Learning Partonomic 3D Reconstruction from Image Collections

Xiaoqian Ruan[1], Pei Yu[2], Dian Jia[1], Hyeonjeong Park[1], Peixi Xiong[3], Wei Tang[1]
[1]University of Illinois Chicago, [2]Microsoft, [3]Intel
{xruan9, djia7, hpark233, tangw}@uic.edu, pei.yu@microsoft.com, peixi.xiong@intel.com

Abstract

Reconstructing the 3D shape of an object from a single-view image is a fundamental task in computer vision. Recent advances in differentiable rendering have enabled 3D reconstruction from image collections using only 2D annotations. However, these methods mainly focus on whole-object reconstruction and overlook object partonomy, which is essential for intelligent agents interacting with physical environments. This paper aims at learning partonomic 3D reconstruction from collections of images with only 2D annotations. Our goal is not only to reconstruct the shape of an object from a single-view image but also to decompose the shape into meaningful semantic parts. To handle the expanded solution space and frequent part occlusions in single-view images, we introduce a novel approach that represents, parses, and learns the structural compositionality of 3D objects. This approach comprises: (1) a compact and expressive compositional representation of object geometry, achieved through disentangled modeling of large shape variations, constituent parts, and detailed part deformations as multi-granularity neural fields; (2) a part transformer that recovers precise partonomic geometry and handles occlusions, through effective part-to-pixel grounding and part-to-part relational modeling; and (3) a 2D-supervised learning method that jointly learns the compositional representation and part transformer, by bridging object shape and parts, image synthesis, and differentiable rendering. Extensive experiments on ShapeNetPart, PartNet, and CUB-200-2011 demonstrate the effectiveness of our approach on both overall and partonomic reconstruction. Code, models, and data are avaliable at https://github.com/XiaoqianRuan1/Partonomic_Reconstruction.

1. Introduction

Reconstructing the 3D shape of an object (*e.g.*, a triangle mesh) from a single-view image is a fundamental task in computer vision with a wide range of applications, including virtual and augmented reality, robotics, and 3D printing. Deep neural networks have achieved excellent performance in single-view 3D reconstruction [12, 39, 54], but their training requires extensive 3D shape and/or pose annotations, which are costly and sometimes impossible to obtain. Thanks to advancement in differentiable rendering, a significant line of recent research [13, 15, 17, 20, 21, 26, 36] has shown promise in learning single-view 3D reconstruction from a collection of images with only 2D annotations, *e.g.*, keypoints [19] and masks [13], following an analysis-by-synthesis paradigm. However, these works focus on reconstructing entire objects and ignore their parts.

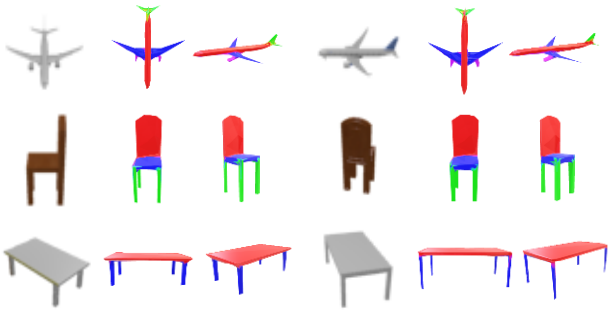

Figure 1. This paper aims at learning partonomic 3D reconstruction from an image collection with only 2D annotations. Our goal is not only to reconstruct the shape of an object from a single-view image but also to decompose the shape into meaningful semantic parts. The first and fourth columns are single-view input images. The other columns are partonomic 3D reconstruction results obtained by our proposed approach.

An object is composed of various parts, which is also known as *partonomy* [44]. Understanding the partonomy of 3D objects is crucial for intelligent agents interacting with the physical world, as it is often the parts that define an object's functionality and affordances. For example, considering an assistant robot serving at home or in an office, it needs to understand that the top of a table can hold other objects when looking for a suitable place to deliver items. Similarly, to assist someone sitting down, a robot must understand that the seat of a chair is for sitting, while the backrest provides support for the user's back. Some 3D recon-

struction methods regress the configurations of predefined primitives, such as cuboids [23, 25, 38, 49] or superquardics [1, 28, 41], from an image to abstract object shapes, but they typically rely on 3D supervision for training. In addition, the geometry of shape and parts is largely restricted by the chosen primitives, resulting in limited expressiveness.

This paper investigates learning partonomic 3D reconstruction from a collection of images, as shown in Figure 1. It means to not only reconstruct the shape of an object from a single-view image but also decomposes the shape into semantic parts. For broader applicability, we assume the only available annotations are 2D object and part masks, and there is no access to any form of 3D annotations (shape, pose, multi-view images, depth, etc) during training.

One naive solution to partonomic 3D reconstruction is to simply extend the self-supervised object reconstruction methods [13, 17, 20, 26, 36] by modeling the part class of each mesh vertex, and comparing rendered and ground truth part masks for learning. However, this approach ignores the intrinsic *part-whole structure* of objects, and thus faces two significant challenges. On one hand, the variability of object shape is further complicated by the challenge of part decomposition, leading to an even larger solution space. On the other hand, the visual input is often deficient and ambiguous, *e.g.*, parts are frequently occluded in a single-view image. As a result, learning a direct mapping from the 2D input image to 3D partonomic reconstruction is extremely difficult, especially with only limited 2D supervision.

To address these challenges, we propose an approach that leverages the part-whole structure of objects, by representing, parsing, and learning the *structural compositionality of 3D objects*. It consists of three main components. *(1) A Compact and Expressive Compositional Representation.* We represent the object geometry as a hierarchical composition of a conditional shape template (CST) and multiple part deformation fields (PDFs), both of which are neural fields but defined at different granularities. The CST models large shape variations and constituent parts, while the PDFs capture the detailed deformation of each part based on finer observations. This decoupling makes the compositional representation both compact and expressive. *(2) Robust Parsing with Part Transformer.* Based on the compositional representation, we introduce a part transformer that reconstructs both the shape and parts of an object from a single-view image. Unlike existing transformers, it employs a set of learnable *part tokens* to gather pixel-level features relevant to each part from the image and model their interactions, thereby effectively recovering part details and handling occluded parts. *(3) 2D-Supervised Learning Partonomic 3D Reconstruction.* Our compositional representation and part transformer are learned end-to-end with only 2D mask supervision, by bridging object shape and parts, image synthesis, and differentiable rendering.

Altogether, our proposed approach is able to infer the partonomy of 3D objects from a single-view image, without costly 3D annotations. Moreover, it is robust to high-dimensionality of object-part geometry and ambiguity of visual input. Our contributions are summarized as follow:

- We investigate partonomic 3D reconstruction from image collections, an important yet largely underexplored task. An effective solution is proposed for this task, by representing, parsing, and learning the structural compositionality of 3D objects.
- We introduce a compact and expressive compositional representation of object geometry, achieved through disentangled modeling of large shape variations, constituent parts, and detailed part deformations as multi-granularity neural fields.
- We propose a part transformer that recovers precise partonomic geometry and handles occlusion, via effective part-to-pixel grounding and part-to-part relation modeling.
- We develop a 2D-supervised learning method that learns our compositional representation and part transformer end-to-end, by bridging object shape and parts, image synthesis, and differentiable rendering.
- Extensive experiments on ShapeNetPart [11], PartNet [35], and CUB-200-2011 [52] demonstrate the effectiveness of our approach.

2. Related Work

Single-view 3D Shape Reconstruction. Reconstructing 3D shape from a single-view RGB image is an ill-posed problem due to the lack of explicit depth information. Previous 3D shape reconstruction methods rely on shape annotations [7, 12, 20, 24, 34, 46, 53–56, 58], which are very costly. Recent single-view 3D shape reconstruction methods address this problem by applying additional supervisions, such as the multi-view images [27, 49, 50, 57], the silhouettes [15, 17, 20], the camera viewpoints [10] and keypoints [19]. CMR [19] estimates three 3D attributes, namely the camera pose, the shape, and the texture. It then minimizes the distance between the rendered masks, images, and keypoints and their corresponding ground truth. U-CMR [13] follows a similar pipeline as CMR and reconstructs 3D shapes by removing the keypoints annotations. Unicorn [36] is the first fully unsupervised 3D shape reconstruction method by predicting four attributes, including the shape, the texture, the pose and the background. AST [17] improves shape reconstruction and texture generation through two transformer arhictectures. However, these methods focus on reconstructing the entire objects and ignore objects' part-whole structures. In contrast, this paper aims to learn partonomic 3D reconstruction by modeling the structural compositionality of 3D objects.

Shape Abstraction. Shape abstraction aims to decompose

3D objects into primitives, such as cuboids [23, 23, 25, 38, 49], superquadrics [1, 28, 41, 42], convexes [8], CSGtree [62, 63] and shape programs [18]. Because of the simplification and limited expressiveness of primitives, recent methods learn effective neural mapping to combine primitives. Neural Parts [43] implements this mapping by an Invertible Neural Network (INN). Other methods apply implicit primitives to boost the capacity [9, 16, 31]. ProGRIP [9] improves the representation capacity by using the implicit functions to represent the parts. An alternative line is part-based modeling, which segments the semantically meaningful parts [48]. Latent Partition Implicit (LPD) [6] represents the whole shape and the parts as Signed Distance Function (SDF) [40]. LPD [6] is able to partition a shape into different numbers of parts. However, this line of research typically relies on 3D shape and/or pose annotations for training.

Part Discovery. The object parts can be discovered by using volumetric cuboids, clustering 3D point clouds, and part priors [48]. Latent Part Discovery (LPD) [59] learns part priors with Part-VAE based on a collection of geometric primitives, reconstructs the parts, and composes the whole object based on these parts. The discovered parts may not be semantically meaningful. LASSIE [60] optimizes the articulated and part shapes based on a shared 3D skeleton and the input images. To enforce the semantic consistency between different instances, LASSIE [60] leverage the features extracted by DINO-ViT [4]. Following LASSIE [60], Hi-LASSIE [61] removes the skeleton templates and estimates the class-specific skeleton automatically. LEPARD [29] maps the input image to a set of primitive parameters and then uses them to reconstruct the 3D articulated shape. However, these works focus on learning articulated shape and motion rather than reconstructing semantic parts of common objects.

3. Method

We propose a novel compositional framework for learning partonomic 3D reconstruction from image collections. It means to not only reconstruct the shape of an object from a single-view image but also decompose shape into semantic parts, without any form of 3D or multi-view supervision. At the core to our framework are new mechanisms for representing (Sec. 3.1), parsing (Sec. 3.2), and learning (Sec. 3.3) the structural compositionality of 3D objects.

3.1. Compositional 3D Object Representation

The 3D geometry of an object instance observed in an image is determined by its pose and shape. The pose includes rotation $\boldsymbol{R} \in \mathtt{SO}(3)$ and translation $\boldsymbol{t} \in \mathbb{R}^3$ w.r.t. the canonical pose, *e.g.*, a chair that is horizontal and facing the camera. The shape is represented as a hierarchical composition of a

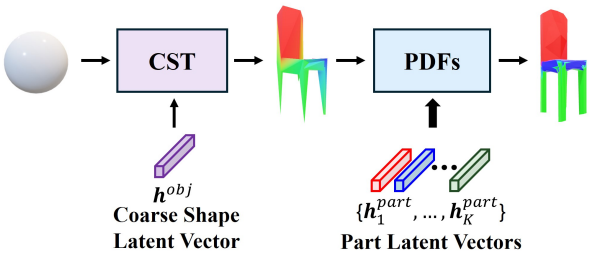

Figure 2. Illustration of the conditional shape template (CST) and part deformation fields (PDFs). The inference of latent vectors will be described in Sec. 3.2.

conditional shape template (CST) and multiple *part deformation fields* (PDFs), as illustrated in Figure 2. Both the CST and PDFs are neural fields defined on the object shape but at different granularities. The CST models large shape variations and constituent parts, while the PDFs model the detailed deformation of each part based on finer observations. This decoupling makes our compositional representation both compact and expressive.

Conditional Shape Template (CST). The CST models large shape variations across instances and constituent parts. Let $\{\boldsymbol{u}_i \in \mathbb{R}^3 : i = 1, \dots, M\}$ denote the vertices of a fixed sphere mesh, which is coarser than the target object mesh. The CST is formulated as:

$$(\bar{\boldsymbol{v}}_i, \bar{\boldsymbol{w}}_i) = \mathsf{MLP}(\boldsymbol{u}_i, \boldsymbol{h}^{\mathsf{obj}}; \boldsymbol{\Theta}^{\mathsf{cst}}), i = 1, \dots, M \quad (1)$$

where $\mathtt{MLP}$ is a multi-layer perceptron, $\boldsymbol{h}^{\mathsf{obj}}$ is a latent vector encoding the coarse shape (inferred in Sec. 3.2), M represents the number of vertices, and Θ^{cst} is parameters. Conditioned on $\boldsymbol{h}^{\mathsf{obj}}$, the CST maps the sphere mesh to the coarse object shape $\{\bar{\boldsymbol{v}}_i \in \mathbb{R}^3 : i = 1, \dots, M\}$ and its (soft) decomposition into K parts: $\{\bar{\boldsymbol{w}}_i \in [0, 1]^K : i = 1, \dots, M\}$.

Part Deformation Fields (PDFs). The PDFs model the detailed deformation of each part and obtain the fine object shape in two steps:

$$\{(\hat{\boldsymbol{v}}_i, \hat{\boldsymbol{w}}_i)\}_{i=1}^N = \mathsf{CSTUpsampling}(\{(\bar{\boldsymbol{v}}_i, \bar{\boldsymbol{w}}_i)\}_{i=1}^M) \quad (2)$$

$$\boldsymbol{v}_i = \hat{\boldsymbol{v}}_i + \sum_k \hat{w}_{i,k} \mathsf{MLP}(\hat{\boldsymbol{v}}_i, \boldsymbol{h}_k^{\mathsf{part}}; \Theta_k^{\mathsf{pdf}}), i = 1, .., N \quad (3)$$

Eq. (2) upsamples the CST via subdivision [32], where $N > M$. It adds a vertex to the midpoint of every edge and connects every pair of added vertices in the same triangle. The location and part assignment of a new vertex are respectively calculated by averaging those of the two original vertices on the corresponding edge, which is differentiable. Based on the upsampled CST, Eq. (3) deforms each part to obtain the final shape, where $\boldsymbol{h}_k^{\mathsf{part}}$ is a part latent vector encoding the kth part's deformation (inferred in Sec. 3.2), and $\boldsymbol{\Theta}_k^{\mathsf{pdf}}$ is parameters of the kth PDF. As each vertex is softly

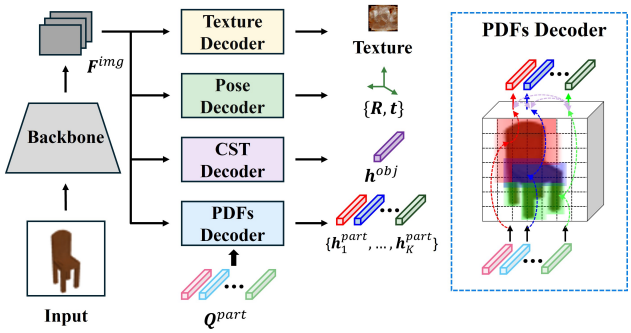

Figure 3. Illustration of the part transformer.

assigned to the parts, we model its deformation as a linear combination of the offsets predicted by all PDFs, weighted by $\hat{\boldsymbol{w}}_i = [\hat{w}_{i,k} : k = 1, \dots, K]$.

Discussion. As the CST is defined on a coarse shape and conditioned on the latent vector that encodes the whole object, it will focus on the shape's major characteristics rather than the subtle ones, and can easily *switch* between different shape subconfigurations that exhibit large variations, *e.g.*, sedan, SUV, and pickup truck. Afterward, it is intuitive for the PDFs to zoom in and deform each part based on finer observations, *i.e.*, part latent vectors. As a result, our compositional representation is both compact and expressive. On one hand, the combinatorial complexity of shape subconfigurations and deformations is disentangled through the CST and PDFs, which are much simpler to model. On the other hand, they are expressive enough to capture both large shape variations and detailed part deformations.

3.2. Part Transformer

Inferring the compositional 3D object representation from an image poses two main challenges. First, recovering the detailed deformation of a part necessitates grounding the part to pixel-level features relevant to it, but the part locations are unknown. Second, parts are frequently occluded in a single-view image. We propose a part transformer to address these challenges.

The input is an image and K learnable D-dimensional part tokens $\boldsymbol{Q}^{\text{part}} \in \mathbb{R}^{K \times D}$ (fixed after training). A convolutional backbone encodes the image into a feature map of height H and width W: $\boldsymbol{F}^{\text{img}} \in \mathbb{R}^{H \times W \times D}$. As illustrated in Figure 3, there are four decoders. We follow previous works [17, 36] to design the pose and texture decoders, which will be detailed in Sec. 4.3. Our CST decoder predicts the coarse shape latent vector $\boldsymbol{h}^{\text{obj}}$ through an MLP. Our PDFs decoder takes as input the feature map $\boldsymbol{F}^{\text{img}}$ and the part tokens $\boldsymbol{Q}^{\text{part}}$ to infer the part latent vectors $\boldsymbol{H}^{\text{part}} = [\boldsymbol{h}_k^{\text{part}} : k = 1, \dots, K]$. It includes cross-attention and self-attention layers:

$$\boldsymbol{Q}^{\text{part}} \leftarrow \phi(\boldsymbol{Q}^{\text{part}}, \boldsymbol{F}^{\text{img}} + \boldsymbol{E}^{\text{pos}}, \boldsymbol{F}^{\text{img}} + \boldsymbol{E}^{\text{pos}}; \boldsymbol{\Theta}^{\text{ca}}) \quad (4)$$

$$\boldsymbol{H}^{\text{part}} \leftarrow \phi(\boldsymbol{Q}^{\text{part}}, \boldsymbol{Q}^{\text{part}}, \boldsymbol{Q}^{\text{part}}; \boldsymbol{\Theta}^{\text{sa}}) \quad (5)$$

$$\begin{aligned} &\phi(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}; \boldsymbol{\Theta} = \{\boldsymbol{W}^{\text{query}}, \boldsymbol{W}^{\text{key}}, \boldsymbol{W}^{\text{value}}\}) \\ &\quad = \text{softmax}(\boldsymbol{Q}\boldsymbol{W}^{\text{query}}(\boldsymbol{K}\boldsymbol{W}^{\text{key}})^T / \sqrt{D})\boldsymbol{V}\boldsymbol{W}^{\text{value}} \end{aligned} \quad (6)$$

where $\boldsymbol{E}^{\text{pos}}$ is the pixel-wise positional embeddings, and $\boldsymbol{W}^{\text{query}}, \boldsymbol{W}^{\text{key}}, \boldsymbol{W}^{\text{value}} \in \mathbb{R}^{D \times D}$ are parameters of the scaled dot-product attention ϕ. Following the common practice in (vision) transformers [51], we use multi-head attention [51], skip connection [14], layer normalization [2], and dropout [47] within the decoder.

Discussion. The core idea of the attention mechanism is to calculate the *alignment* probabilities between each query in $\boldsymbol{Q}$ and a set of keys $\boldsymbol{K}$ and use them to *retrieve* relevant information from the values $\boldsymbol{V}$: a more relevant value will contribute more to updating the query. By taking part tokens as queries and image features as keys and values, Eq. (4) updates each part token by focusing on its relevant region in the image, which is crucial to recovering part deformation. Eq. (5) performs self-attention [51], where queries, keys, and values are from the same entity, *i.e.*, the updated part tokens, and models the relationship between different parts, which handles occluded parts. Based on the compositional representation in Sec. 3.1, the inferred latent vectors $\boldsymbol{h}^{\text{obj}}$ and $\{\boldsymbol{h}_1^{\text{part}}, \dots, \boldsymbol{h}_K^{\text{part}}\}$ respectively determine the CST and PDFs and thus the object shape and parts.

3.3. Learning from 2D Supervision

Our compositional representation and part transformer are learned end-to-end on a collection of images with only 2D object and part mask annotations. We use a differentiable renderer to render the estimated object pose, shape, parts, and texture into an image $\boldsymbol{I}'$, an object mask $\boldsymbol{M}'$, and a part mask $\boldsymbol{P}'$. Our learning objective comprises four terms:

$$\mathcal{L} = \mathcal{L}^{\text{rgb}} + \lambda^{\text{obj}}\mathcal{L}^{\text{obj}} + \lambda^{\text{part}}\mathcal{L}^{\text{part}} + \lambda^{\text{reg}}\mathcal{L}^{\text{reg}} \quad (7)$$

where λ denotes a balancing hyper-parameter.

The image synthesis loss $\mathcal{L}^{\text{rgb}}$ compares the rendered image $\boldsymbol{I}'$ and the input image $\boldsymbol{I}$:

$$\mathcal{L}^{\text{rgb}} = \|\boldsymbol{I}' - \boldsymbol{I}\|_2^2 \quad (8)$$

The object mask loss $\mathcal{L}^{\text{obj}}$ uses the intersection-over-union loss between the rendered object mask $\boldsymbol{M}'$ and the true object mask $\boldsymbol{M}$:

$$\mathcal{L}^{\text{obj}} = 1 - \frac{||\boldsymbol{M} \odot \boldsymbol{M}'||_1}{||\boldsymbol{M} + \boldsymbol{M}' - \boldsymbol{M} \odot \boldsymbol{M}'||_1} \quad (9)$$

The part mask loss $\mathcal{L}^{\text{part}}$ is a pixel-wise multi-class cross-entropy loss comparing the rendered part mask and the

ground truth part mask:

$$\mathcal{L}^{\text{part}} = -\sum_i \boldsymbol{p}_i \cdot \log \boldsymbol{p}'_i \tag{10}$$

where $\boldsymbol{p}_i$ and $\boldsymbol{p}'_i$ are the ground truth and rendered part class vectors at pixel i, and $\cdot$ denotes dot product.

The regularizer $\mathcal{L}^{\text{reg}}$ follows previous works [17, 36] and enforces smoothness and consistency priors on the reconstruction. The detailed formulas can be found in the supplementary material.

4. Experiments

We first introduce the datasets (Sec. 4.1), the evaluation metrics (Sec. 4.2), and the implementation details (Sec. 4.3). Sec. 4.4 shows the comparisons between our proposed method and the state-of-the-art methods. We visualize the partonomic reconstruction based on the real images in Sec. 4.5. Finally, we conduct ablation studies to show the influence of different modules and supervisions in Sec. 4.6. The details of dataset generation and more experimental results are included in the supplementary material.

4.1. Datasets

Among commonly used shape datasets, only two provide 3D part-level annotations: ShapeNetPart [11] and PartNet [35]. Therefore, we use these datasets for quantitative evaluation. Additionally, we present qualitative results on CUB-200-2011 [52], a real-world image dataset without any 3D annotations.

ShapeNetPart. The ShapeNetPart [11] comprises 16,881 3D shapes spanning 16 categories, with each instance annotated into 2 to 5 distinct part labels. In our experiments, we adopt five categories (airplane, car, chair, lamp, and table) that overlap with the rendered images provided by Soft-Ras [30]. Part masks are rendered under the same settings as Soft-Ras [30], with a resolution of 64 $\times$ 64. We adopt the official training and testing splits. Note that ShapeNet-Part [11] has much fewer samples than the commonly used ShapeNet [5], which is detailed in the supplementary material.

PartNet. The PartNet [35] contains 26,671 3D models across 24 categories, with fine-grained, instance-level, and hierarchical 3D part annotations. We focus on five common categories (bottle, bowl, display, knife, and mug) and the coarsest-level parts in our experiments. Part masks are rendered under the same setting as Soft-Ras [30]. We adopt the official training and testing splits.

CUB-200-2011. The CUB-200-2011 [52] is composed of 11,788 images of 200 sub-categories of birds. Each image has detailed annotations, including one class label, 15 part locations, 312 binary attributes, and one bounding box. In our experiments, we use the first 70 categories for training as their 2D part annotations are available.

4.2. Evaluation Metrics

We evaluate our proposed methods using four key metrics: **Chamfer-L_1 distance**, **Part Chamfer-L_1 distance**, **Part Classification Accuracy**, and **Part mIoU (mean Intersection-over-Union)**. The **Chamfer-L_1 distance** [33, 37] is used to assess the overall shape reconstruction. To evaluate part-level reconstruction, we introduce the **Part Chamfer-L_1 distance**, which computes the mean Chamfer-L_1 distance [33, 37] between the predicted shape and the ground truth based on different parts. This metric is defined as $\frac{1}{K}\sum_{k=1}^{K} d_k$, where d_k is the Chamfer-L_1 distance between the prediction and the ground truth of the kth part, and K is the total number of part classes for each object category based on part masks. Additionally, the **Part Classification Accuracy** is computed as the proportion of correctly classified vertices in the 3D space. The **Part mIoU** metric quantifies segmentation performance by averaging the IoU scores across all part classes. Following standard practice [17, 36], we pre-align the predicted shape with the ground truth using a gradient-based variant of the Iterative Closest Point (ICP) algorithm [3], incorporating anisotropic scaling.

4.3. Implementation Details

Since no prior work has addressed the same task as this paper, we construct two strong baselines by extending latest state-of-the-art methods [17, 36] that learn whole object reconstruction from image collections. Concretely, we extend Unicorn [36] and AST [17] by modeling the part class of each mesh vertex, and adding the part rendering loss for learning. We denote these two extensions as Unicorn* and AST*. All models are trained per object category.

The initial sphere mesh in our model consists of 162 vertices and 320 faces, which is subsequently upsampled to a higher resolution, containing 642 vertices and 1280 faces. The resolution of the input images and part mask is set as 64 $\times$ 64. Our backbone architecture is a U-Net [45] which contains 4 encoder layers and 4 decoder layers. The weights for object mask loss, part mask loss, and regularizer are 0.1, 0.1, and 1, respectively. Following the baselines [17, 36], we use a camera multiplex-based pose decoder to generate multiple hypotheses of the object pose and their probability distribution. Our texture decoder follows AST [17], which is based on a transformer architecture, and generates a texture map with 32 $\times$ 32 resolution, and upsampled to 64 $\times$ 64. Training is conducted with the Adam optimizer [22] using a learning rate of 1×10^{-4}.

Method	Chamfer-L_1 ↓						Part Chamfer-L_1 ↓					
	Avg	Airplane	Car	Chair	Lamp	Table	**Avg**	Airplane	Car	Chair	Lamp	Table
Unicorn*	0.249	0.099	0.157	0.243	0.499	0.247	0.446	0.227	0.355	0.425	0.898	0.325
AST*	0.217	0.090	0.151	**0.222**	0.393	0.229	0.459	0.293	0.332	0.417	0.777	0.477
Ours	**0.197**	**0.082**	**0.148**	0.227	**0.340**	**0.189**	**0.379**	**0.183**	**0.308**	**0.410**	**0.683**	**0.311**

Method	Part Classification Accuracy ↑						Part mIoU ↑					
	Avg	Airplane	Car	Chair	Lamp	Table	**Avg**	Airplane	Car	Chair	Lamp	Table
Unicorn*	0.595	0.642	0.263	0.745	0.599	0.727	0.458	0.464	0.161	0.557	0.462	0.646
AST*	0.610	0.602	0.332	0.756	0.610	0.752	0.446	0.410	0.215	0.534	0.523	0.550
Ours	**0.675**	**0.698**	**0.441**	**0.765**	**0.663**	**0.807**	**0.514**	**0.532**	**0.241**	**0.591**	**0.543**	**0.662**

Table 1. **Quantitative results on ShapeNetPart.** Unicorn [36] and AST [17] are state-of-the-art methods for learning whole object reconstruction from image collections. We extend them for partonomic reconstruction.

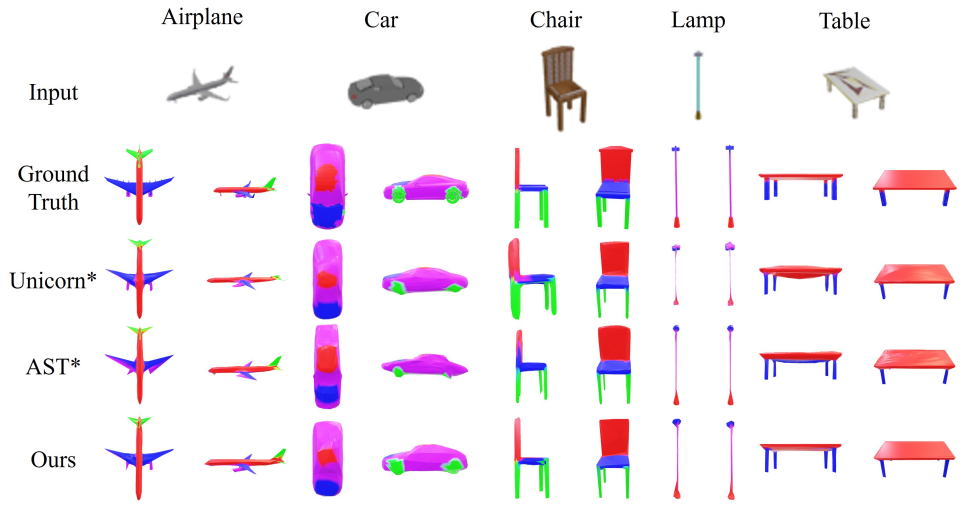

Figure 4. **Qualitative results on ShapeNetPart.** For each input, reconstruction results in two different viewpoints are shown. Unicorn [36] and AST [17] are state-of-the-art methods for learning whole object reconstruction from image collections. We extend them for partonomic reconstruction (Sec. 4.3). More qualitative comparisons can be found in the supplementary material.

4.4. Main Results

Evaluation on ShapeNetPart. Quantitative results are reported in Table 1. This table shows that our proposed method consistently outperforms both Unicorn* and AST* on almost all categories. The results indicate that our proposed method is particularly effective in handling complex shapes and preserving part boundaries, making it a robust choice for high-quality partonomic reconstruction. The lamp category demonstrates the strengths of our proposed method in handling objects with challenging shapes. By achieving the best performance in both overall and partonomic shape, our proposed method shows effectiveness in accurately preserving boundaries and differentiating parts within complex structures.

We visualize and compare the qualitative results of all three methods with five categories, shown in Figure 4. We can see that our proposed method generates better partonomic reconstruction with superior boundary preservation and accurate part distinction. Compared with Unicorn* and AST*, our proposed method consistently maintains clear part segmentation without color bleeding, particularly in challenging areas where parts are closely connected, such as the wings and body of airplane. This robustness in handling diverse object geometries and preserving fine details highlights the method's effectiveness for applications requiring high-quality part segmentation and reconstruction.

Evaluation on PartNet. Table 2 shows the quantitative comparison between our proposed method with the two state-of-the-arts approaches. The comparison suggests that our proposed method is capable to reconstruct better overall shapes along with the parts in most situations. Specifically, our proposed method achieves the lowest average Chamfer-L_1 (0.209) and Part Chamfer-L_1 (0.403), outperforming Unicorn* (0.223 for Chamfer-L_1 and 0.480 for Part Chamfer-L_1), and AST* (0.224 for Chamfer-L_1 and 0.414 for Part Chamfer-L_1), indicating superior accuracy in both whole shape reconstructions and part decompositions.

Method	Chamfer-L_1 ↓						Part Chamfer-L_1 ↓					
	Avg	Bottle	Bowl	Display	Knife	Mug	**Avg**	Bottle	Bowl	Display	Knife	Mug
Unicorn*	0.223	0.218	0.361	0.241	0.0721	0.222	0.480	0.340	0.625	0.688	0.206	0.541
AST*	0.224	0.215	0.359	0.236	0.0772	0.232	0.414	**0.275**	0.439	0.673	0.201	0.483
Ours	**0.209**	**0.205**	**0.336**	**0.228**	**0.0714**	**0.207**	**0.403**	0.291	**0.419**	**0.666**	**0.200**	**0.438**

Table 2. **Quantitative results on PartNet.** Unicorn [36] and AST [17] are state-of-the-art methods for learning whole object reconstruction from image collections. We extend them for partonomic reconstruction (Sec. 4.3).

Method	Chamfer-L_1 ↓						Part Chamfer-L_1 ↓					
	Avg	Airplane	Car	Chair	Lamp	Table	**Avg**	Airplane	Car	Chair	Lamp	Table
Base Model	0.239	0.084	0.148	0.243	0.502	0.216	0.667	0.238	0.414	0.465	1.844	0.376
+Deform	0.228	0.085	0.151	0.228	0.480	0.198	0.633	0.184	0.387	0.411	1.847	0.337
Ours	**0.197**	**0.082**	**0.148**	**0.227**	**0.340**	**0.189**	**0.379**	**0.183**	**0.308**	**0.410**	**0.683**	**0.311**

Table 3. **Quantitative ablation study on ShapeNetPart.** The base model uses the CST representation and predicts a global latent vector for partonomic reconstruction. +Deform means to extend the base model with the PDF representation but predict object and part latent vectors through MLPs.

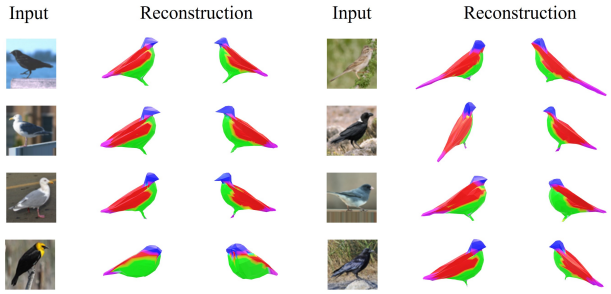

Figure 5. **Qualitative results on CUB-200-2011.** For each input image (seen species on the left and unseen species on the right), our 3D partonomic reconstruction results in two different viewpoints are shown. Our approach can reconstruct diverse bird species consistently with part decomposition, such as body (green), wings (red), head (blue), and tail (purple).

For the mug category, our proposed method achieves a Part Chamfer-L_1 of 0.438, compared to Unicorn* (0.541) and AST* (0.483), reflecting the superior part-level precision. Similarly, for the bowl category, our Chamfer-L_1 is 0.336, which is significantly better than Unicorn* (0.361) and AST* (0.359), showing a notable improvement in global shape accuracy.

4.5. Results on In-the-Wild Images

Figure 5 visualizes the qualitative results (both seen and unseen species) of our model trained on the CUB-200-2011 [52]. Each row corresponds to a sampled instance from the dataset, including both the input bird image and the partonomic reconstruction. These results show our proposed method is capable to reconstruct various bird species with consistent semantic parts, such as body (green), wing (red), head (blue), and tail (purple), across different poses and viewpoints. For example, the first instance (top-left) shows the good representation for the elongated body and slender beak.

We also visualize the sampled partonomic reconstruction from the unseen species (instances on the right in Figure 5). Despite from the unseen species, our proposed method is capable to generate bird-like shapes with coherent semantical segmentation and realistic poses. For example, the fifth instance (top-right) shows our method captures the details, such as the long wing and slightly-raised head.

4.6. Ablation Studies

Component analysis. To validate the effectiveness of our proposed compositional representation and part transformer, we conduct an ablation study by comparing three configurations. All the experiments are based on the same setting in Sec. 4.4. The base model uses the CST representation and predicts a global latent vector for partonomic reconstruction. +Deform means to extend the base model with PDF representations but predict object and part latent vectors through MLPs.

Table 3 and Table 4 show the quantitative ablation study on ShapeNetPart [11] and PartNet [35] respectively. Table 3 demonstrates our proposed method significantly outperforms both the base model and +Deform in terms of the overall and partonomic reconstruction quality, proving the necessity of both our compositional representation and the part transformer. Achieving an average overall Chamfer-L_1 (0.197) and Part Chamfer-L_1 (0.379), our proposed method consistently generates better performance across all categories, particularly in complex shapes, such as lamp.

Figure 6 visualizes the qualitative results of our proposed method compared to the base model and +Deform. Our proposed method generates the most accurate partonomic reconstruction with clear boundary, minimal color bleeding, and close alignment with the ground truth across all categories, especially for complex shapes like airplane and

Method	Chamfer-L_1 ↓						Part Chamfer-L_1 ↓					
	Avg	Bottle	Bowl	Display	Knife	Mug	**Avg**	Bottle	Bowl	Display	Knife	Mug
Base Model	0.241	0.226	0.370	0.311	0.0862	0.213	0.518	0.342	0.617	0.853	0.206	0.573
+Deform	0.229	0.219	0.342	0.291	0.0843	0.209	0.498	0.315	0.609	0.842	0.212	0.511
Ours	**0.209**	**0.205**	**0.336**	**0.228**	**0.0714**	**0.207**	**0.403**	**0.291**	**0.419**	**0.666**	**0.200**	**0.438**

Table 4. **Quantitative ablation study on PartNet.** The base model uses the CST representation and predicts a global latent vector for partonomic reconstruction. +Deform means to extend the base model with the PDF representation but predict object and part latent vectors through MLPs.

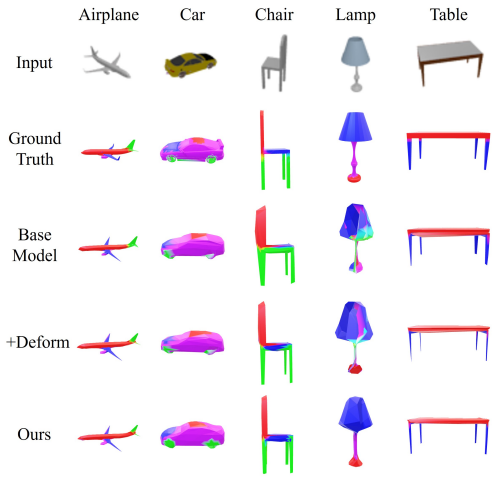

Figure 6. **Qualitative ablation study on ShapeNetPart.** The base model uses the CST representation and predicts a global latent vector for partonomic reconstruction. +Deform means to extend the base model with the PDF representation but predict object and part latent vectors through MLPs.

Mask	Method	**Avg**	Airplane	Car	Chair	Lamp	Table
None	Unicorn	0.267	0.099	0.172	0.279	0.527	0.257
	AST	0.242	0.099	0.167	0.268	0.438	0.238
	Ours	**0.219**	**0.089**	**0.161**	**0.247**	**0.389**	**0.211**
Object	Unicorn	0.258	0.1	0.168	0.255	0.515	0.252
	AST	0.230	0.095	0.162	0.249	0.415	0.230
	Ours	**0.208**	**0.087**	**0.151**	**0.240**	**0.362**	**0.199**
Object & Part	Unicorn	0.249	0.099	0.157	0.243	0.499	0.247
	AST	0.217	0.090	0.151	0.222	0.393	0.229
	Ours	**0.197**	**0.082**	**0.148**	**0.227**	**0.340**	**0.189**

Table 5. **Quantitative comparison using different mask supervisions on ShapeNetPart.** The Chamfer-L_1 performance is reported.

elongated structures like the lamp.

Different mask supervisions. Table 5 shows quantitative results using different mask supervisions on ShapeNetPart [11]. The 3D reconstruction accuracy of all methods decreases with reduced supervision. Our model consistently outperforms Unicorn [36] and AST [17].

Mesh subdivision. Following the Unicorn [36] and AST [17], all models in our experiments output a mesh with 642 vertices. When using PDFs, the initial sphere mesh has 162 vertices, which Eq. (2) subdivides into 642 vertices. Otherwise, the initial sphere mesh is directly set to 642 vertices. The ablation study in Table 6 shows that removing the subdivision in PDFs and using 642 input vertices degrade the Chamfer-L_1 performance from 0.197 to 0.204.

Method	**Avg**	Airplane	Car	Chair	Lamp	Table
w/o subdiv	0.204	0.082	0.163	0.234	0.349	0.192
Ours	**0.197**	**0.082**	**0.148**	**0.227**	**0.340**	**0.189**

Table 6. **Impact of mesh subdivision.** The Chamfer-L_1 performance is reported.

Regions	Method	**Avg**	Airplane	Car	Chair	Lamp	Table
Visible	Base	0.238	0.094	0.149	0.253	0.476	0.217
	Ours	**0.207**	**0.095**	**0.135**	**0.238**	**0.362**	**0.207**
Invisible	Base	0.246	0.091	0.163	0.258	0.502	0.217
	Ours	**0.210**	**0.090**	**0.140**	**0.245**	**0.361**	**0.214**

Table 7. **Quantitative ablation study on visible and invisible regions.** The Chamfer-L_1 performance is reported on ShapeNetPart.

Visible and invisible regions. We identify visible and invisible regions on the ground truth mesh for each image during the data generation process. Table 7 compares the reconstruction accuracies in visible and invisible regions achieved by our model with those of base model (without part deformable and part tokens).

5. Conclusion

This paper aims at learning partonomic 3D reconstruction from collections of images with only 2D annotations. Our goal is to not only reconstruct the shape of an object from a single-view image but also decompose it into meaningful semantic parts. This is an important yet largely underexplored task. To handle the expanded solution space and frequent part occlusions in single-view images, we introduce a novel approach that represents, parses, and learns the structural compositionality of 3D objects. Experimental results on ShapeNetPart [11], PartNet [35], and CUB-200-2011 [52], demonstrate the effectiveness of our method.

Acknowledgements. This work was supported in part by the National Science Foundation (NSF) grant ECCS-2400900, the National Artificial Intelligence Research Resource (NAIRR) Pilot, and Amazon Web Services (AWS) provided through CloudBank.

References

[1] Stephan Alaniz, Massimiliano Mancini, and Zeynep Akata. Iterative superquadric recomposition of 3d objects from multiple views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18013–18023, 2023. 2, 3

[2] Jimmy Lei Ba. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[3] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 5

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[5] 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. 5

[6] Chao Chen, Yu-Shen Liu, and Zhizhong Han. Latent partition implicit with surface codes for 3d representation. In *European Conference on Computer Vision*, pages 322–343. Springer, 2022. 3

[7] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6970–6981, 2020. 2

[8] Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, and Andrea Tagliasacchi. Cvxnet: Learnable convex decomposition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 31–44, 2020. 3

[9] Boyang Deng, Sumith Kulal, Zhengyang Dong, Congyue Deng, Yonglong Tian, and Jiajun Wu. Unsupervised learning of shape programs with repeatable implicit parts. *Advances in Neural Information Processing Systems*, 35:37837–37850, 2022. 3

[10] Shivam Duggal and Deepak Pathak. Topologically-aware deformation fields for single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1536–1546, 2022. 2

[11] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 2, 5, 7, 8

[12] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9785–9795, 2019. 1, 2

[13] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 88–104. Springer, 2020. 1, 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[15] Tao Hu, Liwei Wang, Xiaogang Xu, Shu Liu, and Jiaya Jia. Self-supervised 3d mesh reconstruction from single images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6002–6011, 2021. 1, 2

[16] Xiaoyang Huang, Yi Zhang, Kai Chen, Teng Li, Wenjun Zhang, and Bingbing Ni. Learning shape primitives via implicit convexity regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3642–3651, 2023. 3

[17] Dian Jia, Xiaoqian Ruan, Kun Xia, Zhiming Zou, Le Wang, and Wei Tang. Analysis-by-synthesis transformer for single-view 3d reconstruction. In *European Conference on Computer Vision*, pages 259–277, 2024. 1, 2, 4, 5, 6, 7, 8

[18] R Kenny Jones, Paul Guerrero, Niloy J Mitra, and Daniel Ritchie. Shapecoder: Discovering abstractions for visual programs from unstructured primitives. *ACM Transactions on Graphics (TOG)*, 42(4):1–17, 2023. 3

[19] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 1, 2

[20] Hiroharu Kato and Tatsuya Harada. Learning view priors for single-view 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9778–9787, 2019. 1, 2

[21] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. 1

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] Florian Kluger, Hanno Ackermann, Eric Brachmann, Michael Ying Yang, and Bodo Rosenhahn. Cuboids revisited: Learning robust 3d shape fitting to single rgb images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13070–13079, 2021. 2, 3

[24] Andrey Kurenkov, Jingwei Ji, Animesh Garg, Viraj Mehta, JunYoung Gwak, Christopher Choy, and Silvio Savarese. Deformnet: Free-form deformation network for 3d shape reconstruction from a single image. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 858–866. IEEE, 2018. 2

[25] Jun Li, Kai Xu, Siddhartha Chaudhuri, Ersin Yumer, Hao Zhang, and Leonidas Guibas. Grass: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 2, 3

[26] 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 *Computer Vision–ECCV 2020: 16th European Conference,*

Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, pages 677–693. Springer, 2020. 1, 2

[27] Chen-Hsuan Lin, Chaoyang Wang, and Simon Lucey. Sdf-srn: Learning signed distance 3d object reconstruction from static images. *Advances in Neural Information Processing Systems*, 33:11453–11464, 2020. 2

[28] Di Liu, Xiang Yu, Meng Ye, Qilong Zhangli, Zhuowei Li, Zhixing Zhang, and Dimitris N Metaxas. Deformer: Integrating transformers with deformable models for 3d shape abstraction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14236–14246, 2023. 2, 3

[29] Di Liu, Anastasis Stathopoulos, Qilong Zhangli, Yunhe Gao, and Dimitris Metaxas. Lepard: Learning explicit part discovery for 3d articulated shape reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[30] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 5

[31] Weixiao Liu, Yuwei Wu, Sipu Ruan, and Gregory S Chirikjian. Marching-primitives: Shape abstraction from signed distance function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8771–8780, 2023. 3

[32] Charles Loop. Smooth subdivision surfaces based on triangles. 1987. 3

[33] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 5

[34] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 306–315, 2022. 2

[35] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019. 2, 5, 7, 8

[36] Tom Monnier, Matthew Fisher, Alexei A Efros, and Mathieu Aubry. Share with thy neighbors: Single-view reconstruction by cross-instance consistency. In *European Conference on Computer Vision*, pages 285–303. Springer, 2022. 1, 2, 4, 5, 6, 7, 8

[37] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 5

[38] Chengjie Niu, Jun Li, and Kai Xu. Im2struct: Recovering 3d shape structure from a single rgb image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4521–4529, 2018. 2, 3

[39] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9964–9973, 2019. 1

[40] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 3

[41] Despoina Paschalidou, Ali Osman Ulusoy, and Andreas Geiger. Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10344–10353, 2019. 2, 3

[42] Despoina Paschalidou, Luc Van Gool, and Andreas Geiger. Learning unsupervised hierarchical part decomposition of 3d objects from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1060–1070, 2020. 3

[43] Despoina Paschalidou, Angelos Katharopoulos, Andreas Geiger, and Sanja Fidler. Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3204–3215, 2021. 3

[44] Simone Pribbenow. Meronymic relationships: From classical mereology to complex part-whole relations. In *The semantics of relationships: An interdisciplinary perspective*, pages 35–50. Springer, 2002. 1

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 5

[46] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008. 2

[47] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 4

[48] Shubham Tulsiani, Hao Su, Leonidas J Guibas, Alexei A Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2635–2643, 2017. 3

[49] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2634, 2017. 2, 3

[50] Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Multi-view consistency as supervisory signal for learning shape and pose prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2897–2905, 2018. 2

[51] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[52] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 5, 7, 8

[53] Bram Wallace and Bharath Hariharan. Few-shot generalization for single-image 3d reconstruction via priors. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3818–3827, 2019. 2

[54] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 1

[55] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. *Advances in neural information processing systems*, 30, 2017.

[56] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 646–662, 2018. 2

[57] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2690–2698, 2019. 2

[58] Zhen Xing, Hengduo Li, Zuxuan Wu, and Yu-Gang Jiang. Semi-supervised single-view 3d reconstruction via prototype shape priors. In *European Conference on Computer Vision*, pages 535–551. Springer, 2022. 2

[59] Chun-Han Yao, Wei-Chih Hung, Varun Jampani, and Ming-Hsuan Yang. Discovering 3d parts from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12981–12990, 2021. 3

[60] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. *Advances in Neural Information Processing Systems*, 35:15296–15308, 2022. 3

[61] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4853–4862, 2023. 3

[62] Fenggen Yu, Zhiqin Chen, Manyi Li, Aditya Sanghi, Hooman Shayani, Ali Mahdavi-Amiri, and Hao Zhang. Capri-net: Learning compact cad shapes with adaptive primitive assembly. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11768–11778, 2022. 3

[63] Fenggen Yu, Qimin Chen, Maham Tanveer, Ali Mahdavi Amiri, and Hao Zhang. D ̄2 csg: Unsupervised learning of compact csg trees with dual complements and dropouts. *Advances in Neural Information Processing Systems*, 36: 22807–22819, 2023. 3