

COALESCE: Component Assembly by Learning to Synthesize Connections

Kangxue Yin
Simon Fraser University

Matthew Fisher
Adobe Research

Zhiqin Chen
Simon Fraser University

Vladimir G. Kim
Adobe Research

Siddhartha Chaudhuri
Adobe Research, IIT Bombay

Hao Zhang
Simon Fraser University

Abstract

We introduce COALESCE, the first data-driven framework for component-based shape assembly which employs deep learning to synthesize part connections. To handle geometric and topological mismatches between parts, we remove the mismatched portions via erosion, and rely on a joint synthesis step, which is learned from data, to fill the gap and arrive at a natural and plausible part joint. Given a set of input parts extracted from different objects, COALESCE automatically aligns them and synthesizes plausible joints to connect the parts into a coherent 3D object represented by a mesh. The joint synthesis network, designed to focus on joint regions, reconstructs the surface between the parts by predicting an implicit shape representation that agrees with existing parts, while generating a smooth and topologically meaningful connection. We demonstrate that our method significantly outperforms prior approaches including baseline deep models for 3D shape synthesis, as well as state-of-the-art methods for shape completion.

1. Introduction

Shape assembly by composing component parts has been the dominant paradigm for data-driven 3D modeling and design exploration [32, 11, 4, 50, 37]. The component assembly gathers parts from a shape collection and composes them into a target shape. To date, most research effort has been devoted to finding compatible parts for the assembly [3, 18, 21, 56, 20, 24, 41]. However, these parts still need to be properly aligned and then connected or joined in a natural and coherent fashion, ideally resulting in a high-quality final shape that is suitable for fabrication, simulation, high-resolution rendering, and seamless texturing.

Compared to the higher-level and human-centric tasks, such as part selection [4, 41] and assembly exploration [50], geometrically aligning and joining 3D parts is a much more tedious and unnatural practice for human modelers. A full automation of this low-level task is highly desirable. How-

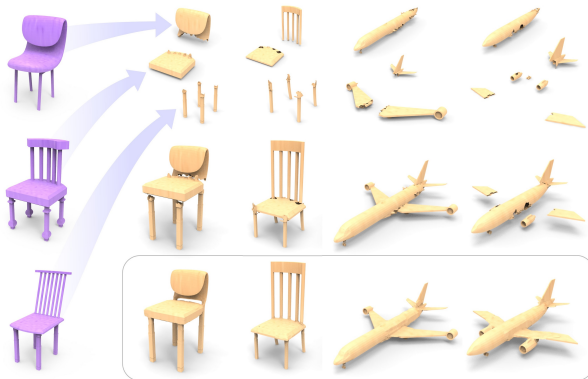


Figure 1: Our deep learning framework, COALESCE, assembles 3D shapes by a learning-based approach to synthesize part connections. Given a set of unaligned, roughly cut parts with possibly mismatched joint regions (top row: all parts come from *different* shapes), the alignment module of our network can learn to arrange these parts into a plausible configuration (middle row). However, this still leaves gaps and geometric/topological mismatches at the joints. After eroding the input parts in advance, our joint synthesis module creates connecting regions, which are stitched to the input to create seamless assemblies (bottom row).

ever, due to the large variability exhibited by the geometry and topology of the shape parts, heuristic approaches such as hole filling and stitching are hardly general and reliable, as they must make varying assumptions about part placement, feature alignment and boundary geometry, or resort to user interactions [22, 8, 15, 53, 29, 40, 38]. Further, nearly all these analytical methods (except that of Huang et al. [15]) are primarily designed for organic shapes with smoothness priors, and do not handle man-made shapes with sharp corners, edges, and rich structural variations.

In this paper, we introduce COALESCE, the first *data-driven* framework for component assembly by learning to *align* shape parts *and synthesize* the connections between them. We stipulate that in the absence of any mathematical formulation of how to connect parts, the most general and

effective way to handle the immensely rich part connection varieties is to *learn* how parts are connected from a large variety of 3D shapes. In addition, the most viable solution to handle geometric or topological mismatches between two parts is to remove the mismatched portions, via an *erosion* process, and then rely on a *learned joint synthesis* step to produce a natural and plausible part connection.

A defining feature of our learning-based part connection framework is that it focuses on *joint* synthesis, rather than learning to reconstruct whole shapes. With proper part alignment, which is also learned, the space of part joints possesses a higher level of predictability compared to that of the whole shapes, making joint synthesis a comparably easier learning problem. Another novel feature of our method, compared to previous works on hole filling and shape completion, is the “retreat (erode) before advance” connection mechanism. It can be regarded as a means of normalization to handle the wide variety of possible part connections.

The main contributions of our work include:

- We tackle an important problem for assembly-based shape modeling: how to best connect parts, which may come from different 3D models and possess varying degrees of geometric and structural incompatibilities.
- An erosion-based part connection mechanism to help reconcile geometric and topological mismatches between parts, and potential part misalignment.
- A novel joint synthesis method to connect input parts into a plausible 3D object, via an implicit surface. As shown in the leftmost chair assembly in Figure 1, our network is able to connect parts into a novel structure which is absent from the parts’ shape origins.

Our networks are trained on the dataset of segmented 3D shapes from Yi et al. [52], which was extracted from ShapeNet [2]. We demonstrate the performance of our “learning to connect” framework on a variety of 3D shapes, both qualitatively and quantitatively, through comprehensive comparative studies.

2. Related work

The literature on modeling shapes from parts is extensive. Hence we mainly discuss geometric stitching with analytical approaches, assembly-based modeling from given parts, and modern neural shape synthesis methods.

Part stitching. Several papers have studied the problem of stitching parts together by connecting a pair of boundary edge loops, where the principal objective is smoothness of the connection. Yu et al. [53] proposed gradient-domain Poisson stitching: after establishing correspondences between the loop vertices, a target continuous gradient field

is set up in the adjacent mesh regions. Sharf et al. [40] proposed a “Soft-ICP” algorithm, which iteratively and non-rigidly brings two overlapping patches (one adjacent to each loop) together until they smoothly blend with each other. Lin et al. [29] create a sketch-guided implicit field to connect two boundary loops. The use of an implicit field to generate a connecting surface is similar to our approach, but the synthesis is entirely guided by user input. Schmidt and Singh [38] join a boundary edge loop to a target surface by creating consistent local parametrizations on the surface as well as in the polygon bounded by the loop. The works of Huang et al. [15] and Lescoat et al. [25] share a more similar goal as COALESCE. However, their methods rely on the existence of an underlying smooth transition between the parts that can be detected via geometric analysis. In contrast to these methods, our approach *learns* how to repair geometrically and topologically dissimilar junctions, generates both low and high curvature regions as indicated by the training data, and co-aligns the assembled parts.

Assembly-based modeling. A variety of methods studied ways to construct shapes from pre-existing parts. The pioneering work of Funkhouser et al. [11] presented tools for searching for database shapes partially matching a user-provided proxy, cutting out a matching part using “intelligent scissors”, and pasting the part into a new model. The majority of subsequent assembly-based methods focused on retrieving and organizing compatible parts to facilitate or automate assembly [23, 3, 18, 20, 42, 50, 10, 19, 41, 37, 39, 47, 7]. These methods did not place particular emphasis on repairing or synthesizing the connections between parts, especially for man-made shapes. Often, parts were simply placed in contact, or slightly overlapped, to appear visually connected, although Kraevoy et al. [23], Chaudhuri et al. [3] and Kalogerakis et al. [20] did create smooth transitions between organic parts using methods described above.

The SCORES system of Zhu et al. [57] shares some aspects of our method: it attempted to realign a crude input arrangement of parts in order to resolve incompatibilities and misalignments using a recursive neural network. However, this work focused on *structural* repair. Only affine transforms were applied to input parts, and no stitching was performed at joints. Duncan et al. [8] presented a method to deform corresponding joint regions in a shape collection to resemble each other, so that parts could be seamlessly interchanged. However, this required substantial modifications to the input parts, and forced all joints between parts from the same two semantic categories to look identical. In general, complicated joints, especially in man-made shapes, remain an open problem.

Generative models of 3D geometry. In contrast to the assembly-based methods, which *reuse* existing shape parts, substantial recent research has focused on *synthesizing* the

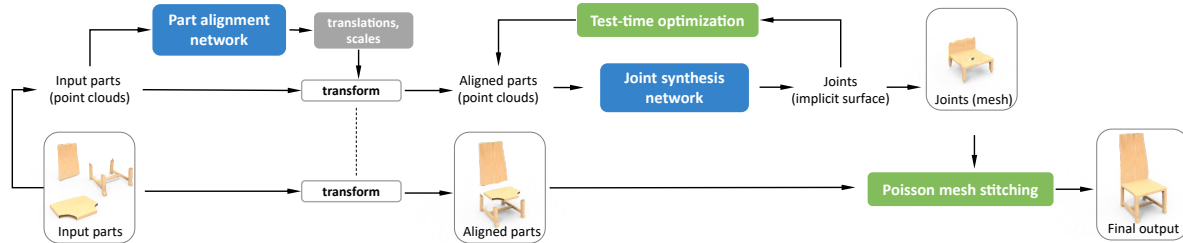


Figure 2: An overview of COALESCE, our component-wise shape assembly. Input mesh parts, with the joint areas eroded to create room for joint synthesis, are first transformed by a part alignment network. Then, a joint synthesis network is employed to synthesize a joint region as an implicit field between each pair of aligned parts. The implicit field is then polygonized and finally Poisson stitching is applied to the input parts and the synthesized joints to produce the final mesh output. Note: dashed lines indicate that the set of part transforms is shared between the point cloud (top) and mesh (bottom) branches.

part geometry itself. Neural generative models employ various representations for the output shape, including point clouds [1], voxel grids [48, 55], octrees [44], deformations of a template mesh [13], 3D embeddings of 2D atlases [14], or combinations of the above [34]. Recent techniques model shapes as continuous implicit functions, where the network learns to map a 3D point to an occupancy value [31] or a clamped signed distance field [35]. We employ a variant of these methods [5] in our joint synthesis, as they demonstrate high visual quality and can capture complex topological variations of shapes. Since these implicit functions are defined everywhere in the 3D volume, it is easier for us to incorporate them into consistency losses that ensure the synthesized joint aligns with the input geometry, without expensive closest point sampling. Some methods explicitly model part layout in addition to generating part geometry [16, 27, 43, 26, 12, 49, 33, 28, 17, 51], but none of them explicitly focus on joints between parts.

Neural generative models have been commonly used in encoder-decoder architectures; they encode input that is partial or of another modality, such as images or depth scans, and output full 3D shapes. The output is evaluated via reconstructive [9] or discriminative [46] losses to train the model. Hence, a naive solution to joint synthesis could employ a shape completion approach [6], where the system would encode unstitched parts, and decode the entire shape. Unfortunately, this approach does not preserve the input parts or provide joints of the highest quality, since a lot of the network’s capacity is allocated to representing the entire shape, as we show in our experiments. In contrast, our method focuses specifically on synthesizing the joint regions, in a part- and global shape-aware way.

3. Method

Given a set of input parts extracted from *different* objects, our network automatically aligns them and synthesizes plausible joints, connecting the parts into a coherent mesh object. Our pipeline is visualized in Figure 2.

The input to our pipeline is a set of labeled parts, where

each part is represented by a triangle mesh. We uniformly sample points on the input parts, and erode the joint regions on each part to ensure that there is sufficient room for joint synthesis and to avoid structural mismatching between the crudely segmented parts. Specifically, for each point on an input part, we compute its distance to the segmentation boundary and if the distance is less than a threshold τ , the point will be excluded from the part.

We assume that the input parts are in a canonical orientation, and normalize the parts to the same scale. The alignment network takes normalized parts as inputs and predicts translation and scaling to create structurally sound composition of parts (Section 3.1). Next, the joint synthesis network takes the aligned parts as inputs and synthesizes connections between them as an implicit surface (Section 3.2). Note that implicit representation is especially suitable for joint synthesis, since the network is able to focus on the joint area and allocate its capacity to the geometry near the joints. To make the output implicit surface closely conform to the input part boundaries, we jointly refine the predicted spatial transformations for the input parts and the weights of the joint synthesis network to minimize a part-joint matching loss. After that, the implicit surface produced by the joint synthesis network is discretized into a voxel grid and converted into a mesh using marching cubes [30]. Finally, we develop a Poisson-based approach, a variant of Poisson blending, to seamlessly stitch the synthesized joint meshes with the original meshes of the aligned input parts. The details of the approach is described in the supplementary material.

3.1. Part alignment network

Given a set of input parts from different objects, we sample points for each of them and normalize the resulting point clouds. After that, we use a part alignment network to assemble them into a shape by applying predicted translation and scaling to each part. As shown in Figure 3(a), the part alignment network contains PointNet++[36] networks that compute shape features for the input parts. Given the fea-

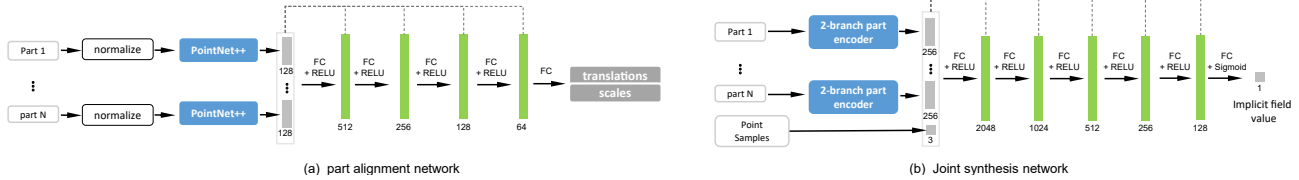


Figure 3: The architectures of our part alignment network (a), and our joint synthesis network (b), where the input parts are represented by point clouds. Dashed lines denote skip connections.

tures of the parts, we use a fully connected regression network with skip connections to produce the spatial transformation that aligns the parts. The features for missing object parts are simply set as zero vectors. We train the part alignment network with parts belonging to the same objects. Our loss function is the Earth Mover’s Distance between the translated and scaled parts and the ground truth. More details of the training process and the architecture of the PointNet++ used in the part alignment network are provided in the supplementary material.

3.2. Joint synthesis network

Our joint synthesis network is a core component of the framework. It takes point clouds of the aligned parts as input and reconstructs the joints as implicit surfaces that connect the parts into a plausible object. We show in Figure 3(b) the architecture of our joint synthesis network. The network is designed as an encoder-decoder network, where each input part is assigned a two-branch point cloud encoder. The first branch of the part encoder is a PointNet++ network for extracting overall shape features from the part. The second branch is another PointNet++ network for extracting localized shape features from the points that are close to the joint area. We represent each part with 2048 points, and consider the 512 points that are closest to the joint area as the input to the second branch for localized shape features. The detailed architecture of the two-branch part encoder is provided in the supplementary material.

The network uses an IM-NET [5] to reconstruct implicit surfaces for the joints. Given an object category, for each semantic part segment, we use the two-branch part encoder to extract a feature vector from the part. If a part does not exist in the object, we use a zero vector instead. The feature vectors of N parts are then concatenated into a single vector to be fed into the IM-NET decoder.

The implicit decoder requires point samples in the 3D space surrounding the shape we aim to reconstruct. When we sample points with inside/outside status for training the IM-decoder, to make the network focus on the joint area, we densely sample points around the joints and sparsely sample points away from the joints as shown in Figure 4. Specifically, we use 16384 sampled points for each shape: 80% of the points are within the joint volume after 5 steps of volumetric inflation in a 256^3 voxel grid, and 10% are

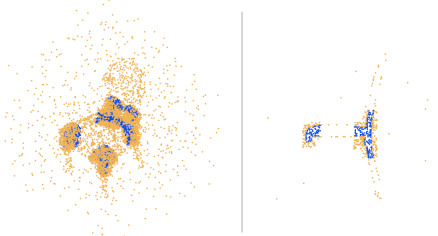


Figure 4: Visualization of our point sampling for training the IM-decoder. Dark yellow dots denote points outside the shape, blue dots denote points inside the shape.

outside the shape but close to the surface of the shape, and the remaining 10% are random samples outside the shape. By sampling training points in this fashion, we only force the IM-decoder to reconstruct the joint area rather than the entire object.

The loss function of the joint synthesis network is defined by two terms:

$$L_{\text{joint}} = L_{\text{mse}} + \alpha L_{\text{match}}$$

where α is a balancing weight set to 0.2 in all experiments. The first term L_{mse} is the original loss term in IM-NET which is defined as:

$$L_{\text{mse}} = \frac{1}{|S|} \sum_{p \in S} |f(p) - F(p)|^2$$

where S is the set of point samples, $f(p)$ is the output value from our network for point sample $p \in S$. $F(p)$ is the ground-truth label of point p .

The term L_{match} is to encourage the synthesized joint to match the open boundaries of the input parts. To define L_{match} , we first select a set of 1024 points $\mathcal{N}(J)$ that are on the surface of the input parts and close to the joint. If the implicit surface tightly matches the open boundaries of the input parts, for each point $p \in \mathcal{N}(J)$, if we move it along its normal direction, the expected implicit function value on its new position is 0 (outside). If we move it in the inverse direction of its normal direction, the expected value is 1 (inside). Thus we define L_{match} as:

$$L_{\text{match}} = \frac{1}{2|\mathcal{N}(J)|} \sum_{p \in \mathcal{N}(J)} |f(p + \lambda n)|^2 + |f(p - \lambda n) - 1|^2$$

where n is the unit normal vector of p , and λ is a small constant with a default value of 0.005.

The joint synthesis network is trained by segmented parts of 3D objects. To make room for the network to learn to synthesize the joints connecting input parts, we erode the parts along the segmentation boundaries. As discussed in section 1, another effect of the erosion is to help eliminating topological mismatches among input parts from different objects. We first pre-train the two-branch point cloud encoders for 100 epochs by pairing it with a fully connected decoder [1] for reconstructing 3D shape parts. Afterwards, we pair the pretrained point cloud encoders with the IM-decoder, and train the IM-decoder with the loss function $L_{\text{joint}} = L_{\text{mse}}$ for 80 epochs. Finally, we train the encoders and the IM-decoder together with the loss function $L_{\text{joint}} = L_{\text{mse}} + \alpha L_{\text{match}}$ for another 80 epochs. More details on our training procedures and training time can be found in the supplementary material.

3.3. Test-time optimization

It is critical that the synthesized joints align closely with the open boundaries of the parts to create a realistic model. While our feed-forward joint synthesis network attempts to maximize boundary consistency, to further improve the results, we employ test-time optimization to explicitly minimize any boundary discontinuities. In this phase, we jointly optimize the scaling factors $\{s\}$, translations $\{t\}$ of parts, and the parameters W of the IM-decoder in the joint synthesis network for a given test input. We define the objective function for this optimization as an L_1 version of the part-joint matching term:

$$h(\{s\}, \{t\}, W) = \frac{1}{2|N(J)|} \sum_{p \in N(J)} |f(p + \lambda n)| + |f(p - \lambda n) - 1|$$

where $N(J)$ is a set of 1024 points close to the joint, n is the unit normal at p , and λ is a small constant set to 0.005.

Minimizing the objective function h with respect to $\{s\}$, $\{t\}$ requires a significantly different step size compared to minimizing the objective function with respect to network parameters W . This motivated our choice to alternately optimize for part alignments and joint synthesis. We first fix the parameters of the IM-decoder and optimize $\{s\}$, $\{t\}$ by minimizing the objective function h with an Adam optimizer with step size 0.002. After that, we fix $\{s\}$, $\{t\}$ and optimize the parameters of the IM-decoder W with step size 0.0001. We keep repeating the above two steps for a fixed number of iterations (25, in our implementation). As we show in Section 4.2, the test-time optimization refines the positions and scales of the parts and the decoding result such that they match more tightly.

3.4. Poisson mesh stitching

After test-time optimization, the joint region represented by implicit function is converted into a polygon mesh by Marching Cubes. A Poisson-based approach is developed to seamlessly stitch the synthesized mesh with the aligned

parts. The Poisson mesh stitching has three phases: (i) finding corresponding loops between the input parts and the synthesized joints, (ii) removing redundant regions from the joint mesh and (iii) Poisson mesh blending. Please find the description of these steps in the supplementary material.

4. Experiment and results

We evaluate the utility of COALESCE for the part assembly problem. In the following, we discuss our data pipeline, ablation studies for various design choices, and comparisons to alternative approaches. More visual results can be found in the supplementary material. Unless stated otherwise (e.g., in ablations), we use the same network architecture with the same hyper-parameters and the same Poisson mesh stitching post-processing for all experiments.

4.1. Dataset

We use three categories of objects from the segmented shapes provided by Yi et al. [52]: 3,746 chairs, 2,690 air-planes, and 184 mugs. We split each category into training (80%) and test (20%) sets. We then create eroded part-wise point clouds as described in Section 3. For our erosion process we set the erosion radius to $\tau = 0.05$ of the bounding box diameter for chairs and mugs and $\tau = 0.025$ for air-planes. Since there are very few mugs, we train the joint synthesis network for the mug dataset for 80 more epochs.

4.2. Ablation study

Qualitative study. As shown in Figure 5, the inputs to our part alignment network are crudely segmented parts from *different* objects. Figure 5b shows that the results of our network may be inaccurate. Similarly, the geometry created by the forward pass through the joint synthesis network might not align tightly with the input parts, e.g., the last row. Our test-time optimization addresses these issues and further refines the alignment and the joint geometry (Figure 5c) to produce the final result (Figure 5d).

A naive approach to assemble parts from different objects is to learn to align them directly, without erosion or joint synthesis. We train a part alignment network that directly aligns input parts to get the assembly, and show the results in Figure 5e. We see several artifacts emerging from this naive pipeline. First, there could be a topological or geometric mismatch, such as armrests of a chair falling outside the range of the seat (see second row) or the cup of a mug having holes that are larger than the size of the handle (see last row). Second, segmented parts might have imperfect boundaries (see first row). Note how our erosion and synthesis approach solves these issues (Figure 5d).

Another alternative is to treat it as a shape completion problem, which is commonly addressed with encoder-decoder architectures. We train a neural network that encodes point clouds of aligned parts, and then decodes them

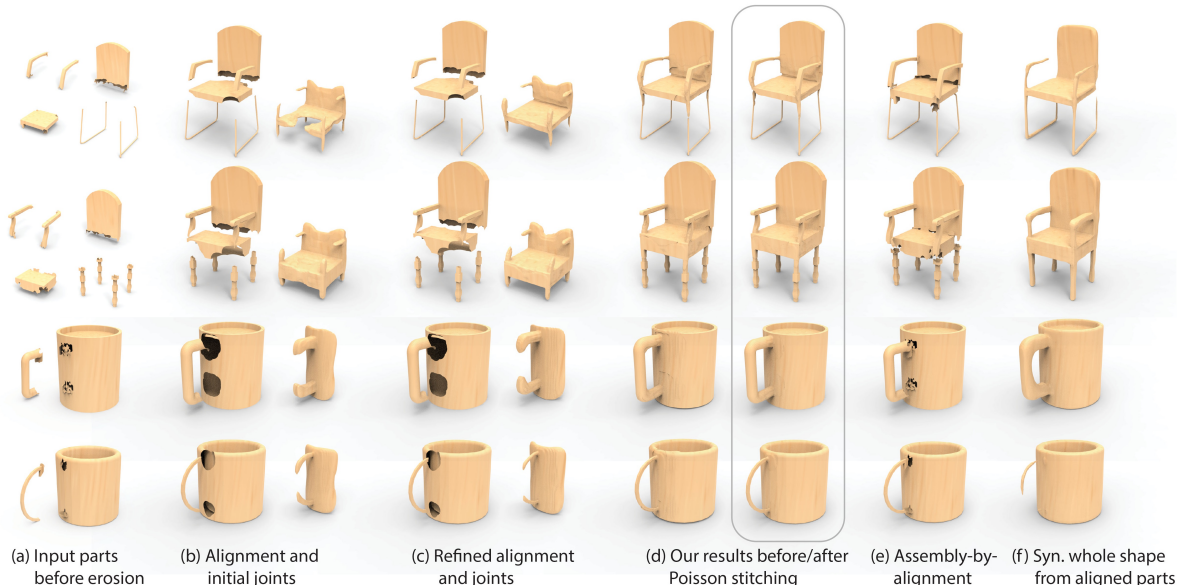


Figure 5: We evaluate the significance of different components of our framework by ablation. Given the input parts (a), we show results that only involve part alignments (b), refined alignment with test-time optimization (c), and the effect of Poisson stitching (d). We also demonstrate the disadvantages of naive approaches, such as assembling input parts without any joint synthesis (e), or directly synthesizing the entire shape (f) from aligned parts.

	Chamfer Dis. ($\times 10^3$)			Light Field Dis.		
	plane	chair	mug	plane	chair	mug
Synthesize entire shape	0.164	0.738	0.804	3265.2	3054.3	1117.8
Syn. entire shape + code opt.	0.173	0.472	0.646	2991.0	2374.8	885.1
Ours (before test-time opt.)	0.073	0.225	0.297	1148.3	1251.6	374.3
Ours (after test-time opt.)	0.071	0.191	0.300	1072.9	951.5	276.9
Ours (after Poisson blending)	0.067	0.181	0.274	895.4	894.6	196.2

Table 1: Quantitative ablation study. We compute the errors of our method against ground-truth shapes on different pipeline stages including before and after test-time optimization, and after Poisson blending.

into a full shape via IM-decoder (with the identical architecture to our joint synthesis network). As we can see in Figure 5f, the results of this approach might not preserve the geometry of the input parts and often have more artifacts near the joints due to the absence of joint focus.

Quantitative study. In addition to the qualitative study in Figure 5, we also did a quantitative ablation study in Table 1. In the study, we feed the eroded input parts from the same test shape to our network, and measure the chamfer distance (*lower is better*) and light field distance (LFD: *lower is better*) between the results and the ground-truth test shapes. The quantitative evaluation results show that the test-time optimization module and the Poisson blending module can improve the quality of the network output. We also evaluate the results of a network that directly synthesizes the entire shape from the aligned parts (first row in Table 1), and optimizes the latent code to minimize a matching loss that is similar to $h(\{s\}, \{t\}, W)$ but on the entire shape rather than just joints (second row in Table 1).

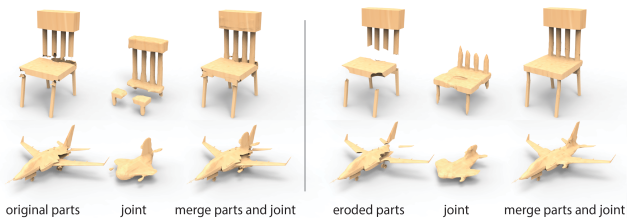


Figure 6: Comparison of two networks that synthesize part connections from original parts before erosion (left) and from eroded parts (right). Incorporating erosion better resolves geometric and topological mismatches.

Part erosion. To show the necessity of part erosion, we train a neural network to synthesize joints from *uneroded* parts, with the same architecture and configuration as our network described in Section 3. We compare the results of directly synthesizing joints from uneroded parts with our results in Figure 6. Note that the variant without erosion cannot successfully resolve the mismatches at part junctions.

4.3. Comparison to Baselines

Although there are no prior works on learning joint synthesis, we can compare to shape completion methods, such as AtlasNet [14] and PCN [54]. We directly use the publicly available source codes of these methods and train the networks on individual categories following the provided instructions. We also tested a customized version of AtlasNet that only outputs surfaces in the joint regions, similar to our joint synthesis network. We denote this version of

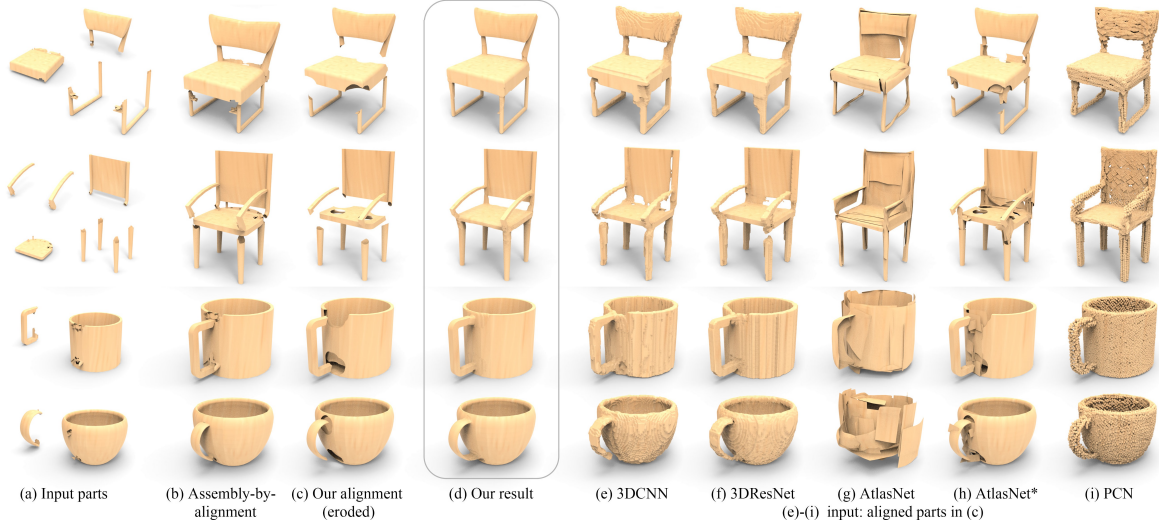


Figure 7: Qualitative comparisons with baseline methods, which took the parts aligned by our method (c) as input.

AtlasNet as AtlasNet* in Figure 7h and Tables 2, 3, 4.

We also consider two voxel-based shape completion baselines that can be applied to our problem: 3DCNN and 3DResNet. In the two baselines, we use architectures that perform shape completion akin to image translation tasks. See the supplementary for more details. Unlike encoder-decoder architectures such as AtlasNet and PCN, these methods do not have a tight bottleneck layer, and they only modify inputs in a relatively local manner. Thus, they do not have the challenge of reconstructing the entire shape from a compact latent code, and might be better at preserving the geometry of the original parts. When testing, we use marching cubes to obtain the meshes from the output voxels.

The qualitative comparison of our method to the baselines is provided in Figure 7. The input parts in Figure 7 are from different objects (Figure 7a), and naively aligning them (Figure 7b) results in a mesh with a large number of boundary artifacts. We feed eroded parts (Figure 7c) as input to our method and various baselines. Our method produces high quality seamless meshes (Figure 7d). Voxel translation techniques (3DCNN, 3DResNet) suffer from discretization artifacts and often produce poor joints (Figure 7e,f). Encoder-decoder architectures (AtlasNet, PCN) tend to significantly change the input geometry (Figure 7g,i). While modified AtlasNet* preserves the original geometry by construction, the synthesized joints do not seamlessly merge with the rest of the geometry (Figure 7h).

To quantitatively evaluate our joint synthesis method, we take eroded input parts from the same test shape, and use that test shape as the ground truth. We report chamfer distance and light field distance as error metrics in Table 2. Our results have the lowest errors in both measures. Although the customized AtlasNet* has errors close to ours in terms of chamfer distances, it does not have the same visual quality as our results, which is more evident from LFD.

	Chamfer Distance ($\times 10^3$)			Light Field Distance		
	plane	chair	mug	plane	chair	mug
AtlasNet	0.117	0.402	0.885	2402.1	2482.4	2227.9
AtlasNet*	0.067	0.205	0.277	1049.4	1316.8	455.2
PCN	0.095	0.308	0.506	-	-	-
3DCNN	0.151	0.268	0.440	2464.6	1446.2	615.8
3DResNet	0.141	0.262	0.421	2086.2	1313.2	426.9
Ours	0.067	0.181	0.274	895.4	894.6	196.2

Table 2: Quantitative results of different methods when the input parts are from the same object and aligned with respect to the ground truth positions. Chamfer Distance is multiplied by 10^3 . Light Field Distance cannot be applied to PCN since PCN generates point clouds rather than meshes.

	Chamfer Distance ($\times 10^3$)			Light Field Distance		
	plane	chair	mug	plane	chair	mug
AtlasNet	0.200	0.525	0.948	3462.5	2691.9	2240.4
AtlasNet*	0.092	0.258	0.288	1716.5	1517.9	485.5
PCN	0.173	0.489	0.583	-	-	-
3DCNN	0.199	0.330	0.495	2896.6	1563.2	655.4
3DResNet	0.179	0.319	0.485	2265.7	1414.0	449.2
Ours	0.080	0.222	0.287	1302.2	1039.9	198.7

Table 3: Quantitative results of different methods when the input parts are warped by a sine function.

	Chamfer Distance ($\times 10^3$)			Light Field Distance		
	plane	chair	mug	plane	chair	mug
AtlasNet	0.313	0.666	1.394	4158.8	3035.8	2843.1
AtlasNet*	0.109	0.255	0.302	2264.6	1835.1	737.2
PCN	0.274	0.512	0.989	-	-	-
3DCNN	0.219	0.314	0.486	3682.1	1570.3	719.8
3DResNet	0.177	0.304	0.476	2362.7	1403.7	505.9
Ours	0.077	0.190	0.292	1247.9	1063.0	231.9

Table 4: Quantitative results of different methods when the input parts are randomly scaled and translated.

To evaluate the robustness of our method and the baselines to part misalignment, we warp the test input parts and the ground-truth shape globally along the y-axis with a sine transform: $y \leftarrow y + 0.02 \sin(4\pi z + f)$, where y and z are the coordinates of a 3D point (x, y, z) , f is a per-shape random

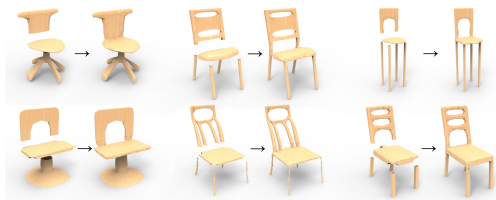


Figure 8: Qualitative results of our method trained on chairs from ShapeNet but tested on chairs from COSEG.



Figure 9: Our method can produce reasonable results even when a few parts, e.g., a chair back, are absent.

number sampled uniformly from $[-\pi, \pi]$, and 0.02 is the constant amplitude. The sine transform creates non-rigid shearing on the input parts, and moves open boundaries of the parts up/down with different amount of translations such that some misalignment is created. We report the quantitative performance of our method and the baselines in Table 3.

To evaluate the robustness of our method and the baselines to translation and scaling of the test shapes, we apply a global random scaling and translation to the input parts, with a uniformly-sampled scale in the range $[0.9, 1.1]$, and uniformly-sampled displacements in the range $[-0.04, 0.04]^3$. Table 4 shows the results under such random scaling and translation, demonstrating that our method is the most robust.

4.4. Generalizability

We also tested our network on the chair category of the COSEG [45] dataset. We do not re-train our network to fit the dataset. Instead, we use the network trained on the chair category from ShapeNet, and test it on COSEG to demonstrate how well it generalizes to novel datasets. Figure 8 shows some results to illustrate such generalizability.

5. Discussion, limitation, and Future Work

We presented COALESCE, the first neural network-based pipeline for aligning and seamlessly connecting parts in a 3D shape assembly. In interactive modeling scenarios, it is hard to perfectly align parts, and even harder (and often impossible) to make topologically and geometrically compatible cuts in adjacent parts for the purpose of seamless joining. COALESCE erodes crudely cut joint regions to prepare them for connections, corrects alignment errors in the assembly, and generates seamless connecting patches filling in the eroded joints, accounting for geometric and topological mismatches. Experiments show that our approach significantly outperforms alternatives, and is

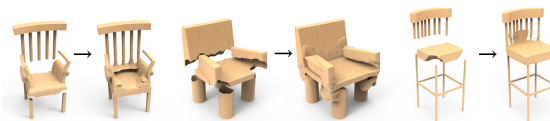


Figure 10: Examples of failure cases by our approach.

a promising addition to the assembly-based modeling toolbox, applying modern neural synthesis techniques to address a need largely neglected by prior work. Our method is also potentially useful for other applications, such as part-based 3D scanning of objects.

By simply setting the latent codes of absent parts as zero vectors, our network can be trained for objects with different part configurations. In the first two rows of Figure 7, we show that the same network trained with ShapeNet chairs can synthesize joints for both armed and armless chairs. The same network can also predict joints for a stool, or fill the holes on a mug when the handle is not provided, as shown in Figure 9. However, as one may expect, our network can fail to produce a natural part connection when the geometric and/or topological incompatibility between the parts becomes too significant or goes beyond what may be learned from data. Figure 10-left shows one such case, where erosion cannot sufficiently alleviate the part mismatch. Significant part incompatibility can also cause inaccurate alignment and lead to failure; see Figure 10-middle. Further, our method is limited by the capability of the PointNet++ encoders and the IM-decoder. When the input parts have complex thin structures, our method may not be able to synthesize detailed structures, e.g. Figure 10-right. In addition, the current version of our network may not generalize well to handle fine-level parts. One reason is that our architecture requires the latent codes from different parts to be concatenated. As the part count increases, our network will require more neurons, weights and GPU memory. Furthermore, fine-level part structures are much less consistent across different objects. Our network may not have sufficient capacity to adapt.

Our method is only a first step and still has several limitations, which suggest ripe avenues for future work. First, like most learning-based methods for shape analysis, COALESCE is still category-specific, and relies on predictable, labeled part layouts and joint locations. Also, our method has been trained on parts which are in a canonical orientation and compatibly scaled. Furthermore, we use heuristic erosion of the input parts, but we believe it is possible to extend this approach to learn a natural erosion operation.

Acknowledgements This research was supported in part by grants from NSERC and gifts from Adobe Research. We thank members of SFU GrUVi lab and Adobe Research for helpful discussions. Part of this work was done while Kangxue Yin was an intern at Adobe Research.

References

- [1] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas. Learning representations and generative models for 3d point clouds. In *ICML*, 2018. 3, 5
- [2] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An information-rich 3D model repository. *CoRR*, abs/1512.03012, 2015. 2
- [3] S. Chaudhuri, E. Kalogerakis, L. Guibas, and V. Koltun. Probabilistic reasoning for assembly-based 3d modeling. *ACM Trans. Graph.*, 30(4):35:1–35:10, 2011. 1, 2
- [4] S. Chaudhuri and V. Koltun. Data-driven suggestions for creativity support in 3D modeling. *ACM Trans. Graph.*, 29(6):183:1–9, 2010. 1
- [5] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, pages 5939–5948, 2019. 3, 4
- [6] A. Dai, C. R. Qi, and M. Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *CVPR*, 2017. 3
- [7] A. Dubrovina, F. Xia, P. Achlioptas, M. Shalah, R. Groskot, and L. Guibas. Composite shape modeling via latent space factorization. In *ICCV*, 2019. 2
- [8] N. Duncan, L.-F. Yu, and S.-K. Yeung. Interchangeable components for hands-on assembly based modelling. *ACM Trans. Graph.*, 35(6):234:1–234:14, 2016. 1, 2
- [9] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017. 3
- [10] L. Fan and P. Wonka. A probabilistic model for exteriors of residential buildings. *ACM Trans. Graph.*, 35(5):155:1–155:13, 2016. 2
- [11] T. Funkhouser, M. Kazhdan, P. Shilane, P. Min, W. Kiefer, A. Tal, S. Rusinkiewicz, and D. Dobkin. Modeling by example. *ACM Trans. Graph.*, 23(3):652–663, 2004. 1, 2
- [12] L. Gao, J. Yang, T. Wu, H. Fu, Y.-K. Lai, and H. Zhang. SDM-NET: Deep generative network for structured deformable mesh. *ACM Trans. Graph.*, 38(6), 2019. 3
- [13] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry. 3D-CODED: 3D correspondences by deep deformation. *ECCV*, 2018. 3
- [14] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation. *CVPR*, 2018. 3, 6
- [15] H. Huang, M. Gong, D. Cohen-Or, Y. Ouyang, F. Tao, and H. Zhang. Field-guided registration for feature-conforming shape composition. *ACM Trans. Graph. (SIGGRAPH Asia)*, 31(6):Article 179, 2012. 1, 2
- [16] H. Huang, E. Kalogerakis, and B. Marlin. Analysis and synthesis of 3D shape families via deep-learned generative models of surfaces. In *SGP*, 2015. 3
- [17] J. Huang, G. Zhan, Q. Fan, K. Mo, L. Shao, B. Chen, L. Guibas, and H. Dong. Generative 3d part assembly via dynamic graph learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3
- [18] A. Jain, T. Thormählen, T. Ritschel, and H.-P. Seidel. Exploring shape variations by 3D-model decomposition and part-based recombination. In *Eurographics*, 2012. 1, 2
- [19] P. Jaiswal, J. Huang, and R. Rai. Assembly-based conceptual 3D modeling with unlabeled components using probabilistic factor graph. *Comput. Aided Des.*, 74(C):45–54, 2016. 2
- [20] E. Kalogerakis, S. Chaudhuri, D. Koller, and V. Koltun. A probabilistic model for component-based shape synthesis. *ACM Trans. Graph. (SIGGRAPH)*, 31(4), 2012. 1, 2
- [21] V. G. Kim, W. Li, N. J. Mitra, S. Chaudhuri, S. DiVerdi, and T. Funkhouser. Learning part-based templates from large collections of 3D shapes. In *SIGGRAPH*, 2013. 1
- [22] V. Kraevoy, D. Julius, and A. Sheffer. Shuffler: Modeling with interchangeable parts. In *Pacific Graphics*, 2007. 1
- [23] V. Kraevoy, D. Julius, and A. Sheffer. Model composition from interchangeable components. In *Pacific Graphics*, pages 129–138, 2007. 2
- [24] H. Laga, M. Mortara, and M. Spagnuolo. Geometry and context for semantic correspondences and functionality recognition in man-made 3d shapes. *ACM Trans. Graph.*, 32(5):150, 2013. 1
- [25] T. Lescoat, P. Memari, J.-M. Thiery, M. Ovsjanikov, and T. Boubekeur. Connectivity-preserving smooth surface filling with sharp features. In *PG (Short Papers)*, pages 7–13, 2019. 2
- [26] J. Li, C. Niu, and K. Xu. Learning part generation and assembly for structure-aware shape synthesis. *arXiv preprint arXiv:1906.06693*, 2019. 3
- [27] J. Li, K. Xu, S. Chaudhuri, E. Yumer, H. Zhang, and L. Guibas. GRASS: Generative Recursive Autoencoders for Shape Structures. *ACM Trans. Graph.*, 36(4):52:1–52:14, 2017. 3
- [28] Y. Li, K. Mo, L. Shao, M. Sung, and L. Guibas. Learning 3d part assembly from a single image. *European conference on computer vision (ECCV 2020)*, 2020. 3
- [29] J. Lin, X. Jin, C. C. Wang, , and K.-C. Hui. Mesh composition on models with arbitrary boundary topology. *IEEE Trans. Visualization & Computer Graphics*, 14:653–665, 2008. 1, 2
- [30] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In *ACM Trans. Graph. (SIGGRAPH)*, volume 21, pages 163–169. ACM, 1987. 3
- [31] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *CVPR*, 2019. 3
- [32] N. Mitra, M. Wand, H. Zhang, D. Cohen-Or, and M. Bokeloh. Structure-aware shape processing. In *Eurographics State-of-the-art Report (STAR)*, 2013. 1
- [33] K. Mo, P. Guerrero, L. Yi, H. Su, P. Wonka, N. Mitra, and L. Guibas. StructureNet: Hierarchical graph networks for 3D shape generation. In *SIGGRAPH Asia*, 2019. 3
- [34] S. Muralikrishnan, V. G. Kim, M. Fisher, and S. Chaudhuri. Shape unicode: A unified shape representation. *CVPR*, 2019. 3
- [35] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 3
- [36] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *NeurIPS*, 2017. 3

- [37] D. Ritchie, S. Jobalia, and A. Thomas. Example-based authoring of procedural modeling programs with structural and continuous variability. *Comp. Graph. For.*, 37(2):401–413, 2018. 1, 2
- [38] R. Schmidt and K. Singh. Drag-and-drop surface composition. Technical Report CSRG-604, University of Toronto, 2009. 1, 2
- [39] N. Schor, O. Katzier, H. Zhang, and D. Cohen-Or. CompoNet: Learning to generate the unseen by part synthesis and composition. In *ICCV*, 2019. 2
- [40] A. Sharf, M. Blumenkrantz, A. Shamir, and D. Cohen-Or. SnapPaste: An interactive technique for easy mesh composition. In *Pacific Graphics*, 2006. 1, 2
- [41] M. Sung, H. Su, V. G. Kim, S. Chaudhuri, and L. Guibas. ComplementMe: Weakly-supervised component suggestions for 3D modeling. *ACM Trans. Graph. (SIGGRAPH Asia)*, 2017. 1, 2
- [42] J. Talton, L. Yang, R. Kumar, M. Lim, N. Goodman, and R. Měch. Learning design patterns with Bayesian grammar induction. In *UIST*, pages 63–74, 2012. 2
- [43] H. Wang, N. Schor, R. Hu, H. Huang, D. Cohen-Or, and H. Huang. Global-to-local generative model for 3d shapes. In *SIGGRAPH Asia*, page 214. ACM, 2018. 3
- [44] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong. Adaptive O-CNN: A Patch-based Deep Representation of 3D Shapes. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 37(6), 2018. 3
- [45] Y. Wang, S. Asafi, O. Van Kaick, H. Zhang, D. Cohen-Or, and B. Chen. Active co-analysis of a set of shapes. *SIGGRAPH Asia*, 31(6):165, 2012. 8
- [46] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NeurIPS*, 2016. 3
- [47] R. Wu, Y. Zhuang, K. Xu, H. Zhang, and B. Chen. PQ-NET: A generative part seq2seq network for 3d shapes. In *CVPR*, 2020. 2
- [48] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *CVPR*, 2015. 3
- [49] Z. Wu, X. Wang, D. Lin, D. Lischinski, D. Cohen-Or, and H. Huang. Sagnet: Structure-aware generative network for 3d-shape modeling. *SIGGRAPH Asia*, 38(4):91:1–91:14, 2019. 3
- [50] K. Xu, H. Zhang, D. Cohen-Or, and B. Chen. Fit and diverse: Set evolution for inspiring 3D shape galleries. *ACM Trans. Graph.*, 31(4):57:1–10, 2012. 1, 2
- [51] J. Yang, K. Mo, Y.-K. Lai, L. J. Guibas, and L. Gao. Dsmnet: Disentangled structured mesh net for controllable generation of fine geometry. *arXiv preprint arXiv:2008.05440*, 2020. 3
- [52] L. Yi, V. G. Kim, D. Ceylan, I.-C. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, and L. Guibas. A scalable active framework for region annotation in 3D shape collections. *SIGGRAPH Asia*, 35(6):1–12, 2016. 2, 5
- [53] Y. Yu, K. Zhou, D. Xu, X. Shi, H. Bao, B. Guo, and H.-Y. Shum. Mesh editing with poisson-based gradient field manipulation. *ACM Trans. Graph.*, 23(4):644–651, 2004. 1, 2
- [54] W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert. Pcn: Point completion network. In *3DV*, 2018. 6
- [55] M. E. Yumer and N. J. Mitra. Learning semantic deformation flows with 3D convolutional networks. In *ECCV*, 2016. 3
- [56] Y. Zheng, D. Cohen-Or, and N. J. Mitra. Smart variations: Functional substructures for part compatibility. *Comp. Graph. For.*, 32(2), 2013. 1
- [57] C. Zhu, K. Xu, S. Chaudhuri, R. Yi, and H. Zhang. SCORES: Shape composition with recursive substructure priors. *ACM Trans. Graph.*, 37(6), 2018. 2