Pacific Graphics 2020
E. Eisemann, A. Jacobson, and F.-L Zhang
(Guest Editors)

Volume 39 (2020), Number 7

Weakly Supervised Part-wise 3D Shape Reconstruction from Single-View RGB Images

Chengjie Niu*, Yang Yu*, Zhenwei Bian, Jun Li and Kai Xu†

National University of Defense Technology, China

Abstract

In order for the deep learning models to truly understand the 2D images for 3D geometry recovery, we argue that single-view reconstruction should be learned in a part-aware *and* weakly supervised *manner. Such models lead to more profound interpretation of 2D images in which part-based parsing and assembling are involved. To this end, we learn a deep neural network which takes a single-view RGB image as input, and outputs a 3D shape in parts represented by 3D point clouds with an array of 3D part generators. In particular, we devise two levels of generative adversarial network (GAN) to generate shapes with both correct part shape and reasonable overall structure. To enable a self-taught network training, we devise a differentiable projection module along with a self-projection loss measuring the error between the shape projection and the input image. The training data in our method is unpaired between the 2D images and the 3D shapes with part decomposition. Through qualitative and quantitative evaluations on public datasets, we show that our method achieves good performance in part-wise single-view reconstruction.*

CCS Concepts

• ***Computing methodologies*** → ***Reconstruction;*** *Shape representations; Point-based models;* • ***Computer systems organization*** → *Neural networks;*

1. Introduction

The fast advancement of deep learning has greatly boosted the performance of single-view reconstruction [CXG*16, GFRG16, KHM17,FSG17], which is otherwise extremely difficult as a highly ill-posed problem. Most existing approaches have so far performed 3D reconstruction in a *holistic* fashion. Through supervised learning with plenty of image-shape training pairs, the existing models essentially learn a highly nonlinear map from 2D images to 3D shapes. Despite the notable success made along this line of research, it is still arguable whether these models do learn how to interpret the 2D images for 3D reconstruction [TRR*19].

We advocate that single-view reconstruction, when being performed in a *part-aware* manner, might lead to more profound understanding of the 2D images since part parsing and part assembling are involved. The recently proposed Im2Struct [NLX18] is a typical example of part-based single-view reconstruction, where a recursive neural network is learned to convert an image into a hierarchy of part bounding boxes. This method, however, is still strongly supervised; it requires large amount of training pairs of images and box structures. To better approach the true object understanding in single-view reconstruction, we believe that part-wise reconstruction should be learned in an *weakly supervised* fashion.

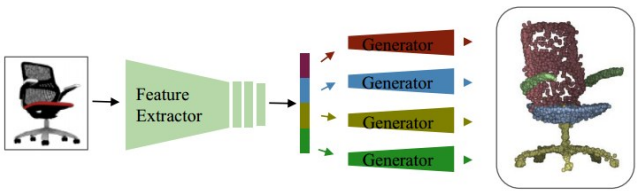

Figure 1: *Given a single-view image, our method is able to recover the corresponding 3D shape as an assembly of semantic parts (shown with distinct colors).*

In this work, we attempt to tackle the problem of weakly-supervised, part-wise 3D reconstruction from single-view RGB images. In particular, we learn a deep neural network which takes a single-view RGB image as input, and outputs a 3D shape in parts represented by 3D point clouds as showed in Figure 1. The network is composed of a 2D feature extractor and an array of 3D part generators. The part generators are trained with generative adversarial network (GAN) [GPAM*14] for 3D point clouds. We devise

* Chengjie Niu and Yang Yu are joint first authors.
† Kai Xu is the corresponding author (kevin.kai.xu@gmail.com)

two levels of discriminators to discriminate whether a part or the whole shape is real or fake. The combination of local and global GANs helps generate shapes with both correct part shape and reasonable overall structure. By splitting the part generation according to semantic part labels, the generated parts naturally come with semantic labels. To enhance the part-wise generation, we utilize a purity loss which minimizes the mutual overlap between the parts generated by different generators.

Furthermore, we realize a self-taught network training via devising a differentiable projection module [ID18] which is able to project the reconstructed 3D point cloud into a 2D image. This way, a self-projection loss can be measured between the projected shape and the input image. Since the projection module is differentiable, the loss can be backpropagated through it to optimize part-wise generators and then 2D feature extractor. This makes the entire network end-to-end trainable. The only supervision required is part decomposition of the training 3D shapes which, however, do not need to be paired with the input 2D images. This makes our learning weakly supervised.

Through extensive qualitative and quantitative evaluations on the ShapeNet [CFG*15] dataset, we show that our method achieves good performance in part-wise single-view reconstruction. We also demonstrate that our method supports shape interpolation and part-based crossover between two shapes reconstructed from images.

In summary, our work makes the following contributions:

- We propose a deep neural network for reconstruction of 3D shape in parts with semantic labels from single-view RGB images in a weakly supervised manner.
- We devise two levels of generative adversarial network(GAN) on point clouds to generate shape with both correct part shape and reasonable overall structure.
- We conduct extensive evaluation on the public dataset ShapeNet to demonstrate the effectiveness of our method.

The rest of this paper is organized as follows. Section 2 reviews the closely related literature. Section 3 provides an overview of the proposed architecture. Section 4 elaborates the details of our method. Section 5 presents the experiments, evaluation and application of our method. Finally, Section 6 concludes the work and discusses the future work.

2. Related Work

3D shape reconstruction. It is fundamental research to reconstruct 3D shape from its 2D projections or 2.5D information in computer vision or graphics. With the development of deep learning and 3D dataset [CFG*15, YLZ*19, MZC*19], there is an increasing growth in 3D shape reconstruction work. Below we briefly introduce the related methods based on deep learning.

The method in [WWX*17] generates shapes in two stages. They separately train the 2.5D sketch estimation and 3D shape estimation components firstly. Then the network is fined-tuned on real images. Zhang et al. [ZZZ*18] propose an algorithm to capture more generic and class-agnostic shape priors through the 2.5D representations of visible surfaces to generate 3D shapes by just giving 2D images. It is a smart and common idea to use one intermediate representation (e.g. 2.5D) to connect the 2D and 3D. Instead of using any intermediate representations, we prefer using the shape priors to help better understand the feature of 2D images. Niu et al. [NLX18] recursively generate 3D shapes based on relationships between parts. This method is consistent with our spirit, we reconstruct the 3D shape from the perspective of the part. Wu et al. [WZZ*18] implements volumetric convolutional networks with adversarially learned shape priors to make the generated shape realistic. In contrast, we used the part priors in point clouds to ensure the reconstructed model reasonable in semantic parts. Tulsiani et al. [TZEM17] study the consistency between multi-view observations using differentiable ray consistency to show single-view reconstruction. [ID18] presents an approach to reconstruct high-fidelity shapes and poses from single images. They apply a differentiable projection operator to get a 2D projection from the given point set and camera pose. Differentiable operation enables learning point clouds without explicit 3D supervision. Our method builds on this approach, however, we use 3D shapes unpaired with input images to train adversarial network to get priors of parts and overall shapes. With the recent development of deep implicit surface representation [MON*19, PFS*19, CZ19, CTZ20, LSCL19], several works study differentiable rendering on implicit surfaces to realize single-view reconstruction [XWC*19, WS20, NMOG20, YAL20].

Part-aware shape generation. To our knowledge, our work is the first to reconstruct 3D shape in point clouds with semantic labels from single images. Since [FKS*04] first proposes “Modeling by Example", which picks parts from a shape repository and then sticks them together. This idea has been employed to various tasks and has achieved outstanding performance [CKGK11, KCKK12, XZCOC12].

In the deep learning era, Li et al. [LXC*17] pioneered part-based shape generation based on a deep recursive neural network. Recently, [LNX19] uses a part-aware deep generative network to model 3D shape variations which is composed of an array of per-part VAE-GANs and use a part assembly module to correlate and assemble the parts into a plausible structure. Schor et al. [SKZCO19] treat a shape as a (re-)composable set of parts, so the key idea is to synthesize shapes by varying both the shape parts and their compositions. However, they both need to pre-train autoencoders for the shape parts, which is not an effective and efficient way in realistic usage. The authors of [ZCOAM14, HFWH15, GYW*19] utilize the spatial arrangement of parts to generate shape. Paschalidou et al. [PvGG20] propose a structure-aware representation goes beyond part-level geometry and focus on global relations by predicting a binary tree of primitives in an unsupervised manner. However, it cannot construct the tree in a flexible way causing the airplane body split as two parts without reasonable semantic information at the first partition. Chen et al. [CYF*19] propose a branched autoencoder to generate semantic labels using different branches.

Sung et al. [SSK*17] propose to learn a deep network for part selection and assembly. Zhu et al. [ZXC*18] develop a recursive neural network to assemble two substructures into a structurally coherent 3D shape. Li et al. [LMS*20] introduce a single-image-guided 3D part assembly problem and using a two-module pipeline to

Figure 2: *Reconstruction results of the proposed pipeline with known camera pose. RGB images are the input, the results are colored according to the part label: seat (olive), back (purple), armrests (blue) and legs (green).*

solve it by inferring part relationships. Different from our method, this method requires image and a set of part point clouds as inputs. Wu et al. [WZX*20] propose seq-to-seq learning for part-based reconstruction from single-view images.

Wang et al. [WSH*18] take a global-to-local (G2L) generative model to synthesize 3D man-made shapes in volumetric representation. They use global discriminator to distinguish between the real and the generated shapes, while the local discriminators are focusing on the individual local parts. Global-to-local generative models have achieved great success in 2D area [ISSI17, HZLH17]. We adopt this method in point cloud representation to ensure the generated local semantic parts and the whole shape are plausible.

3. Overview

3.1. Problem Statement

The goal of our work is to reconstruct 3D shape with semantic labels from single-view RGB images. The 3D shape is represented as a point set $S = \{P_l\}$ with $l = 1, ..., L$ and $P_l = \{(x_i, y_i, z_i)\}_{i=1}^{K}$. K is a predefined constant and l is the label of shape parts. We use K = 2000 by default which is sufficient to present the shape part with high resolution.

3.2. Approach

Weakly supervised learning to reconstruct 3D shapes from single-view RGB images is challenging, especially when semantic labels are required to be predicted together. To address this problem, we design a deep neural network which consists of three modules: feature extractor, part-wise shape generator and camera pose predictor. For each shape category with L semantic parts, the part-wise shape generator contains L local GANs and one global GAN to generate shape with both correct part structure and reasonable overall shape.

The network architecture of our model is shown in Figure 3. During the training stage, assuming two RGB images x_1 and x_2 are taken as input, which are captured from two different views towards the same object. In fact, x_1 can be the same one as x_2, we prefer using two different views to better describe this method. $\hat{f}_1$ and $\hat{f}_2$ are the features extracted from image x_1 and image x_2 respectively. $\hat{f}_1$ is split to L code segments and fed to the L generators to recover semantic shape parts $\hat{S}_1 = \{\hat{S}_l\}$. While $\hat{f}_2$ is passed to camera pose predictor to estimate the view point $\hat{Q}_2$. Inspired by the work [ID18], we empoly a differentiable projection P to render the assembled shapes $\hat{S}_1$ according to the view point $\hat{Q}_2$, and get a 2D silhouette $\hat{v} = P(\hat{S}_1, \hat{Q}_2)$. Then the difference between the silhouette images $\hat{v}$ and the ground truth is measured as the projection loss.

The part-wise generator is trained to learn how to generate reasonable shape part respectively. To train each shape part generator, the corresponding 3D shape parts are taken as real samples for each local GAN to form the adversarial loss. The global GAN aims to further improve the part connectivity and overall quality by exploiting the whole shape priors. It should be noticed that both the 3D complete shapes and shape parts are not required to be paired with the input image, which offers great flexibility for training data preparing. The purity loss, designed to measure the consistency of labels in a local region, is introduced to prevent the overlapping between different semantic parts. In reference time, the method can predict 3D shape with semantic labels given just one single-view image.

4. Method and Implementation Details

4.1. Network Architecture Details

Our network consists of three modules: feature extractor, part-wise shape generator and camera pose predictor. The 2D feature extractor takes RGB image with size $228 * 228 * 3$ as input and contains 7 layers totally. Each of the first 4 layers contains one convolutional layer with a stride of 2, one batch normalization layer and followed by a LeakyReLU activation function. Then two fully connected layers are added to extract the global feature of image as a 1024 vector. After this, the feature extractor is split into two branches for shape generation and pose prediction respectively. The first branch uses one MLP to transfer the 1024 latent vector into a $256 * L$ vector, where L is the number of semantic labels of the object category. The other branch outputs 1024 vector as camera pose feature instead of $256 * L$ vector.

The part-wise shape generator contains L generators for different semantic parts individually (e.g., back, seat, leg and armrest for chair), which consists of three fully connected layers, each is followed by one LeakyReLU except the last layer by Tanh. The last

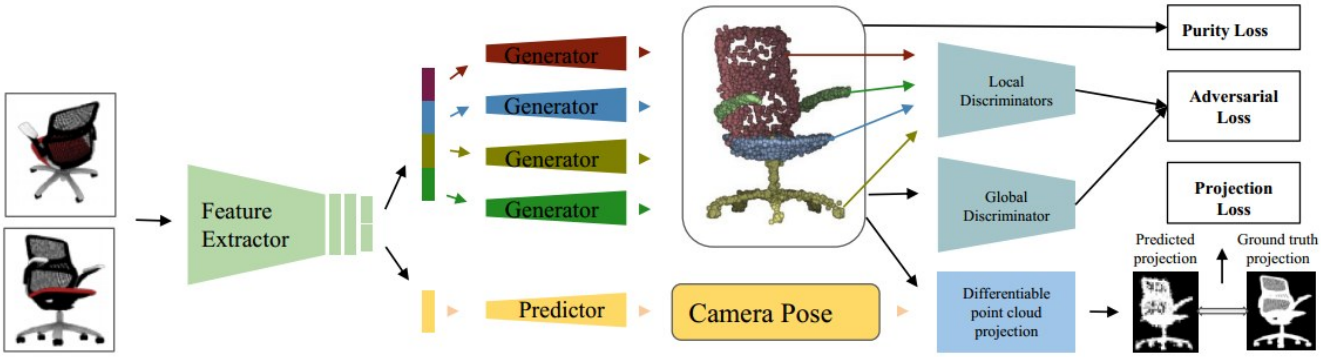

Figure 3: *An overview of our network architecture. Given two views of the same object, we predict the corresponding shape with semantic information and the camera pose. Then we use L part discriminators and one global discriminator to ensure the shape and part are reasonable. At the same time, we use a differentiable projection to generate the view of the predicted shape from the predicted camera pose. The purity loss can ensure the semantic part clean.*

fully connected layer outputs $N*3$ points, where $N = 2000$ by default is the number of points of each semantic part. Finally, the L generated shape parts are assembled in orders to form a reasonable shape with semantic labels. The assembly operation here we used is just concatenate the outputs of part generators for simplification. Furthermore, we use global GAN to adjust the global consistency.

The camera pose predictor contains three fully connected layers and outputs a quaternion to present the camera pose. For local and global discriminators, we choose to use PointNet [QSMG17] to extract the feature of reconstructed shape parts and the overall shape and output true or false to present its judgement. For simplicity, we do not try PointNet++ [QYSG17] or other more complex network to learn the feature of the point and make judgement, although it may give more accurate judgment in theory.

4.2. Differentiable Point Clouds

Given a predicted point cloud $S = \{p_i, s_i\}_i^N$, with a camera pose Q, the differentiable point clouds projection P can generate a view $v = P(S,Q)$. N is the number of points, s_i is the scale and and $p_i = (x_i, y_i, z_i)$ is the position of the point p.

Differentiable point clouds. Firstly, we random sample some points from S to speed up the projection process without compromising quality. Because there are few meaningful points of S obtained at the beginning of training, as the number of iterations in the training process increases, S is getting closer and closer to the ground truth, and the number of sampling points keeps increasing. Then we convert the camera pose Q to the corresponding projective transformation T_Q. Then we use $\hat{p}_i = T_Q * p_i$ to compute the position of points on the standard coordinate frame. After the transformation operation, it needs to be discretized into a 3D volume with resolution $D_1 \times D_2 \times D_3$ by using trilinear interpolation and be smoothed via 3D convolutions. Note that the third index corresponds to the projection axis and $[1, D_3]$ represents the range between the closest and furthest to the camera.

Nextly, A differentiable ray tracing formulation proposed by [TEM18] is applied after the scaling operation by multiply s_i . In order to prevent that the foreground points disturb the sign from occluded points during the projection process, the ray termination probabilities r can be calculated from the occupancies o by the formulation:

$$r_{k_1,k_2,k_3} = \begin{cases} o_{k_1,k_2,k_3} \prod_{u=1}^{k_3-1} (1 - o_{k_1,k_2,u}), & \text{if } k_3 \leq D_3 \\ \prod_{u=1}^{D_3} (1 - o_{k_1,k_2,u}), & \text{if } k_3 = D_3 + 1 \end{cases} \tag{1}$$

In the end, based on the ray termination probabilities we obtained, we can project the volume to 2D plane by the formulation:

$$v_{k_1,k_2} = \sum_{k_3=1}^{D_3+1} r_{k_1,k_2,k_3} y_{k_1,k_2,k_3}, \tag{2}$$

where y is a parameter being set manually, here we use $y_{k_1,k_2,k_3} = 1 - \delta_{k_3,D_3+1}$ to get the silhouette.

Projection loss. The predicted view can be synthesized by rendering the reconstructed point cloud according to the estimated camera pose: $\hat{v}_{i,j} = P(\hat{S}_i, \hat{Q}_j)$. Then we define the projection loss as the difference between the projection of the reconstructed shape and the ground truth:

$$L_{proj}(\hat{S}, \hat{Q}) = \frac{1}{N m_i} \sum_{i=1}^{N} \sum_{j_1,j_2=1}^{m_i} \left\| \hat{v}^i_{j_1,j_2} - v^i_{j_2} \right\|_2^2, \tag{3}$$

where N means the number of objects in one category and m_i means the total view numbers of the i-th object. The projection loss measures the difference between all pairs of the rendered image and the ground truth.

4.3. Local and Global GANs

We train local and global GANs based on the architecture of DCGAN proposed by [RMC15] to generate shape with both correct

part structure and reasonable overall shape. There are L local GANs focusing on the individual semantic parts recovery, where L denotes the number of semantic parts in one category. All the local GANs share the same architecture as well as the global GAN. The global GAN is aiming to learn the distribution of generated shapes and the real samples. The shape and part priors provided by the GANs ensure the semantic part and whole shape plausible and help to generate high quality shape in the form of point clouds. The adversarial loss consists of L local adversarial loss and one global adversarial loss.

Implementation details. For a single local GAN, it is composed of one generator and one discriminator. Each semantic part generator is the corresponding part generator of the local GAN. The input is a 256-vector and the outputs are K points to represent the semantic part of the shape. Each part generator consists of three fully connected layers followed by one LeakyReLU except the last layer was followed by Tanh. PointNet is the backbone of the discriminator to extract the feature of the point cloud and output the binary number to judge the predicted semantic part is real or fake by using LogSoftmax method to classify it. Intuitively, the L local GANs are independent mutually. So they are parallel to each other in the training time. As for the global GAN, the generator is an array of the L local generators and the whole shape which is assembled by all the semantic parts should be viewed as the output. The assembly unit can use and fine-tune the composition module of [SKZCO19] or simply concatenate the outputs of part generators. We use the method of concatenation as assembly unit for simplification and use global GAN to adjust the overall shape. The global discriminator has the same architecture as the local discriminators except that the first layer is adjusted to take $K * L$ points as input. We train the local GANs and global GAN simultaneously.

Adversarial loss. Adversarial loss is designed to use the whole shape and semantic segmented part priors. For each GAN, the generator and the discriminator are worked in competition, until the discriminator cannot distinguish the difference between the real and the fake. The adversarial loss for each GAN is set as [RMC15]. We use NLLLoss as a loss function to compare predicted label with the ground truth label in implementation. The adversarial loss in this paper between different GANs is weighted as follows:

$$L_{adv} = L_{global} + \sum_{i=1}^{L} w_i L_{local}^i, \tag{4}$$

where w_i is the weight of the i-th adversarial loss, we set $w_i = 1$ by default. L is the number of the local GANs.

4.4. Purity Loss

In order to obtain semantic segmented shape, it is important to make each part segmented distinctly. Purity loss is devised to prevent the mutual overlap between the parts generated by different generators. This part is motivated by the method proposed by [WSH*18]. Unlike they calculated the purity loss on 3D regular grid, we extend it on 3D point clouds for our part-wise shape reconstruction.

The 2D version of the meaning of purity loss is shown in Figure 4. For one point in point cloud, we calculate the sum of L_1 distance between its one-hot label and its n nearest neighbors. In Figure 4, for the red points (labeled as 0), we calculate the sum of L_1 distance between it and the other n (we set $n = 4$ by default) blue points (labeled as 1) as its purity loss.

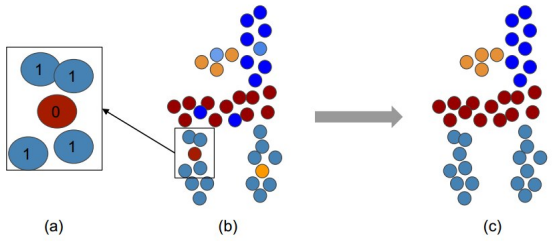

Figure 4: *A 2D illustration of the purity loss. (a) shows points with different parts are mixed. Red points should represent seat. However, one red point goes into the blue part which should be leg. The purified result is shown in (c).*

The purity loss is to compute the sum of Mean Squared Error between the one hot labels of point and its adjacent points. It can be summarized as follows:

$$L_{purity} = \sum_{i=1}^{K} \sum_{j=1}^{n} \left\| l_{adj}^{i,j} - l_i \right\|_1, \tag{5}$$

where l_i denotes the label of the i-th point, $l_{adj}^{i,j}$ defined as the label of the j-th adjacent point of the i-th point and K is the number the points of the semantic part, $K = 2000$ by default.

In summary, our generator loss in the architecture is defined as follows:

$$L_{gen} = w_{pro} L_{proj} + w_{adv} L_{adv} + w_{pur} L_{purity}. \tag{6}$$

Intuitively, the contribution of projection loss, adversarial loss and purity loss to the generator loss depends on w_{pro}, w_{adv} and w_{pur} respectively. we set $w_{pro} = 0.05$, $w_{adv} = 1$ and $w_{pur} = 0.05$ in our experiment.

5. Experimental Evaluation

5.1. Dataset

The dataset used to train our framework is divided into two types. The first type is 2D image-silhouette pairs of objects in different views. The pair dataset is a subset of ShapeNet 3D dataset [CFG*15], derived from the dataset in [ID18]. The image-silhouette pairs are rendered from ShapeNet models, and we just use 3 categories in it: chair, airplane and car to train our model and use the same train/test splits as [ID18]. The second is segmented semantic shape parts derived from the point cloud dataset with part annotations produced by [YKC*16]. These 3D shape parts are unpaired with the 2D image-silhouette dataset. We divide point clouds with part annotations into individual segmented parts with the same number points (Chair: back, seat, legs and armrests, Airplane: body, tail, wings and engine, Car: body, wheels, hood and roof).

Data augmentation For tasks based on part-wise generation,

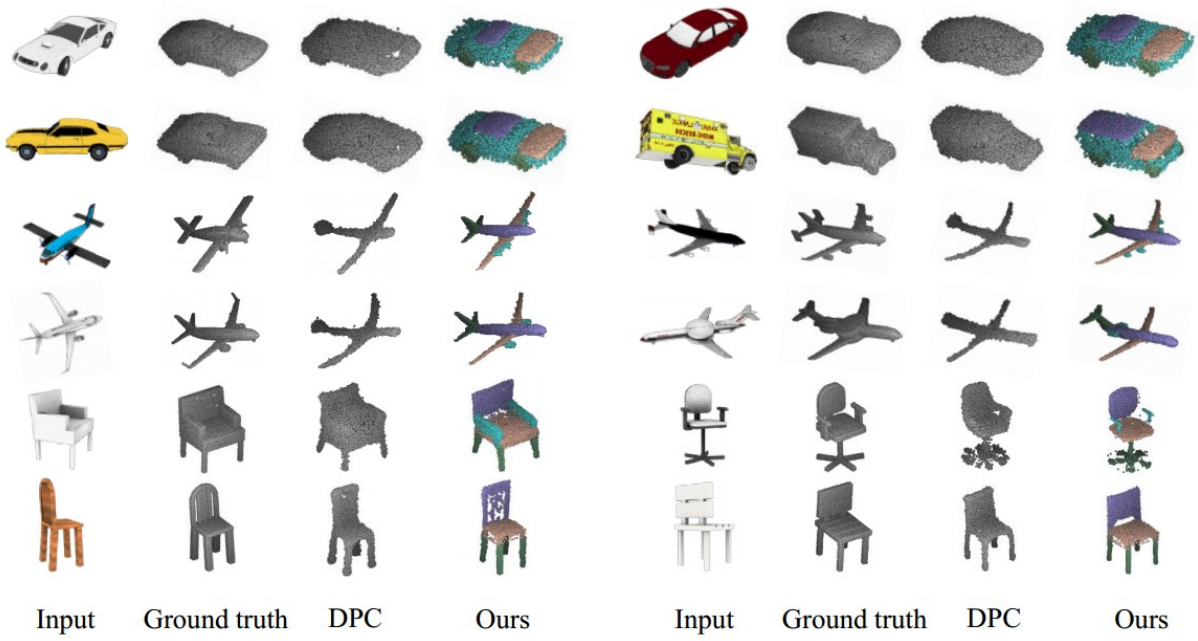

Figure 5: *Shape reconstruction from one single RGB input image. All the shapes are shown at the best view on screen. The ground truth shape and the shape predicted by DPC are without semantic labels. The shape we predicted is colored to intuitively present the segmented parts and more detailed structure is preserved by our method.*

there is an obvious problem to be dealt with. How to represent a missing part? We use zero to indicate the missing part. Therefore the local discriminators need to know the zero data is true for the part. In order to make the training successfully. we randomly put some zero data into the segmented semantic shape part dataset to represent the missing part. If one semantic part dose not exist, it is shown as coordinate origin zero.

5.2. Evaluation Metrics

We use the Chamfer Distance and the Earth Mover's Distance as our evaluation metrics for shape reconstruction. The Chamfer Distance is defined as Equation 7. For each point in one point set, the Chamfer Distance first finds the nearest point in the other set and get the average squared distance of them in the point set. For the other point set, the computation is as well. The Chamfer Distance requires both distances to measure the difference between two point sets. The Earth Mover's Distance is defined as Equation 8. It calculates the minimum distance of changing one point set into another point set by a bijective function.

$$d_{CD}(S_1,S_2) = \frac{1}{|S_1|}\sum_{p_1\in S_1}\min_{p_2\in S_2}||p_1-p_2||_2 + \frac{1}{|S_2|}\sum_{p_2\in S_2}\min_{p_1\in S_1}||p_2-p_1||_2 \tag{7}$$

	CD		EMD	
	DPC	Ours	DPC	Ours
Airplane	3.50	**2.72**	1.53	**1.26**
Car	2.98	**2.41**	**1.06**	1.34
Chair	**4.15**	4.57	1.31	**1.27**
Mean	3.55	**3.24**	1.30	**1.29**

Table 1: *Quantitative comparison in Chamfer Distance (CD) and Earth Mover's Distance (EMD). The smaller value is better in both evaluation measures. We present the Chamfer Distance and Earth Mover's Distance multiplied by 100.*

$$d_{EMD}(S_1,S_2) = \frac{1}{|S_1|}\min_{\phi:S_1\to S_2}\sum_{p_1\in S_1}||p_1-\phi(p_1)||_2, \tag{8}$$

where $\phi : S_1 \to S_2$ is a bijection.

5.3. Shape Prediction with Ground-truth Pose

We evaluate our network with known pose on test images both on qualitative and quantitative aspects. Figure 2 shows some reconstructed results. The input images are shown on the top row and the bottom line is the reconstructed 3D shapes. Unlike other 3D shape reconstruction on single image works, our reconstructed shapes with semantic labels to indicate different parts. We illustrate the semantic information with different colors.

We compare with DPC [ID18], which can reconstruct point

	CD		EMD	
	DPC	Ours	DPC	Ours
Airplane	3.91	**3.18**	2.21	**1.64**
Car	3.47	**2.42**	**1.28**	2.11
Chair	**4.30**	4.63	3.55	**3.14**
Mean	3.89	**3.41**	2.35	**2.30**

Table 2: *Quantitative comparison of generated shape without known camera pose in Chamfer Distance (CD) and Earth Mover's Distance (EMD).*

clouds from one single image and achieve state of the art results. However, DPC cannot generate semantic information and we render their results as the same gray as ground truth. For comparison, we use the same training data and the same resolution as DPC with known camera pose. Figure 5 provides a qualitative comparison. The results show our method achieves a more detailed structure and more accurate reconstruction than DPC. The position of engine, the diverse tails in airplane and the round wheels in car all demonstrate the structure details in shape parts, since our local GANs put much attention on the the part structure reconstruction. Shape semantic information also gives us a more intuitive display of different functions of parts. To demonstrate the quantitative results, we use Chamfer Distance and Earth Mover's Distance to express it in Table 1. Our results perform better than DPC in airplane categories both in Chamfer Distance and Earth Mover's Distance. We contribute it to that our model can transform all the predicted points into reasonable position. However, there can be some border points floating using DPC, since the airplane is nearly flat compare to other categories, it can hard congress the border points into the airplane body. For chair category, our method is a little worst than DPC in Chamfer Distance. That is because our model reconstructs the shapes and the parts plausibly, some rare parts or shapes cannot be predicted very real to the input image.

5.4. Shape Prediction without Known Pose

Our approach can generate shape from a single-view image, no matter without any ground truth shape or pose information. We drop the unrealistic setting of having the pose supervision and predict the shape without known camera pose from the input single view image. We compare with the results produced by [ID18] in the same setting of having no pose constrain. To better express the quality of the generated results, we report the shape in two different views on screen in Figure 6. From the global aspect, the results of our approach are much more realistic compared to DPC. That is attributed to our global GAN focusing on generating a plausible whole shape. In the local view, we observe that our semantic parts are segmented clearly in the shape. The rotation of wheels in the car is nearly the same as the ground truth and such details may miss in the results of DPC. The quantitative results are shown in Table 2. Our results are better than DPC in both metrics in general. During the weakly supervised training time, the GANs provide much support to generate plausible shapes.

5.5. Ablation Study

We conduct ablation studies to demonstrate the necessity of the various losses in our method. The projection part of our method is the critical connection between input image and shape reconstruction, therefore it is not removable.

(1) Removing local discriminators and purity loss. The purity loss is based on the part results. So the purity loss is meaningless without local adversarial loss. Local discriminators enable each part generator to explicitly observe its semantic feature. After removing the local discriminators, we feed the output of part generators to the global discriminator and then the ablated network cannot get semantic part feature.

(2) Removing purity loss. The unit learns to prevent the mutual overlap between the parts and make each part clean. By removing it, the results may be suboptimal.

(3) Removing global discriminator. This unit is used to make the result looks reasonably from global aspect. By comparison, we directly remove the global discriminator and therefore the overall consistency of the results may be somewhat lacking.

Figure 9 compares the results of all the ablated networks. *Firstly*, the greatest impact is caused by the removal of the local discriminators. The results are almost impossible to capture the features of the input image. What's more, each part generator outputs one global shape without any semantic information. Without local discriminators, the ablated network is adjusted by the projection loss and the global adversarial loss which means part generators cannot be trained to get semantic information. *Secondly*, the results of the removal of purity loss network show some floating points in space and overlap between different semantic parts demonstrating the necessary of the purity loss to control the reconstructed part clean. Finally, the results of removing global discriminator module are suboptimal on the global aspect. There are some disconnection between the adjacent parts without global discriminator.

5.6. Comparison on Single-view Reconstruction

We provide comparison on single-view reconstruction with both state-of-the-art methods and different view settings of our approach as shown in Figure 10. In general, all of the methods: 3D-R2N2 [CXG*16], OccNet [MON*19], DISN [XWC*19] can predict the shape with a single-view image as input. However, they cannot predict the semantic information of the shape. We have to admit that the representation of implicit surfaces can present better visual impression, like OccNet and DISN. In training time, we can use not only single-view setting but also multi-view for consistency constraints. Because of local and global GANs, our method uses the part and shape prior to limit the overall consistency of predicted shape to avoid that the generated shape only looks reasonable from the view of the input single image. The results of single-view training are comparable to those existing methods and have semantic labels of the shape. The results of multi-view training are slightly better than the results of single-view training because of the consistency constraints of different views.

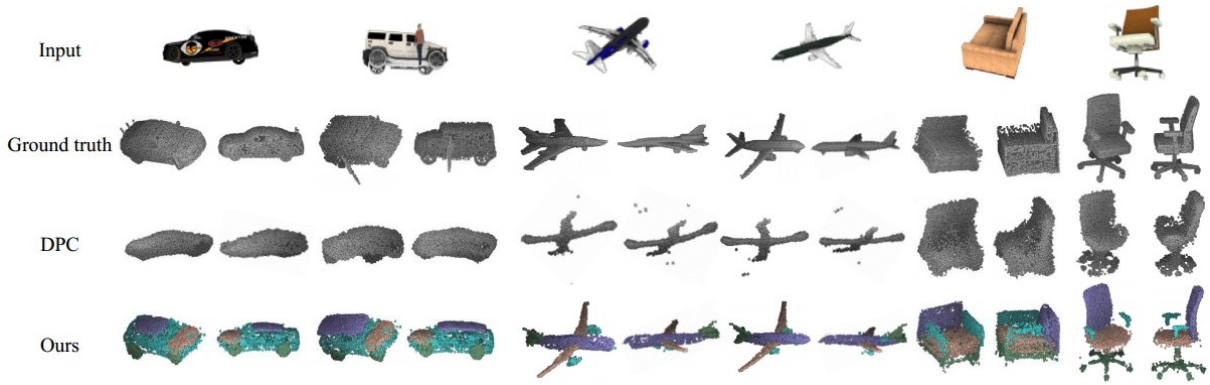

Figure 6: *Shape prediction without Ground-truth pose. The top row shows the input single image, the next three rows present Ground truth, DPC results and our results successively. The shapes are displayed in two different views for each input image.*

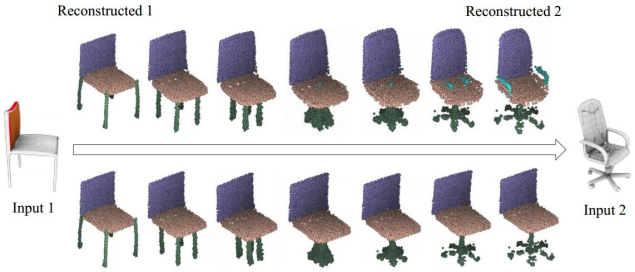

Figure 7: *Interpolating between one pair of chairs at the left and right. The full shape interpolation is shown on the top row, and the leg part interpolation is provided underneath.*

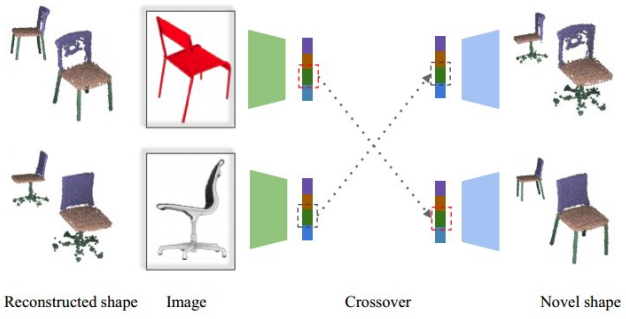

Figure 8: *Crossover between two latent space encoded from two different images. We use dashed boxes to indicate the leg latent vector in different colors. The red box corresponds to the red chair image and the gray one matches the gray swivel chair image.*

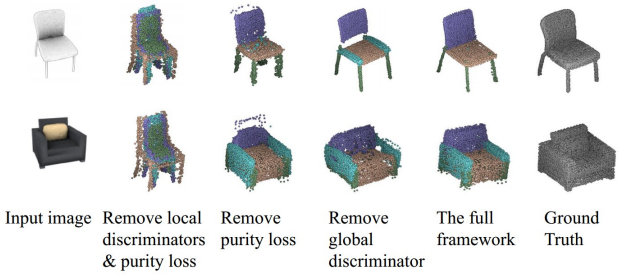

Figure 9: *Qualitative results of ablation study.*

Semantic Part	back	seat	leg	armrest
Correct Rate	71.76%	81.49%	51.78%	84.89%

Table 3: *The discriminator accuracy in discriminating the unseen data.*

5.7. Reconstruction Study

To demonstrate that our method does actually perform shape reconstruction instead of shape recognition, we make some deformation on the semantic part training dataset as input to the local discriminators. The deformation on the semantic part dataset has two types of chair. One type is the stretch in three dimensions to make it longer or shorter in different dimensions. Second type is the translation on different dimensions to make it higher or lower in different dimensions in space. The judgment of the unseen input data should be almost 0, if the ability that the discriminator learns is recognition. However, the correct rate is above fifty percent on each the semantic part as showed in Table 3. The correct rate on leg dataset is lower than other categories but also above fifty percent. The reason is that the complexity of the leg may cause the transformation on it to be incorrect. Obviously, from the experimental results, we can see that our method is actually to learn the characteristics of the object to be reconstructed, instead of memorizing the objects that have been seen.

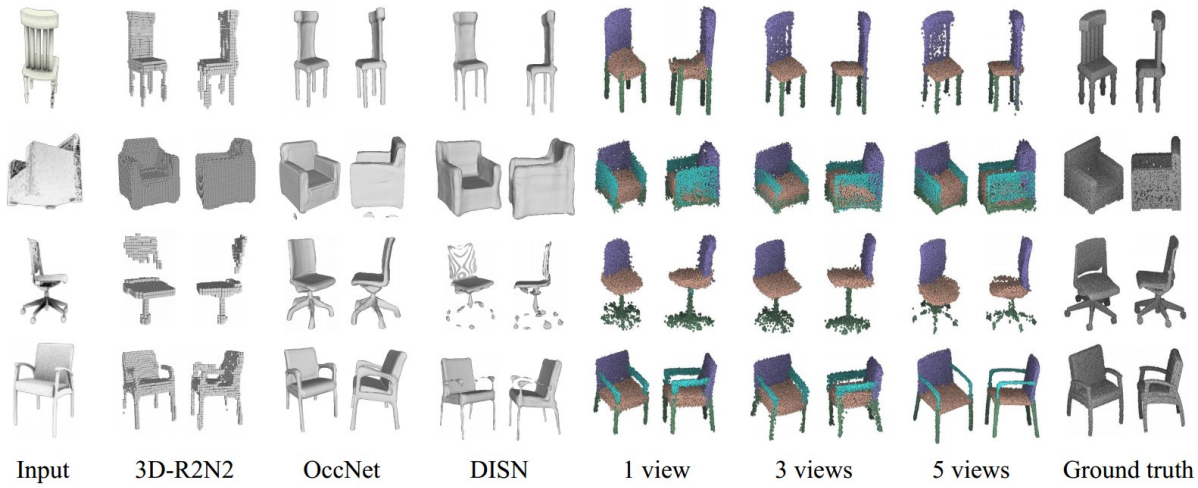

Figure 10: *Single-view reconstruction results of various methods. The fifth to seventh columns with semantic information are the results of our method. Different views means different view settings during the training time.*

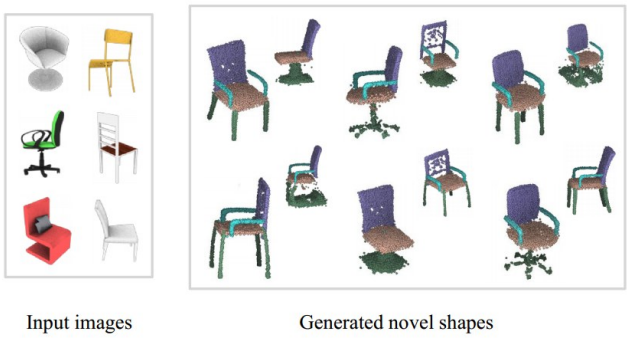

Figure 11: *Results of shape evolution. 12 chairs evolved from 6 input images using our approach.*

5.8. Latent Space Interpolation

Unlike other Shape-from-X works, our work can not only reconstruct the 3D shape from one single image, but also predict the semantic labels of each point. Our results has the same property of segmented shapes, which can be obtained by making operation on latent space vector, like the results produced by [DXA*19] and [WWL*18]. In our model, we encode the image feature into different part latent space, and each latent space can be decoded by the corresponding part generator. After that, we use one trained assembly module to make it consistency. Here we can make the latent space interpolation both on the whole shape and each semantic part of the shape. Figure 7 illustrates the shape and part interpolation results. We choose two generated shapes reconstructed from two different images individually, then make interpolation between the two latent space to express the deformation from one to another in global and local aspect.

5.9. Part-aware Shape DNA

The predicted shape depends on the input RGB image and the connecting point between shape and image is the latent space encoded from the input image. The latent vector can be regarded as a "shape DNA": each latent vector defines a part of the shape uniquely. Such shape DNA can be used to generate diverse shape world by crossover or mutation similar to [XZCOC12]. We show the crossover process in Figure 8. The novel shapes are like the results of two reconstructed shapes exchange their legs, but crossover makes it in the latent space. In Figure 11, it illustrates 12 novel chairs by making crossover or mutation of the latent space encoded from 6 input images.

6. Conclusion

As an weakly supervised single-view shape prediction work, our method not only achieves good performance on the public dataset ShapeNet, but also gives the shape correct part structure. Naturally, the predicted shape in parts with semantic labels can augment the segmented 3D shape dataset and the semantic information can be used in many areas, e.g., robots can take action by analyzing the function of each semantic part from the input images. However, there is one limitation in our method. Local GAN may cause the predicted part plausible but not realistic for some rare structure, although it can help reconstruct the shape from the part aspect and preserve more detailed structure. For the future work, we plan to study how to extend our model and use it to reconstruct implicit shape by using fewer parameters.

Acknowledgement

We thank the anonymous reviewers for their valuable comments. This work was supported in part by National Key Research and Development Program of China (2018AAA0102200), NSFC (61572507, 61532003, 61622212, 61902419).

References

[CFG*15] 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 preprint arXiv:1512.03012* (2015). 2, 5

[CKGK11] CHAUDHURI S., KALOGERAKIS E., GUIBAS L., KOLTUN V.: Probabilistic reasoning for assembly-based 3d modeling. In *ACM Transactions on Graphics (TOG)* (2011), vol. 30, ACM, p. 35. 2

[CTZ20] CHEN Z., TAGLIASACCHI A., ZHANG H.: Bsp-net: Generating compact meshes via binary space partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 45–54. 2

[CXG*16] 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 *European conference on computer vision* (2016), Springer, pp. 628–644. 1, 7

[CYF*19] CHEN Z., YIN K., FISHER M., CHAUDHURI S., ZHANG H.: Bae-net: Branched autoencoder for shape co-segmentation. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 8490–8499. 2

[CZ19] CHEN Z., ZHANG H.: Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 5939–5948. 2

[DXA*19] DUBROVINA A., XIA F., ACHLIOPTAS P., SHALAH M., GUIBAS L.: Composite shape modeling via latent space factorization. *arXiv preprint arXiv:1901.02968* (2019). 9

[FKS*04] FUNKHOUSER T., KAZHDAN M., SHILANE P., MIN P., KIEFER W., TAL A., RUSINKIEWICZ S., DOBKIN D.: Modeling by example. In *ACM transactions on graphics (TOG)* (2004), vol. 23, ACM, pp. 652–663. 2

[FSG17] FAN H., SU H., GUIBAS L. J.: 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* (2017), pp. 605–613. 1

[GFRG16] GIRDHAR R., FOUHEY D. F., RODRIGUEZ M., GUPTA A.: Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision* (2016), Springer, pp. 484–499. 1

[GPAM*14] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. In *Advances in neural information processing systems* (2014), pp. 2672–2680. 1

[GYW*19] GAO L., YANG J., WU T., YUAN Y.-J., FU H., LAI Y.-K., ZHANG H.: Sdm-net: Deep generative network for structured deformable mesh. *ACM Transactions on Graphics (TOG) 38*, 6 (2019), 1–15. 2

[HFWH15] HUANG S.-S., FU H., WEI L.-Y., HU S.-M.: Support substructures: Support-induced part-level structural representation. *IEEE transactions on visualization and computer graphics 22*, 8 (2015), 2024–2036. 2

[HZLH17] HUANG R., ZHANG S., LI T., HE R.: Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 2439–2448. 3

[ID18] INSAFUTDINOV E., DOSOVITSKIY A.: Unsupervised learning of shape and pose with differentiable point clouds. In *Advances in Neural Information Processing Systems* (2018), pp. 2802–2812. 2, 3, 5, 6, 7

[ISSI17] IIZUKA S., SIMO-SERRA E., ISHIKAWA H.: Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG) 36*, 4 (2017), 1–14. 3

[KCKK12] KALOGERAKIS E., CHAUDHURI S., KOLLER D., KOLTUN V.: A probabilistic model for component-based shape synthesis. *ACM Transactions on Graphics (TOG) 31*, 4 (2012), 55. 2

[KHM17] KAR A., HÄNE C., MALIK J.: Learning a multi-view stereo machine. In *Advances in neural information processing systems* (2017), pp. 365–376. 1

[LMS*20] LI Y., MO K., SHAO L., SUNG M., GUIBAS L.: Learning 3d part assembly from a single image. *arXiv preprint arXiv:2003.09754* (2020). 2

[LNX19] LI J., NIU C., XU K.: Learning part generation and assembly for structure-aware shape synthesis. *arXiv preprint arXiv:1906.06693* (2019). 2

[LSCL19] LIU S., SAITO S., CHEN W., LI H.: Learning to infer implicit surfaces without 3d supervision. In *Advances in Neural Information Processing Systems* (2019), pp. 8295–8306. 2

[LXC*17] LI J., XU K., CHAUDHURI S., YUMER E., ZHANG H., GUIBAS L.: Grass: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics (TOG) 36*, 4 (2017), 1–14. 2

[MON*19] MESCHEDER L., OECHSLE M., NIEMEYER M., NOWOZIN S., GEIGER A.: Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 4460–4470. 2, 7

[MZC*19] MO K., ZHU S., CHANG A. X., YI L., TRIPATHI S., GUIBAS L. J., SU H.: Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 909–918. 2

[NLX18] NIU C., LI J., XU K.: Im2struct: Recovering 3d shape structure from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 4521–4529. 1, 2

[NMOG20] NIEMEYER M., MESCHEDER L., OECHSLE M., GEIGER A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 3504–3515. 2

[PFS*19] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 165–174. 2

[PvGG20] PASCHALIDOU D., VAN GOOL L., GEIGER A.: Learning unsupervised hierarchical part decomposition of 3d objects from a single rgb image. *arXiv preprint arXiv:2004.01176* (2020). 2

[QSMG17] QI C. R., SU H., MO K., GUIBAS L. J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 652–660. 4

[QYSG17] QI C. R., YI L., SU H., GUIBAS L. J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems* (2017), pp. 5099–5108. 4

[RMC15] RADFORD A., METZ L., CHINTALA S.: Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015). 4, 5

[SKZCO19] SCHOR N., KATZIR O., ZHANG H., COHEN-OR D.: Componet: Learning to generate the unseen by part synthesis and composition. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 8759–8768. 2, 5

[SSK*17] SUNG M., SU H., KIM V. G., CHAUDHURI S., GUIBAS L.: Complementme: weakly-supervised component suggestions for 3d modeling. *ACM Transactions on Graphics (TOG) 36*, 6 (2017), 1–12. 2

[TEM18] TULSIANI S., EFROS A. A., MALIK J.: Multi-view consistency as supervisory signal for learning shape and pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 2897–2905. 4

[TRR*19] TATARCHENKO M., RICHTER S. R., RANFTL R., LI Z., KOLTUN V., BROX T.: What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 3405–3414. 1

[TZEM17] TULSIANI S., ZHOU T., EFROS A. A., MALIK J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 2626–2634. 2

[WS20] WU Y., SUN Z.: Dfr: Differentiable function rendering for learning 3d generation from images. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 241–252. 2

[WSH*18] WANG H., SCHOR N., HU R., HUANG H., COHEN-OR D., HUANG H.: Global-to-local generative model for 3d shapes. In *SIGGRAPH Asia 2018 Technical Papers* (2018), ACM, p. 214. 3, 5

[WWL*18] WU Z., WANG X., LIN D., LISCHINSKI D., COHEN-OR D., HUANG H.: Structure-aware generative network for 3d-shape modeling. *arXiv preprint arXiv:1808.03981* (2018). 9

[WWX*17] WU J., WANG Y., XUE T., SUN X., FREEMAN B., TENENBAUM J.: Marrnet: 3d shape reconstruction via 2.5 d sketches. In *Advances in neural information processing systems* (2017), pp. 540–550. 2

[WZX*20] WU R., ZHUANG Y., XU K., ZHANG H., CHEN B.: Pq-net: A generative part seq2seq network for 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 829–838. 3

[WZZ*18] WU J., ZHANG C., ZHANG X., ZHANG Z., FREEMAN W. T., TENENBAUM J. B.: Learning shape priors for single-view 3d completion and reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 646–662. 2

[XWC*19] XU Q., WANG W., CEYLAN D., MECH R., NEUMANN U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *Advances in Neural Information Processing Systems* (2019), pp. 492–502. 2, 7

[XZCOC12] XU K., ZHANG H., COHEN-OR D., CHEN B.: Fit and diverse: set evolution for inspiring 3d shape galleries. *ACM Transactions on Graphics (TOG) 31*, 4 (2012), 57. 2, 9

[YAL20] YARIV L., ATZMON M., LIPMAN Y.: Universal differentiable renderer for implicit neural representations. *arXiv preprint arXiv:2003.09852* (2020). 2

[YKC*16] YI L., KIM V. G., CEYLAN D., SHEN I., YAN M., SU H., LU C., HUANG Q., SHEFFER A., GUIBAS L., ET AL.: A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (TOG) 35*, 6 (2016), 210. 5

[YLZ*19] YU F., LIU K., ZHANG Y., ZHU C., XU K.: Partnet: A recursive part decomposition network for fine-grained and hierarchical shape segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 9491–9500. 2

[ZCOAM14] ZHENG Y., COHEN-OR D., AVERKIOU M., MITRA N. J.: Recurring part arrangements in shape collections. In *Computer Graphics Forum* (2014), vol. 33, Wiley Online Library, pp. 115–124. 2

[ZXC*18] ZHU C., XU K., CHAUDHURI S., YI R., ZHANG H.: Scores: Shape composition with recursive substructure priors. *ACM Transactions on Graphics (TOG) 37*, 6 (2018), 1–14. 2

[ZZZ*18] ZHANG X., ZHANG Z., ZHANG C., TENENBAUM J., FREEMAN B., WU J.: Learning to reconstruct shapes from unseen classes. In *Advances in Neural Information Processing Systems* (2018), pp. 2257–2268. 2