vessel	All
	✓	✓	✓	56.7	34.4	56.0	68.3	43.1	34.8	47.2	59.9	50.6	48.5	41.1	42.7	53.4	49.0
✓		✓	✓	57.2	36.2	60.7	72.2	44.1	39.8	48.2	63.6	51.9	49.6	43.3	51.5	55.1	51.8
✓	✓		✓	57.1	35.8	**61.4**	73.7	45.1	39.4	48.5	63.7	52.7	49.3	43.9	52.5	54.9	52.2
✓	✓	✓		57.1	36.0	61.0	74.1	45.2	39.7	48.5	**63.8**	53.0	49.7	43.9	52.2	55.1	52.3
✓	✓	✓	✓	**57.3**	**37.3**	60.9	**75.2**	**45.5**	**40.8**	**49.6**	63.3	**54.5**	**50.1**	**44.3**	**52.7**	**56.2**	**52.9**

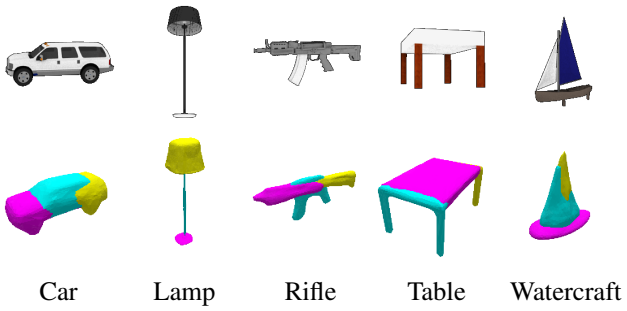

Figure 5: **Generalization across classes.** We show sample inputs (top) and LPD results (bottom) of different ShapetNet classes.

Table 2: **Voxel IoU results on the ShapeNet dataset [1].** We compare our method with the state-of-the-art single-view supervised and 3D supervised approaches.

Method	Supervision	Airplane	Car	Chair	All
SIF [12]	3D shapes	53.0	65.7	38.9	49.9
OccNet [38]	3D shapes	57.1	73.7	50.1	57.1
CvxNet [5]	3D shapes	59.8	67.5	49.1	56.7
HPD [41]	3D shapes	52.9	70.2	52.6	58.0
SoftRas [34]	Single-view	52.2	65.7	40.4	46.9
VPL [22]	Single-view	53.1	70.1	45.4	51.3
LPD (ours)	Single-view	**57.3**	**75.2**	**45.5**	**52.9**

4.1. Results on ShapeNet

We first conduct experiments with the synthetic dataset provided by Kar *et al*. [21], which contains 43,784 objects in 13 classes from ShapeNet [1]. Each sample includes a 3D CAD model, 20 camera viewpoints for rendering, and the corresponding rendered images at a resolution of 224×224 pixels. We use the same training/validation/testing splits as the original dataset. The training images are augmented by random shuffling the RGB channels and horizontal flipping. We only use one view per object for training and evaluate on all the 20 views in the test set (with independent single-view reconstruction on each view). The ground-truth 3D shapes are only used for testing. We show some qualitative results of our method and other baselines in Figure 4. More part reconstruction results on car, lamp, rifle, table, and watercraft samples are shown in Figure 5 to demonstrate that our method generalizes well across diverse object classes.

Ablations on the Proposed Models. We perform ablation studies on the proposed method by removing each component at a time. As shown in Table 1, removing part prior learning causes a significant drop (3.9%) on the overall reconstruction accuracy. It shows that the part prior provided by Part-VAE effectively improves the generalization of the discovered parts. Without the use of adversarial learning and color reconstruction, we also observe lower voxel IoU by a clear margin (0.6-1.1%).

Comparisons with the State-of-the-arts. Table 2 shows performance comparisons against the state-of-the-art methods. With the single-view training setting, our method performs favorably against the existing approaches. When compared to the 3D-supervised methods, our model achieves competitive results on many object classes even though we use weaker single-view supervision. Among the evaluated methods, SIF [12] and CvxNet [5] can subdivide an object into fine-grained regions, and HPD [41] performs hierarchical part reasoning. However, their shape representations require stronger supervision from 3D ground truths to produce faithful part shapes. The qualitative results in Figure 4 demonstrate that our model discovers more faithful and consistent parts compared to the baseline methods.

Ablations on Part Representations. We further compare 3D reconstruction methods that use different part shape constraints. A majority of existing methods represent objects with a single mesh [23, 50, 14, 22, 34, 18] and allow each shape to be fully deformable by predicting the vertex deformations directly. On the other end of the spectrum, most part-reasoning approaches represent parts with primitive shapes like cuboids [48] or super-quadratic surfaces [41]. Our method lies within these two extremes and enables variable degree of freedom by adjusting the latent dimension of the part shape embedding. Table 3 shows the quantitative comparisons between part representations like free-form meshes, cuboid reconstructions, and our Part-VAE embedding via different evaluation metrics. In addition to 3D voxel IoU, we calculate the 2D re-projection IoU, 2D structural similarity (SSIM), and Chamfer distance (CD) between the point sets sampled from 3D volumes. The

Table 3: **Quantitative evaluations on the ShapeNet dataset [1] using different metrics.** LPD allows part reasoning and achieves higher accuracy in terms of all metrics.

Method	Part	2D IoU ↑	SSIM ↑	CD ↓	Voxel IoU ↑
SoftRas [34]		80.5	86.7	4.76	46.9
VPL [22]		81.0	88.5	2.65	51.3
Free-form	✓	81.1	87.9	3.83	48.1
Cuboids	✓	72.5	67.3	6.12	39.7
LPD (ours)	✓	**83.6**	**91.0**	**2.37**	**52.9**

Table 4: **Voxel IoU results with different number of parts k.** Note that the models are trained and tested on a single class.

Method	k	Airplane	Car	Chair
SoftRas [34]	1	54.1	69.5	43.1
VPL [22]	1	54.6	74.1	45.3
LPD (ours)	1	54.5	74.3	45.0
LPD (ours)	2	55.4	**76.1**	45.9
LPD (ours)	3	55.6	75.5	**46.6**
LPD (ours)	6	**55.9**	75.2	46.4

results show that LPD achieves a better trade-off between the degree of deformation and shape regularization among the part-based methods. To observe how our method adapts to different object classes, we perform single-class training with different number of parts k on airplane, car, and chair images. As shown in Table 4, the optimal number of parts k varies across object classes. This suggests that each class have a distinct underlying part configuration to optimally represent the object shapes. Note that LPD achieves higher accuracy than other methods on all three classes with more than one part. Our main reconstruction model is class-agnostic and we use the same number of parts for all classes, and yet the performance could be further improved if it is optimized for each class separately.

4.2. Results on PartNet

To evaluate the quality of the discovered parts, we compare our results with the labeled parts in PartNet dataset. The dataset contains hierarchical part annotations of several ShapeNet models. We collect 111 chair samples that are in both the ShapeNet and PartNet testing sets, then combine the annotated part models into chair back, seat, and base at the coarsest level. Note that we do not use any 3D part supervision for training, so the PartNet annotations are not ground-truths but a reference. Since the discovered parts are not semantically labeled, we manually associate our parts to the closest corresponding PartNet annotations. We report the quantitative voxel IoU in Table 5 and the qualitative results in Figure 6. Compared to the baseline without part prior and other representations, our method discovers more faithful parts and achieves considerably higher IoU with respect to PartNet annotations.

Table 5: **Voxel IoU results on the PartNet [39] chair samples.**

Method	Back	Seat	Base	Avg
Free-form	16.8	19.5	10.3	15.5
Cuboids	22.3	23.4	10.7	18.8
LPD (ours)	**30.4**	**46.0**	**16.2**	**30.9**

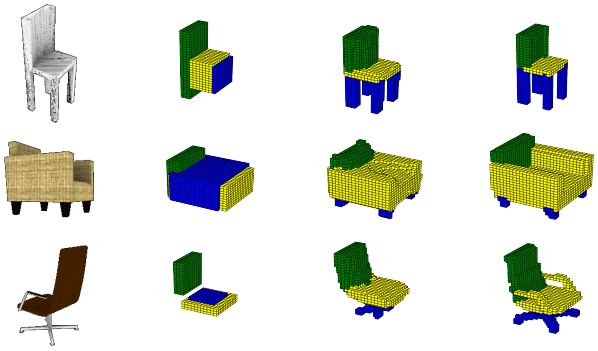

Figure 6: **Qualitative results on the PartNet dataset [39].** We show the voxelized 3-part results of our method and a cuboid baseline. Each part is specified with a color: chair back→green, seat→yellow, base→blue. Our method discovers faithful and consistent parts from diverse objects that are relatively closer to the pseudo-GT part annotations in PartNet.

4.3. Part Interpolation and Generation.

In addition to shape reconstruction, we demonstrate two applications of Part-VAE: part-interpolation and random shape generation. Since the object parts discovered by our models are consistent across instances, one can swap or interpolate parts to create new 3D objects. In Figure 7, we linearly interpolate the latent encodings (u_1, u_2) of two objects from different categories as: $u = \lambda u_1 + (1 - \lambda)u_2$. Compared to the baseline without part prior, our method deforms each object part smoothly and results in more realistic shapes. The Part-VAE can also be used as a generative model to create novel shapes. Specifically, we fit a Gaussian Mixture Model (GMM) with k components on the latent shape vectors of class-specific images. By sampling random vectors from individual GMM distribution, we can generate k random parts and combine them into a new 3D shape. In Figure 8, we show some randomly generated shapes of chairs and airplanes using the Part-VAE trained on the ShapeNet dataset.

4.4. Results on Pascal 3D+

We also evaluate the proposed method on the real-world images from the Pascal 3D+ dataset [51] processed by Tulsiani *et al*. [49]. It consists of images in Pascal VOC [7], annotations of 3D models, silhouettes, and viewpoints in Pascal 3D+ [51], and additional images in ImageNet [45] with silhouettes and viewpoints automatically annotated by [30]. This dataset is more challenging due to the complicated ob-

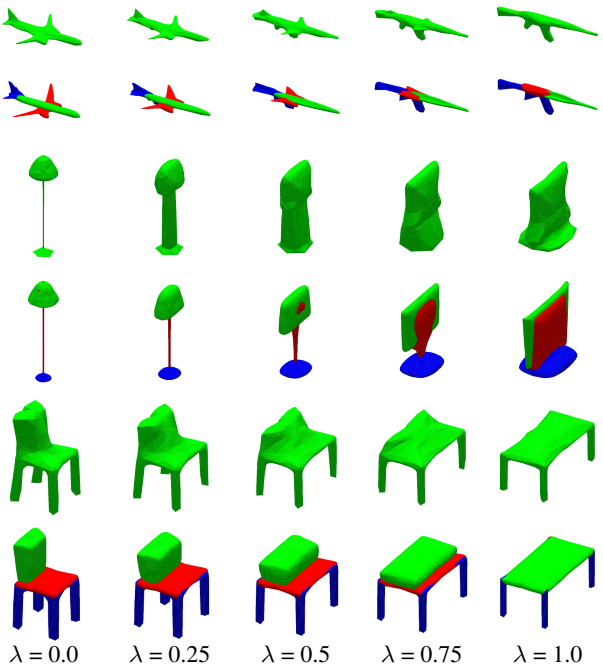

Figure 7: **Cross-category interpolation.** We perform interpolation on ShapeNet airplane-rifle, lamp-display, and chair-table. We show the VPL [22] results (mesh interpolation) in rows 1, 3, 5 and LPD results (latent interpolation) in rows 2, 4, 6.

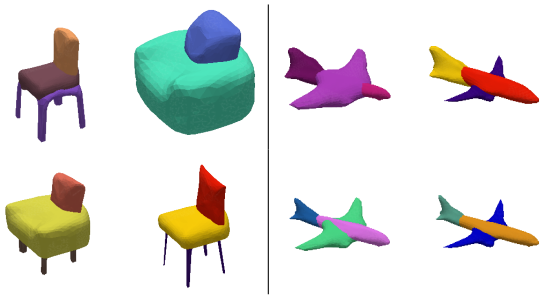

Figure 8: **Random shape generation of chairs and airplanes.** We fit a GMM model on the latent shape vectors and generate random parts by sampling from individual GMM components.

ject shapes, image background, occlusions, and noisy silhouette annotations. We train and evaluate our models with image resolution of 224×224. The quantitative and qualitative results are shown in Table 6 and Figure 9, respectively. Despite the challenges, our method discovers consistent object parts and achieves higher reconstruction accuracy than the state-of-the-art approaches.

5. Concluding Remarks

In this work, we propose LPD to discover 3D parts from single-view image collections. By learning a part prior with Part-VAE, we demonstrate that each part can be deformed to fit a realistic object shape while constrained to have simple

Table 6: **Voxel IoU results on the Pascal 3D+ dataset [51].**

Method	Part	Aeroplane	Car	Chair	Avg
SoftRas [34]		46.4	67.6	29.1	47.7
VPL [22]		47.5	67.9	30.4	48.6
Free-form	✓	47.0	68.5	28.7	48.0
Cuboids	✓	37.1	60.7	18.9	38.9
LPD (ours)	✓	**48.2**	**69.1**	**31.0**	**49.4**

Figure 9: **Part and color reconstruction results on the Pascal 3D+ dataset [51].** Despite that the dataset contains complicated 3D objects in a realistic scene, our method is able to discover consistent parts and effectively reconstruct the objects shapes.

geometry. With the goal to compose an object with simple parts, our reconstruction model automatically learns a latent part configuration. In turn, the discovered parts can alleviate shape ambiguity and improve the quality of full object reconstruction. Extensive experimental results show that LPD can discover faithful parts from diverse object classes, and the parts are consistent across different instances within a same category. Furthermore, we achieve the state-of-the-art reconstruction accuracy in a single-view training setting. Our work opens up the possibilities to learn, infer, and manipulate object parts without the need for any ground-truth part labels or 3D shape supervision.

Acknowledgement. This work is supported in part by the NSF CAREER Grant #1149783.

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[2] Wenzheng Chen, Jun Gao, Huan Ling, Edward J Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *arXiv preprint arXiv:1908.01210*, 2019.

[3] Zhiqin Chen, Kangxue Yin, Matthew Fisher, Siddhartha Chaudhuri, and Hao Zhang. Bae-net: Branched autoencoder for shape co-segmentation. In *ICCV*, pages 8490–8499, 2019.

[4] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, pages 628–644, 2016.

[5] Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, and Andrea Tagliasacchi. Cvxnet: Learnable convex decomposition. In *CVPR*, pages 31–44, 2020.

[6] Theo Deprelle, Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. Learning elementary structures for 3d shape generation and matching. *arXiv preprint arXiv:1908.04725*, 2019.

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[8] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, pages 605–613, 2017.

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

[10] Lin Gao, Jie Yang, Tong Wu, Yu-Jie Yuan, Hongbo Fu, Yu-Kun Lai, and Hao Zhang. Sdm-net: Deep generative network for structured deformable mesh. *TOG*, 38(6):1–15, 2019.

[11] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *CVPR*, pages 4857–4866, 2020.

[12] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *ICCV*, pages 7154–7164, 2019.

[13] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[14] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *ICCV*, pages 9785–9795, 2019.

[15] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *ECCV*, pages 88–104, 2020.

[16] Zekun Hao, Hadar Averbuch-Elor, Noah Snavely, and Serge Belongie. Dualsdf: Semantic shape manipulation using a two-level representation. In *CVPR*, pages 7631–7641, 2020.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] Paul Henderson, Vagia Tsiminaki, and Christoph H Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *CVPR*, pages 7498–7507, 2020.

[19] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *CVPR*, pages 869–878, 2019.

[20] Eldar Insafutdinov and Alexey Dosovitskiy. Unsupervised learning of shape and pose with differentiable point clouds. In *NeurIPS*, pages 2802–2812, 2018.

[21] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *NeurIPS*, pages 365–376, 2017.

[22] Hiroharu Kato and Tatsuya Harada. Learning view priors for single-view 3d reconstruction. In *CVPR*, pages 9778–9787, 2019.

[23] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, pages 3907–3916, 2018.

[24] Yuki Kawana, Yusuke Mukuta, and Tatsuya Harada. Neural star domain as primitive representation. *arXiv preprint arXiv:2010.11248*, 2020.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[28] K L Navaneet, Priyanka Mandikal, Varun Jampani, and Venkatesh Babu. Differ: Moving beyond 3d reconstruction with differentiable feature rendering. In *CVPR Workshops*, pages 18–24, 2019.

[29] Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, Nicu Sebe, et al. Motion-supervised co-part segmentation. *arXiv preprint arXiv:2004.03234*, 2020.

[30] Ke Li, Bharath Hariharan, and Jitendra Malik. Iterative instance segmentation. In *CVPR*, pages 3659–3667, 2016.

[31] Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. Synthesizing 3d shapes from silhouette image collections using multi-projection generative adversarial networks. In *CVPR*, pages 5535–5544, 2019.

[32] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *ECCV*, pages 677–693, 2020.

[33] Yichen Li, Kaichun Mo, Lin Shao, Minhyuk Sung, and Leonidas Guibas. Learning 3d part assembly from a single image. In *ECCV*, pages 664–682, 2020.

[34] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *CVPR*, pages 7708–7717, 2019.

[35] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37, 2016.

[36] Tiange Luo, Kaichun Mo, Zhiao Huang, Jiarui Xu, Siyu Hu, Liwei Wang, and Hao Su. Learning to group: A bottom-up framework for 3d part discovery in unseen categories. *arXiv preprint arXiv:2002.06478*, 2020.

[37] Priyanka Mandikal, Navaneet KL, and R Venkatesh Babu. 3d-psrnet: Part segmented 3d point cloud reconstruction from a single image. In *ECCV*, pages 0–0, 2018.

[38] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, pages 4460–4470, 2019.

[39] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *CVPR*, pages 909–918, 2019.

[40] KL Navaneet, Priyanka Mandikal, Mayank Agarwal, and R Venkatesh Babu. Capnet: Continuous approximation projection for 3d point cloud reconstruction using 2d supervision. In *AAAI*, pages 8819–8826, 2019.

[41] Despoina Paschalidou, Luc Van Gool, and Andreas Geiger. Learning unsupervised hierarchical part decomposition of 3d objects from a single rgb image. In *CVPR*, pages 1060–1070, 2020.

[42] Despoina Paschalidou, Ali Osman Ulusoy, and Andreas Geiger. Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In *CVPR*, pages 10344–10353, 2019.

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019.

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015.

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[47] Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Multi-view consistency as supervisory signal for learning shape and pose prediction. In *CVPR*, pages 2897–2905, 2018.

[48] Shubham Tulsiani, Hao Su, Leonidas J Guibas, Alexei A Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *CVPR*, pages 2635–2643, 2017.

[49] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, pages 2626–2634, 2017.

[50] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, pages 52–67, 2018.

[51] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, pages 75–82, 2014.

[52] Kohei Yamashita, Shohei Nobuhara, and Ko Nishino. 3d-gmnet: Learning to estimate 3d shape from a single image as a gaussian mixture. *arXiv preprint arXiv:1912.04663*, 2019.