DeFormer: Integrating Transformers with Deformable Models for 3D Shape Abstraction from a Single Image

Di Liu[1], Xiang Yu[2], Meng Ye[1], Qilong Zhangli[1], Zhuowei Li[1], Zhixing Zhang[1], Dimitris N. Metaxas[1]
[1]Rutgers University [2]Amazon Prime Video

Abstract

Accurate 3D shape abstraction from a single 2D image is a long-standing problem in computer vision and graphics. By leveraging a set of primitives to represent the target shape, recent methods have achieved promising results. However, these methods either use a relatively large number of primitives or lack geometric flexibility due to the limited expressibility of the primitives. In this paper, we propose a novel bi-channel Transformer architecture, integrated with parameterized deformable models, termed DeFormer, to simultaneously estimate the global and local deformations of primitives. In this way, DeFormer can abstract complex object shapes while using a small number of primitives which offer a broader geometry coverage and finer details. Then, we introduce a force-driven dynamic fitting and a cycle-consistent re-projection loss to optimize the primitive parameters. Extensive experiments on ShapeNet across various settings show that DeFormer achieves better reconstruction accuracy over the state-of-the-art, and visualizes with consistent semantic correspondences for improved interpretability.

1. Introduction

Accurate 3D shape abstraction with semantically meaningful parts is an active research field in computer vision for decades. It can be applied to many downstream tasks, such as shape reconstruction [47, 9, 53, 58, 68, 49, 50, 56, 43], object segmentation [31, 51, 37, 21, 73, 39, 41, 27, 38, 7, 71, 20, 42, 40, 16, 28, 45, 19], shape editing [69, 24] and re-targeting [15, 23, 70]. Due to the large success of deep neural networks (DNNs), a series of learning-based works [56, 54, 64, 55, 14] propose to decompose an object shape into primitives and use the deformed primitives to represent the target shape. The primitive-based methods usually interpret a shape as a union of simple parts (*e.g.*, cuboids, spheres, or superquadrics), offering interpretable abstraction of a shape target.

To achieve high accuracy of shape reconstruction, existing methods require joint optimization of a number of primitives, which sometimes do not accurately correspond to the object parts and therefore limit the interpretability of the reconstructions [56, 14, 55] (see Fig. 1). To this end, using a small number of primitives to abstract complex shapes becomes a trend in recent research [55]. However, the dilemma lies in that using fewer primitives usually results in sub-optimal reconstruction accuracy due to their reduced representation power, while using more primitives lowers the interpretability and requires higher computational costs.

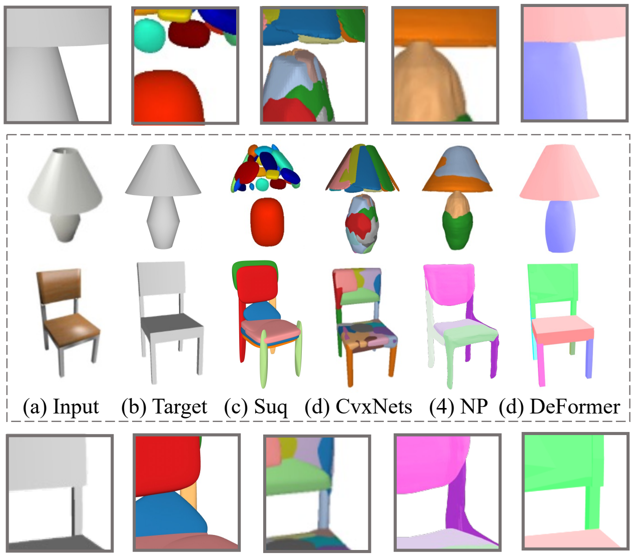

(a) Input (b) Target (c) Suq (d) CvxNets (4) NP (d) DeFormer

Figure 1: DeFormer uses a small number of primitives to abstract a 3D shape from a 2D image with better accuracy and part correspondence. Taking “lamp” and “chair” as examples, we compare to Suq [56], CvxNets [14], and Neural Parts (NP) [55] with ∼20, 25, and 5 primitives, respectively, while ours applies 2 primitives for lamps (1 for shade and 1 for body) and 6 primitives for chairs (4 for legs, 1 for seat and 1 for back).

Physics-based deformable models (DMs) [48, 62] are well known for their strong abstraction ability in shape representation, and have been successfully applied to various complex shape modeling applications. DMs leverage a physical modeling framework to predict global and local deformations of primitives, in which force-driven dynamic fitting across the data and the generalized latent space are used to jointly minimize the divergence between the deformed primitives and the target shapes. Although DMs can offer strong representation power for shape abstraction, a main concern is that they require handcrafted parametric initialization for each specific shape abstraction, which limits their usage to general and automated shape modeling.

To address the aforementioned limitations, we propose a bi-channel Transformer combined with deformable models, termed *DeFormer*, to leverage the superior interpretability from DMs and overcome the parametric initialization limitation by taking advantage of the universal approximation capabilities [29, 30] of deep neural networks. Moreover, we leverage general superquadric primitives with global deformations as our primitive formulation, which offer a broader shape coverage and improve abstraction accuracy. To further enhance the shape coverage of the proposed DeFormer, we employ a diffeomorphic mapping that preserves shape topology to predict local deformations for finer details beyond the coverage of global deformations.

To improve the primitive parameter optimization, we introduce "external force" during training, to minimize the divergence between the deformed primitives and target shapes. This allows us to further use kinematic modeling for more flexible transformations across the data space, the generalized latent space, and the projected image space for improved robust training. To guarantee the training convergence, we leverage a cycle-consistent re-projection loss to achieve consistency between the reconstructed shapes, with the projected image and the original image as the input, respectively. Extensive experiments across several settings show that DeFormer outperforms the state-of-the-art (SOTA) with fewer primitives on the core thirteen shape categories of *ShapeNet*.

Our main contributions are summarized as follows:

- To the best of our knowledge, DeFormer is the first work that integrates Transformers with deformable models for accurate shape abstraction. We show that our novel learning formulation achieves better abstraction ability using a small number of primitives with a broader shape coverage.
- A force-driven dynamic fitting loss combined with a cycle-consistent re-projection regularization is introduced for effective and robust model training.
- Extensive experiments show that our method achieves better reconstruction accuracy and improved semantic consistency compared to the state-of-the-art.

2. Related Work

3D shape reconstruction can be categorized into several mainstreams. (1) Voxel-based methods [11, 67, 18, 33, 9] leverage voxels to capture 3D geometries. These methods usually require large memory and computation resources. Some methods reduce the memory cost [46, 60, 25], but the complexity of these frameworks increases significantly. (2) Point Cloud-based methods [17, 57, 1, 32, 63] require less computation but additional post-processing to address the lack of surface connectivity for mesh generation. (3) Mesh-based [34, 66, 52, 9] can generate smooth shape surfaces, but they do not offer part-level decomposition of the shape. (4) Implicit function-based methods [47, 9, 53, 58, 68, 49, 50] can also achieve high reconstruction accuracy of the shape, but they require heavy post-processing to obtain the final mesh. (5) Primitive-based methods [64, 56, 54, 26, 14, 55] represent object shapes by deforming a number of primitives, each of which is explicitly described by a set of shape-related parameters (*e.g.*, scaling, squareness, tapering) whose properties are described in the following.

Primitive-based Shape Abstraction. Since our approach is primitive-based we thus focus on the most relevant primitive-based methods. Tulsiani *et al.* employ a union of cuboids to abstract object shapes [64], while in [56, 54], Paschalidou *et al.* extend cuboids to superquadrics which provide extra geometric flexibility of the primitive. Other primitive shapes such as spheres [26] and convexes [14] have also been investigated. The accuracy of these methods highly depends on the typically large number of primitives. Following this, Neural Parts [55] employ an Invertible Neural Network [2] with reduced number of primitives to improve the performance. However, the primitives in Neural Parts do not often correspond to the object parts (especially those without clearly identified boundaries), thus resulting in reduced interpretability. To address these limitations, we propose DeFormer with a small number of primitives for more accurate shape abstraction. We leverage deformable models to parameterize the primitives and guarantee the consistent correspondence between the primitives and the target shape, which significantly improves the abstraction ability. Another close work Pix2Mesh [66] is mesh-based, and applies a single template for shape deformation. But it lacks explicit correspondence during deformation with the use of graph unpooling layers and may yield invalid mesh (*e.g.*, self-intersecting mesh). In contrast, our deformation is diffeomorphic, which can preserve the topology of the primitive shape without breaking the connectivity in the mesh.

Implicit Function-based Methods. This set of methods mainly leverage implicit functions (*i.e.*, level-sets) to directly estimate the signed distance function [8, 9, 53, 58, 15, 47, 22, 68, 36, 49, 50, 72, 61]. While they achieve high reconstruction accuracy, they usually need post-processing (*e.g.*, marching cubes) to recover the shape surface. In contrast, primitive-based methods seek to decompose a target shape into parts and also decompose each part into explicit shape-related parameters (*e.g.*, scaling, squareness, tapering, bending), which contribute to the understanding of

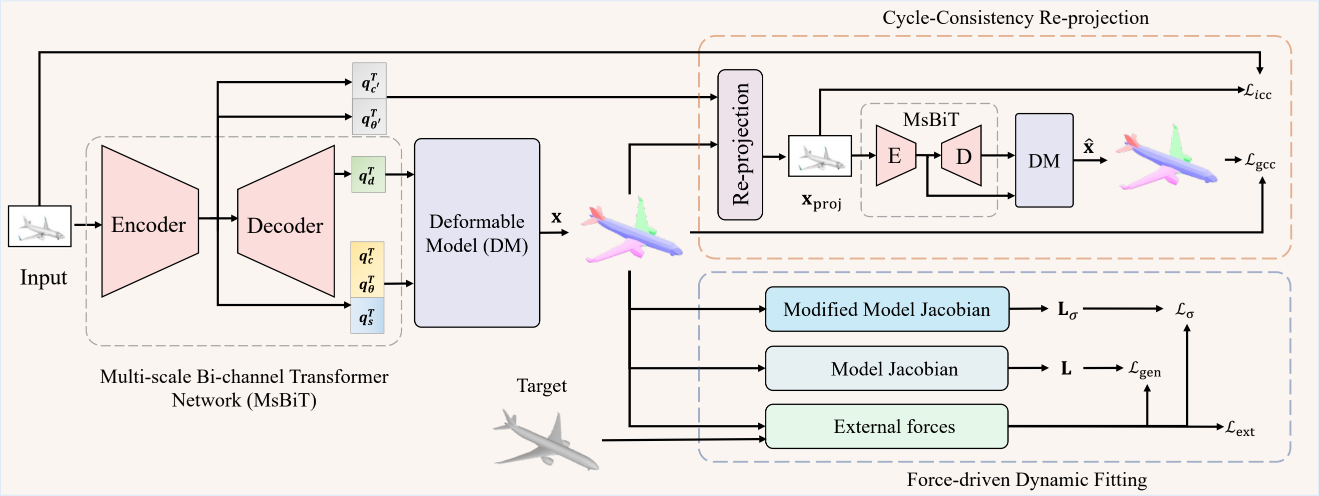


Figure 2: DeFormer overview. Given an input image $\mathcal{X}$, a *Multi-scale Bi-channel Transformer Network (MsBiT)* is proposed to hierarchically map $\mathcal{X}$ to a set of camera- and shape-related parameters that describe P deformed primitives. The primitive parameters $\mathbf{q}_c$, $\mathbf{q}_\theta$, $\mathbf{q}_s$, $\mathbf{q}_d$, are passed through the deformable models to give a shape reconstruction $\mathbf{x}$. To optimize the reconstruction, we employ a *Force-driven Dynamic Fitting* module to minimize the forces applied to the primitives. To prevent overfitting to the training data, we propose a *Cycle-Consistency Re-projection* loss for further regularization.

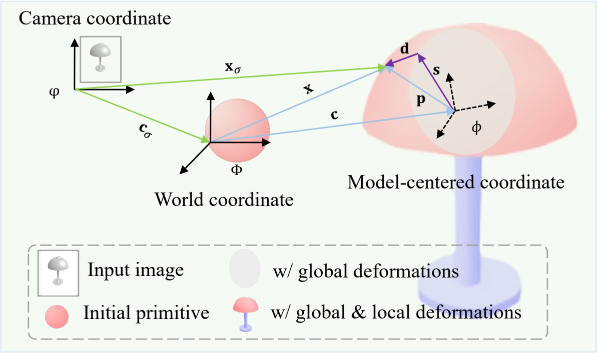


Figure 3: Generalized DeFormer geometry with perspective projection. It enables flexible transformations among the world coordinate Φ, the model-centered coordinate ϕ, and the camera coordinate φ.

primitive deformation. These shape-related parameters enable the explicit modeling for each shape part and provide semantic consistency among shapes.

3. Approach

DeFormer (Fig. 2) targets learning a set of primitives parameterized by DMs (Sec. 3.1) given a single image as input. Each primitive is represented by a group of shape-related parameters $\mathbf{q}$ which are estimated by the proposed *Multi-scale Bi-channel Transformer Network (MsBiT)* (Sec. 3.2). A *Force-driven Dynamic Fitting* module is introduced (Sec. 3.3) to minimize the forces applied onto the primitives. To further improve the modeling accuracy, we propose a novel *cycle-consistent re-projection* loss in Sec. 3.4 to regularize the estimated primitive deformations.

3.1. Geometry and Primitive Formulation

Canonical Geometry. Following the physics-based deformable models [62, 48], DeFormer assumes each individual primitive is a closed surface with a model-centered coordinate ϕ. As shown in Fig. 3, given a point k on the primitive surface, its location $\mathbf{x} = (x, y, z)$ *w.r.t.* the world coordinate Φ is

$$\mathbf{x} = \mathbf{c} + \mathbf{R}\mathbf{p} = \mathbf{c} + \mathbf{R}(\mathbf{s} + \mathbf{d}), \tag{1}$$

where $\mathbf{c}$ and $\mathbf{R}$ represent the primitive translation and rotation *w.r.t.* Φ; $\mathbf{p}$ denotes the relative position of the point k on the primitive surface *w.r.t.* ϕ, which includes global deformation $\mathbf{s}$ and local deformation $\mathbf{d}$.

Generalized Geometry with Perspective Projection. We seek to abstract the object shape in the world coordinate Φ given a single image $\mathcal{X}$, where $\mathcal{X}$ is in the camera reference frame φ (See Fig. 3). Theoretically, the camera has a certain relative orientation corresponding to $\mathcal{X}$. To enable robust 3D shape abstraction that matches the 2D observation, we integrate the camera parameters into our current geometry for accurate camera pose estimation. Let $\mathbf{x}_\sigma = (x_\sigma, y_\sigma, z_\sigma)$ be the location of the point k on the primitive surface *w.r.t.* the camera frame φ. Similar to Eq. (1) we denote:

$$\mathbf{x}_\sigma = \mathbf{c}_\sigma + \mathbf{R}_\sigma \mathbf{x}, \tag{2}$$

where $\mathbf{c}_\sigma$ and $\mathbf{R}_\sigma$ are the translation and rotation of the camera frame φ *w.r.t.* the world coordinate frame Φ.

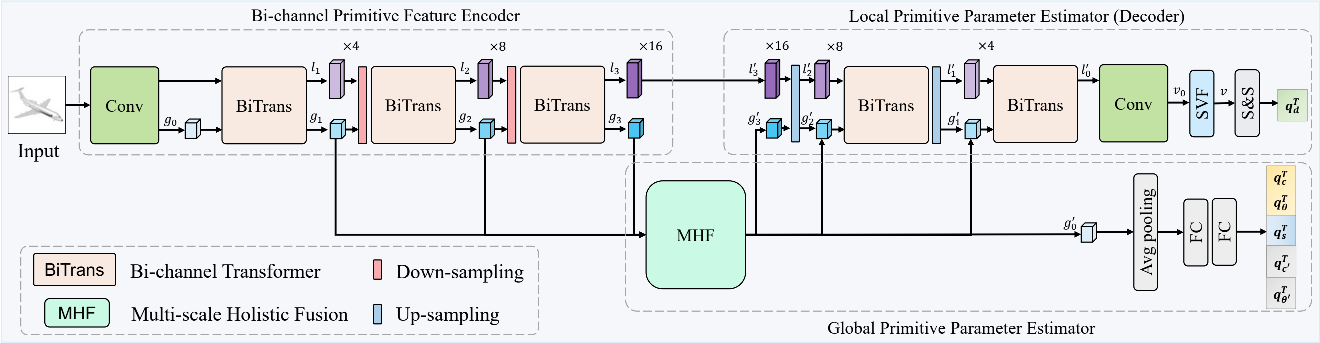


Figure 4: The architecture of Multi-scale Bi-channel Transformer Network (MsBiT). Given an input image, MsBiT hierarchically predicts a set of primitive parameters. The *Bi-channel Primitive Feature Encoder* first maps the input into two feature branches for global and local representations, g_i and l_i, respectively. The encoded holistic maps g_i are then passed through the *Global Primitive Parameter Estimator* which outputs the final global shape-related parameters $\mathbf{q}_c$, $\mathbf{q}_\theta$, $\mathbf{q}_{c'}$, $\mathbf{q}_{\theta'}$, $\mathbf{q}_s$. We employ the *Local Primitive Parameter Estimator* to collect the encoded features and pass them through a diffeomorphic mapping for the final estimation of local deformation $\mathbf{q}_d$.

In summary, we decompose the transformations between the primitive and the target shape as camera translation $\mathbf{c}_\sigma$, camera rotation $\mathbf{R}_\sigma$, primitive translation $\mathbf{c}$, primitive rotation $\mathbf{R}$, global deformations $\mathbf{s}$ and local deformations $\mathbf{d}$.

Primitive Parameterization. We use superquadric surfaces with global and local deformations to represent the primitives due to their broad geometry coverage. We follow the original formulation of superquadrics in [5, 62, 56] and develop a modified version by introducing global tapering and bending deformations as well as diffeomorphic local deformations. We provide a detailed formulation of our primitives in *suppl. material*.

3.2. Multi-scale Bi-channel Transformer Network

Bi-channel Transformer (BiTrans) Module. Since the local receptive field of CNNs limits the modeling of global primitive features [65, 59], we employ Transformers to collect long-range dependencies for the prediction of the global primitive parameters. Geometrically, the primitive parameters related to the translation, rotation and global deformations have a holistic point of view to preserve the most salient primitive features, which makes the all-to-all attention in the Multi-head self-attention (MHSA) [65] highly redundant. To address this, we propose a Bi-channel Transformer (BiTrans) with two channels: 1) a low-dimensional feature map g_i (blue) to preserve the holistic information, and 2) a conventional feature map l_i (purple) to embed the local non-rigid information. The local deformation map l_i is firstly projected to $Q/K/V$ with depth-wise separable convolution [10]. We employ 1×1 convolution to project g_i with a much smaller size to $\overline{Q}/\overline{K}/\overline{V}$ to avoid any additional noise introduced in the depth-wise separable convolution padding. Due to the symmetry of the query and key dot product, we achieve the cross-attention map by transposing the dot product matrix to aggregate the global and local information of the primitive:

$$\begin{aligned}(l_i^j, g_i^j) &= \mathrm{BiTrans}(l_i^{j-1}, g_i^{j-1}) \\ &= (\mathrm{softmax}(\frac{Q\overline{K}^\top}{\sqrt{d}})\overline{V}, \mathrm{softmax}(\frac{\overline{Q}K^\top}{\sqrt{d}})V),\end{aligned} \tag{3}$$

where l_i^j and g_i^j are the "Bi-channel" j-th layer outputs.

Global Primitive Parameter Estimator. We employ an estimator with an average pooling and two fully connected layers (Fig. 4) to map the embedded holistic features g'_0 to global parameters $\mathbf{q}_c$, $\mathbf{q}_\theta$, $\mathbf{q}_{c'}$, $\mathbf{q}_{\theta'}$, $\mathbf{q}_s$ corresponding to $\mathbf{c}$, $\mathbf{R}$, $\mathbf{c}_\sigma$, $\mathbf{R}_\sigma$, $\mathbf{s}$, respectively. Specifically, $\mathbf{q}_c = \mathbf{c}$, $\mathbf{q}_{c'} = \mathbf{c}_\sigma$. $\mathbf{q}_\theta$ and $\mathbf{q}_{\theta'}$ are two four-dimensional quaternions related to $\mathbf{R}$ and $\mathbf{R}_\sigma$ defined in [62]. $\mathbf{q}_s = (a, \epsilon, t, b)$ determines the scaling a, squareness ϵ, tapering t and bending b parameters of each primitives.

Local Primitive Parameter Estimator. To capture the finer shape details beyond the coverage of global deformations, we employ a diffeomorphic mapping to estimate the local non-rigid deformations $\mathbf{q}_d = \mathbf{d}$. Since the deformation with diffeomorphism is differentiable and invertible [3, 13], it guarantees one-to-one mapping and preserves topology during the non-rigid deformations of the primitives. Specifically, given the local features l'_0 from the MsBiT decoder, we first use a convolution stem to map l'_0 to a vector field v_0, and then map v_0 to a stationary velocity field (SVF) v using a Gaussian smoothing layer. We follow [3, 13, 12, 4] and employ an Euler integration with a scaling and squaring layer (S&S) to obtain the final local deformation $\mathbf{q}_d$.

3.3. Force-driven Dynamic Fitting

Similar to DMs, the primitives of DeFormer are able to dynamically deform to fit the target shape under the influence of external forces. Following the principle of virtual work[1], we express the energy of the primitive as $\mathcal{E}_f = \int f^\top d\mathbf{x}$. f denotes the external force which measures how well the primitives are deformed to fit the target shape in data space (*i.e.*, point-wise difference between the primitive surface and the target shape). When the primitive is far from the target, the force is large; vice versa. To optimize $\mathcal{E}_f$ we designate three specific loss terms as follows.

External Model Loss $\mathcal{L}_{\text{ext}}$. We first employ the external model loss $\mathcal{L}_{\text{ext}}$ to minimize the external forces applied to the p-th primitive, f^p, as:

$$\mathcal{L}_{\text{ext}} = \frac{1}{P}\sum_{p=1}^{P} f^p = \frac{\gamma}{P}\sum_{p=1}^{P}\mathcal{D}(\mathcal{M}_p, \mathcal{T}), \tag{4}$$

where γ is a constant modeling the strength of f^p and $\mathcal{D}(\cdot)$ is the distance function that measures the difference between the points $\mathbf{x}_m^p$ on the p-th deformed primitive $\mathcal{M}_p$ and the points τ_n on the target shape $\mathcal{T}$. Specifically, we employ a bi-directional Chamfer Distance (CD) for $\mathcal{D}(\cdot)$, denoted as:

$$\begin{aligned}\mathcal{D}(\mathcal{M}_p, \mathcal{T}) = &\frac{1}{|\mathcal{M}_p|}\sum_{m=1}^{|\mathcal{M}_p|}\min_{\tau_n \in \mathcal{T}} \|\mathbf{x}_m^p - \tau_n\|_2^2 \\ &+ \frac{1}{|\mathcal{T}|}\sum_{n=1}^{|\mathcal{T}|}\min_{\mathbf{x}_m^p \in \mathcal{M}_p} \|\tau_n - \mathbf{x}_m^p\|_2^2.\end{aligned} \tag{5}$$

Generalized Model Loss $\mathcal{L}_{\text{gen}}$. $\mathcal{L}_{\text{ext}}$ can be viewed as a standard loss term for shape reconstruction, which, however, only controls the surface points of the predicted primitives with loose constraints. In addition, we seek to regularize the prediction by constraining each sub-transformation (*i.e.*, primitive translation $\mathbf{c}$, primitive rotation $\mathbf{R}$, global $\mathbf{s}$ and local deformations $\mathbf{d}$) during dynamic fitting. Inspired by the kinematics of DMs [62], we achieve this by converting the forces computed in data space to the generalized forces in the generalized latent space. Specifically, the kinematics are computed by $d\mathbf{x} = \mathbf{L}d\mathbf{q}$, where $\mathbf{L}$ is the Model Jacobian matrix that includes the Jacobians for $\mathbf{c}$, $\mathbf{R}$, $\mathbf{s}$, $\mathbf{d}$ [48]. $\mathbf{q}$ is the group of parameters controlling these sub-transformations. Then $\mathcal{E}_f$ is expressed as:

$$\mathcal{E}_f = \int f^\top d\mathbf{x} = \int f^\top \mathbf{L}d\mathbf{q} = \int f_q d\mathbf{q}, \tag{6}$$

where f_q is the generalized force that measures the corresponding parameter-wise difference for each sub-transformation in the generalized latent space. It is based on the Model Jacobian $\mathbf{L} = [\mathbf{I}, \mathbf{B}, \mathbf{RJ}, \mathbf{R}]$, where $\mathbf{R}$ the rotation matrix, $\mathbf{B} = \partial\mathbf{Rp}/\partial\mathbf{q}_\theta$ a rotation-related matrix, and $\mathbf{J}$ the Jacobian matrix [62]. Then f_q is derived as:

$$\begin{aligned}f_q = f^\top\mathbf{L} &= [f^\top, f^\top\mathbf{B}, f^\top\mathbf{RJ}, f^\top\mathbf{R}] \\ &= [f_c^\top, f_\theta^\top, f_s^\top, f_d^\top],\end{aligned} \tag{7}$$

where f_c, f_θ, f_s and f_d are the generalized force terms for the four sub-transformations $\mathbf{c}$, $\mathbf{R}$, $\mathbf{s}$ and $\mathbf{d}$, respectively. This shows how the generalized forces f_q are related to the external force f^p using the Model Jacobian matrix $\mathbf{L}$.

Therefore, to regularize each sub-transformation during dynamic fitting, we employ a generalized model loss $\mathcal{L}_{\text{gen}}$ that minimizes f_q:

$$\begin{aligned}\mathcal{L}_{\text{gen}} &= \sum_{p=1}^{P}((f_c^p)^\top + (f_\theta^p)^\top + (f_s^p)^\top + (f_d^p)^\top) \\ &= \sum_{p=1}^{P}((f^p)^\top + (f^p)^\top\mathbf{B} + (f^p)^\top\mathbf{RJ} + (f^p)^\top\mathbf{R}).\end{aligned} \tag{8}$$

Note that our formulation of f^p is a scalar approximation of the external force and not a vector. However, by minimizing each point-wise CD, we actually observe an approximated optimization result, in the sense of minimizing each point-wise force leading to the minimized joint force.

Image Model Loss $\mathcal{L}_\sigma$. In addition to jointly optimize the primitive surface points using $\mathcal{L}_{\text{ext}}$ and the sub-transformations for the four primitive deformations using $\mathcal{L}_{\text{gen}}$, we also seek to optimize the sub-transformations for the camera translation $\mathbf{c}_\sigma$ and camera rotation $\mathbf{R}_\sigma$. Similarly, we achieve this by converting f to the forces in the projected image space according to the generalized geometry with perspective projection presented in Sec. 3.1. Specifically, Given a point on the primitive surface with location $\mathbf{x}_\sigma = (x_\sigma, y_\sigma, z_\sigma)$, its corresponding point on the projected image is expressed as $\mathbf{x}_{\text{proj}} = (x_{\text{proj}}, y_{\text{proj}})$, where $x_{\text{proj}} = x_\sigma\mathcal{F}/z_\sigma$, $y_{\text{proj}} = y_\sigma\mathcal{F}/z_\sigma$ with $\mathcal{F}$ a constant to represent the focal length of the camera. By taking the time derivative, we obtain $d\mathbf{x}_{\text{proj}} = \mathbf{P}d\mathbf{x}_\sigma$, where

$$\mathbf{P} = \begin{bmatrix}\mathcal{F}/z_\sigma & 0 & -x_\sigma\mathcal{F}/z_\sigma^2 \\ 0 & \mathcal{F}/z_\sigma & -y_\sigma\mathcal{F}/z_\sigma^2\end{bmatrix}. \tag{9}$$

Given Eq. (2) and the kinematics $d\mathbf{x} = \mathbf{L}d\mathbf{q}$ [48], we obtain:

$$d\mathbf{x}_{\text{proj}} = \mathbf{P}d\mathbf{x}_\sigma = \mathbf{P}d(\mathbf{c}_\sigma + \mathbf{R}_\sigma\mathbf{x}) = \mathbf{P}\mathbf{R}_\sigma d\mathbf{x}. \tag{10}$$

The above Eq. (10) allows us to modify the Model Jacobian matrix $\mathbf{L}$ in DMs with $\mathbf{L}_\sigma = \mathbf{P}\mathbf{R}_\sigma\mathbf{L}$ for our generalized deformable model geometry with perspective projection, where $\mathbf{L}_\sigma$ is named the Modified Model Jacobian matrix. By replacing the $\mathbf{L}$ in Eq. (7) with $\mathbf{L}_\sigma$, we obtain $f_\sigma = f^\top\mathbf{P}\mathbf{R}_\sigma\mathbf{L}$, which allows us to transform the external forces f to the projected image forces f_σ. Similar to Eq. (4), the projected image model loss $\mathcal{L}_\sigma$ computed using f_σ is then derived as:

$$\mathcal{L}_\sigma = \frac{1}{P}\sum_{p=1}^{P} f_\sigma^p = \frac{1}{P}\sum_{p=1}^{P}(f^p)^\top\mathbf{P}\mathbf{R}_\sigma\mathbf{L}. \tag{11}$$

[1] In mechanics, virtual work is the total work done by the applied forces of a mechanical system as it moves through a set of virtual displacements.

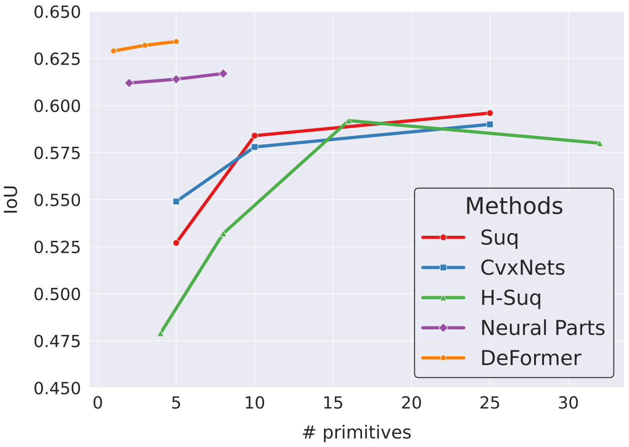


Figure 5: Analysis of accuracy *w.r.t.* the number of primitives used. We focus on comparing to Neural Parts, with three data points, showing using a small number of primitives (<10) to achieve better reconstruction accuracy.

By combining the illustrated losses related to the external, generalized, and projected image forces together, we summarize the dynamic fitting loss as:

$$\mathcal{L}_f = \mathcal{L}_{\text{ext}} + \mathcal{L}_{\text{gen}} + \mathcal{L}_{\sigma}. \tag{12}$$

3.4. Cycle-Consistent Re-projection

To prevent network overfitting on the training data, we apply a differentiable re-projection module. As shown in Fig. 2, given the reconstructed primitive, we employ a differentiable renderer [44] to re-project it onto the image domain using the predicted camera-related parameters $\mathbf{q}_{c'}$, $\mathbf{q}_{\theta'}$. Then, by sending it to the network again, we expect DeFormer to have the same shape reconstruction as $\mathbf{x}$. This process is formulated as a cycle-consistency regularization:

$$\mathcal{L}_{\text{gcc}} = \frac{1}{P}\sum_{p=1}^{P} \hat{f}^p = \frac{\hat{\gamma}}{|P|}\sum_{p=1}^{P} \mathcal{D}(\hat{\mathcal{M}}_p, \mathcal{M}_p), \tag{13}$$

where $\hat{\gamma}$ is the strength of the pseudo external forces $\hat{f}^p$, and $\hat{\mathcal{M}}_p$ denotes the p-th re-reconstructed primitive given the projected image $\mathbf{x}_{\text{proj}}$ as input. If the above re-reconstruction is optimized, the projected image $\mathbf{x}_{\text{proj}}$ should also match the original input $\mathcal{X}$. Therefore, we employ the image-level cycle-consistency loss $\mathcal{L}_{\text{icc}}$ to minimize the difference between $\mathbf{x}_{\text{proj}}$ and $\mathcal{X}$:

$$\mathcal{L}_{\text{icc}} = \frac{1}{P}\sum_{p=1}^{P} \hat{f}_{\sigma}^p = \frac{\hat{\gamma}}{|P|}\sum_{p=1}^{P} \min \|\mathbf{x}_{\text{proj}}^p - \mathcal{X}\|_2^2, \tag{14}$$

where $\hat{f}_{\sigma}^p$ is the pseudo image force and $\mathbf{x}_{\text{proj}}^p$ is the image projected from the p-th primitive $\mathbf{x}^p$. Together with the dynamic fitting loss $\mathcal{L}_f$, we obtain the overall optimization objective as:

$$\mathcal{L} = \mathcal{L}_f + \mathcal{L}_{\text{gcc}} + \mathcal{L}_{\text{icc}}. \tag{15}$$

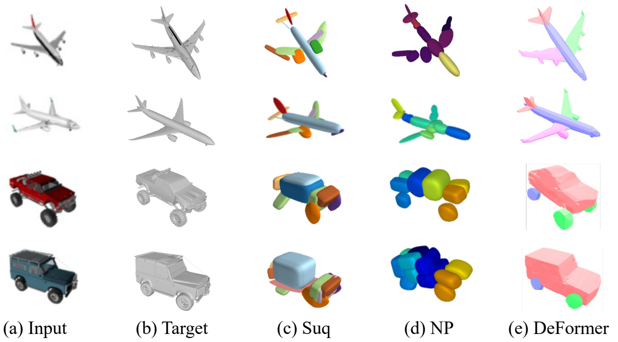


Figure 6: Abstraction visualization compared to superquadrics-based methods, including Suq [56] and H-Suq [54] with ∼20 primitives. In contrast, our model yields more accurate reconstructions with significantly fewer primitives (4 for airplanes and 3 for cars).

4. Experiments

Datasets. We evaluate on *ShapeNet* [6], a richly-annotated, large-scale dataset of 3D shapes. A subset of *ShapeNet* including 50k models and 13 major categories are used in our experiments. We use the rendered views from 3D-R2N2 [11], and their training and testing split setting, which was the seminal work in the literature, and the setting has been utilized by most of the following-up papers.

Baselines. Since DeFormer lies in the primitive-based mainstream with explicit representation, we mainly compare it to primitive-based methods, *i.e.*, Suq [56] and H-Suq [54] that both use superquadrics, CvxNets [14] using convexes, and NP [55] using spheres. For non-primitive methods, we compare to SIF [22], P2M [66], and OccNet [47] which, however, lack explicit understanding of part correspondence.

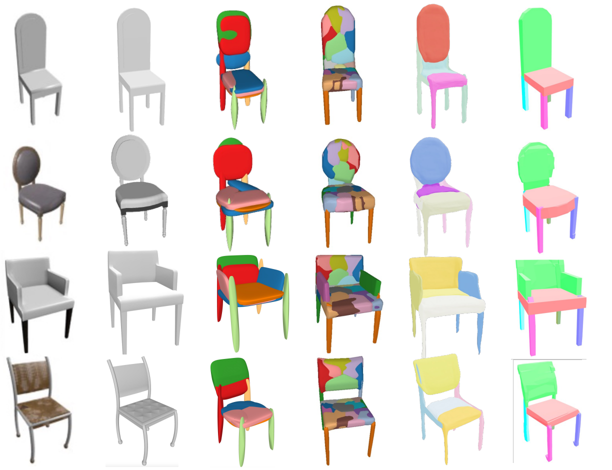
4.1. Implementation Details

Throughout the training, Adam [35] is employed for optimization and the learning rate is initialized as 10^{-4}. We use a batch size of 32 and train the model for 300 epochs. All experiments are implemented with PyTorch and run on a Linux system with eight Nvidia A100 GPUs. Assuming input image size $H \times W$ and d the token dimension, compared to CNNs complexity $\mathcal{O}(k^2 HWd^2)$ with convolution kernel size k, our BiTrans complexity is $\mathcal{O}(4HWd^2 + 2(HW)^2 d)$, which achieves the same order of complexity as CNNs.

Similar to [14, 55, 56], for each shape category with a certain number of primitives used, we train a separate

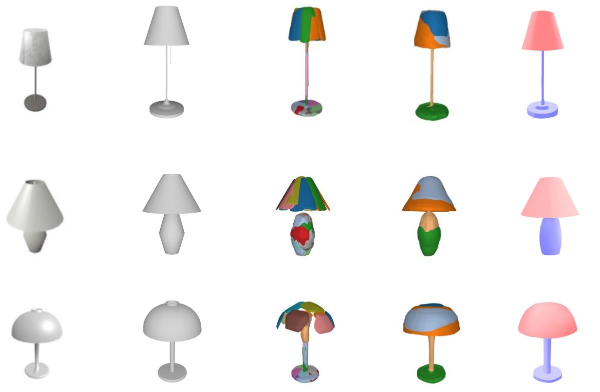
Table 1: Reconstruction results on the thirteen categories of *ShapeNet*. We evaluate DeFormer (4) against P2M [66], SIF [22] (50), OccNet [47], Suq [56] ($\leq$ 64), CvxNets [14] (25), H-Suq [54] ($\leq$ 64), and NP [55] (5). The Abs. Gain shows an absolute improvement to the second best. Numbers in $(\cdot)$ indicates the number of primitives used.

Category	IoU (↑)									Chamfer-L_1 (↓)								
	P2M	SIF	OccNet	Suq	CvxNets	H-Suq	NP	DeFormer	Abs. Gain	P2M	SIF	OccNet	Suq	CvxNets	H-Suq	NP	DeFormer	Abs. Gain
airplane	0.420	0.530	0.571	0.456	0.598	0.529	0.611	**0.641**	3.0%	0.187	0.167	0.147	0.122	0.093	0.175	0.089	**0.072**	0.017
bench	0.323	0.333	0.485	0.202	0.461	0.437	0.502	**0.528**	2.6%	0.201	0.261	0.155	0.114	0.133	0.153	0.108	**0.087**	0.021
cabinet	0.664	0.648	0.733	0.110	0.709	0.658	0.681	**0.717**	0.4%	0.196	0.233	0.167	0.087	0.102	0.087	0.083	**0.074**	0.009
car	0.552	0.657	0.737	0.650	0.675	0.702	0.719	**0.729**	1.0%	0.180	0.161	0.159	0.117	0.103	0.141	0.127	**0.093**	0.010
chair	0.396	0.389	0.501	0.176	0.491	0.526	0.532	**0.551**	1.9%	0.265	0.380	0.228	0.138	0.337	0.114	0.107	**0.089**	0.018
display	0.490	0.491	0.471	0.200	0.576	0.633	0.646	**0.653**	0.7%	0.239	0.401	0.278	0.106	0.223	0.137	0.098	**0.087**	0.011
lamp	0.323	0.260	0.371	0.189	0.311	0.441	0.402	**0.442**	**4.0%**	0.308	1.096	0.479	0.189	0.795	0.169	0.153	**0.141**	0.012
speaker	0.599	0.577	0.647	0.136	0.620	0.660	0.693	**0.715**	2.2%	0.285	0.554	0.300	0.132	0.462	0.108	0.128	**0.092**	0.016
rifle	0.402	0.463	0.474	0.519	0.515	0.435	0.537	**0.540**	0.3%	0.164	0.193	0.141	0.127	0.106	0.203	0.189	**0.089**	0.017
sofa	0.613	0.606	0.680	0.122	0.677	0.693	0.712	**0.729**	1.7%	0.212	0.272	0.194	0.106	0.164	0.128	0.107	**0.088**	0.018
table	0.395	0.372	0.506	0.180	0.473	0.491	0.531	**0.546**	1.5%	0.218	0.454	0.189	0.110	0.358	0.122	0.102	**0.081**	0.021
phone	0.661	0.658	0.720	0.185	0.719	0.770	0.810	**0.822**	1.2%	0.149	0.159	0.140	0.112	0.083	0.149	0.076	**0.064**	0.012
vessel	0.397	0.502	0.530	0.471	0.552	0.570	0.605	**0.630**	2.5%	0.212	0.208	0.218	0.125	0.173	0.178	0.119	**0.096**	0.023
Average	0.480	0.499	0.571	0.277	0.567	0.580	0.614	**0.634**	1.8%	0.216	0.349	0.215	0.122	0.245	0.143	0.114	**0.089**	**0.025**



(a) Input (b) Target (c) Suq (d) CvxNets (4) NP (d) DeFormer

Figure 7: Abstraction visualization on chairs compared to primitive-based methods, including Suq [56], CvxNets [14], NP [55] with ∼20, 25 and 5 primitives, respectively. Ours applies 6 primitives (4 legs, 1 seat, and 1 back) and achieves better part consistency.



(a) Input (b) Target (c) CvxNets (d) NP (e) DeFormer

Figure 8: Abstraction visualization on lamps compared to SOTA primitive-based methods, including CvxNets [14] and NP [55] with 25 and 5 primitives, respectively. Ours applies 2 primitives (1 head and 1 base) and achieves better part consistency.

model. We draw 2k random sample points from the surface of each target mesh as ground truth, and we sample 1k points from each generated primitive for shape reconstruction. During the evaluation, we uniformly sample 100k points on the target/predicted mesh to compute the volumetric Intersection over Union (IoU) and the Chamfer-$L1$ distance (CD). We empirically set the weights for the dynamic fitting loss in Eq. (12) as 0.5, 0.3, and 0.2, respectively. Similarly, we set the balance factors for the joint loss in Eq. (15) as 0.6, 0.2, and 0.2, respectively, for best performance. Ablation study for losses is provided in Tab. 2. For the estimation of local deformations, we follow [3, 13, 12, 4] and use $T = 7$ scaling and squaring steps.

4.2. Representation Power

We first report the results of reconstruction accuracy *w.r.t.* the number of primitives P in Fig. 5. Our method shows consistently better IoU regardless of the number of primitives used. We further see that the reconstruction curve of DeFormer saturates fast when the number of primitives increases. This is due to the broad geometric coverage of the proposed primitive formulation where a small number of primitives are sufficient for the optimal shape abstraction. Moreover, to qualitatively demonstrate the representation superiority of our primitive formulation, we compare to Suq [56] and H-Suq [54] with ∼ 20 primitives, which also use superquadrics in Fig. 6. We train DeFormer with fewer

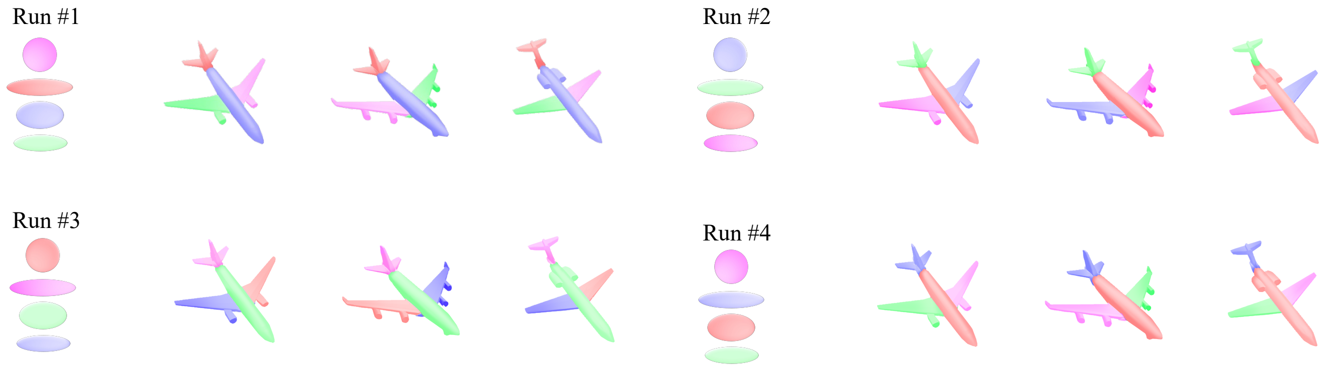


Figure 9: Illustration of semantic consistency. We set four different random seeds. For each seed, we observe consistent part correspondence (*e.g.*, left-wing, tail, body and right-wing denoted as the same color) across the three "airplane" instances.

primitives (4 for airplanes and 3 for cars) and obtain better reconstruction accuracy and semantic consistency.

4.3. Reconstruction Accuracy

We quantitatively evaluate the reconstruction performance against a number of SOTAs in Tab. 1. Following their settings we train Suq [56] and H-Suq [54] with a maximum of 64 primitives. For CvxNets [14] and SIF [22] we report results with 25 primitives and 50 elements, respectively. For NP [55] and DeFormer, we use 5 and 4 primitives, respectively. Note that P2M [66] and the implicit function-based methods OccNet [47] and SIF [22] are not directly comparable with the primitive-based methods, due to their lack of shape abstraction ability. Nevertheless, we observe from Tab. 1 that DeFormer outperforms all the SOTA results with on average 1.8% IoU accuracy improvement and 2.5% less Chamfer-L_1 distance. We provide a qualitative comparison in Fig. 7 and Fig. 8.

4.4. Ablation Study

Semantic Consistency. We investigate the ability of DeFormer to decompose 3D shapes into semantically consistent parts using different primitive initializations. Specifically, we train with four different random seeds on the airplane category and observe in Fig. 9 that the reconstructions preserve similar semantic parts for each seed.

Loss Components. In Tab. 2 using the "leave-one-out" way, each of the loss terms is highlighted and demonstrated to be a uniquely effective component within our overall loss term. Another observation is that training without $\mathcal{L}_{\text{ext}}$ results in a severe performance drop. The cycle-consistency losses $\mathcal{L}_{\text{gcc}}$ and $\mathcal{L}_{\text{icc}}$ provide key self-supervision for unreasonable reconstruction correction.

Settings					IoU (↑)		
$\mathcal{L}_{\text{ext}}$	$\mathcal{L}_{\text{gen}}$	$\mathcal{L}_{\sigma}$	$\mathcal{L}_{\text{gcc}}$	$\mathcal{L}_{\text{icc}}$	car	airplane	chair
✗	✓	✓	✓	✓	0.718	0.633	0.547
✓	✗	✓	✓	✓	0.723	0.641	0.550
✓	✓	✗	✓	✓	0.729	0.648	0.556
✓	✓	✓	✗	✓	0.731	0.647	0.552
✓	✓	✓	✓	✗	0.733	0.652	0.559
✓	✓	✓	✗	✗	0.721	0.638	0.545
✓	✓	✓	✓	✓	**0.729**	**0.641**	**0.551**

Table 2: Ablation studies on loss terms. We report the average IoU on the major three categories of *ShapeNet*.

5. Conclusion

We propose a novel bi-channel Transformer integrated with deformable models, termed DeFormer, to jointly predict global and local deformations for 3D shape abstraction. DeFormer achieves improved semantic correspondences thanks to the diffeomorphic mapping for shape estimation. Moreover, we leverage the force-driven dynamic fitting and the cycle-consistent re-projection loss to effectively optimize the shape parameters. Extensive experiments demonstrate our method achieves superior reconstruction performance and semantic consistency. Future work will consider more primitive formulations and global deformations for more general shape abstraction scenarios.

Acknowledgments

This research has been partially funded by research grants to D. Metaxas through NSF: IUCRC CARTA 1747778, 2235405, 2212301, 1951890, 2003874, and NIH-5R01HL127661.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 2

[2] Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W Pellegrini, Ralf S Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks. *arXiv preprint arXiv:1808.04730*, 2018. 2

[3] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer, 2006. 4, 7

[4] John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007. 4, 7

[5] Alan H Barr. Superquadrics and angle-preserving transformations. *IEEE Computer graphics and Applications*, 1(1):11–23, 1981. 4

[6] 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. 6

[7] Qi Chang, Zhennan Yan, Mu Zhou, Di Liu, Khalid Sawalha, Meng Ye, Qilong Zhangli, Mikael Kanski, Subhi Al'Aref, Leon Axel, et al. Deeprecon: Joint 2d cardiac segmentation and 3d volume reconstruction via a structure-specific generative method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 567–577. Springer, 2022. 1

[8] Zhiqin Chen, Kangxue Yin, Matthew Fisher, Siddhartha Chaudhuri, and Hao Zhang. Bae-net: Branched autoencoder for shape co-segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8490–8499, 2019. 2

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 1, 2

[10] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 4

[11] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016. 2, 6

[12] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018. 4, 7

[13] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236, 2019. 4, 7

[14] 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. 1, 2, 6, 7, 8

[15] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10286–10296, 2021. 1, 2

[16] Matthias Eisenmann, Annika Reinke, Vivienn Weru, Minu Dietlinde Tizabi, Fabian Isensee, Tim J Adler, Patrick Godau, Veronika Cheplygina, Michal Kozubek, Sharib Ali, et al. Biomedical image analysis competitions: The state of current participation practice. *arXiv preprint arXiv:2212.08568*, 2022. 1

[17] 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

[18] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*, pages 402–411. IEEE, 2017. 2

[19] Yunhe Gao, Zhuowei Li, Di Liu, Mu Zhou, Shaoting Zhang, and Dimitris N Meta. Training like a medical resident: Universal medical image segmentation via context prior learning. *arXiv preprint arXiv:2306.02416*, 2023. 1

[20] Yunhe Gao, Mu Zhou, Di Liu, Zhennan Yan, Shaoting Zhang, and Dimitris N Metaxas. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. *arXiv preprint arXiv:2203.00131*, 2022. 1

[21] Chuanbin Ge, Di Liu, Juan Liu, Bingshuai Liu, and Yi Xin. Automated recognition of arrhythmia using deep neural networks for 12-lead electrocardiograms with fractional time–frequency domain extension. *Journal of Medical Imaging and Health Informatics*, 10(11):2764–2767, 2020. 1

[22] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7154–7164, 2019. 2, 6, 7, 8

[23] Oshri Halimi, Or Litany, Emanuele Rodola, Alex M Bronstein, and Ron Kimmel. Unsupervised learning of dense shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4370–4379, 2019. 1

[24] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Yuxiao Chen, Di Liu, Qilong Zhangli, et al. Improving negative-prompt inversion via proximal guidance. *arXiv preprint arXiv:2306.05414*, 2023. 1

[25] Christian Häne, Shubham Tulsiani, and Jitendra Malik. Hierarchical surface prediction for 3d object reconstruction. In

2017 International Conference on 3D Vision (3DV), pages 412–420. IEEE, 2017. 2

[26] Zekun Hao, Hadar Averbuch-Elor, Noah Snavely, and Serge Belongie. Dualsdf: Semantic shape manipulation using a two-level representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7631–7641, 2020. 2

[27] Ali Hatamizadeh, Debleena Sengupta, and Demetri Terzopoulos. End-to-end trainable deep active contour models for automated image segmentation: Delineating buildings in aerial imagery. In *European Conference on Computer Vision*, pages 730–746. Springer, 2020. 1

[28] Xiaoxiao He, Chaowei Tan, Bo Liu, Liping Si, Weiwu Yao, Liang Zhao, Di Liu, Qilong Zhangli, Qi Chang, Kang Li, et al. Dealing with heterogeneous 3d mr knee images: A federated few-shot learning method with dual knowledge distillation. *arXiv preprint arXiv:2303.14357*, 2023. 1

[29] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991. 2

[30] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989. 2

[31] Ping Hu, Bing Shuai, Jun Liu, and Gang Wang. Deep level sets for salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2300–2309, 2017. 1

[32] Li Jiang, Shaoshuai Shi, Xiaojuan Qi, and Jiaya Jia. Gal: Geometric adversarial loss for single-view 3d-object reconstruction. In *Proceedings of the European conference on computer vision (ECCV)*, pages 802–816, 2018. 2

[33] Danilo Jimenez Rezende, SM Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images. *Advances in neural information processing systems*, 29, 2016. 2

[34] 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. 2

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[36] Manyi Li and Hao Zhang. D2im-net: Learning detail disentangled implicit fields from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10246–10255, 2021. 2

[37] Zhuowei Li, Long Zhao, Zizhao Zhang, Han Zhang, Di Liu, Ting Liu, and Dimitris N Metaxas. Steering prototype with prompt-tuning for rehearsal-free continual learning. *arXiv preprint arXiv:2303.09447*, 2023. 1

[38] Di Liu, Yunhe Gao, Qilong Zhangli, Zhennan Yan, Mu Zhou, and Dimitris Metaxas. Transfusion: Multi-view divergent fusion for medical image segmentation with transformers. *arXiv preprint arXiv:2203.10726*, 2022. 1

[39] Di Liu, Chuanbin Ge, Yi Xin, Qin Li, and Ran Tao. Dispersion correction for optical coherence tomography by the stepped detection algorithm in the fractional fourier domain. *Optics express*, 28(5):5919–5935, 2020. 1

[40] Di Liu, Jiang Liu, Yihao Liu, Ran Tao, Jerry L Prince, and Aaron Carass. Label super resolution for 3d magnetic resonance images using deformable u-net. In *Medical Imaging 2021: Image Processing*, volume 11596, page 1159628. International Society for Optics and Photonics, 2021. 1

[41] Di Liu, Yi Xin, Qin Li, and Ran Tao. Dispersion correction for optical coherence tomography by parameter estimation in fractional fourier domain. In *2019 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 674–678. IEEE, 2019. 1

[42] Di Liu, Zhennan Yan, Qi Chang, Leon Axel, and Dimitris N Metaxas. Refined deep layer aggregation for multi-disease, multi-view & multi-center cardiac mr segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 315–322. Springer, 2021. 1

[43] Di Liu, Long Zhao, Yunhe Gao, Qilong Zhangli, Ting Liu, and Dimitris N Metaxas. Deep physics-based deformable models for efficient shape abstractions. 2022. 1

[44] 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. 6

[45] Carlos Martín-Isla, Víctor M Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J Fulton, Tewodros Weldebirhan Arega, Kumaradevan Punithakumar, et al. Deep learning segmentation of the right ventricle in cardiac mri: The m&ms challenge. *IEEE Journal of Biomedical and Health Informatics*, 2023. 1

[46] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 922–928. IEEE, 2015. 2

[47] 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. 1, 2, 6, 7, 8

[48] Dimitris N Metaxas. *Physics-based deformable models: applications to computer vision, graphics and medical imaging*, volume 389. Springer Science & Business Media, 2012. 1, 3, 5

[49] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4743–4752, 2019. 1, 2

[50] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5379–5389, 2019. 1, 2

[51] 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. 1

[52] 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. 2

[53] 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. 1, 2

[54] 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. 1, 2, 6, 7, 8

[55] 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. 1, 2, 6, 7, 8

[56] 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. 1, 2, 4, 6, 7, 8

[57] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2

[58] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 1, 2

[59] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 4

[60] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE international conference on computer vision*, pages 2088–2096, 2017. 2

[61] Konstantinos Tertikas, Despoina Paschalidou, Boxiao Pan, Jeong Joon Park, Mikaela Angelina Uy, Ioannis Emiris, Yannis Avrithis, and Leonidas Guibas. Partnerf: Generating part-aware editable 3d shapes without 3d supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[62] Demetri Terzopoulos and Dimitri Metaxas. Dynamic 3 d models with local and global deformations: deformable superquadrics. *IEEE Transactions on pattern analysis and machine intelligence*, 13(7):703–714, 1991. 1, 3, 4, 5

[63] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 2

[64] 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. 1, 2

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[66] 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. 2, 6, 7, 8

[67] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 2

[68] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[69] Kaizhi Yang, Xiaoshuai Zhang, Zhiao Huang, Xuejin Chen, Zexiang Xu, and Hao Su. Movingparts: Motion-based 3d part discovery in dynamic radiance field. *arXiv preprint arXiv:2303.05703*, 2023. 1

[70] Meng Ye, Mikael Kanski, Dong Yang, Qi Chang, Zhennan Yan, Qiaoying Huang, Leon Axel, and Dimitris Metaxas. Deeptag: An unsupervised deep learning method for motion tracking on cardiac tagging magnetic resonance images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7261–7271, 2021. 1

[71] Qilong Zhangli, Jingru Yi, Di Liu, Xiaoxiao He, Zhaoyang Xia, Haiming Tang, He Wang, Mu Zhou, and Dimitris Metaxas. Region proposal rectification towards robust instance segmentation of biological images. *arXiv preprint arXiv:2203.02846*, 2022. 1

[72] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3d shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1429–1439, 2021. 2

[73] Chuhang Zou, Ersin Yumer, Jimei Yang, Duygu Ceylan, and Derek Hoiem. 3d-prnn: Generating shape primitives with recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 900–909, 2017. 1