Unsupervised Pose-aware Part Decomposition for Man-made Articulated Objects

Yuki Kawana[1], Yusuke Mukuta[1,2], and Tatsuya Harada[1,2]

[1]The University of Tokyo [2]RIKEN

Abstract. Man-made articulated objects exist widely in the real world. However, previous methods for unsupervised part decomposition are unsuitable for such objects because they assume a spatially fixed part location, resulting in inconsistent part parsing. In this paper, we propose PPD (unsupervised Pose-aware Part Decomposition) to address a novel setting that explicitly targets man-made articulated objects with mechanical joints, considering the part poses in part parsing. As an analysis-by-synthesis approach, We show that category-common prior learning for both part shapes and poses facilitates the unsupervised learning of (1) part parsing with abstracted part shapes, and (2) part poses as joint parameters under single-frame shape supervision. We evaluate our method on synthetic and real datasets, and we show that it outperforms previous works in consistent part parsing of the articulated objects based on comparable part pose estimation performance to the supervised baseline.

1 Introduction

Our daily life environments are populated with man-made articulated objects, ranging from furniture and household appliances such as drawers and ovens to tabletop objects such as eyeglasses and laptops. Humans are capable of recognizing such objects by decomposing them into simpler semantic parts based on part kinematics. Researchers have shown that even very young infants learn to group objects into semantic parts using the location, shape, and kinematics as a cue [38, 37, 42], even from a single image [34, 18]. Although humans can naturally achieve such reasoning, it is challenging for machines, particularly in the absence of rich supervision.

3D part-level understanding of shapes and poses from a single frame observation has wide range of applications in computer vision and robotics. Learning to represent complex target shapes with simpler part components as a generative approach enables applications such as structure modeling [25, 33] and unsupervised 3D part parsing [39, 29, 7, 28]. The previous unsupervised approaches have mainly focused on non-articulated objects. Because they exploit the consistent part location as a cue to group shapes into semantic parts, these approaches are unsuitable for decomposing articulated objects when considering the kinematics of *dynamic part locations*. For part pose, modeling kinematic structures as joint parameters has various applications, such as motion planning in robotic manipulation [1] and interaction with environment in augmented reality [4]. There

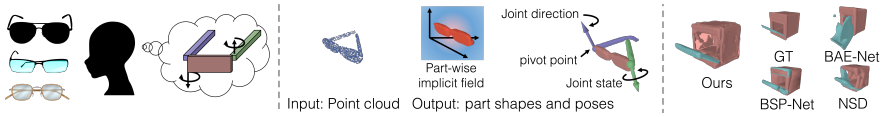

Fig. 1. (Left) Even through independent observations, infants can build a mental model of the articulated object for part parsing based on its kinematics. (Middle) Likewise, we propose an unsupervised generative method that learns to parse the single-frame, unstructured 3D data of articulated objects and predict the part-wise implicit fields as abstracted part shapes as well as their part poses as joint parameters. (Right) Our approach outperforms the previous works in consistent part parsing for man-made articulated objects.

exists a large body of works for discriminative approaches dedicated to man-made articulated objects for part pose estimation in addition to part segmentation. However, they require explicit supervision, such as segmentation labels and joint parameters [11, 1, 44, 41, 20]. Removing the need for such expensive supervision has been an important step toward more human-like representation learning [2].

In this study, as a novel problem setting, we investigate the unsupervised part decomposition task for man-made articulated objects with mechanical joints, considering part poses as *joint parameters*, in an *unsupervised fashion*. Specifically, we consider the revolute and prismatic parts with one degree-of-freedom joint state because they cover most of the kinematic types that common man-made articulated objects have [41, 1, 24]. This task aims to learn consistent part parsing as a generative shape abstraction approach for man-made articulated objects with various part poses from single-frame shape observation. An overview is shown in Figure 1. Recent part decomposition studies have focused on novel part shape representations for *shape reconstruction*. In contrast, we focus on *part parsing* and *part pose modeling* as a first step to expand the current generative part decomposition's applications to man-made articulated objects in novel ways, such as part pose consistent segmentation and part pose estimation as joint parameter prediction. To realize the task, we identify the two challenges; (1) for pose-aware part decomposition, the model must consider the kinematics between possibly distant shapes to group them as a single part and (2) has to disentangle the part poses from shape supervision. A comparison with previous studies is presented in Table 1.

To address these challenges, we propose PPD (unsupervised Pose-aware Part Decomposition) that takes an unsegmented, single-frame point cloud with various underlying part poses as an input. PPD predicts abstracted part-wise shapes transformed using the estimated joint parameters as the part poses. We train PPD as an autoencoder using single-frame shape supervision. PPD employs category-common decoders to capture category-specific rest-posed part shapes and joint parameters. Learning to transform the rest-posed shapes properly disentangles shape and pose, and (2) constraining the position of the parts by the joint parameters forces shapes in distant space that share the same kinematics to be recovered as the same part. We also propose a series of losses to regularize the learning process. Furthermore, we employ non-primitive-based part shape

	Part segmentation	Joint parameter estimation	Generative	Unsupervised
ANSCH [20]	✓	✓		
A-SDF [26]	✓		✓	
Nueral Parts [28]	✓		✓	✓
Ours	✓	✓	✓	✓

Table 1. Overview of the previous works. We regard a method as unsupervised if the checked tasks can be learned only via shape supervision during training.

representation and utilize deformation by part poses to induce part decomposition, in contrast to previous works that employ primitive shapes and rely on its limited expressive power as an inductive bias.

Our contributions are summarized as follows: (1) We propose a novel unsupervised generative part decomposition method for man-made articulated objects based on part kinematics. (2) We show that the proposed method learns disentangled part shape and pose: a non-primitive-based implicit field as part shape representation and the joint parameters as the part poses, using single-frame shape supervision. (3) We also demonstrate that the proposed method outperforms previous generative part decomposition methods in terms of semantic capability (parsimonious shape representation, consitent part parsing and interpretability of recovered parts) and show comparable part pose estimation performance to the supervised baseline.

2 Related works

Unsupervised part decomposition. Existing unsupervised generative part decomposition studies mostly assume non-articulated objects in which the part shapes are in a fixed 3D location [39, 27, 8, 7, 9, 15], or also targeting human body and hand shapes without considering part pose [28]. They induce part decomposition by limiting the expressive power of the shape decoders by employing learnable primitive shapes. Closest work of ours is BAE-Net [8], whose main focus is consistent part parsing by generative shape abstraction. It also employs a non-primitive-based implicit field as the part shape representation, similar to ours. However, it still limits the expressive power of the shape decoder using MLP with only three layers. In contrast, our approach assumes parts to be dynamic with the consistent kinematics and induces part decomposition through rigid transformation of the reconstructed part shapes with the estimated part poses to make the decomposition pose-aware.

Articulated shape representation. A growing number of studies have tackled the reconstruction of category-specific, articulated objects with a particular kinematic structure, such as the human body and animals. Representative works rely on the use of category-specific template models as the shape and pose prior

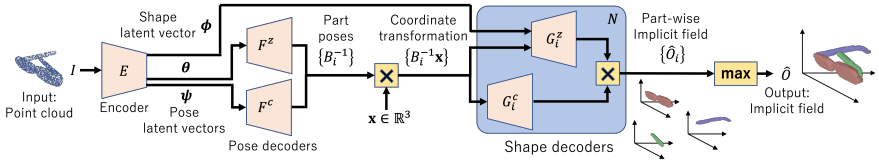

Fig. 2. Model overview. PPD infers implicit field $\hat{O}$ based on part poses $\{B_i\}$ and part-wise implicit fields $\{\hat{O}_i\}$ given input point cloud I. The category-common decoders F^c and $\{G_i^c\}$ capture part pose biases and part shape priors in constant latent vectors. Instance-dependent decoders F^z and $\{G_i^z\}$ model input specific components. Constraining the instance-dependent decoders by the category-common biases and the priors in the proposed approach realizes unsupervised part decomposition and joint parameter learning. Note we shorthand $\{*_i\}$ to denote an ordered set $\{*_i\}_{i=1}^N$ for brevity.

[21, 46, 3, 45, 19]. Another body of works reconstruct target shapes without templates, such as by reconstructing a part-wise implicit field given a part pose as an input [10] or focusing on non-rigid tracking of the seen samples [5]. The recent work [26] targets man-made articulated objects and supervised part shape reconstruction. In contrast, our approach focuses on man-made articulated objects with various kinematic structures. Our approach learns the part shapes and poses during training, without any part label and pose information either as supervision or input, and is applicable to unseen samples.

Part pose estimation. In discriminative approaches, a number of studies have focused on the inference of part poses as joint parameters [20, 41, 1] targeting man-made articulated objects. These approaches require expensive annotations, such as part labels and ground-truth joint parameters. Moreover, they require category-specific prior knowledge of the kinematic structure. In contrast, our model is based on generative approach and is category agnostic. Moreover, it only requires shape supervision during training. A recent work [13] assumes an unsupervised setting where multi-frame, complete shape point clouds are available for both input and supervision signals during training and inference. Whereas our approach assumes a single-frame input and shape supervision, it also works with partial shape input during inference. Note that, in this study, the purpose of part pose estimation is, as an auxiliary task, to facilitate consistent part parsing. It is not our focus to outperform the state-of-the-art supervised approaches in part pose estimation.

3 Methods

In our approach, the goal is to represent an articulated object as a set of semantically consistent part shapes based on their underlying part kinematics. We represent the target object shape as an implicit field that can be evaluated at an arbitrary point $\mathbf{x} \in \mathbb{R}^3$ in 3D space as $O : \mathbb{R}^3 \rightarrow [0, 1]$, where $\{\mathbf{x} \in \mathbb{R}^3 \mid O(\mathbf{x}) = 0\}$ defines the outside of the object, $\{\mathbf{x} \in \mathbb{R}^3 \mid O(\mathbf{x}) = 1\}$ the inside, and $\{\mathbf{x} \in \mathbb{R}^3 \mid O(\mathbf{x}) = 0.5\}$ the surface. Given a 3D point cloud $I \in \mathbb{R}^{P \times 3}$

of P points as an input, we approximate the object shape using a composite implicit field $\hat{O}$ that is decomposed into a collection of N parts. The i-th part has an implicit field $\hat{O}_i(\mathbf{x} \mid I)$ as part shape and part pose $B_i \in \mathrm{SE}(3)$. We ensure that O is approximated as $O(\mathbf{x}) \approx \hat{O}(\mathbf{x} \mid I, \{B_i\}_{i=1}^N)$ through the losses.

An overview of PPD is shown in Figure 2. PPD employs an autoencoder architecture, and is trained under single category setting. Given a point cloud I, the encoder derives the disentangled shape latent vector $\boldsymbol{\phi} \in \mathbb{R}^m$ and the two pose latent vectors $\boldsymbol{\theta} \in \mathbb{R}^n$ and $\boldsymbol{\psi} \in \mathbb{R}^o$. Category-common pose decoder F^c captures joint parameter biases given $\boldsymbol{\psi}$. Instance-dependent pose decoder F^z models residual joint parameters to the biases given $\boldsymbol{\theta}$. The part-wise category-common shape decoder G_i^c captures category-common shape prior. Given $\boldsymbol{\phi}$ and conditioned by G_i^c, instance-dependent shape decoder G_i^z infers residual shape details of the target shape to decode a part-wise implicit field $\hat{O}_i$. We discuss the details about F^z and F^c in Section 3.1, and G_i^z and G_i^c in Section 3.2.

3.1 Part pose representation

We characterize part pose B_i by its part kinematic type and joint parameters. Each part kinematic type $y_i \in \{\text{fixed}, \text{prismatic}, \text{revolute}\}$ is manually set as a hyperparameter. The joint parameters consist of the joint direction $\mathbf{u}_i \in \mathbb{R}^3$ with the unit norm and joint state $s_i \in \mathbb{R}^+$. Additionally, the "revolute" part has the pivot point $\mathbf{q}_i \in \mathbb{R}^3$. We refer to the joint direction and pivot point as the joint configuration. For the "fixed" part, we set B_i as an identity matrix because no transformation is applied. For the "prismatic" part, we define $B_i = T(s_i \mathbf{u}_i)$, where $T(\cdot)$ represents a homogeneous translation matrix given the translation in $\mathbb{R}^3$, and s_i and $\mathbf{u}_i$ represent the translation amount and direction, respectively. For the "revolute" part, we set $B_i = T(\mathbf{q}_i)R(s_i, \mathbf{u}_i)$, where $R(\cdot)$ denotes a homogeneous rotation matrix given the rotation representation, and s_i and $\mathbf{u}_i$ represent the axis-angle rotation around the axis $\mathbf{u}_i$ by angle s_i. In human shape reconstruction methods using template shape, its pose is initialized to be close to the real distribution to avoid the local minima [14, 19]. Inspired by these approaches, we parametrize the joint direction as $[\mathbf{u}_i; 1] = R(\mathbf{r}_i)[\mathbf{e}_i; 1]$, where $\mathbf{e}_i$ is a constant directional vector with the unit norm working as the initial joint direction as a hyperparameter and $\mathbf{r}_i \in \mathbb{R}^3$ represents the Euler-angle representation working as a residual from the initial joint direction $\mathbf{e}_i$. This allows us to manually initialize the joint direction in a realistic distribution through $\mathbf{e}_i$ by initializing $\mathbf{r}_i = \mathbf{0}$. Figure 4 illustrates the joint parameters.

Based on our observations, we assume that the joint configuration has a category-common bias, while the joint state strongly depends on each instance. This is because the location of each part and the entire shape of an object can constrain the possible trajectory of the parts, which is defined by the joint configuration. To illustrate this idea, we propose to decompose the joint configuration into a category-common bias term and an instance-dependent residual term denoted as $\mathbf{r}_i = \mathbf{r}_i^c + \mathbf{r}_i^z$ and $\mathbf{q}_i = \mathbf{q}_i^c + \mathbf{q}_i^z$, respectively. We employ the category-common pose decoder $F^c(\mathrm{qt}(\boldsymbol{\psi}))$, which outputs $\{\mathbf{r}_i^c \mid i \in \mathbb{A}^p\}$ and $\{\mathbf{q}_i^c \mid i \in \mathbb{A}^r\}$, where $\mathbb{A}^p = \{i \in [N] \mid y_i \neq \text{fixed}\}$, $\mathbb{A}^r = \{i \in [N] \mid y_i = \text{revolute}\}$,

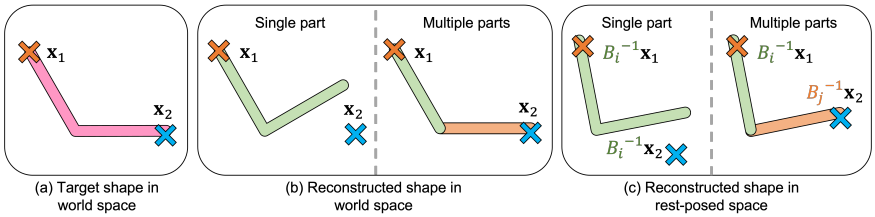

Fig. 3. Illustration of part decomposition induction. "Single part" indicates that the model is degenerated to use only a single part to reconstruct the whole target shape. "Multiple parts" indicates that part decomposition is correctly induced. In (b), the "single part" model misclassifies query point $\mathbf{x}_2$ as outside, in contrast to $\mathbf{x}_1$. As shown in (c), a single part pose $\{B_i\}$ cannot correctly transform both query points inside the rest-posed shape. The "multiple parts" model successfully classifies both query points using different part poses per part. Minimizing the reconstruction loss incentivizes the model to use multiple parts and appropriate part types for $\{B_i\}$.

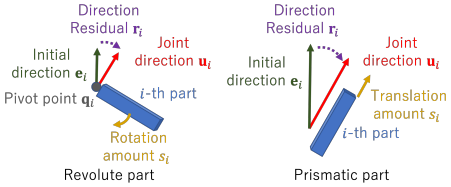

Fig. 4. Geometric relationship between the joint parameters.

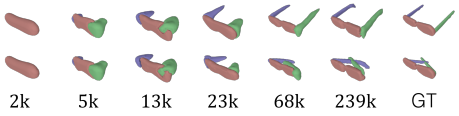

Fig. 5. Visualization of the training process. The numbers in the figure show the training steps.

$\boldsymbol{\psi}$ denotes a pose latent vector, and $\mathrm{qt}(\cdot)$ is a latent vector quantization operator following VQ-VAE [32]. The operator $\mathrm{qt}(\cdot)$ outputs the nearest constant vector to the input latent vector $\boldsymbol{\psi}$ among the N_{qt} candidates. Instead of using a single constant vector, the model selects a constant vector among multiple constant vectors to capture the discrete, multi-modal category-common biases. We also employ an instance-dependent pose decoder $F^z(\boldsymbol{\theta})$ that outputs $\{s_i \mid i \in \mathbb{A}^p\}$, $\{\mathbf{r}_i^z \mid i \in \mathbb{A}^p\}$, and $\{\mathbf{q}_i^z \mid i \in \mathbb{A}^r\}$. We constrain the possible distribution of the joint configuration around the category-common bias by the loss function explained in Section 3.3. This constraint incentivizes the model to reconstruct the instance-dependent shape variation by the joint state, which constrains the part location along the joint direction. This kinematic constraint biases the model to represent the shapes having the same kinematics with the same part. The previous works [15, 9, 29] do not impose such a constraint on the part localization, thus learned part decomposition is not necessarily consistent under different poses.

3.2 Part shape representation

We propose a non-primitive-based part shape representation that is decomposed into the category-common shape prior and instance-dependent shape details. We employ MLP-based decoders to model a part-wise implicit field. We capture the category-common shape prior using the category-common shape decoder $G_i^c(\mathbf{x})$. Because G_i^c does not take a latent vector from the encoder, it learns an input-independent, rest-posed part shape template as the category-common

shape prior. We also employ an instance-dependent shape decoder $G_i^z(\mathbf{x} \mid \boldsymbol{\phi})$ to capture the additional instance-dependent shape details conditioned with the shape prior. We formulate a part-wise implicit field $\hat{O}_i$ as follows:

$$\hat{O}_i(\mathbf{x} \mid I) = \sigma(G_i^z(\mathbf{x}, \boldsymbol{\phi})\hat{O}_i^c(\mathbf{x})) \tag{1}$$

where $\sigma(\cdot)$ represents the sigmoid function and $\hat{O}_i^c(\mathbf{x}) = \sigma(G_i^c(\mathbf{x}))$. For brevity, we omit I in $\hat{O}_i$ and simply denote it as $\hat{O}_i(\mathbf{x})$. Given the part poses $\{B_i\}$ as part-wise locally rigid deformation, we formulate $\hat{O}$ as the composition of $\{\hat{O}_i\}$ defined as $\hat{O}(\mathbf{x} \mid I, \{B_i\}) = \max_i\{\hat{O}_i(B_i^{-1}\mathbf{x})\}$. As in the piecewise rigid model of [10], coordinate transformation $B_i^{-1}\mathbf{x}$ realizes locally rigid deformation by B_i of the part-wise implicit field by querying the rest-posed indicator. Note that, although we set the maximum number of parts N, the actual number of parts used for reconstruction can change; it is possible that some parts do not contribute to the reconstruction because of the max operation or simply because $\hat{O}_i < 0.5$ for all 3D locations.

In Equation 1, we experimentally found that conditioning G_i^z by $\hat{O}_i^c$ through multiplication rather than addition effectively prevents G_i^z from deviating largely from G_i^c. This conditioning induces the unsuperivsed part decomposition. We illustrate the idea in Figure 3. Considering reconstructing the target shape by single i-th part, since the multiplication makes it difficult to output shapes that deviating largely from the category-common prior shape, the large shape variations of target shapes are expressed by B_i regarded as the global pose of the reconstructed shape. However, the large shape variations in target shapes are due to the various local poses of multiple part shapes. Therefore, the large shape variations of target shapes cannot be expressed only by the single part and its part pose B_i. Thus, as an inductive bias of the unsupervised part decomposition, the model is incentivized to use a composition of multiple parts to express the shape variations due to various local part poses. We visualize the learning process in Figure 5. First, the model learns high indicator values in spatial locations of static parts with high probabilities of space occupancy in any instance. Next, part decomposition is induced to accommodate various target shapes' part poses, generating multiple dynamic parts. Indicator values in the spatial locations with less displacement by different part poses (e.g., near pivot points of revolute parts) first exceed the iso-surface threshold. Then, the model simultaneously optimizes part pose estimation and shape reconstruction during training as an analysis-by-synthesis approach.

3.3 Training losses

Shape losses. To learn the shape decoders, we minimize the reconstruction loss using the standard binary cross-entropy loss (BCE) defined as:

$$\mathcal{L}_{\text{reconstruction}} = \lambda_{\text{reconstruction}}\text{BCE}(\hat{O}, O) + \lambda^c_{\text{reconstruction}}\text{BCE}(\hat{O}^c, O) \tag{2}$$

where $\hat{O}^c(\mathbf{x} \mid B) = \max_i\{\hat{O}_i^c(B_i^{-1}\mathbf{x})\}$, and $\lambda_{\text{reconstruction}}$ and $\lambda^c_{\text{reconstruction}}$ are the loss weights. The second term in Equation 2 is essential for stable training;

it facilitates fast learning of $\{G_i^c\}$, so that $\{G_i^z\}$ can be correctly conditioned in the early stage of the training process. Moreover, because we consider the locally rigid deformation of the shape, the volumes of the shape before and after the deformation should not be changed by the intersection of parts; we formulate this constraint as follows:

$$\begin{aligned}\mathcal{L}_{\text{volume}} = \lambda_{\text{volume}} \Bigg(& \mathbb{E}_{\mathbf{x}} \Big[\text{ReLU}(\max_i \{G_i^z(B_i^{-1}\mathbf{x}, \boldsymbol{\phi})\}) \Big] \\ & - \mathbb{E}_{\mathbf{x}} \Big[\text{ReLU}(\max_i \{G_i^z(\mathbf{x}, \boldsymbol{\phi})\}) \Big] \Bigg)^2 \end{aligned} \tag{3}$$

Joint parameter losses. For the joint parameters $\mathbf{q}_i$ and $\mathbf{r}_i$, we prevent an instance-dependent term from deviating too much from the bias term, we regularize them by the loss:

$$\mathcal{L}_{\text{deviation}} = \lambda_{\text{deviation}} \left(\frac{1}{N^r} \sum_{i \in \mathbb{A}^r} \|\mathbf{q}_i^z\| + \frac{1}{N^p} \sum_{i \in \mathbb{A}^p} \|\mathbf{r}_i^z\| \right) \tag{4}$$

where $N^r = |\mathbb{A}^r|$, $N^p = |\mathbb{A}^p|$, and $\lambda_{\text{deviation}}$ is the loss weight. Moreover, we propose a novel regularization loss that constrains the pivot point with the implicit fields. We assume that the line in 3D space, which consists of the pivot point and joint direction, passes through the reconstructed shape. The joint should connect at least two parts, which means that the joint direction anchored by the pivot point passes through at least two reconstructed parts. We realize this condition as follows:

$$\mathcal{L}_{\text{pivot}} = \frac{\lambda_{\text{pivot}}}{N^r} \sum_{i \in \mathbb{A}^r} \left(\min_{\mathbf{x} \in \mathbb{S}_{\text{GT}}} \|\mathbf{q}_i - \mathbf{x}\| + \frac{1}{2} \Big(\min_{\mathbf{x} \in \mathbb{S}_i} \|\mathbf{q}_i - \mathbf{x}\| + \min_{\mathbf{x} \in \mathbb{S}_{i,j}} \|\mathbf{q}_i - \mathbf{x}\| \Big) \right) \tag{5}$$

where $\mathbb{S}_{\text{GT}} = \{\mathbf{x} \in \mathbb{R}^3 \mid O(\mathbf{x}) = 1\}$, $\mathbb{S}_i = \{\mathbf{x} \in \mathbb{R}^3 \mid \hat{O}_i(B_i^{-1}\mathbf{x}) > 0.5\}$, $\mathbb{S}_{i,j} = \{\mathbf{x} \in \mathbb{R}^3 \mid \hat{O}_j(B_j^{-1}\mathbf{x}) > 0.5,\ j \in \mathbb{A}^r \setminus i\}$, and λ_{pivot} is the loss weight. Note that $\mathcal{L}_{\text{pivot}}$ is self-regularizing and not supervised by the ground-truth part segmentation. See supplementary material for an illustration of $\mathcal{L}_{\text{pivot}}$ and further details. To reflect the diverse part poses, we prevent the joint state s_i from degenerating into a static state. In addition, to prevent multiple decomposed parts from representing the same revolute part, we encourage the pivot points to be spatially spread. We realize these requirements by the loss defined as:

$$\mathcal{L}_{\text{variation}} = \frac{1}{N^p} \sum_{i \in \mathbb{A}^p} \left(\frac{\lambda_{\text{variation}^{\text{s}}}}{\text{std}(s_i)} + \lambda_{\text{variation}^{\text{q}}} \sum_{j \in \mathbb{A}^r \setminus i} \exp\Big(- \frac{\|\mathbf{q}_i - \mathbf{q}_j\|}{v} \Big) \right) \tag{6}$$

where $\text{std}(\cdot)$ denotes the batch statistics of the standard deviation, v is a constant that controls the distance between pivot points, and $\lambda_{\text{variation}^{\text{s}}}$ and $\lambda_{\text{variation}^{\text{q}}}$ are the loss weights. Lastly, following the loss proposed in [32], the pose latent vector $\boldsymbol{\psi}$ is optimized by the loss:

$$\mathcal{L}_{\text{quantization}} = \|\boldsymbol{\psi} - \text{sg}(\text{qt}(\boldsymbol{\psi}))\| \tag{7}$$

where sg denotes an operator stopping gradient on the backpropagation.

Adversarial losses. Inspired by human shape reconstruction studies [6, 30], we employ the adversarial losses from WGAN-GP [12] to regularize the shape and pose in the realistic distribution. The losses are defined as:

$$\begin{aligned}\mathcal{L}_{\text{discriminator}} = \lambda_{\text{discriminator}}\Big(&\mathbb{E}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_g}[D(\tilde{\boldsymbol{x}})] - \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_r}[D(\boldsymbol{x})]\Big) \\ &+ \mathbb{E}_{\hat{\boldsymbol{x}}\sim\mathbb{P}_{\hat{\boldsymbol{x}}}}[(\|\nabla_{\hat{\boldsymbol{x}}}D(\hat{\boldsymbol{x}})\| - 1)^2]\end{aligned} \tag{8}$$

$$\mathcal{L}_{\text{generator}} = -\lambda_{\text{generator}}\mathbb{E}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_g}[D(\tilde{\boldsymbol{x}})] \tag{9}$$

where $D(\cdot)$ is a discriminator; $\tilde{\boldsymbol{x}}$ is a sample from the reconstructed shapes $\mathbb{P}_g$ transformed by the estimated joint configuration and randomly sampled joint state $\tilde{s}_i \sim \text{Uniform}(0, h_i)$, with the maximum motion amount h_i treated as a hyperparameter; $\boldsymbol{x}$ is a sample from the ground-truth shapes $\mathbb{P}_r$; $\hat{\boldsymbol{x}}$ is a sample from $\mathbb{P}_{\hat{\boldsymbol{x}}}$, which is a set of randomly and linearly interpolated samples between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}$; and $\lambda_{\text{generator}}$ and $\lambda_{\text{discriminator}}$ are the loss weights. As an input to D, we concatenate the implicit field and corresponding 3D points to create a 4D point cloud, following [17].

3.4 Implementation details

We use the Adam solvers [16] with a learning rate of 0.0001 with a batch size of 18 to optimize the sum of the losses: $\mathcal{L}_{\text{reconstruction}} + \mathcal{L}_{\text{volume}} + \mathcal{L}_{\text{quantization}} + \mathcal{L}_{\text{deviation}} + \mathcal{L}_{\text{pivot}} + \mathcal{L}_{\text{variation}} + \mathcal{L}_{\text{generator}}$ and the discriminator loss $\mathcal{L}_{\text{discriminator}}$, respectively. We use the complete shape point cloud with 4096 points sampled from the surface of the target shape as input, unless otherwise noted. We use 4096 coordinate points and their corresponding indicator values for the ground-truth implicit field. We set the loss weights as follows: $\lambda_{\text{reconstruction}} = 0.01$, $\lambda^{\text{c}}_{\text{reconstruction}} = 0.001$, $\lambda_{\text{deviation}} = 0.1$, $\lambda_{\text{pivot}} = 100$, $\lambda_{\text{variation}^{\text{s}}} = 0.1$, $\lambda_{\text{variation}^{\text{q}}} = 0.01$, $\lambda_{\text{volume}} = 1000$, $\lambda_{\text{generator}} = 0.65$, and $\lambda_{\text{discriminator}} = 0.35$. We set $v = 0.01$ in $\mathcal{L}_{\text{variation}}$ and $N_{qt} = 4$ for $\text{qt}(\cdot)$. For h_i in $\mathcal{L}_{\text{discriminator}}$, we set to $\frac{\pi}{2}$ and 0.4 the "revolute" and "prismatic" parts, respectively. Note that we experimentally found that it does not constrain the model to predict s_i larger than h_i to reconstruct the target shape. Because we do not impose any geometric constraints on the part shapes, we set the number of parts for each part kinematics y_i as its maximum number in the datasets plus an additional one part for over-parameterization. The detail of the datasets is explained in Section 4. We set $N = 8$, which consists of one "fixed" part, three "revolute" parts, and four "prismatic" parts. We use the same hyperparameter for all categories, without assuming the category-specific knowledge. During the training, the max operation is substituted with LogSumExp for gradient propagation to each shape decoder. See supplementary material for further training details.

Network architecture. We use the PointNet [31]-based architecture from [23] as an encoder E and the one from [35] as a discriminator D. Our shape decoders $\{G_i^c\}$ and $\{G_i^z\}$ are MLP with sine activation [36] for a uniform activation magnitude suitable for propagating gradients to each shape decoder. For the category-common pose decoder F^c, we use separate networks of MLP for each kind of output variables. For the instance-dependent pose decoder F^z, we employ MLP with a single backbone having multiple output branches. The detailed network architecture can be found in supplementary material.

4 Experiments

Datasets. In our evaluation, we follow the recent part pose estimation studies targeting man-made articulated objects for the synthetic datasets and the object categories covering various part kinematics: oven, eyeglasses, laptop, and washing machine categories from Motion dataset [40], and the drawer category from SAPIEN dataset [41]. Each category has a fixed number of parts with the same kinematic structure. We generate 100 instances with different poses per sample, generating 24k instances in total. We divide the samples into the training and test sets with a ratio of approximately 8:2. We also normalize the side length of samples to 1, following [23]. Further details can be found in supplementary material. To verify the transferability of our approach trained on synthetic data to real data, we use the laptop category from RBO dataset [22] and Articulated Object Dataset [24], which is the intersecting category with the synthetic dataset.

Baselines. We compare our method with the state-of-the-art unsupervised generative part decomposition methods with various characteristics: BAE-Net [7] (non-primitive-based part shape representation), BSP-Net [7] (primitive-based part shape representation with part localization by 3D space partitioning), NSD [15] and Neural Parts [28] denoted as NP (primitive-based part shape representation with part localization in $\mathbb{R}^3$). For BSP-Net, we train up to 32^3 grids of the implicit field instead of 64^3 grids in the original implementation to match those used by other methods. For NSD and Neural Parts, we replace its image encoder with the same PointNet-based encoder in our approach. For the part pose estimation, we use NPCS [20] as the supervised baseline. NPCS performs part-based registration by iterative rigid-body transformation, which is a common practice in articulated pose estimation of rigid objects. See supplementary material for further training details of the baselines.

Metrics. For the quantitative evaluation of the consistent part parsing as a part segmentation task, we use the standard label IoU, following the previous studies [8, 7, 9, 15]. As our method is unsupervised, we follow the standard initial part labeling procedure using a training set to assign each part a ground-truth label for evaluation purposes following [9, 15]. A detailed step can be found in supplementary material. For the part pose evaluation, we evaluate the 3D motion

	Drawer	Eye-glasses	Oven	Laptop	Washing machine	mean	# of parts
BAE [8]	6.25*	11.11*	73.06	25.11*	80.30	39.17	1.42/8
BSP [7]	66.31	**70.69**	81.65	76.68	87.92	76.65	27.50/256
NSD [15]	38.39	42.11	74.67	74.44	89.11	63.75	10
NP [28]	60.57	64.69	**85.41**	86.23	74.65	74.31	5
Ours	**74.73**	66.18	82.07	**86.81**	**95.15**	**80.99**	4.16/8

Table 2. Part segmentation performance in label IoU. Higher is better. The starred numbers indicate the failure of part decomposition and that only one recovered part represents the entire shape. The average and the predefined maximum numbers of recovered parts or primitives are shown before and after the slash, in the last column.

flow of the deformation from the canonical pose to the predicted pose as the endpoint error (EPE) [43], which is a commonly used metric for pose estimation of articulated objects [40, 5]. We scale it by 100 in experiment results.

4.1 Semantic capability

We evaluate the semantic capability of our approach in part parsing. As part decomposition approaches aim to learn 3D structure reasoning with *as small a number of ground-truth labels as possible*, it is preferable to obtain the initial manual annotations with as *few numbers of shapes* as possible. This requirement is essential for articulated objects, which have diverse shape variations owing to the different articulations. As our approach is part pose consistent, we only need a minimal variety of instances for the initial manual labeling. To verify this, we evaluate the part segmentation performance using only the canonically posed (joint states were all zero) samples in the training set. See supplementary material for further studies on pose variation for the initial annotation.

The evaluation results are shown in Table 2. Our model uses a much smaller number of parts than BSP-Net [7]; however, it still performs the best. This shows that our model is more parsimonious, and each part has more semantic meaning in part parsing. The segmentation results are visualized in Figure 6. To eliminate differences in the number of parts and primitives for each method, Table 3 shows the result when each method's maximum number of parts and primitives is aligned to $N = 8$. Our method outperforms the previous works by a large margin. For visualization procedure and additional results of our part segmentation result, see supplementary material.

We also visualize the generated part shapes in Figure 7. We can see that a single part shape successfully reconstructs the complex target shape, such as disconnected shapes that a single primitive shape cannot express. Also, our part shapes are more semantic and interpretable than the previous works. This demonstrates the advantage of using non-primitive-based part shape representation. As we can see in the improved part segmentation performance, our approach realizes semantically more consistent part decomposition without a complicated mechanism such as grouping primitive shapes based on part kinematics.

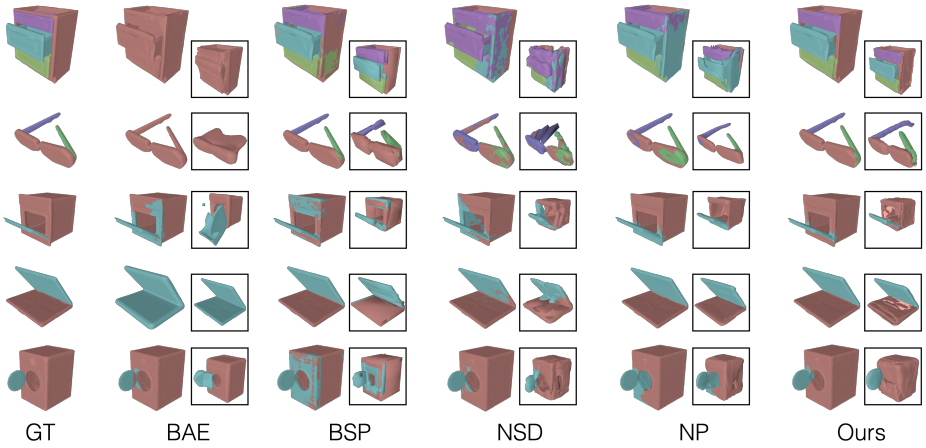

Fig. 6. Visualization of the part segmentation. Reconstructed shape in mesh is shown inside a box. The same color indicates the same segmentation part.

	Label IoU ↑
BAE [8]	39.17
BSP [7]	66.79
NSD [15]	59.46
NP [28]	70.71
Ours	**80.99**

Table 3. Label IoU with the aligned number of primitives and parts for all methods ($N = 8$).

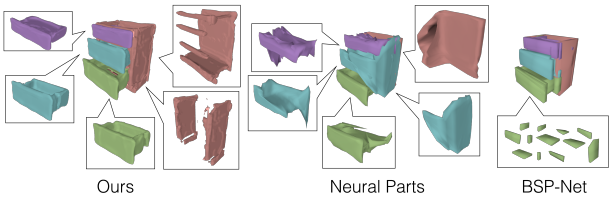

Fig. 7. Visualization of parts and primitives. The boxes represent the parts or primitives used to reconstruct the semantic parts.

Disentanglement between the part shapes and poses. Because our approach disentangles shape supervision into part shapes and poses, it realizes pose-aware part decomposition. To verify the learned disentanglement, we visualize the interpolation results of part shapes and joint states as part poses in Figure 8. In the middle row, we show the shape interpolation between the source and the target while fixing the joint state s_i of the source to maintain the same part pose. The shape is smoothly deformed from the source to the target maintaining the original pose. In the bottom row, we interpolate the joint state s_i between the source and the target; the joint state changes from the source to the target maintaining the shape identity of the source shape. Our model successfully disentangles the part shapes and poses, unlike previous methods as shown in the top row.

4.2 Part pose estimation

To validate whether the predicted part decomposition is based on the reasonable part pose estimation, we quantitatively evaluate part pose estimation. Because we train our model without specifying a canonically posed shape, we use

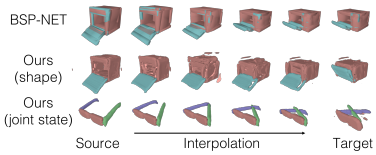

Fig. 8. Interpolation in terms of disentangled part shapes and joint states as part pose.

	Drawer	Eye-glasses	Oven	Laptop	Washing machine	mean
NPCS [20] (Supervised)	1.598	1.087	2.702	0.751	1.594	1.546
Ours (Unsupervised)	3.452	2.631	3.360	2.546	2.529	2.903

Table 4. Part pose estimation performance in EPE. Lower is better. NPCS is trained with ground-truth for both part labels and part-wise rigid-body transformations as part pose, offering an upper bound for our unsupervised approach.

	$\mathcal{L}_{\text{volume}}$	$\mathcal{L}_{\text{deviation}}$	$\mathcal{L}_{\text{pivot}}$	$\mathcal{L}_{\text{variation}}$	$\mathcal{L}_{\text{generator}}$	VQ	CS	CP	Full
Label IoU ↑	72.20	73.21	74.27	65.29	70.14	72.78	55.67	71.35	**80.99**
EPE ↓	4.362	6.628	9.250	6.676	7.276	10.772	8.827	7.219	**2.988**

Table 5. Ablation study of the losses and the proposed components: VQ, CP and CS indicates disabling the use of multiple constant vectors introduced in Section 3.1, the category-common pose decoder, and the category-common shape decoders, respectively. "Full" means using all the losses and the components.

the part pose transformations between the target instance and the canonically posed instance of the same sample as the estimated part pose to align with the prediction of the supervised baseline, NPCS [20]. Note that NPCS assumes that part segmentation supervision and ground-truth of part-wise rigid-body transformations as part pose are available during training, and part kinematic type per part is known, which we do not assume. Therefore, NPCS offers an upper bound for our unsupervised approach. We present the evaluation results in Table 4. Our method is comparable with NPCS, with the same order of performance. Note again that we are not attempting to outperform supervised pose estimation methods; rather, we aim to show that our unsupervised approach can decompose parts based on reasonable part pose estimation. See supplementary material for further discussion on part pose estimation.

4.3 Ablation studies

We evaluate the effect of the proposed losses, the multiple constant vectors for multi-modal category-common pose bias learning, and the category-common decoders on part segmentation and part pose estimation. We disable each loss and component one at a time. We only use the corresponding instance-dependent decoder(s) when disabling the category-common decoders for pose and shape. The results are shown in Table 5. Enabling all losses and the components performs the best. Particularly, disabling the category-common shape decoders significantly degrades both label IoU and EPE. This indicates that learning category-common shape prior is essential to perform proper part decomposition and to facilitate part pose learning, which is the core idea of this study.

	Label IoU ↑	EPE ↓
Complete	80.99	2.903
Depth	80.65	3.203

Table 6. Comparison between the point cloud input types: complete shape and depth map.

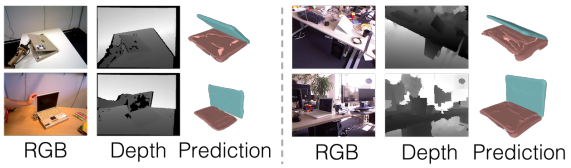

Fig. 9. Real depth map input. (Left) RBO dataset [22] and (Right) Articulated Object Dataset [24].

4.4 Depth map input and real data

Because PPD's decoders do not assume a complete shape as an input, it works with depth map input. Following BSP-Net [7], we train a new encoder that takes a depth map captured from various viewpoints as a partial point cloud and replace the original encoder. We minimize the mean squared error between the output latent vectors of the original and the new encoders so that their output are close for the same target shape. The results are shown in Table 6. The depth map input performs comparably to the complete point cloud input. We also verify that our model trained on synthetic depth maps reasonably generalizes to real data, as shown in Figure 9.

5 Conclusion

We propose a novel unsupervised generative part decomposition method, PPD, for man-made articulated objects considering part kinematics. We show that the proposed method learns the disentangled representation of the part-wise implicit field as the decomposed part shapes and the joint parameters of each part as the part poses. We also show that our approach outperforms previous generative part decomposition methods in terms of semantic capability and show comparable part pose estimation performance with the supervised baseline.

As shown in qualitative results, our generative method achieves reasonable part shape reconstruction reflecting target shape variations sufficient to induce part decomposition and challenging joint parameter learning. As a limitation, our method currently fails to capture details of the target shapes up to the primitive-based previous works [15, 7], focusing on the shape reconstruction performance rather than part pose consistency. Also, joint parameter learning requires manual initialization of joint direction and part types for each part. The future work will address the above limitation.

Acknowledgements. We would like to thank Atsuhiro Noguchi, Hao-Wei Yeh, Haruo Fujiwara, Qier Meng, Tomu Hirata, Yang Li, and Yusuke Kurose for their insightful feedback. We also appreciate the members of the Machine Intelligence Laboratory for constructive discussion during the research meetings. This work was partially supported by JST AIP Acceleration Research JPMJCR20U3, Moonshot R&D JPMJPS2011, CREST JPMJCR2015, JSPS KAKENHI JP19H01115, and Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo.

References

1. Abbatematteo, B., Tellex, S., Konidaris, G.: Learning to generalize kinematic models to novel objects. In: Proceedings of the Conference on Robot Learning (CoRL). pp. 1289–1299 (2020)
2. Becker, S., Hinton, G.E.: Self-organizing neural network that discovers surfaces in random-dot stereograms. Nature **355**(6356), 161–163 (1992)
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 561–578 (2016)
4. Bonanni, L., Lee, C.H., Selker, T.: Counterintelligence: Augmented reality kitchen. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). vol. 2239, p. 45 (2005)
5. Božič, A., Palafox, P., Zollhöfer, M., Thies, J., Dai, A., Nießner, M.: Neural deformation graphs for globally-consistent non-rigid reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1450–1459 (2021)
6. Chen, C.H., Tyagi, A., Agrawal, A., Drover, D., Stojanov, S., Rehg, J.M.: Unsupervised 3d pose estimation with geometric self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5714–5724 (2019)
7. Chen, Z., Tagliasacchi, A., Zhang, H.: Bsp-net: Generating compact meshes via binary space partitioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 45–54 (2020)
8. Chen, Z., Yin, K., Fisher, M., Chaudhuri, S., Zhang, H.: Bae-net: branched autoencoder for shape co-segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR). pp. 8490–8499 (2019)
9. Deng, B., Genova, K., Yazdani, S., Bouaziz, S., Hinton, G., Tagliasacchi, A.: Cvxnet: Learnable convex decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 31–44 (2020)
10. Deng, B., Lewis, J.P., Jeruzalski, T., Pons-Moll, G., Hinton, G., Norouzi, M., Tagliasacchi, A.: Nasa neural articulated shape approximation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 612–628 (2020)
11. Desingh, K., Lu, S., Opipari, A., Jenkins, O.C.: Factored pose estimation of articulated objects using efficient nonparametric belief propagation. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 7221–7227 (2019). https://doi.org/10.1109/ICRA.2019.8793973
12. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems (NeurIPS). p. 5769–5779 (2017)
13. Huang, J., Wang, H., Birdal, T., Sung, M., Arrigoni, F., Hu, S.M., Guibas, L.J.: Multibodysync: Multi-body segmentation and motion estimation via 3d scan synchronization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7108–7118 (2021)
14. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7122–7131 (2018)
15. Kawana, Y., Mukuta, Y., Harada, T.: Neural star domain as primitive representation. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 7875–7886 (2020)

16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Kleineberg, M., Fey, M., Weichert, F.: Adversarial generation of continuous implicit shape representations. In: Eurographics. pp. 41–44 (2020)
18. Kourtzi, Z., Kanwisher, N.: Activation in human mt/mst by static images with implied motion. Journal of Cognitive Neuroscience **12**(1), 48–55 (2000)
19. Kulkarni, N., Gupta, A., Fouhey, D.F., Tulsiani, S.: Articulation-aware canonical surface mapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 452–461 (2020)
20. Li, X., Wang, H., Yi, L., Guibas, L.J., Abbott, A.L., Song, S.: Category-level articulated object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3706–3715 (2020)
21. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015)
22. Martín-Martín, R., Eppner, C., Brock, O.: The rbo dataset of articulated objects and interactions. arXiv preprint arXiv:1806.06465 (2018)
23. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4460–4470 (2019)
24. Michel, F., Krull, A., Brachmann, E., Ying Yang, M., Gumhold, S., Rother, C.: Pose estimation of kinematic chain instances via object coordinate regression. In: Proceedings of the British Machine Vision Conference (BMVC). pp. 181.1–181.11 (2015)
25. Mo, K., Guerrero, P., Yi, L., Su, H., Wonka, P., Mitra, N.J., Guibas, L.J.: Structedit: Learning structural shape variations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8859–8868 (2020)
26. Mu, J., Qiu, W., Kortylewski, A., Yuille, A., Vasconcelos, N., Wang, X.: A-sdf: Learning disentangled signed distance functions for articulated shape representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13001–13011 (2021)
27. Paschalidou, D., van Gool, L., Geiger, A.: Learning unsupervised hierarchical part decomposition of 3d objects from a single rgb image. In: Proceedings IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
28. Paschalidou, D., Katharopoulos, A., Geiger, A., Fidler, S.: Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3204–3215 (2021)
29. Paschalidou, D., Ulusoy, A.O., Geiger, A.: Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10344–10353 (2019)
30. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7753–7762 (2019)
31. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 652–660 (2017)

32. Razavi, A., Oord, A.v.d., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. arXiv preprint arXiv:1906.00446 (2019)
33. Roberts, D., Danielyan, A., Chu, H., Golparvar-Fard, M., Forsyth, D.: Lsd-structurenet: Modeling levels of structural detail in 3d part hierarchies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5836–5845 (2021)
34. Shirai, N., Imura, T.: Implied motion perception from a still image in infancy. Experimental Brain Research **232**(10), 3079–3087 (2014)
35. Shu, D.W., Park, S.W., Kwon, J.: 3d point cloud generative adversarial network based on tree structured graph convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
36. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 7462–7473 (2020)
37. Slater, A., Morison, V., Town, C., Rose, D.: Movement perception and identity constancy in the new-born baby. British Journal of Developmental Psychology **3**(3), 211–220 (1985)
38. Spelke, E.S., Kestenbaum, R., Simons, D.J., Wein, D.: Spatiotemporal continuity, smoothness of motion and object identity in infancy. British Journal of Developmental Psychology **13**(2), 113–142 (1995)
39. Tulsiani, S., Su, H., Guibas, L.J., Efros, A.A., Malik, J.: Learning shape abstractions by assembling volumetric primitives. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2635–2643 (2017)
40. Wang, X., Zhou, B., Shi, Y., Chen, X., Zhao, Q., Xu, K.: Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8876–8884 (2019)
41. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., et al.: Sapien: A simulated part-based interactive environment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11097–11107 (2020)
42. Xu, F., Carey, S.: Infants' metaphysics: The case of numerical identity. Cognitive psychology **30**(2), 111–153 (1996)
43. Yan, Z., Xiang, X.: Scene flow estimation: A survey. arXiv preprint arXiv:1612.02590 (2016)
44. Yi, L., Huang, H., Liu, D., Kalogerakis, E., Su, H., Guibas, L.: Deep part induction from articulated object pairs. ACM Transactions on Graphics (TOG) **37**(6), 1–15 (2018)
45. Zuffi, S., Kanazawa, A., Berger-Wolf, T., Black, M.J.: Three-d safari: Learning to estimate zebra pose, shape, and texture from images" in the wild". In: Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR). pp. 5359–5368 (2019)
46. Zuffi, S., Kanazawa, A., Jacobs, D.W., Black, M.J.: 3d menagerie: Modeling the 3d shape and pose of animals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6365–6373 (2017)