

Unsupervised 3D Shape Reconstruction by Part Retrieval and Assembly

Xianghao Xu
Brown University
USA

Paul Guerrero
Adobe Research
UK

Matthew Fisher
Adobe Research
USA

Siddhartha Chaudhuri
Adobe Research
India

Daniel Ritchie
Brown University
USA

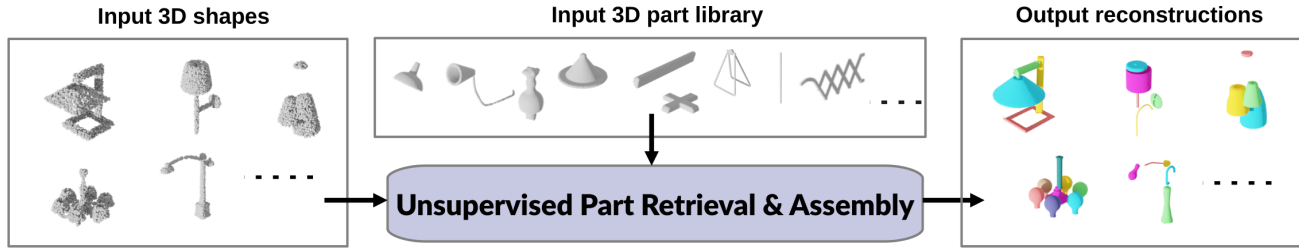


Figure 1. Our system takes target 3D shapes together with a 3D part library as input and outputs a set of retrieved and transformed parts from the part library that recreates the input target shapes.

Abstract

Representing a 3D shape with a set of primitives can aid perception of structure, improve robotic object manipulation, and enable editing, stylization, and compression of 3D shapes. Existing methods either use simple parametric primitives or learn a generative shape space of parts. Both have limitations: parametric primitives lead to coarse approximations, while learned parts offer too little control over the decomposition. We instead propose to decompose shapes using a library of 3D parts provided by the user, giving full control over the choice of parts. The library can contain parts with high-quality geometry that are suitable for a given category, resulting in meaningful decompositions with clean geometry. The type of decomposition can also be controlled through the choice of parts in the library. Our method works via a self-supervised approach that iteratively retrieves parts from the library and refines their placements. We show that this approach gives higher reconstruction accuracy and more desirable decompositions than existing approaches. Additionally, we show how the decomposition can be controlled through the part library by using different part libraries to reconstruct the same shapes.

1. Introduction

The ability to compactly represent a 3D shape as a combination of primitive elements has applications in multiple domains. In computer vision, the ability to automatically

decompose a shape into parts can aid machine perception of the 3D structure of objects, which can in turn help autonomous agents plan how to manipulate such objects. In computer graphics, a combination of primitives can be used as a compressed geometry representation, as a way to abstract, stylize, or edit a 3D shape by allowing users to alter the underlying primitive library. Ideally, a system that performs this kind of shape decomposition should be able to do so without supervision in the form of ground-truth decompositions, as such data is rarely available at scale.

Past research in vision and graphics has studied this unsupervised shape decomposition problem. Initially, researchers sought methods for decomposing 3D shapes into sets of simple parametric primitives, such as cuboids or superquadric surfaces [20, 23, 25, 28]. These methods produce clean, parametric geometry as output, and the choice of primitive type allows a small degree of user control over the decomposition. However, parametric primitives produce only a coarse approximation of the input shape, which may not be desirable in all applications. Thus, more recent work has investigated unsupervised decomposition of shapes into arbitrarily-shaped primitives whose geometries are determined by a neural network [4, 11, 19]. These methods produce a set of “neural primitives” whose union closely approximates the input shape. However, the geometries of these primitives may contain artifacts (e.g. bumps, blobs). Further, these methods offer little to no control over the type of decomposition they produce – the network outputs whatever primitives it thinks are best to reconstruct the input

shape since it lacks access to a supervised part prior.

Is it possible to obtain a decomposition of a 3D shape whose primitives exhibit clean geometry and closely reconstruct the input shape, while also providing more control over the type of decomposition produced? This is possible if, rather than using simple parametric primitives or arbitrary neural primitives, one chooses a middle point between these two extremes: reconstruct an input shape by retrieving and assembling primitives from a *library of pre-defined 3D parts*. This retrieve-and-assemble approach has several advantages. First, the parts in the library can be high-quality 3D meshes, guaranteeing clean geometry as output. Second, a large part library can contain parts that are good geometric matches for different regions of various shapes, meaning that accurate reconstructions of input shapes are possible. Finally, this approach offers a high degree of controllability, as the user can change the part library to produce different decompositions of the same input shape.

In this paper, we present a method for unsupervised decomposition of 3D shapes using a user-defined library of parts. Finding a subset of parts from a large part library which best reconstructs an input shape is a large-scale combinatorial search problem. To make this problem tractable, we represent the library of parts on a continuous manifold by training a part autoencoder. This continuous representation of the part library allows us jointly optimize for the identities and poses of parts which reconstruct the input shape. To escape the many worst of local optimas in this optimization landscape, the algorithm periodically uses its current predicted set of parts to segment the input shape; these segments are then re-encoded into the part feature manifold to produce a new estimate of the parts that best reconstruct the input shape. This data-driven, discontinuous jump in the optimization state is similar to stages from other non-gradient-based algorithms for global optimization or latent variable estimation, including the mean shift step from the mean shift algorithm and the E-step from expectation maximization [2].

Our algorithm can be run independently for any individual target shape, allowing it to work in a “zero-shot” setting. When a larger dataset of related shapes is available, we can also optimize for their part decomposition in advance (a “training” phase) and then perform fast decomposition of a new shape from that category by initializing its decomposition using its nearest neighbor from the “training” set.

We evaluate our algorithm by using it to reconstruct shapes from point clouds, using parts from the PartNet dataset. We compare to the recent Neural Parts unsupervised decomposition system [19] and show that our algorithm produces qualitatively more desirable decompositions that also achieve higher reconstruction accuracy. We demonstrate the control offered by our method by showing how it is possible to reconstruct shapes from one cat-

egory using parts from another (e.g. make a chair out of lamp parts). This also has application for 3D graphics content creation, which we demonstrate by reconstructing target shapes using parts from a modular 3D asset library.

In summary, our contribution is an unsupervised algorithm which retrieves and poses 3D parts to reconstruct input 3D shapes. We will release our code upon publication.

2. Related Work

Our contribution is related to prior work on unsupervised shape decomposition and on modeling by part retrieval and assembly. We do not discuss the considerable body of work on *supervised* decomposition/segmentation of 3D shapes.

Unsupervised shape decomposition: One class of unsupervised shape decomposition method reconstructs shapes using parametric primitives. Several approaches approximate 3D shapes as collections of cuboids [23, 25, 28]. One can obtain slightly more geometric flexibility by using a collection of superquadric surfaces instead of cuboids [20]. These approaches produce clean output geometry and offer some small degree of control over the type of decomposition produced, but their low-degree-of-freedom primitive representation results in a poor fit to the input shape.

Another class of approaches approximates the input shape with a collection of more general polyhedra. Several methods focus on convex polyhedra, either decomposing individual shapes in isolation [1, 9, 16] or training a neural network to produce similar convex decompositions for similar shapes from a category [6]. Another option is to decompose the input shape into pieces which can be represented as generalized cylinders [29]. These approaches produce clean geometry with a better fit to the input shape than parametric primitives allow, but they offer no control over the type of decomposition produced. They also typically need many primitives to fit the input shape well, making them non-compact and not well-suited for shape editing.

Recent research in this space has focused on decomposing shapes using neural primitives. BAE-Net trains an implicit shape representation that uses multiple decoder “heads,” where each head tends to represent the same localized part across many shape instances [4]. Other approaches represent neural parts as star domains [11] or deformed sphere meshes [19]. These approaches produce decompositions that fit the input shape well using a small number of primitives. However, their output geometry can exhibit undesirable artifacts, and they provide no control over the type of decomposition produced. We compare our algorithm to one of these approaches later in the paper and show that ours achieves even better reconstruction accuracy while also producing qualitatively better decompositions.

Modeling by retrieval and assembly: A large body of work in computer graphics has considered computer-assisted or fully-automated 3D modeling via retrieving and assembling pre-existing 3D shapes. Early work in this space focused on geometric heuristics to decide what object parts might be connected to one another [7, 14]. Later, researchers began exploring machine learning methods to suggest relevant parts to add to partial 3D object [3] or to synthesize entire 3D objects by putting parts together [10]. More recently, deep networks have been applied to the problems of suggesting parts [24] and part-based shape synthesis [15, 27]. These methods focus on creating new 3D shapes, rather than approximating/reconstructing existing ones, so they tackle a different problem than we do.

There has also been some work on retrieval-based reconstruction of input 3D shapes. One method learns to retrieve and deform entire shapes from a shape database to best match an input point cloud or image [26]. If the input is structurally distinct from any of the shapes in the database, however, this approach will not perform well. More closely related to our algorithm is the Structure Recovery by Part Assembly system [22], which takes a depth scan and a library of shapes as input and produces a reconstruction of the shape implied by the scan using parts from the shape library. This approach relies heavily on having access to a library of complete shapes drawn from the same semantic category as the input scan; in contrast, our algorithm assumes only a library of isolated parts. As a consequence, our algorithm also supports stylized reconstructions of shapes using e.g. parts from other semantic shape categories.

3. Method

Our system takes as input either a single target 3D shape \mathbf{T} to reconstruct, or a collection of target 3D shapes \mathcal{T} from the same semantic category; as we will show, by using a shape collection our system can take advantage of within-category shape similarity. We assume all 3D target shapes are represented as volumetric point clouds, which can be obtained from various sources (e.g. sampling the interior volume of a mesh or of a learned neural implicit shape representation [5, 18]). In addition to the target shapes, our system also takes as input a library of 3D parts \mathcal{B} that it will use to reconstruct the target shape(s). The output of our system is, for each target shape \mathbf{T} , a collection of transformed parts \mathcal{P} from the part library that approximates the target shape. Figure 2 shows a schematic overview of our approach, and Algorithm 1 provides pseudocode.

Our method begins by pre-training a variational autoencoder on all the parts in the part library (Section 3.1). This VAE’s continuous latent space helps turn the discrete combinatorial search problem of part retrieval into a tractable continuous optimization problem.

The heart of our algorithm is an iterative optimization

process, composed of three phases structured as nested loops. In **Phase I: Part Optimization**, for multiple parts \mathbf{P}_i the algorithm directly optimizes the VAE latent code \mathbf{e}_i , translation \mathbf{t}_i , and rotation around world up-axis r_i such that the parts reconstruct the target shape well (Section 3.2.1). In **Phase II: Part Shift**, the algorithm segments the target shape \mathbf{T} into regions using the optimized parts \mathcal{P} from Phase I and then re-projects them back to the latent space using the pre-trained VAE encoder to form the new initial state for the next iteration of Phase I (Section 3.2.2). By doing this, Phase II helps Phase I to escape from spurious local optima; its operation is analogous to the mode-seeking before of the mean shift algorithm or the expectation maximization algorithm. The algorithm also contains an optional **Phase III: Part Borrowing**, which it can use when a collection of target shapes \mathcal{T} is available (Section 3.2.3). For a given target shape \mathbf{T} that is currently not well-reconstructed by its optimized parts, this phase copies (or borrows) the optimization state $\{(\mathbf{e}_i, \mathbf{t}_i, r_i)\}$ from some other target shape $\mathbf{T}' \in \mathcal{T}$ where \mathbf{T}' is geometrically similar to \mathbf{T} . This phase further helps Phase I to escape from local optima.

To determine the number of parts to use for each shape, our system runs this iterative optimization process with different numbers of parts $k \in \mathcal{K}$, returning the k for which it achieved the best reconstruction (Section 3.5). Finally, it retrieves parts from the part library which are the closest match to the continuous latent space parts produced by the optimization (Section 3.3).

The process described thus far can be run independently for each new shape presented to the system. Alternatively, when given a collection of target shapes \mathcal{T} , the system can run the above optimization procedure on all of them as a preprocess and treat the output as a “training set.” Given a new target shape, it can retrieve the geometric nearest neighbor of this target in its training set, initialize optimization using that part decomposition, and execute a short optimization run to refine the decomposition to better fit the new target shape (Section 3.6). We also considered training a neural network on the “training set” to perform amortized inference of part decompositions, but we found that it did not perform as well as our nearest neighbor retrieval + optimization inference scheme (see supplemental for details).

3.1. Part VAE

To turn the combinatorial search problem of part retrieval into a continuous optimization problem, we construct a continuous latent space of part geometries by training a variational autoencoder (VAE) [13] on all the parts in the input library. Part meshes are first converted into volumetric point clouds by sampling their interior with 512 points. We use a 4-layer PointNet [21] as the encoder network, a 64-dimensional latent space, and a 3 layer MLP decoder which produces point clouds with 512 points. The VAE

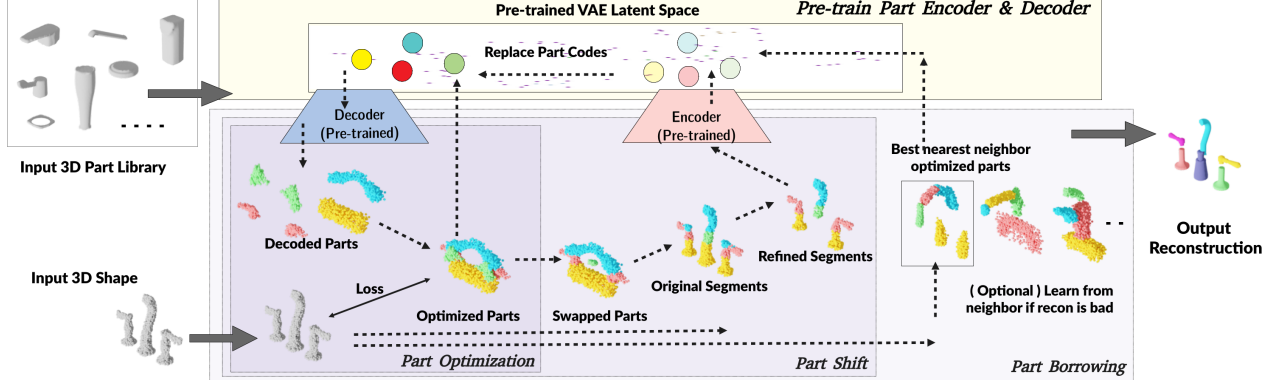


Figure 2. System Overview: Our system takes a target 3D volumetric point cloud \mathbf{T} and a library \mathcal{B} of parts as input and outputs a set of transformed parts \mathcal{P} from the part library which approximate \mathbf{T} . It first pre-trains a variational autoencoder (VAE) to project all parts into a continuous latent space. This allows it to turn the combinatorial part retrieval problem into a continuous optimization problem which proceeds in three phases: **Phase I: Part Optimization**, **Phase II: Part Shift** and **Phase III: Part Borrowing**. Phase I directly optimizes part latent codes \mathbf{e} , translations \mathbf{t} , and rotations \mathbf{r} to reconstruct the target shape. Phase II segments the input target shape using the optimized parts from Phase I and re-projects them back to the latent space. Phase III is an optional phase which borrows good part decompositions from other well-reconstructed similar shapes. When the optimization converges, real parts from the part library are retrieved. Note: This figure shows the example with symmetry constraints, so each part has a reflected duplicate.

is trained using a combination of Chamfer distance reconstruction loss and the standard KL divergence latent space regularization loss. Once trained, the weights of the encoder and decoder are fixed for all the subsequent processes.

3.2. Iterative Part Optimization

Given the part latent space, the system seeks to optimize for the latent codes \mathbf{e}_i and poses $(\mathbf{t}_i, \mathbf{r}_i)$ of a set of parts $\mathcal{P} = \{\mathbf{P}_i | i \in \{1 \dots k\}\}$ such that the latent codes, when decoded and posed, reconstruct the target shape. This optimization proceeds in three stages, organized in a nested loop.

3.2.1 Phase I: Part Optimization

In Phase I, the system directly optimizes the latent codes and poses of the parts via gradient-based optimization. For each target shape \mathbf{T} , this phase starts with k randomly initialized latent codes $\mathbf{e}_i \in \mathbb{R}^d$, k rotation angles about the world up axis $\mathbf{r}_i \in \mathbb{R}$, and k translation vectors $\mathbf{t}_i \in \mathbb{R}^3$. We don't consider scaling in our system, as scaling can warp the geometry of a high-quality part from the part library; this may be unacceptable for some applications. During the optimization, each latent part code \mathbf{e}_i is decoded through the pre-trained decoder to produce a decoded volumetric point cloud \mathbf{D}_i . This point cloud is then rotated and translated according to \mathbf{r}_i and \mathbf{t}_i . The values of these variables are iteratively updated by the gradient acquired from an objective function \mathcal{L} which consists a target reconstruction loss and a part non-overlap collision loss, i.e. $\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{overlap}}$.

Algorithm 1 Iterative Part Optimization

```

1: Input
2:   Target shape  $\mathbf{T}$ , Other target shapes  $\mathcal{T}$ , Part library  $\mathcal{B}$ 
3:   Possible numbers of parts  $\mathcal{K}$ 
4: Output
5:   Retrieved and assembled parts  $\mathcal{P}$  for target  $\mathbf{T}$ 
6: procedure
7:   Encoder, Decoder  $\leftarrow$  PreTrain( $\mathcal{B}$ )
8:   for  $k \in \mathcal{K}$  do ▷ executed in parallel
9:      $\mathbf{e}_i \leftarrow \text{rand}()$  ▷ part latent code,  $i \in k$ 
10:     $\mathbf{t}_i \leftarrow \text{rand}()$  ▷ part translation,  $i \in k$ 
11:     $\mathbf{r}_i \leftarrow \text{rand}()$  ▷ part rotation,  $i \in k$ 
12:    for  $i_3 \in n_3$  do
13:      for  $i_2 \in n_2$  do
14:        for  $i_1 \in n_1$  do
15:           $\mathbf{e}_i, \mathbf{t}_i, \mathbf{r}_i \leftarrow \text{Optimize}(\mathbf{e}_i, \mathbf{t}_i, \mathbf{r}_i, \mathbf{T})$ 
16:        end for
17:         $\mathbf{e}_i, \mathbf{t}_i, \mathbf{r}_i \leftarrow \text{Shift}(\mathbf{e}_i, \mathbf{t}_i, \mathbf{r}_i, \mathbf{T})$ 
18:      end for
19:       $\mathbf{e}_i, \mathbf{t}_i, \mathbf{r}_i \leftarrow \text{Borrow}(\mathbf{e}_i, \mathbf{t}_i, \mathbf{r}_i, \mathbf{T}, \mathcal{T})$ 
20:    end for
21:     $\mathbf{D}_i^k \leftarrow \text{Decode}(\mathbf{e}_i)$ 
22:     $\mathbf{D}_i^k \leftarrow \text{Pose}(\mathbf{D}_i^k, \mathbf{t}_i, \mathbf{r}_i)$ 
23:     $\mathbf{P}_i^k \leftarrow \text{Retrieve}(\mathbf{D}_i^k, \mathcal{B})$ 
24:     $\mathcal{P}^k \leftarrow \bigcup_i \mathbf{P}_i^k$ 
25:  end for
26:  return ChooseK( $\bigcup_{k \in \mathcal{K}} \mathcal{P}^k$ )
27: end procedure

```

Target Reconstruction Loss: To approximate the input shape, a collection of decoded part volumetric point clouds $\mathcal{D} = \bigcup_i \mathbf{D}_i$ should match the volumetric point cloud of the

target shape \mathbf{T} . We use chamfer distance between these two volumetric point clouds to measure this matching:

$$\mathcal{L}_{\text{recon}} = d_{\text{chamfer}}(\mathcal{D}, \mathbf{T})$$

Part Overlap Penalty Loss: The optimized parts should not only cover the target shape; they also should not overlap with each other. Since decoded parts can have complex geometry, it would be non-trivial and time consuming to compute a bounding proxy for them for differentiable collision checking. We thus designed a collision penalty term based on the volumetric point cloud representation. The collision penalty between two decoded parts \mathbf{D}_a and \mathbf{D}_b is computed by summing point-to-point pairwise distances between all points in those parts. If the distance $\|\mathbf{p}_i - \mathbf{p}_j\|_2$ between a pair of points $\mathbf{p}_i \in \mathbf{D}_a, \mathbf{p}_j \in \mathbf{D}_b$ is below a threshold $\tau = 0.1$, an overlap penalty is added as $\tau - \|\mathbf{p}_i - \mathbf{p}_j\|_2$, otherwise the penalty is zero. The final overlap penalty is the average penalty across all point pairs.

$$\mathcal{L}_{\text{overlap}} = \frac{1}{|\mathbf{D}_a||\mathbf{D}_b|} \sum_{\mathbf{p}_i \in \mathbf{D}_a} \sum_{\mathbf{p}_j \in \mathbf{D}_b} \max(0, \tau - \|\mathbf{p}_i - \mathbf{p}_j\|_2)$$

3.2.2 Phase II: Part Shift

The optimization problem solved by Phase I is non-convex with many local minima. In particular, it can be sensitive to different initializations of part latent codes and poses. These local minima manifest as regions of the target shape which are uncovered by any part. Phase II of the algorithm is designed to help the optimization escape such local optima. It does so by shifting the optimized parts \mathbf{D}_i from Phase I around the target to capture more regions in the target shape.

Given the optimized latent code \mathbf{e}_i and pose (\mathbf{t}_i, r_i) for each part, the algorithm produces the decoded part point cloud \mathbf{D}_i via the pre-trained decoder and poses it according to (\mathbf{t}_i, r_i) . As shown in Figure 3 2nd column, these parts may have become stuck in a local optimum, missing large regions of the target shape. Thus, instead of using the decoded parts directly, the algorithm instead uses them to segment the target point cloud \mathbf{T} into a set of segments $\mathcal{S} = \bigcup_i \mathbf{S}_i$, where \mathbf{S}_i consists of all points from \mathbf{T} which are closer to \mathbf{D}_i than to any other decoded part (Figure 3, 4th column). This guarantees that each every point in the target point cloud is assigned to one of the optimized parts.

However, the shapes of these segments may be implausible; for example, the green segment in the Figure 3 bottom row example is divided into two connected components, which is implausible for a single physical part. Ideally, we want each part to consist of a single connected component, and for those parts to cover the target point cloud. Thus, for each segment, the algorithm discards all but one of its connected components (Figure 3 column 6). Specifically, it chooses the connected component which is farthest from

any of the other segments, so as to focus on the most distinct part of the target point cloud covered by this segment. We have found that this connected component selection step performs most consistently when the segmented target point cloud \mathcal{S} is first filtered to remove points which were very well-covered by one of the original decoded parts \mathbf{D}_i (i.e. this helps the algorithm shift each segment to include the less-well-covered regions of the target shape). Specifically, the algorithm computes the distance between each point in the target point cloud \mathbf{T} and its closest point in the part \mathbf{D}_i to which it is assigned, and discards the points with the smallest $p = 30\%$ of these distances (Figure 3 column 5).

We now have one connected component \mathbf{C}_i for each segment \mathbf{S}_i of the segmented target point cloud \mathcal{S} . The algorithm normalizes the pose of each connected component by translating its centroid to the origin and rotating it such that the axes of its minimum volume bounding box align with the world axes. This normalized part is re-encoded to the part latent space to produce a new part latent code \mathbf{e}'_i . This code, along with the inverse of the pose normalization transformation (\mathbf{t}'_i, r'_i) , become the optimization variables for the next round of Phase I (Figure 3 column 7).

Finally, we have found that it can be helpful to add an additional part rearrangement step to the beginning of Phase II (Figure 3 column 3). This step directly replaces one of the optimized parts \mathbf{D}_i with the segment $\mathbf{S}_i \in \mathcal{S}$ which is least covered by any of the parts (call this \mathbf{S}^-), before proceeding with the rest of Phase II. Specifically, the algorithm iterates over all optimized parts \mathbf{D}_i and tries to replace each part with the \mathbf{S}^- . The algorithm accepts whichever replacement results in the greatest improvement in coverage of the target points by the resulting set of parts (if any does).

3.2.3 Phase III: Part Borrowing

If the system is given a collection of target shapes \mathcal{T} as input, this opens up the possibility of using similarities between target shapes to further improve part decompositions and escape from local optima. In this optional phase, the system identifies all target shapes \mathbf{T} that are currently not well-reconstructed and sets their optimization variables $\{(\mathbf{e}_i, \mathbf{t}_i, r_i)\}$ to the values of those copied from other geometrically-similar target shapes that are well-reconstructed. The algorithm performs this update on the subset of target shapes \mathcal{T}_{bad} whose reconstruction error is in the worst 60% of all target shapes \mathcal{T} . To quickly identify geometrically-similar targets, we pre-compute a target-to-target distance matrix \mathbf{M} , where M_{ij} is the chamfer distance between target shapes \mathbf{T}_i and \mathbf{T}_j . For each $\mathbf{T} \in \mathcal{T}_{\text{bad}}$, the algorithm iterates over the $m = 5$ nearest neighbors of \mathbf{T} in $\mathcal{T}/\mathcal{T}_{\text{bad}}$ according to the distance matrix \mathbf{M} and uses their current optimization variables $\{(\mathbf{e}_i, \mathbf{t}_i, r_i)\}$ to reconstruct \mathbf{T} . If the resulting reconstruction error is within the

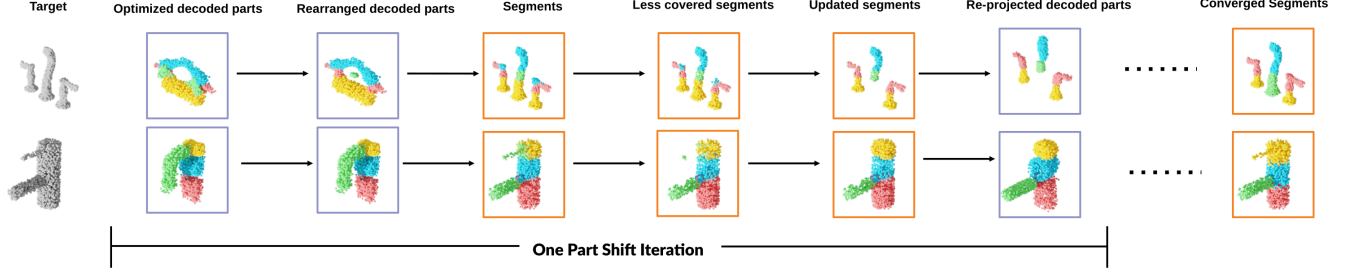


Figure 3. Phase II Part Shift. Given decoded and posed parts $\{\mathbf{D}_i\}$ (2nd column), the algorithm segments the target point cloud (1st column) \mathbf{T} based on which part each point is closest to (4th column). It then discards some of the points which were very close to their assigned part (5th column) and then retains only one connected component for each segment—whichever is farthest away from any other segment (6th column). These segments are finally re-encoded into the part latent space to produce the input for the next round of Phase I optimization (7th column shows what these parts look like when decoded). When the entire optimization procedure has converged, the output segments from this process will be the same as the original segments (last column). Purple boxes denote decoded part point clouds; orange boxes denote segmentations of the target point cloud using those decoded parts.

best 10% of reconstruction errors across \mathcal{T} , then these variable values become the new values for \mathbf{T} for the next iteration of optimization Phase I. Otherwise, \mathbf{T} 's variables are re-initialized to a random state.

3.3. Final Part Retrieval

The output from the above three phases is the decoded and posed part volumetric point clouds $\{\mathbf{D}_i\}$. To convert these point clouds into actual parts from our part library, the algorithm first segments the target shape \mathbf{T} into segments $\mathcal{U} = \bigcup_i \mathbf{U}_i$, where \mathbf{U}_i consists of all points from \mathbf{T} which are closer to \mathbf{D}_i than to any other decoded part (as in Phase II). For each segment \mathbf{U}_i , it then considers the q nearest neighbors of \mathbf{U}_i in the part library \mathcal{B} by Chamfer distance. We could use our pre-trained part latent space to accelerate this nearest neighbor search using standard Euclidean approximate nearest neighbor methods. However, for fairer comparison with methods which do not build such a latent space, in our experiments we set $q = |\mathcal{B}|$, i.e. a linear scan across the entire part library. The algorithm then optimizes poses (\mathbf{t}, \mathbf{r}) for each of these q candidate parts and takes the one which achieves minimal volumetric Chamfer distance between the part and the target shape segment as the final retrieved part. Figure 4 visualizes this process.

3.4. Leveraging Symmetry

Many manufactured shapes exhibit symmetries; our algorithm takes advantage of this property to improve its reconstructions. As a preprocess, we automatically detect the bilateral symmetry plane (if there is one) of each target shape. If a target shape \mathbf{T} has a bilateral symmetry, then whenever the algorithm decodes and poses a part point cloud \mathbf{D}_i , it also produces a symmetric copy of this part \mathbf{D}_{i+k} by duplicating \mathbf{D}_i and reflecting it about the symmetry plane. We disable collision checking between pairs of symmetric parts in Phase I. At the beginning of Phase II,

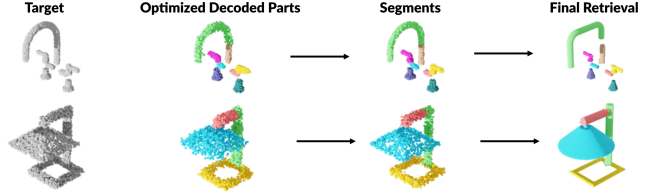


Figure 4. The final part retrieval phase. The target shape point cloud is segmented into regions based on which points are closest to which optimized part point cloud. Then, for each segment, the poses of q parts from the part library are optimized to fit the segment. The part that matches the segment best is retrieved and used in the final part decomposition.

if two symmetric parts are in contact, then the algorithm merges them together into one part. The same merge step is also applied before performing final part retrieval.

3.5. Choosing the Number of Parts

Due to the structural complexity of the input shapes, different numbers of parts k can lead to the best reconstruction for different shapes. Thus, our algorithm runs iterative multi-phase part optimization for different numbers of parts $k \in \mathcal{K} = [2, 4, 6, 8, 10]$. Then for each target shape \mathbf{T} , it selects the k whose optimized parts \mathcal{P}^k give the minimum value of a penalty function $d_{\text{chamfer}}(\mathcal{P}^k, \mathbf{T}) + \alpha|\mathcal{P}^k|$. This penalty function is a linear combination of a reconstruction error term and a complexity penalty term (which prefers fewer parts, all else being equal). We set $\alpha = 1.5^{-4}$ in our experiments. Figure 5 shows examples of target shapes reconstructed with different numbers of parts.

3.6. Amortized Inference

If our system receives a set of target shapes \mathcal{T} as input, then the optimized part decompositions for these shapes can facilitate amortized inference on new target shapes. Given

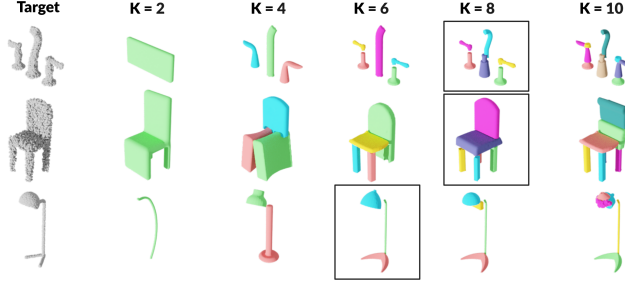


Figure 5. Choosing the number of parts k . The value of k which achieves the best balance between reconstruction quality and complexity (i.e. number of parts) is chosen. Note: some reconstructions for a particular value of k may contain fewer than k parts because symmetric pairs of parts are merged if they are in contact.

a never-before-seen target shape \mathbf{T} , the system first finds its nearest neighbor in \mathcal{T} , runs another round of optimization Phase I with \mathbf{T} 's variables initialized to those of the retrieved nearest neighbor, and finally retrieves parts for \mathbf{T} from the part library. This short optimization is considerably faster than optimizing for \mathbf{T} from scratch and is also more likely to avoid spurious local optima.

4. Results & Evaluation

We evaluate our method by using it to reconstruct target shapes from three categories from the PartNet dataset [17]: Chair, Lamp and Faucet. For each category, we split the shapes into source shapes (from which we construct part libraries), training shapes, and testing shapes. For each category, the sizes of the source set / part library / training set / test set are: Chair (601/1766/1000/120), Lamp (601/1985/1000/120), Faucet (151/780/250/100). For all three categories, 100 randomly selected shapes from the training set are used when evaluating training set performance. The part libraries are built using the ground-truth part segmentations from PartNet.

On a machine with an Intel(R) i9-9900K CPU and a NVIDIA RTX 2080 Ti GPU, our algorithm takes around 12 to 20 seconds on average to optimize the part decomposition for one shape from scratch. In amortized inference mode, it takes around 5 to 8 seconds. For Phase I optimization, we use the Adam optimizer [12] with learning rate 0.008.

Comparison to baselines: In this section, we compare our method against several baselines in terms of how well their part decompositions reconstruct the target shape. We evaluate the reconstruction quality using two metrics: **Surface Chamfer Distance (SCD)** and **Volume Chamfer Distance (VCD)**, i.e. chamfer distance evaluated between surface point cloud and volume point cloud shape representations, respectively. We compare against the following base-

lines:

- **Joint Learning of 3D Shape Retrieval and Deformation (JRD)** [26] learns to reconstruct shapes by retrieving and deforming 3D shapes from a database. It requires complete 3D shapes, whereas ours requires only parts. Since our method does not deform parts, we compare against a version of JRD with deformation disabled.
- **Neural Parts (NP)** [19] learns to reconstruct shapes by deforming a set of sphere meshes. In their paper, the authors focus on reconstruction from images; we swap out their image encoder with a 3D encoder (a volumetric CNN) for fair comparison to our method, which receives 3D data as input. Also motivated by fair comparison to our method, we primarily evaluate NP by using its predicted meshes to retrieve parts from our part library (using the same logic from Section 3.3).
- **Brute Force (BF)** is a naive baseline which randomly samples k parts from the part library and optimizes their poses to best reconstruct the target shape. For fair comparison, we let this process run iteratively for as long as our method took to run, and then take the best result.

Table 1 shows the results of comparing our method to JRD and NP. Our method outperforms both across all metrics. JRD can only retrieve whole shapes, which may not structurally match the target shape. The gap between our method and NP is largest on more geometrically-complex categories such as Faucet and Lamp, where knowledge of the irregular potential part geometries via the part latent space helps our method. Figure 7 show qualitative results for NP and Ours. See the supplemental material for JRD results and other comparison experiments.

We also conducted an experiment evaluating the impact of part output representations on reconstruction quality, comparing our method to NP and BF. The representations are: using the predicted geometry from the method directly (**Direct Recon**); retrieving parts which best match the predicted geometry (**Direct Retrieval**); retrieving parts which best match the segmentation of the target shape induced by the predicted geometry (**Segment Retrieval**). For Ours, the predicted geometry is the part point clouds output by the pre-trained decoder; for NP, it is the part meshes output by their network; for BF, it is a point sampling of the randomly-retrieved parts. Table 2 shows quantitative results, and Figure 6 shows qualitative results. For BF, Direct Recon and Direct Retrieval are quite bad, as BF is unlikely to find a good part decomposition by random sampling; it performs better for Segment Retrieval, as the segments can happen to correspond to an actual part in the part library. NP and Ours give reliable results across all output formats, with ours consistently outperforming NP.

Category	Method	Train (SCD) ↓	Train (VCD) ↓	Test (SCD) ↓	Test (VCD) ↓
Lamp	NP	0.349	0.204	0.390	0.195
	Ours	0.307	0.163	0.303	0.163
Faucet	NP	0.326	0.171	0.370	0.174
	Ours	0.256	0.135	0.288	0.134
Chair	JRD	0.746	0.448	0.669	0.397
	NP	0.495	0.233	0.547	0.240
	Ours	0.470	0.219	0.539	0.234
Average	JRD	0.746	0.448	0.669	0.397
	NP	0.390	0.203	0.436	0.203
	Ours	0.344	0.172	0.377	0.177

Table 1. Comparing our method to two baselines, JRD and NP, on reconstructing shapes from three PartNet categories. CD values are multiplied by 100. JRD is evaluated only on Chair because it does not provide data for the other categories.

Method	Direct Recon↓	Direct Retrieval↓	Segment Retrieval↓
BF (VCD)	0.680	0.750	0.182
NP (VCD)	0.128	0.207	0.153
Ours (VCD)	0.114	0.142	0.124

Table 2. Comparing our method with Brute Force (BF) and Neural Parts (NP) with different final decomposition output formats. CD values are multiplied by 100.

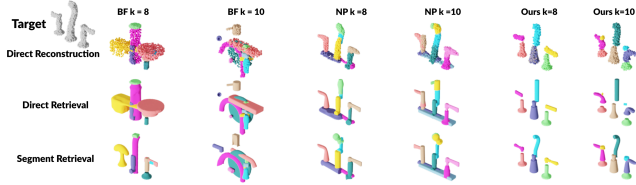


Figure 6. Comparing our method with Brute Force (BF) and Neural Parts (NP) with different final decomposition output formats. NP’s direct reconstruction format is a mesh; here we visualize it as a point cloud via sampling.

Cross-category reconstruction: Since our method does not require assembled shapes as training data, it reconstruct shapes out of different sets of parts than those from which they were originally built. Figure 8 shows examples of reconstructing chairs out of lamp and faucet parts, respectively. See the supplemental material more for results.

Ablation study: Table 3 shows the results of ablating different optimization phases for the Faucet category. Using all phases gives the best performance; Phase II matters more than Phase III. See the supplement for additional ablations.

5. Conclusion

We presented an unsupervised algorithm which retrieves and places 3D parts from a given part library to reconstruct 3D target shapes. Our approach uses a continuous relaxation of this combinatorial problem via a pre-trained part

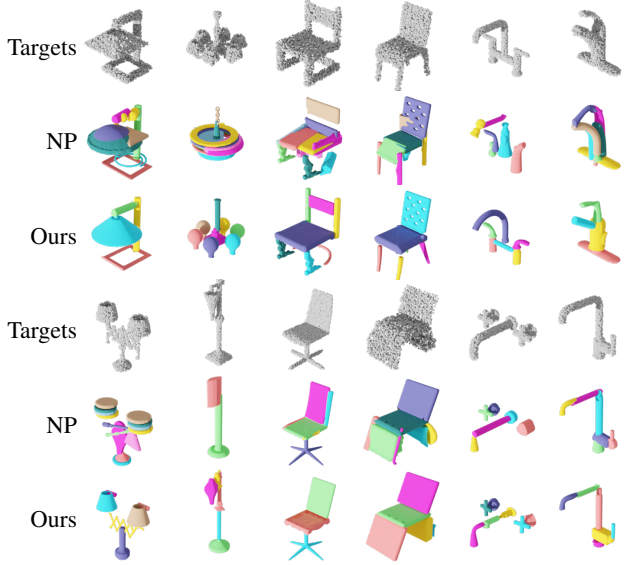


Figure 7. Neural Parts (NP) vs. Ours on reconstructing Lamps, Chairs and Faucets

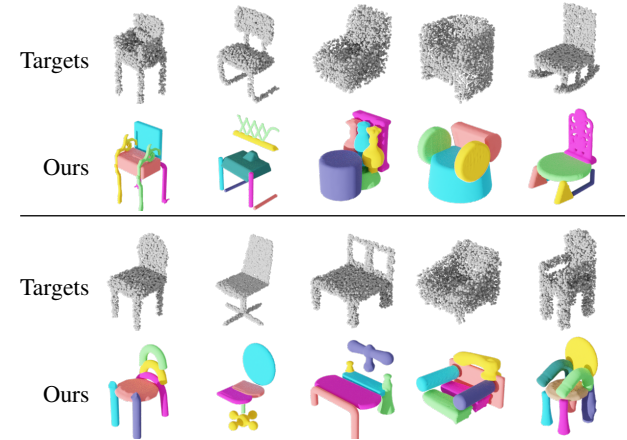


Figure 8. Cross Category Reconstruction: reconstructing Chairs from Lamp parts (top) or Faucet parts (bottom)

Method	SCD↓	VCD↓
Phase I	0.337	0.175
Phase I + Phase III	0.328	0.171
Phase I + Phase II	0.282	0.144
Phase I + Phase II + Phase III	0.255	0.134

Table 3. Ablation for phase components in our system. CD values are multiplied by 100.

latent space, plus a multi-phase direct optimization scheme to fit part shape and poses while avoiding many of the worst local optimas. Experimental evaluation shows that our approach outperforms strong baselines for retrieval-based reconstruction and unsupervised part decomposition. Potential directions for improvement include introducing phys-

ical priors into the optimization to make reconstructions more physically plausible and supporting incomplete point clouds as input.

References

- [1] Shmuel Asafi, Avi Goren, and Daniel Cohen-Or. Weak convex decomposition by lines-of-sight. *Computer Graphics Forum*, 32(5):23–31, 2013. 2
- [2] Miguel A. Carreira-Perpinan. Gaussian mean-shift is an em algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776, 2007. 2
- [3] Siddhartha Chaudhuri, Evangelos Kalogerakis, Leonidas Guibas, and Vladlen Koltun. Probabilistic reasoning for assembly-based 3D modeling. *ACM Trans. Graph.*, 30(4), July 2011. 3
- [4] Zhiqin Chen, Kangxue Yin, Matthew Fisher, Siddhartha Chaudhuri, and Hao Zhang. Bae-net: Branched autoencoder for shape co-segmentation. *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. 1, 2
- [5] 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. 3
- [6] Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, and Andrea Tagliasacchi. Cvxnet: Learnable convex decomposition. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2
- [7] Thomas Funkhouser, Michael Kazhdan, Philip Shilane, Patrick Min, William Kiefer, Ayellet Tal, Szymon Rusinkiewicz, and David Dobkin. Modeling by example. In *ACM SIGGRAPH 2004 Papers*, 2004. 3
- [8] Jingwei Huang, Hao Su, and Leonidas Guibas. Robust watertight manifold surface generation method for shapenet models. *arXiv preprint arXiv:1802.01698*, 2018. 11
- [9] Oliver Van Kaick, Noa Fish, Yanir Kleiman, Shmuel Asafi, and Daniel Cohen-Or. Shape segmentation by approximate convexity analysis. *ACM Transactions on Graphics (TOG)*, 34(1):1–11, 2014. 2
- [10] Evangelos Kalogerakis, Siddhartha Chaudhuri, Daphne Koller, and Vladlen Koltun. A probabilistic model for component-based shape synthesis. *ACM Trans. Graph.*, 31(4), July 2012. 3
- [11] Yuki Kawana, Yusuke Mukuta, and Tatsuya Harada. Neural star domain as primitive representation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NeurIPS '20*, Red Hook, NY, USA, 2020. Curran Associates Inc. 1, 2
- [12] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR 2015*, 2015. 7
- [13] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014. 3
- [14] Vladislav Kreavoy, Dan Julius, and Alla Sheffer. Model composition from interchangeable components. In *Proceedings of the 15th Pacific Conference on Computer Graphics and Applications*, PG '07, page 129–138, USA, 2007. IEEE Computer Society. 3
- [15] Jun Li, Kai Xu, Siddhartha Chaudhuri, Ersin Yumer, Hao Zhang, and Leonidas Guibas. GRASS: Generative recursive autoencoders for shape structures. *ACM Trans. Graphics (Proc. SIGGRAPH)*, 36(4), 2017. 3
- [16] Jyh-Ming Lien and Nancy M Amato. Approximate convex decomposition of polyhedra and its applications. *Computer Aided Geometric Design*, 25(7):503–522, 2008. 2
- [17] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *CVPR*, 2019. 7
- [18] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3
- [19] Despoina Paschalidou, Angelos Katharopoulos, Andreas Geiger, and Sanja Fidler. Neural parts: Learning expressive 3D shape abstractions with invertible neural networks. In *CVPR*, 2021. 1, 2, 7
- [20] Despoina Paschalidou, Ali Osman Ulusoy, and Andreas Geiger. Superquadrics revisited: Learning 3D shape parsing beyond cuboids. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2
- [21] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 3
- [22] Chao-Hui Shen, Hongbo Fu, Kang Chen, and Shi-Min Hu. Structure recovery by part assembly. *ACM Trans. Graph.*, 31(6), Nov. 2012. 3
- [23] Chunyu Sun, Qianfang Zou, Xin Tong, and Yang Liu. Learning adaptive hierarchical cuboid abstractions of 3D shape collections. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 38(6), 2019. 1, 2
- [24] Minhyuk Sung, Hao Su, Vladimir G. Kim, Siddhartha Chaudhuri, and Leonidas Guibas. ComplementMe: Weakly-supervised component suggestions for 3D modeling. *SIGGRAPH Asia*, 2017. 3
- [25] Shubham Tulsiani, Hao Su, Leonidas J. Guibas, Alexei A. Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *CVPR*, 2017. 1, 2
- [26] Mikaela Angelina Uy, Vladimir G. Kim, Minhyuk Sung, Noam Aigerman, Siddhartha Chaudhuri, and Leonidas Guibas. Joint learning of 3D shape retrieval and deformation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 7
- [27] Kai Wang, Paul Guerrero, Vladimir Kim, Siddhartha Chaudhuri, Minhyuk Sung, and Daniel Ritchie. The shape part slot machine: Contact-based reasoning for generating 3D shapes from parts. In *ECCV*, 2022. 3
- [28] Kaizhi Yang and Xuejin Chen. Unsupervised learning for cuboid shape abstraction via joint segmentation from tpoint clouds, 2021. 1, 2
- [29] Yang Zhou, Kangxue Yin, Hui Huang, Hao Zhang, Minglun Gong, and Daniel Cohen-Or. Generalized cylinder decomposition. *ACM TOG*, 34(6), nov 2015. 2