

CompoNet: Learning to Generate the Unseen by Part Synthesis and Composition

Nadav Schor¹ Oren Katzir¹ Hao Zhang² Daniel Cohen-Or¹
 Tel Aviv University¹ Simon Fraser University²

nadav.schor@cs.tau.ac.il orenkatzir@mail.tau.ac.il haoz@cs.sfu.ca dcor@tau.ac.il

“For object recognition, the visual system decomposes shapes into parts, . . . , parts with their descriptions and spatial relations provide a first index into a memory of shapes.”

— Hoffman & Richards [18]

Abstract

Data-driven generative modeling has made remarkable progress by leveraging the power of deep neural networks. A reoccurring challenge is how to enable a model to generate a rich variety of samples from the entire target distribution, rather than only from a distribution confined to the training data. In other words, we would like the generative model to go beyond the observed samples and learn to generate “unseen”, yet still plausible, data. In our work, we present *CompoNet*, a generative neural network for 2D or 3D shapes that is based on a part-based prior, where the key idea is for the network to synthesize shapes by varying both the shape parts and their compositions. Treating a shape not as an unstructured whole, but as a (re-)composable set of deformable parts, adds a combinatorial dimension to the generative process to enrich the diversity of the output, encouraging the generator to venture more into the “unseen”. We show that our part-based model generates richer variety of plausible shapes compared with baseline generative models. To this end, we introduce two quantitative metrics to evaluate the diversity of a generative model and assess how well the generated data covers both the training data and unseen data from the same target distribution. Code is available at <https://github.com/nschor/CompoNet>.

1. Introduction

Learning generative models of shapes and images has been a long standing research problem in visual computing. Despite the remarkable progress made, an inherent and reoccurring limitation still remains: a generative model is often only as good as the given training data, as it is always trapped or bounded by the empirical distribution of the observed data. More often than not, what can be observed

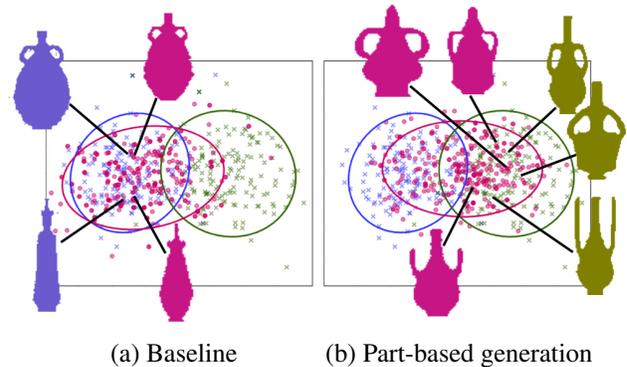


Figure 1: *CompoNet*, our part-based generative model (b) covers the “unseen” data significantly more than the baseline (a). Generated data (pink dots) by two methods are displayed over training data (purple crosses) and unseen data (green crosses) from the same target distribution. The data is displayed via PCA over a classifier feature space, with the three distributions summarized by ellipses, for illustration only. A few samples of training, unseen, and generated data are displayed to reveal their similarity/dissimilarity.

is not sufficiently expressive of the true target distribution. Hence, the generative power of a learned model should not only be judged by the plausibility of the generated data as confined by the training set, but also its *diversity*, in particular, by the model’s ability to generate plausible data that is sufficiently far from the training set. Since the target distribution which encompasses both the observed and unseen data is unknown, the main challenge is how to effectively train a network to learn to generate the “unseen”, without making any compromising assumption about the target distribution. Due to the same reason, even evaluating the generative power of such a network is a non-trivial task.

We believe that the key to generative diversity is to enable more drastic changes, i.e., non-local and/or structural transformations, to the training data. At the same time, such changes must be within the confines of the target data distribution. In our work, we focus on generative modeling of 2D or 3D *shapes*, where the typical modeling constraint is

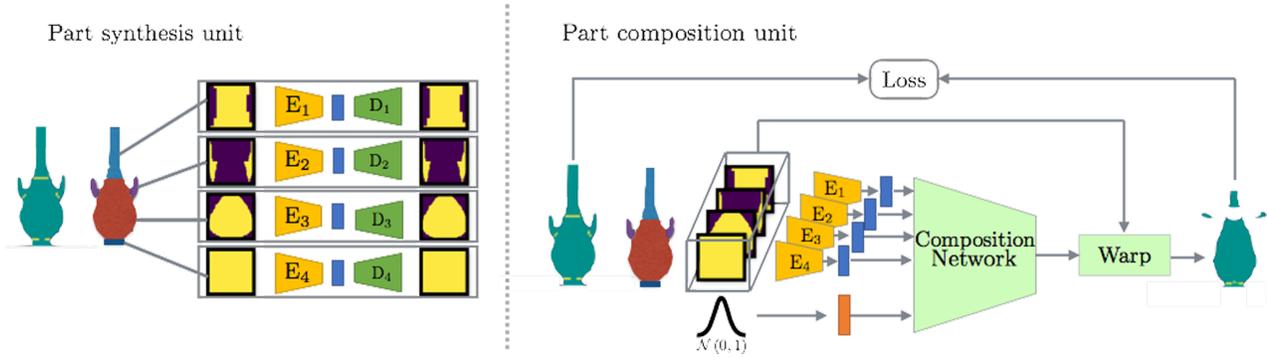


Figure 2: The two training units of CompoNet, our part-based generative model: The first unit, the *part synthesis unit*, consists of parallel generative AEs; an independent AE for each semantic part of the shape. The second unit, the *part composition unit*, learns to compose the encoded parts. We use the pre-trained part encoders from the part synthesis unit. Then, a noise vector z , is concatenated to the parts’ latent representation and fed to the composition network, which outputs transformation parameters per part. The parts are then warped and combined to generate the entire input sample.

to produce shapes belonging to the same category as the exemplars, e.g., chairs or vases. We develop a generative deep neural network based on a *part-based* prior. That is, we assume that shapes in the target distribution are composed of parts, e.g., chair backs or airplane wings. The network, coined CompoNet, is designed to synthesize novel parts, *independently*, and compose them to form a complete shape.

It is well-known that object recognition is intricately tied to reasoning about parts and part relations [18, 43]. Hence, building a generative model based on varying parts and their compositions, while respecting category-specific part priors, is a natural choice and also facilitates grounding of the generated data to the target object category. More importantly, treating a shape as a (re-)composable set of parts, instead of a whole entity, adds a *combinatorial* dimension to the generative model and improves its diversity. By synthesizing parts independently and then composing them, our network enables both part variation and novel combination of parts, which induces non-local and more drastic shape transformations. Rather than sampling only a single distribution to generate a whole shape, our generative model samples both the geometric distributions of individual parts and the combinatorial varieties arising from part compositions, which encourages the generative process to venture more into the “unseen”, as shown in Figure 1.

While the part-based approach is generic and not strictly confined to specific generative network architecture, we develop a *generative autoencoder* to demonstrate its potential. Our generative AE consists of two parts. In the first, we learn a distinct part-level generative model. In the second stage, we concatenate these learned latent representation with a random vector, to generate a new latent representation for the entire shape. These latent representations are fed into a conditional parts compositional network, which is based on a spatial transformer network (STN) [22].

We are not the first to develop deep neural networks for part-based modeling. Some networks learn to compose images [29, 3] or 3D shapes [23, 6, 53], by combining *existing* parts sampled from a training set or provided as input to the networks. In contrast, our network is fully generative as it learns both novel part synthesis and composition. Wang *et al.* [44] train a generative adversarial network (GAN) to produce semantically segmented 3D shapes and then refine the part geometries using an autoencoder network. Li *et al.* [28] train a VAE-GAN to generate structural hierarchies formed by bounding boxes of object parts and then fill in the part geometries using a separate neural network. Both works take a coarse-to-fine approach and generate a rough 3D shape holistically from a noise vector. In contrast, our network is trained to perform both part synthesis and part composition (with noise augmentation); see Figure 2. Our method also allows the generation of more *diverse parts*, since we place less constraints per part while holistic models are constrained to generate all parts at once.

We show that the part-based CompoNet produces plausible outputs that better cover unobserved regions of the target distribution, compared to baseline approaches, e.g., [1]. This is validated over random splits of a set of shapes belonging to the same category into a “seen” subset, for training, and an “unseen” subset. In addition, to evaluate the generative power of our network relative to baseline approaches, we introduce two quantitative metrics to assess how well the generated data covers both the training data and the unseen data from the same target distribution.

2. Background and Related Work

Generative neural networks. In recent years, generative modeling has gained much attention within the deep learning framework. Two of the most commonly used deep gen-

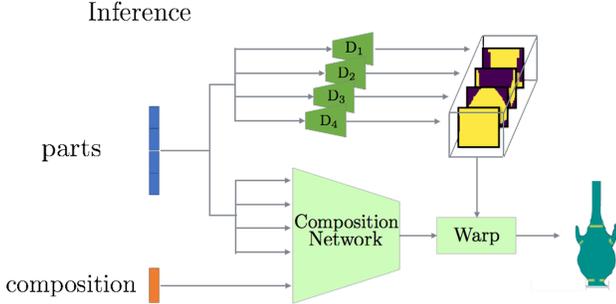


Figure 3: Novel shape generation at inference time. We randomly sample the latent spaces for shape parts and part composition. Using the pre-trained part-decoders and the composition network, we generate novel parts and then warp them to produce a coherent whole shape.

erative models are variational auto-encoders (VAE) [25] and generative adversarial networks (GAN) [15]. Both methods have made remarkable progress in image and shape generation problems [47, 21, 37, 54, 45, 49, 44].

Many works are devoted to improve the basic models and their training. In [16, 31, 4], new cost functions are suggested to achieve smooth and non-vanishing gradients. Sohn *et al.* [41] and Odena *et al.* [35] proposed conditional generative models, based on VAE and GAN, respectively. Hoang *et al.* [17] train multiple generators to explore different modes of the data distribution. Similarly, MIX+GAN [2] uses a mixture of generators to improve diversity of the generated distribution, while a combination of multiple discriminators and a single generator aims at constructing a stronger discriminator to guide the generator. GMAN [10] explores an array of discriminators to boost generator learning. Some methods [20, 30, 51] use a global discriminator together with multiple local discriminators.

Following PointNet [36], a generative model that works directly on point clouds was proposed. Achlioptas *et al.* [1] proposed an AE+GMM generative model for point clouds, which is considered state-of-the-art.

Our work is orthogonal to these methods. We address the case where the generator is unable to generate other valid samples since they are not well represented in the training data. We show that our part-based priors can assist the generation process and extend the generator’s capabilities.

Learning-based shape synthesis. Li *et al.* [28] present a top-down and structure-oriented approach for 3D shape generation. They learn symmetry hierarchies [46] of shapes with an autoencoder and generate variations of these hierarchies using an VAE-GAN. The nodes of the hierarchies are independently instantiated with parts. However, these parts are not necessarily connected and their aggregation does not form a coherent connected shape. In our work, the shapes are generated coherently as a whole, and special care

is given to the inter-parts relation and their connectivity.

Most relevant to our work is the shape variational auto-encoder by Nash and Williams [34], where a point-cloud based autoencoder is developed to learn a low-dimensional latent space. Then novel shapes can be generated by sampling vectors in the learned space. Like our method, the generated shapes are segmented into semantic parts. In contrast however, they require a one-to-one dense correspondence among the training shapes, since they represent the shapes as an order vector. Their autoencoder learns the overall (global) 3D shapes with no attention to the local details. Our approach pays particular attention to both the generated parts and their composition.

Inverse procedural modeling aims to learn a generative procedure from a given set of exemplars. Some recent works, e.g., [38, 52, 39], have focused on developing neural models, such as autoencoders, to generate the shape synthesis procedures or programs. However, current inverse procedural modeling methods are not designed to generate unseen data that are away from the exemplars.

Assembly-based synthesis. The early and seminal work of Funkhouser *et al.* [14] composes new shapes by retrieving relevant shapes from a repository, extracting shape parts, and gluing them together. Many follow-up works [5, 42, 7, 23, 48, 24, 12, 19] improve the modeling process with more sophisticated techniques that consider the part relations or shape structures, e.g., employing Bayesian networks or modular templates. We refer to recent surveys [32, 33] for an overview of these and related works.

In the image domain, recent works [29, 3] develop neural networks to assemble images or scenes from existing components. These works utilized an STN [22] to compose the components to a coherent image/scene. In our work, an STN is integrated as an example for prior information regarding the data generation process. In contrast to previous works, we first synthesize parts using multiple generative AEs and then employ an STN to compose the parts.

Recent concurrent efforts [9, 27] also propose deep neural networks for shape modeling using a part-based prior, but on voxelized representations. Dubrovina *et al.* [9] encode shapes into a factorized embedding space, where shape composition and decomposition become simple linear operations on the embedding coordinates, allowing both shape reconstruction and part exchange. While this work was not going after generative diversity, the network of Li *et al.* [27] also combines part generation with assembly. Their results reinforce our premise that shape generation using part synthesis and composition does improve diversity, which is measured using inception scores in their work.

3. Method

In this section, we present CompoNet, our generative model which learns to synthesize shapes that can be represented as a composition of distinct parts. At training time, every shape is pre-segmented to its semantic parts, and we assume that the parts are independent of each other. Thus, every combination of parts is valid, even if the training set may not include it. As shown in Figure 2, CompoNet consists of two units: a generative model of parts and a unit that combines the generated parts into a global shape.

3.1. Part synthesis unit

We first train a generative model that estimates the marginal distribution of each part separately. In the 2D case, we use a standard VAE as the part generative model, and train an individual VAE for each semantic part. Thus, each part is fed into a different VAE and is mapped onto a separate latent distribution. The encoder consists of several convolutional layers followed by Leaky-ReLU activation functions. The final layer of the encoder is a fully connected layer producing the latent distribution parameters. Using the reparameterization trick, the latent distribution is sampled and decoded to reconstruct each individual input part. The decoder mirrors the encoder network, applying a fully connected layer followed by transposed convolution layers with ReLU non-linearity functions. In the 3D case, we borrow an idea from Achlioptas *et al.* [1], and replace the VAE with an AE+GMM, where we approximate the latent space of the AE by using a GMM. The encoder is based on PointNet [36] architecture and the decoder consists of fully-connected layers. The part synthesis process is visualized in Figure 2, part synthesis unit.

Once the part synthesis unit is trained, the part encoders are fixed, and are used to train the part composition unit.

3.2. Parts composition unit

This unit composes the different parts into a coherent shape. Given a shape and its parts, where missing parts are represented by null shapes (i.e., zeros), the pre-trained encoders encode the corresponding parts (marked in blue in Figure 2). At training time, these generated codes are fed into a composition network which learns to produce transformation parameters per part (scale and translation), such that the composition of all the parts forms a coherent complete shape. The loss measures the similarity between the input shape and the composed shape. We use Intersection-over-Union (IoU) as our metric in the 2D domain, and Chamfer distance for the 3D domain, where the Chamfer distance is given by

$$d_C(Q, P) = \sum_{q \in Q} \min_{p \in P} (q - p)^2 + \sum_{p \in P} \min_{q \in Q} (p - q)^2, \quad (1)$$

where P and Q are point clouds which represent the 3D shapes. Note that the composition network yields a set of affine (similarity) transformations, which are applied on the input parts, and does not directly synthesize the output.

The composition network does not learn the composition based solely on part codes, but also relies on an input noise vector. This network is another generative model on its own, generating the scale and translation from the noise, conditioned on the codes of the semantic parts. This additional generative model enriches the variation of the generated shapes, beyond the generation of the parts.

3.3. Novel shape generation

At inference time, we sample the composition vector from a normal distribution. In the 2D case, since we use VAEs, we sample the part codes from normal distribution as well. For 3D, we sample the code of each part from its GMM distribution, randomly sampling one of the Gaussians. When generating a new shape with a missing part, we use the embedding of the part’s null vector, and synthesize the shape from that compound latent vector; see Figure 3. We feed each section of the latent vector representing a part to its associated pre-trained decoder from the part synthesis unit, to generate novel parts. In parallel, the entire shape representation vector is fed to the composition network to generate scale and translation parameters for each part. The synthesized parts are then warped according to the generated transformations and combined to form a novel shape.

4. Architecture and implementation details

The backbone architecture of our part based synthesis is an AE: VAE for 2D and AE+GMM for 3D.

4.1. Part-based generation

2D Shapes. The input parts are assumed to have a size of $64 \times 64 \times 1$. We denote $C(k)$ ($TC(k)$) as a 2D convolution (transpose convolution) layer with k filters of size 5×5 and stride 2, followed by batch normalization and a leaky-ReLU (ReLU) activation. A fully-connected layer with k outputs is denoted by $L(k)$. The encoder takes a 2D part as input and has the structure of $C(8) - C(16) - C(32) - C(64) - L(10)$. The decoder mirrors the encoder as $L(1024) - TC(32) - TC(16) - TC(8) - TCS(1)$, where in the last layer, TCS , we omitted batch normalization and replaced ReLU by a Sigmoid activation. The output of the decoder is equal in size to the 2D part input, $(64 \times 64 \times 1)$. We use an Adam optimizer with learning rate $= 2e^{-4}$, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The batch size is set to 64.

3D point clouds. Our input parts are assumed to have a fixed number of point for each part. Different parts can vary in number of points, but this becomes immutable once

training has started. We used 400 points per part. We denote MP as a feature-wise max-pooling layer and $1DC(k)$ as a 1D convolution layer with k filters of size 1 and stride 1, followed by a batch normalization layer and a ReLU activation function. The encoder takes a part with 400×3 as input. The encoder structure is $1DC(64) - 1DC(64) - 1DC(64) - 1DC(128) - 1DC(64) - MP$. The decoder consist of fully-connected layers. We denote $L(k)$ to be a fully-connected layer with k output nodes, followed by a batch normalization layer and a ReLU activation function. The decoder takes a latent vector of size 64 as input. The decoder structure is $L(256) - L(256) - LC(400 \times 3)$, where in the last layer, LC , we omitted the batch-normalization layer and the ReLU activation function. The output of the decode is equal in size to the input (400×3). For each AE, we use a GMM with 20 Gaussians, to model their latent space distribution. We use an Adam optimizer with learning rate = 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is set to 64.

4.2. Part composition

2D. The composition network encodes each semantic part by the associated pre-trained VAE encoder, producing a 10-dim vector for each part. The composition noise vector is set to be 8-dim. The part codes are concatenated together with the noise, yielding a 48-dim vector. The composition network structure is $L(128) - L(128) - L(16)$. Each fully connected layer is followed by a batch normalization layer, a ReLU activation function, and a Dropout layer with keep rate of 0.8, except for the last layer. The last layer outputs a 16-dim vector, four values per part. These four values represent the scale and translation in the x and y axes. We use the grid generator and sampler, suggested by [22], to perform differential transformation. The scale is initialized to 1 and the translation to 0. We use a per-part IoU loss and an Adam optimizer with learning rate = 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is set 64.

3D. The composition network encodes each semantic part by the associated pre-trained AE encoder, producing a 64-dim vector for each part. The composition noise vector is set to size 16. The parts codes are concatenated together with the noise vector, yielding a 272-dim vector. The composition network structure is $L(256) - L(128) - L(24)$. Each fully connected layer is followed by a batch normalization layer and a ReLU activation function, except for the last layer. The last layer outputs a 24-dim vector, six values per part. These six values represent the scale and translation in the x , y and z axes. The scale is initialized to 1 and the translation to 0. We then reshape the output vector to match an affine transformation matrix:

$$M = \begin{bmatrix} s_x & 0 & 0 & t_x \\ 0 & s_y & 0 & t_y \\ 0 & 0 & s_z & t_z \end{bmatrix} \quad (2)$$

The task of performing an affine transformation on point clouds is easy, we simply concatenate 1 to each point $(x, y, z, 1)$ and multiply the transformation matrix with each point. We use Chamfer distance loss and an Adam optimizer with learning rate = 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is set 64.

5. Results and evaluation

In this section, we analyze the results of applying our generative approach to 2D and 3D shape collections.

5.1. Datasets

Projected COSEG. We used the COSEG dataset [40] which consists of 300 vases, segmented to four different semantic labels: top, handle, body and base (each vase may or may not contain any of these parts). Similar to the projection procedure in [13], each vase is projected from the main view to constitute a collection of 300 silhouettes of size 64×64 , where each semantic part is stored in a different channel. In addition, we create four sets, one per part. The parts are normalized by finding their axis-aligned bounding box and stretching it to a 64×64 resolution.

Shape-Net. For 3D data, we chose to demonstrate our method on point clouds taken from ShapeNet part dataset [50]. We chose to focus on two categories: chairs and airplanes. Point clouds, compared to 3D voxels, enable higher resolution while keeping the model complexity relatively low. Similar to the 2D case, each shape is divided into its semantic parts (Chair: legs, back, seat and arm-rests, Airplane: tail, body, engine and wings). We first normalize each shape to the unit square. We require an equal number of points N in each point cloud, thus, we randomly sample each part to $N = 400$ points. If a part consists of $M < N$ points, we randomly duplicate $N - M$ of its points (since our non-local operation preforms only max global pooling, the duplication of points has no affect on the embedding of the shape). This random sampling process occurs every epoch. For consistency between the shape and its parts, we first normalize the original parts to the unit square, and only then sample (or duplicate) the same points that were selected to generate the complete sampled shape.

Seen and Unseen splits. To properly evaluate the diversity of our generative model, we divide the resulting collections into two subsets: (i) training (*seen*) set and (ii) *unseen* set. The term *unseen* emphasizes that unlike the nominal division into train and test sets, the unseen set is not represented well in the training set. Thus, there exists an unbridgeable gap, by an holistic approach, between the unseen set and the training set. To avoid bias during evaluation, we preform several random splits for each seen-unseen split percentage (e.g., 15%-85% seen-unseen in the 3D case; see

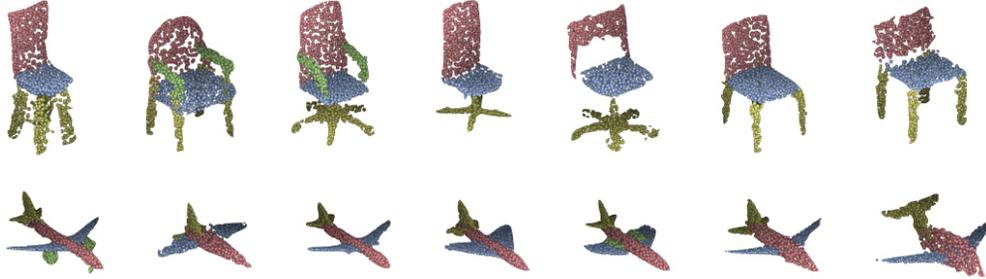


Figure 4: Representative samples of 3D shapes generated by our part-based generative network, CompoNet.



Figure 5: A gallery of 10 randomly sampled vases generated by CompoNet (top row) and below, their 3-nearest-neighbors from the training set, based on pixel-wise Euclidean distance. One can observe that the generated vases are different from their nearest neighbors.

Table 1). In the 2D case, since the dataset is much smaller, we used 50%-50% split between the training and unseen sets. In both cases, the unseen set is used to evaluate the ability of a model to generate diverse shapes.

5.2. Baselines

For 2D shapes, we use a naive model - a one-channel VAE. Its structure is identical to the part VAE with a latent space of 48-dim. We feed it with a binary representation of the data (a silhouette) as input. We use an Adam optimizer with learning rate $= 2e^{-4}$, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The batch size is set to 64. In the 3D case, we use two baselines; (i) a WGAN-GP [16] for point clouds and (ii) an AE+GMM model [1], which generates remarkable 3D point cloud results. We train the baseline models using our 400-points per part data set (1600×3 per shape). We use [1] official implementation and parameters, which also includes the WGAN-GP implementation.

5.3. Qualitative evaluation

We evaluate our network, CompoNet, on 2D data and 3D point clouds. Figure 4 shows some generated 3D results. Unlike other naive approaches, we are able to generate versatile samples beyond the empirical distribution. In order to visualize the versatility, we present the nearest neighbor

of the generated samples in the training set. As shown in Figure 5, for the 2D case, samples generated by our generative approach differ from the closest training samples. In Figure 6 we also compare this qualitative diversity measure with the baseline [1], showing that our generated samples are more distinct from their nearest neighbors in the training set, compared to those generated by the baseline. In the following sections, we quantify this attribute. More generated results can be found in the supplementary material.

5.4. Quantitative evaluation

We quantify the ability of our model to generate realistic unseen samples using two novel metrics. To evaluate our model, we use 5,000 randomly sampled shapes from our trained model and from the baselines.

k -set-coverage. We define the k -set-coverage of set A by set B as the percentage P of shapes from A which are one of the k -nearest-neighbors of some shape in B . Thus, if the set B is similar only to a small part of set A , the k -Set-Coverage will be small and vice-versa. In our case, we calculate the nearest neighbors using Chamfer distance. In Figure 7, we compare the k -set-coverage of the unseen set and the training set by our generated data and by the baseline [1] generated data. It is clear that the baseline covers the training better, since most of its samples lie close to it. However, the unseen set is covered poorly by the baseline, for all k , while, our method, balances between generating seen samples and unseen samples.

Diversity. We develop a second measure to quantify the generated unseen data, which relies on a trained classifier to distinguish between the training and the unseen set. Then, we measure the percentage of generated shapes which are classified as belonging to the unseen set. The classifier architecture is a straight-forward adaption of the encoder from the part synthesis unit of the training process, followed by fully connected layers which classify between unseen and train sets (see supplementary file for details).

Table 1 shows some classification results for generated vases, chairs, and airplanes by our method and two baselines. We can observe that when the seen set is relatively

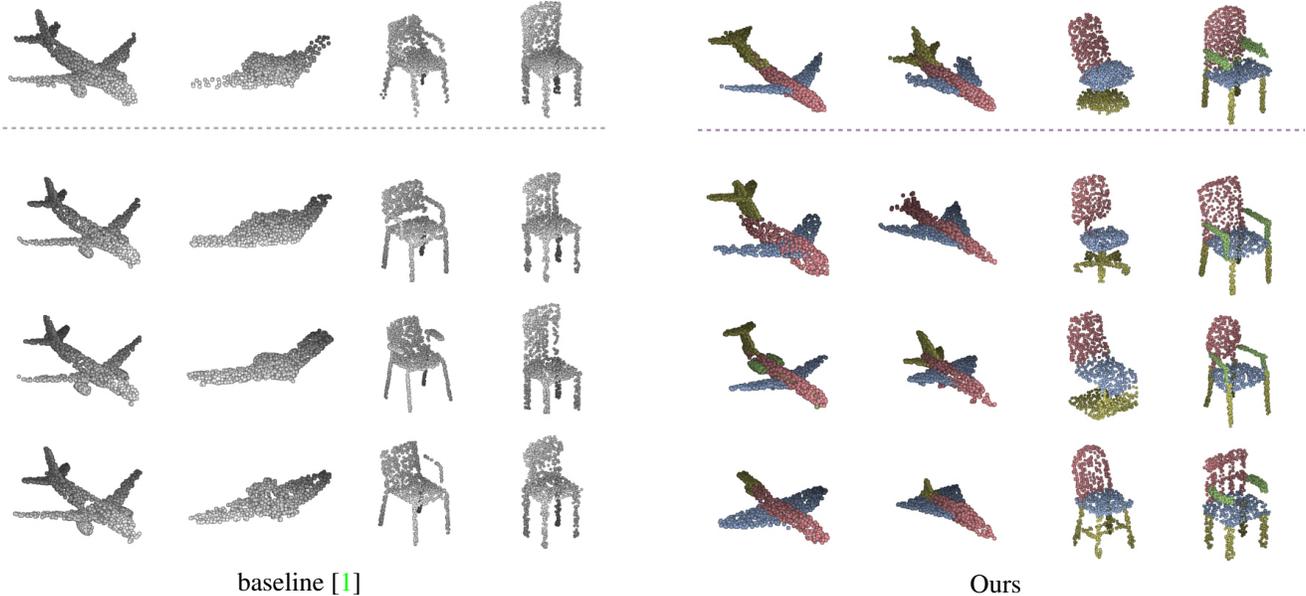


Figure 6: Qualitative comparison of diversity. The generated data of both methods (first row) is realistic. However, searching for the nearest neighbors of the generated data in the training set (rows two to four) reveals that our method (right side) exhibits more diversity compared to the baseline [1] (left side). Please note that the baseline’s shapes are presented in gray to emphasize that the baseline generates the entire shape, which is not segmented to semantic parts.

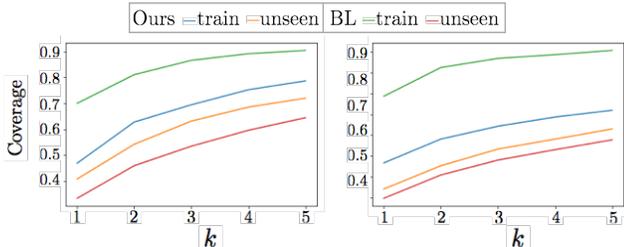


Figure 7: k -set-coverage comparison for chairs (left) and airplanes (right) point clouds sets. We generate an equal number of samples in both our method and the baseline [1] and calculate the k -set-coverage of both the training set and unseen set by them. While the baseline covers the training set almost perfectly, it has lower coverage of the unseen set. Our method CompoNet balances between generating samples similar to the training set and the unseen set.

small, e.g., 5% or 15% of the total, our model clearly performs better than the baselines in terms of generative diversity, as exhibited by the higher levels of coverage over the unseen set. However, as the seen set increases in size, e.g., to 30%, the difference between our method and the baselines becomes smaller. We believe that this trend is not a reflection that our method starts to generate less diverse samples, but rather that the unseen set is becoming more similar to the seen set, hence less diverse itself.

To visualize the coverage of seen/unseen regions by the

generated samples, we use the classifier’s embedding (the layer before the final fully-connected layer) and reduce its dimension by projecting it onto the 2D PCA plane, as shown in Figures 1 and 8. The training and unseen sets have overlap in this representation, reflecting data which is similar between the two sets. While both methods are able to generate unseen samples in the overlap region, the baseline samples are biased toward the training set. In contrary, our generated samples are closer to the unseen.

JSD. The Jensen-Shannon divergence is a distance measure between two probability distributions and is given by

$$\text{JSD}(S||T) = \frac{1}{2}D(S||M) + \frac{1}{2}D(T||M), \quad (3)$$

where S, T are probability distributions, $M = \frac{1}{2}(S + T)$ and D is the KL-divergence [26]. Following [1] we define the occupancy probability distribution for a set of point clouds by counting the number of points lying within each voxel in a regular voxel grid. Assuming the point clouds are normalized and axis-aligned, the JSD between two such probabilities measure the degree to which two point cloud sets occupy similar locations. Thus, we calculate the occupancy matrix for the unseen set and compare it to the occupancy matrix of samples generated by our method and the baselines. The results are summarized in Table 2 and clearly show that our generated samples are closer to the unseen set structure. The number of voxels used is 28^3 , as in [1] and the size of all point clouds sets is equal.

Category	Vase	Chair	Airplane
VAE	0.3±0.08	-	-
WGAN-GP	-	0.67±0.02	0.66±0.03
AE+GMM [1]	-	0.61±0.03	0.67±0.12
Ours	0.46±0.08	0.76±0.06	0.8±0.02

(a) 2D vase: 50%-50% (seen-unseen); 3D: 15%-85% (seen-unseen)

Category	WGAN-GP	AE+GMM [1]	Ours
Chair	0.75±0.05	0.57±0.06	0.87±0.05
Airplane	0.76±0.07	0.43±0.1	0.86±0.04

(b) 5%-95% (seen-unseen)

Category	WGAN-GP	AE+GMM [1]	Ours
Chair	0.54±0.07	0.63±0.07	0.55±0.01
Airplane	0.61±0.07	0.52±0.12	0.65±0.1

(c) 30%-70% (seen-unseen)

Table 1: Comparing generative diversity between baselines and CompoNet. We report the percentage of generated samples that are classified as belonging to the unseen set, averaged over five random splits, for three split percentages.

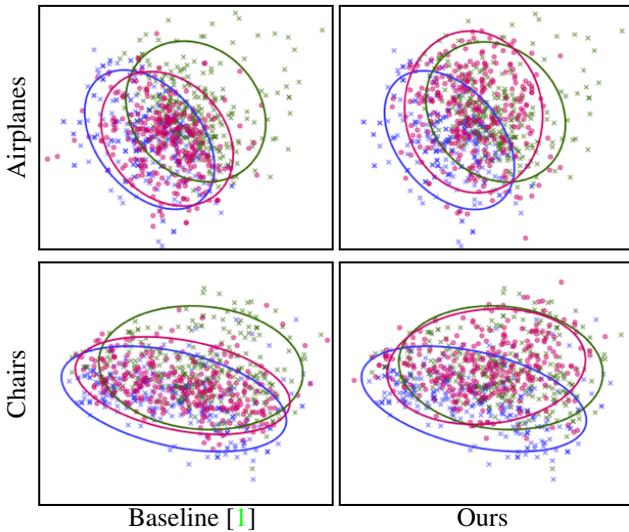


Figure 8: A visualization of the classifiers’ feature spaces for 3D data. The classifier is trained to distinguish between the training set (purple crosses) and the unseen set (green crosses). The two sets are clearly visible in the resulting space. While the baseline method [1] (pink dots) generates samples similar to the seen set, CompoNet (pink dots) generates samples in the unseen region.

6. Conclusion, limitation, and future work

We believe that effective generative models should strive to venture more into the “unseen” data of a target distribution, beyond the observed exemplars from the training set.

Category	WGAN-GP	AE+GMM [1]	Ours
Chair	0.3±0.02	0.19±0.006	0.02±0.003
Airplanes	0.32±0.016	0.14±0.007	0.07±0.013

Table 2: The JSD distance between the unseen set and the generated samples from CompoNet and the baselines, averaged over five random seen-unseen splits.

Covering both the seen and the unseen implies that the generated data is both *fit* and *diverse* [48]. Fitness constrains the generated data to be close to data from the target domain, both the seen and the unseen. Diversity ensures that the generated data is not confined only to the seen data.

We have presented a generic approach for “fit-n-diverse” shape modeling based on a part-based prior, where a shape is not viewed as an unstructured whole but as the result of a coherent composition of a set of parts. This is realized by CompoNet, a novel deep generative network composed of a part synthesis unit and a part composition unit. Novel shapes are generated via inference over random samples taken from the latent spaces of shape parts and part compositions. Our work also contributes two novel measures to evaluate generative models: the k -set-coverage and a diversity measure which quantifies the percentage of generated data classified as “unseen” vs. data from the training set.

Compared to baseline approaches, our generative network demonstrates superiority, but still somewhat limited diversity, since the generative power of the part-based approach is far from being fully realized. Foremost, an intrinsic limitation of our composition mechanism is that it is still “in place”: it does not allow changes to part structures or feature transfers between different part classes. For example, enabling a simple symmetric switch in part composition would allow the generation of right hand images when all the training images are of the left hand.

CompoNet can be directly applied for generative modeling of organic shapes. But in terms of plausibility, such shapes place a more stringent requirement on coherent and smooth part connections, an issue that our current method does not account for. Perfecting part connections can be a post process. Learning a deep model for the task is worth pursuing for future work. Our current method is also limited by the spatial transformations allowed by the STN during part composition. As a result, we can only deal with man-made shapes without part articulation.

As more immediate future work, we would like to apply our approach to more complex datasets, where parts can be defined during learning. In general, we believe that more research shall focus on other generation related prior information, besides parts-based priors. Further down the line, we envision that the fit-n-diverse approach, with generative diversity, will form a baseline for *creative modeling* [8], po-

tentially allowing part exchanges across different object categories. This may have to involve certain perceptual studies or scores, to judge creativity. The compelling challenge is how to define a generative neural network with sufficient diversity to cross the line of being creative [11].

References

- [1] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. J. Guibas. Learning representations and generative models for 3d point clouds. *arXiv preprint arXiv:1707.02392*, 2017. 2, 3, 4, 6, 7, 8
- [2] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *Proc. Int. Conf. on Machine Learning*, volume 70, pages 224–232, 2017. 3
- [3] S. Azadi, D. Pathak, S. Ebrahimi, and T. Darrell. Compositional gan: Learning conditional image composition. *arXiv preprint arXiv:1807.07560*, 2018. 2, 3
- [4] D. Berthelot, T. Schumm, and L. Metz. Began: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 3
- [5] M. Bokeloh, M. Wand, and H.-P. Seidel. A connection between partial symmetry and inverse procedural modeling. *ACM Transactions on Graphics (TOG)*, 29(4):104:1–104:10, 2010. 3
- [6] S. Chaudhuri, E. Kalogerakis, L. Guibas, and V. Koltun. Probabilistic reasoning for assembly-based 3D modeling. *ACM Trans. on Graphics*, 30(4):35:1–35:10, 2011. 2
- [7] S. Chaudhuri and V. Koltun. Data-driven suggestions for creativity support in 3D modeling. *ACM Trans. on Graphics*, 29(6):183:1–183:10, 2010. 3
- [8] D. Cohen-Or and H. Zhang. From inspired modeling to creative modeling. *The Visual Computer*, 32(1):1–8, 2016. 8
- [9] A. Dubrovina, F. Xia, P. Achlioptas, M. Shalah, and L. Guibas. Composite shape modeling via latent space factorization. *arXiv preprint arXiv:1901.02968*, 2019. 3
- [10] I. Durugkar, I. Gemp, and S. Mahadevan. Generative multi-adversarial networks. *arXiv preprint arXiv:1611.01673*, 2016. 3
- [11] A. M. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone. CAN: creative adversarial networks, generating “art” by learning about styles and deviating from style norms. *CoRR*, abs/1706.07068, 2017. 9
- [12] N. Fish, M. Averkiou, O. van Kaick, O. Sorkine-Hornung, D. Cohen-Or, and N. J. Mitra. Meta-representation of shape families. *ACM Trans. on Graphics*, 33(4):34:1–34:11, 2014. 3
- [13] N. Fish, O. van Kaick, A. Bermano, and D. Cohen-Or. Structure-oriented networks of shape collections. *ACM Transactions on Graphics (TOG)*, 35(6):171, 2016. 5
- [14] T. Funkhouser, M. Kazhdan, P. Shilane, P. Min, W. Kiefer, A. Tal, S. Rusinkiewicz, and D. Dobkin. Modeling by example. *ACM Trans. on Graphics*, 23(3):652–663, 2004. 3
- [15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3
- [16] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5769–5779, 2017. 3, 6
- [17] Q. Hoang, T. D. Nguyen, T. Le, and D. Phung. Multi-generator generative adversarial nets. *arXiv preprint arXiv:1708.02556*, 2017. 3
- [18] D. D. Hoffman and W. A. Richards. Parts of recognition. *Cognition*, pages 65–96, 1984. 1, 2
- [19] H. Huang, E. Kalogerakis, and B. Marlin. Analysis and synthesis of 3d shape families via deep-learned generative models of surfaces. In *Computer Graphics Forum*, volume 34, pages 25–38. Wiley Online Library, 2015. 3
- [20] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Trans. on Graphics*, 36(4):107:1–107:14, 2017. 3
- [21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3
- [22] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 2, 3, 5
- [23] E. Kalogerakis, S. Chaudhuri, D. Koller, and V. Koltun. A probabilistic model for component-based shape synthesis. *ACM Trans. on Graphics*, 31(4):55:1–55:11, 2012. 2, 3
- [24] 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. *ACM Trans. on Graphics*, 32(4):70:1–70:12, 2013. 3
- [25] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proc. Int. Conf. on Learning Representations*, 2014. 3
- [26] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951. 7
- [27] J. Li, C. Niu, and K. Xu. Learning part generation and assembly for structure-aware shape synthesis. *arXiv preprint arXiv:1906.06693*, 2019. 3
- [28] J. Li, K. Xu, S. Chaudhuri, E. Yumer, H. Zhang, and L. Guibas. Grass: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics (TOG)*, 36(4):52, 2017. 2, 3
- [29] C. Lin, E. Yumer, O. Wang, E. Shechtman, and S. Lucey. Stgan: Spatial transformer generative adversarial networks for image compositing. In *Computer Vision and Pattern Recognition, 2018. CVPR 2018. IEEE Conference on*, pages –. –, 2018. 2, 3
- [30] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. *arXiv preprint arXiv:1804.07723*, 2018. 3
- [31] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2813–2821. IEEE, 2017. 3
- [32] N. Mitra, M. Wand, H. Zhang, D. Cohen-Or, and M. Bokeloh. Structure-aware shape processing. *Com-*

- puter Graphics Forum (Eurographics State-of-the-art Report)*, pages 175–197, 2013. 3
- [33] N. Mitra, M. Wand, H. R. Zhang, D. Cohen-Or, V. Kim, and Q.-X. Huang. Structure-aware shape processing. In *SIGGRAPH Asia 2013 Courses*, pages 1:1–1:20, 2013. 3
- [34] C. Nash and C. K. I. Williams. The shape variational autoencoder: A deep generative model of part-segmented 3D objects. *Computer Graphics Forum*, 36(5):1–12, 2017. 3
- [35] A. Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016. 3
- [36] C. R. Qi, H. Su, K. Mo, and L. 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, 4
- [37] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 3
- [38] D. Ritchie, A. Thomas, P. Hanrahan, and N. D. Goodman. Neurally-guided procedural models: Amortized inference for procedural graphics programs using neural networks. In *NIPS*, pages 622–630, 2016. 3
- [39] G. Sharma, R. Goyal, D. Liu, E. Kalogerakis, and S. Maji. CSGNet: Neural shape parser for constructive solid geometry. In *CVPR*, June 2018. 3
- [40] O. Sidi, O. van Kaick, Y. Kleiman, H. Zhang, and D. Cohen-Or. Unsupervised co-segmentation of a set of shapes via descriptor-space spectral clustering. *ACM Trans. on Graphics*, 30(6):Article 126, 2011. 5
- [41] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015. 3
- [42] J. Talton, L. Yang, R. Kumar, M. Lim, N. Goodman, and R. Měch. Learning design patterns with bayesian grammar induction. In *Proc. ACM Symp. on User Interface Software and Technology*, pages 63–74, 2012. 3
- [43] D. W. Thompson. *On Growth and Form*. Dover reprint of 1942 2nd ed., 1992. 2
- [44] H. Wang, N. Schor, R. Hu, H. Huang, D. Cohen-Or, and H. Hui. Global-to-local generative model for 3d shapes. *ACM Transactions on Graphics (TOG)*, 37(6):00, 2018. 2, 3
- [45] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, pages 318–335. Springer, 2016. 3
- [46] Y. Wang, K. Xu, J. Li, H. Zhang, A. Shamir, L. Liu, Z.-Q. Cheng, and Y. Xiong. Symmetry hierarchy of man-made objects. *Computer Graphics Forum*, 2011. 3
- [47] 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 *Advances in Neural Information Processing Systems (NIPS)*, pages 82–90, 2016. 3
- [48] K. Xu, H. Zhang, D. Cohen-Or, and B. Chen. Fit and diverse: Set evolution for inspiring 3D shape galleries. *ACM Trans. on Graphics*, 31(4):57:1–57:10, 2012. 3, 8
- [49] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *Proc. Euro. Conf. on Computer Vision*, pages 776–791, 2016. 3
- [50] 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. *ACM Trans. on Graphics*, 35(6):210:1–210:12, 2016. 5
- [51] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, pages 5505–5514, 2018. 3
- [52] M. E. Yumer, P. Asente, R. Mech, and L. B. Kara. Procedural modeling using autoencoder networks. In *ACM UIST*, pages 109–118, 2015. 3
- [53] C. Zhu, K. Xu, S. Chaudhuri, R. Yi, and H. Zhang. SCORES: Shape composition with recursive substructure priors. *ACM Transactions on Graphics*, 37(6), 2018. 2
- [54] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *Proc. Euro. Conf. on Computer Vision*, pages 597–613, 2016. 3

A. Supplementary Material

A.1. Network Architectures

Part synthesis The architectures of the part synthesis generative autoencoders, for both 3D and 2D cases, are listed in Table 3 and Table 4 respectively. We used the following standard hyper-parameters to train the 3D (2D) model: Adam optimizer, $\beta_1 = 0.9(0.5)$, $\beta_2 = 0.999$, learning rate = $0.001(2e^{-4})$, batch size = 64.

Operation	Kernel	Strides	Feature maps	Act. func.
Encode x : 400x3 point cloud \rightarrow 64-dim feature vector f_x				
1D conv.	1x64	1	400x64	Relu
1D conv.	1x64	1	400x64	Relu
1D conv.	1x64	1	400x64	Relu
1D conv.	1x128	1	400x128	Relu
1D conv.	1x64	1	400x64	Relu
Max Pooling	-	-	128	-
Decode f_x : 64-dim feature vector \rightarrow 400x3 point cloud x'				
Linear	-	-	256	Relu
Linear	-	-	256	Relu
Linear	-	-	400x3	-

Table 3: Layers of the 3D part generative AE shown in order from input to output.

Operation	Kernel	Strides	Feature maps	Act. func.
Encode x : 64x64x1 input shape \rightarrow 10-dim feature vector f_x				
Conv.	5x5x8	2x2	32x32x8	l-Relu
Conv.	5x5x16	2x2	16x16x16	l-Relu
Conv.	5x5x32	2x2	8x8x32	l-Relu
Conv.	5x5x64	2x2	4x4x64	l-Relu
2xLinear	-	-	10	-
Decode f_x : 10-dim feature vector \rightarrow 64x64x1 shape x				
Linear	-	-	1024 (=4x4x64)	Relu
Trans. conv.	5x5x32	2x2	8x8x32	Relu
Trans. conv.	5x5x16	2x2	16x16x16	Relu
Trans. conv.	5x5x8	2x2	32x32x8	Relu
Trans. conv.	5x5x1	2x2	64x64x1	Sigmoid

Table 4: Layers of the 2D part VAE shown in order from input to output. After the last conv. layer we have two parallel linear layers for computing the mean and std of the VAE.

Parts composition The architectures of the parts composition units are listed in Table 5 and Table 6, for the 3D and 2D cases respectively. We used the following standard hyper-parameters to train the 3D and 2D models: Adam optimizer, $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate = 0.001, batch size = 64.

Operation	Feature maps	Act. func.
Comp. net x : 64xC+16 feature vector \rightarrow 6XC comp. vector f_x		
Linear	256	Relu
Linear	128	Relu
Linear	6xC	-

Table 5: Layers of the 3D parts composition network shown in order from input to output. Where C is the number of parts. The input vector is a concatenation of C feature vectors of size 64 with a noise vector of size 16.

Operation	Feature maps	Act. func.
Comp. net x : 10xC+8 feature vector \rightarrow 4XC comp. vector f_x		
Linear	128	Relu
Linear	128	Relu
Linear	4xC	-

Table 6: Layers of the 2D parts composition network shown in order from input to output. Where C is the number of parts. The input vector is a concatenation of C feature vectors of size 10 with a noise vector of size 8.

A.2. More results

In this section, we present additional results of CompoNet for the three categories. For the Chair and Airplane categories we have randomly sampled 80 shapes from the 5,000 we have generated for the quantitative metrics; see Figure 9 and Figure 10 respectively. For the Vases, since the output is smaller, we sampled 300 shapes from the 1,024 we generated for the quantitative metrics; see Figure 11. Furthermore, we present additional interpolation results from CompoNet on the 3D categories; see Figure 12 and Figure 14 for linear interpolations, Figure 13 and Figure 15 for part-by-part interpolations.

A.3. Comparison to baseline

In this section, we present randomly picked generated results from CompoNet and the baseline on the Chair and Airplane categories. For each generated shape, we present its three nearest neighbors based on the Chamfer distance. We present the results side by side to emphasize the power of CompoNet in generating novel shapes in comparison to the baseline; see Figure 16 for chairs and Figure 17 for airplanes.

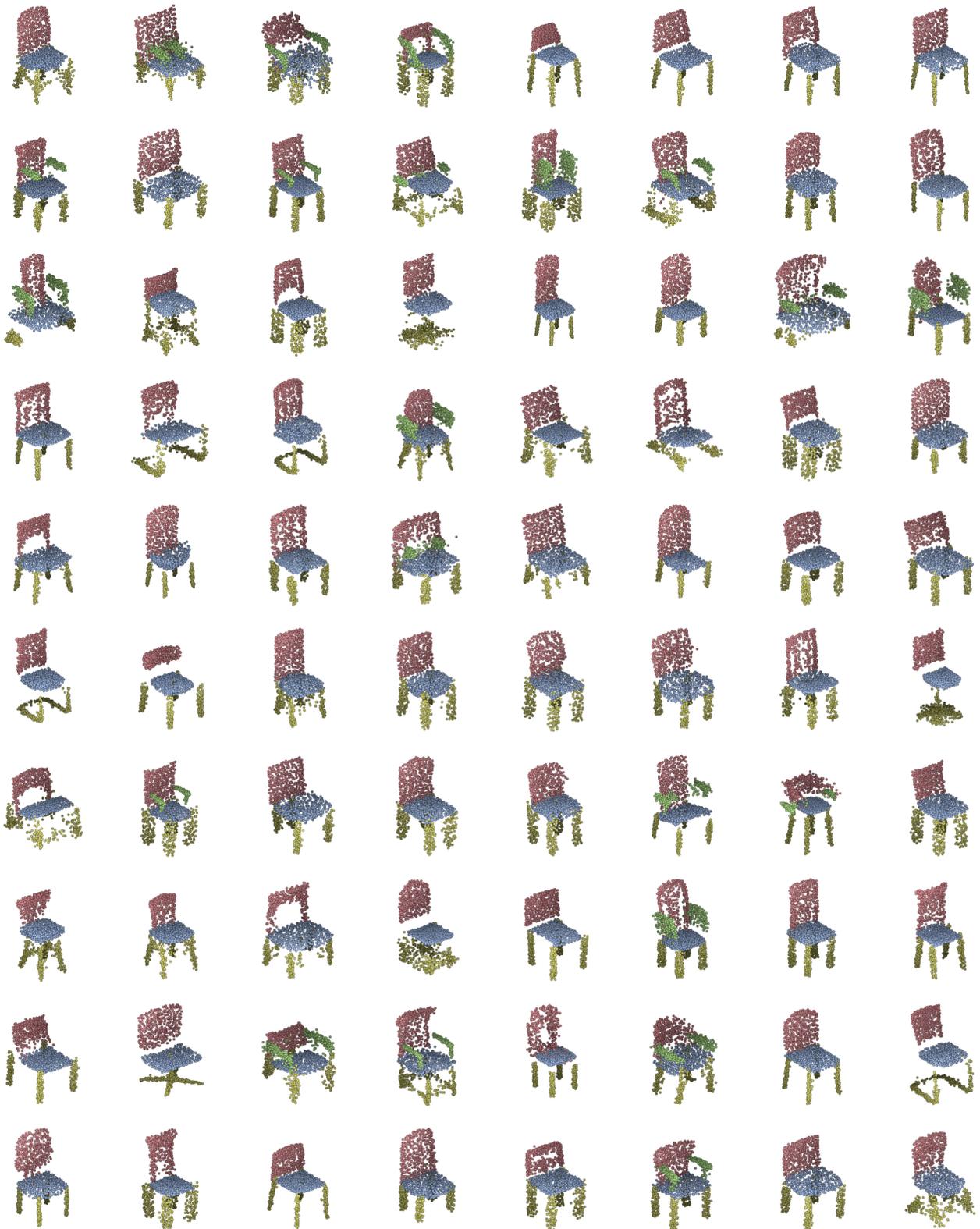


Figure 9: Chair gallery. We present 80 randomly sampled chairs from the 5,000 which were generated by CompoNet.

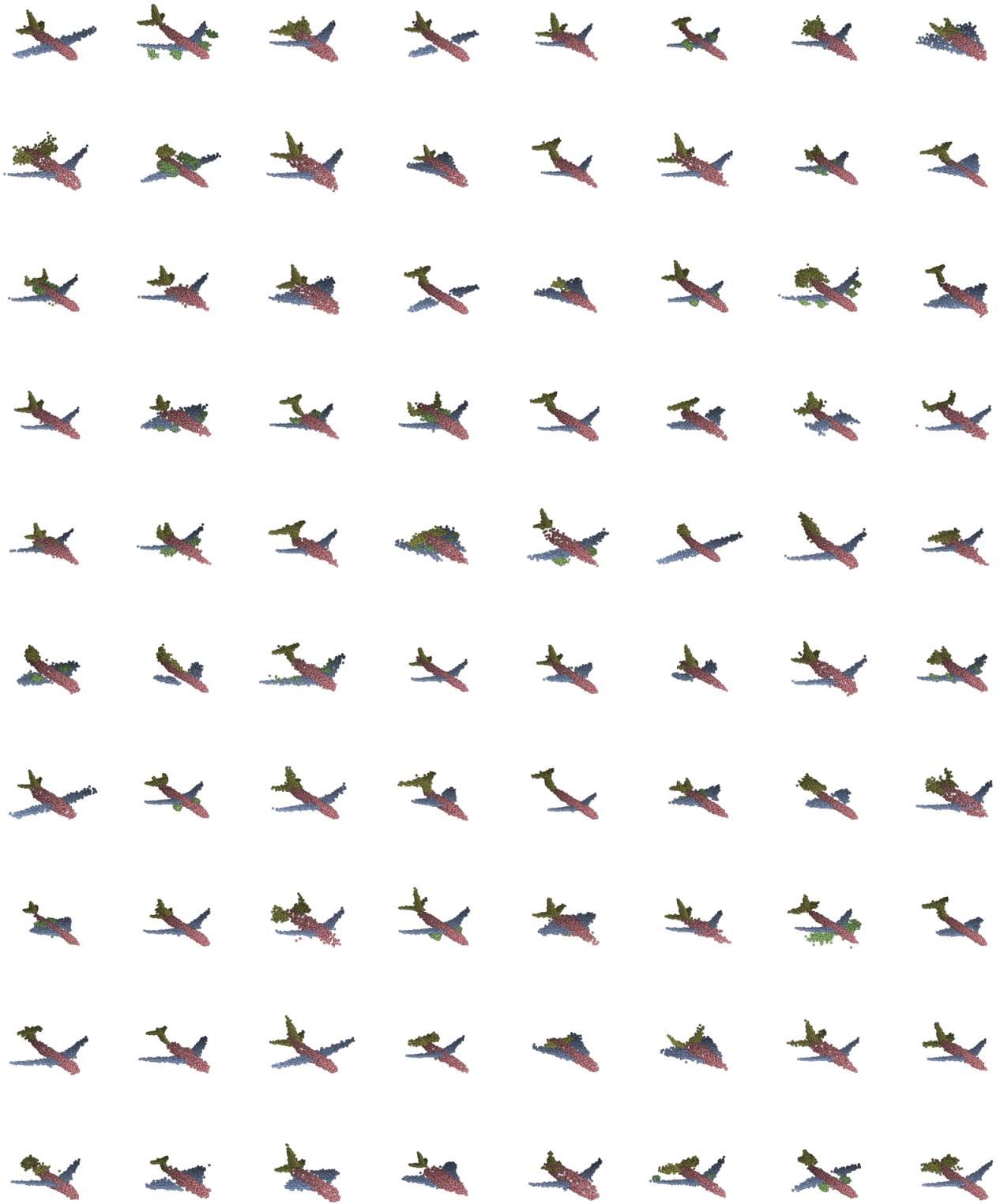


Figure 10: Airplane gallery. We present 80 randomly sampled airplanes from the 5,000 which were generated by CompoNet.

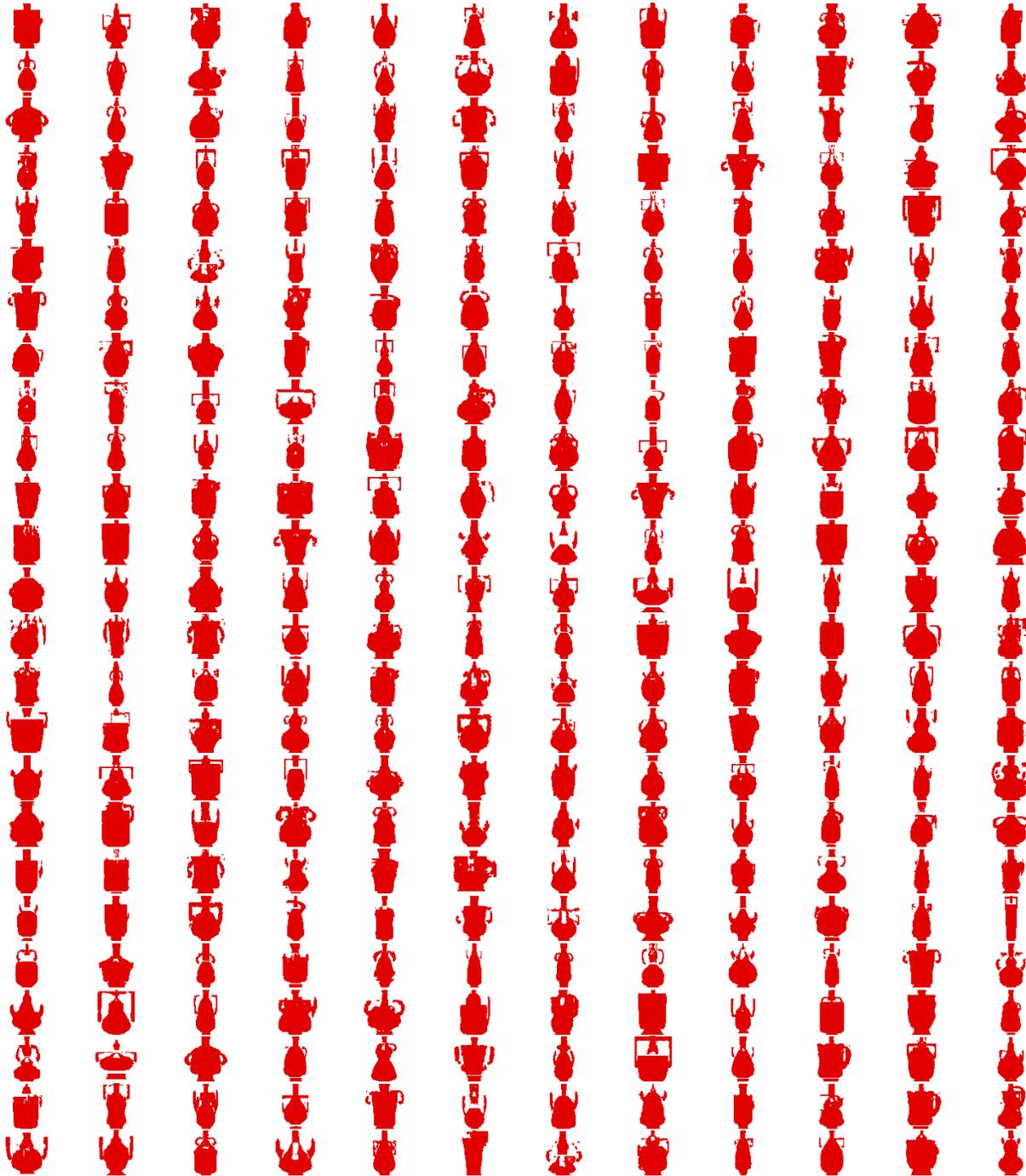


Figure 11: Vases gallery. We present 300 randomly sampled vases from the 1,024 which were generated by CompoNet.

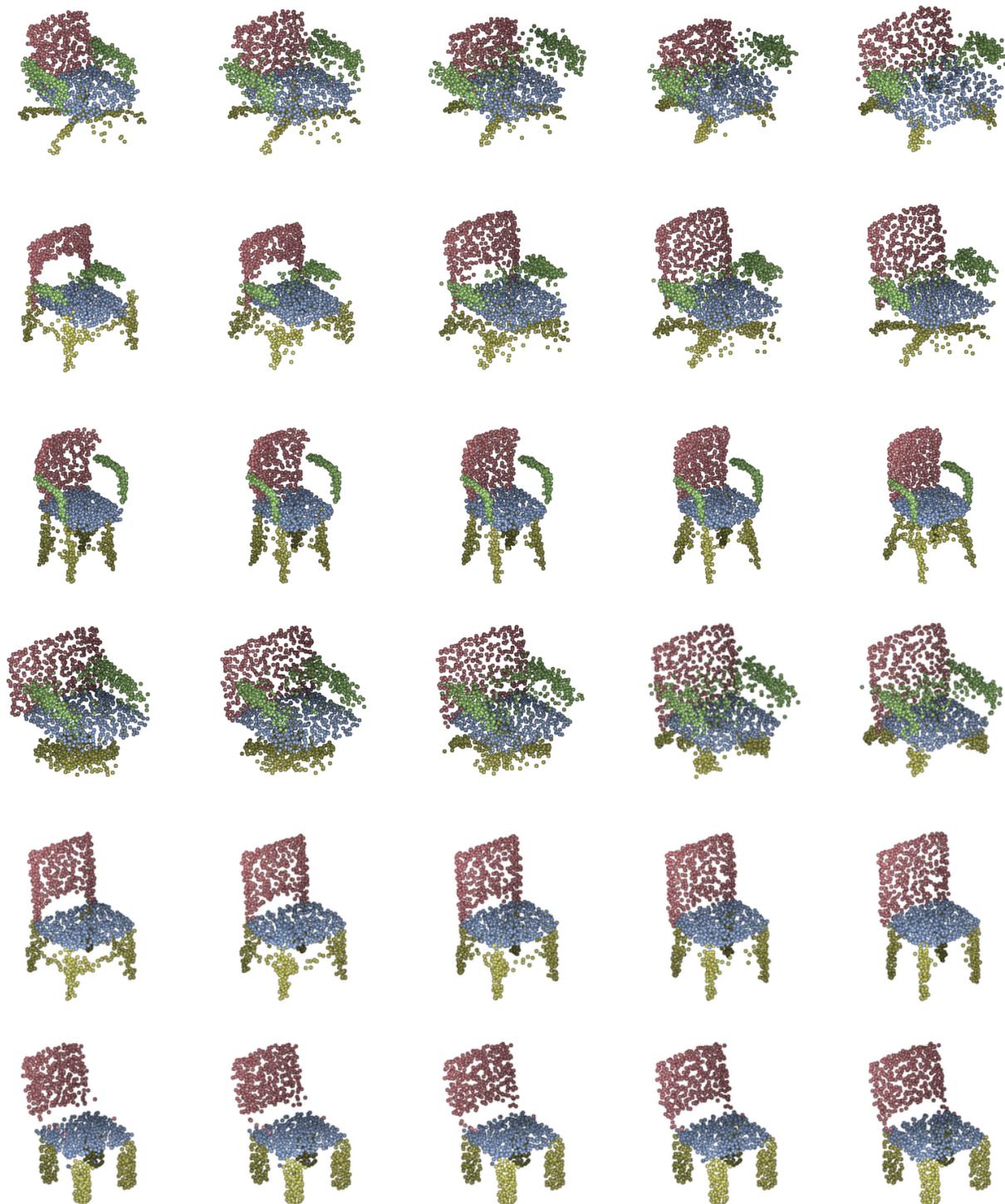


Figure 12: Complete latent space interpolation (parts and composition) for the chair category. The shape's code on the left is linearly interpolated to the shape's code on the right.

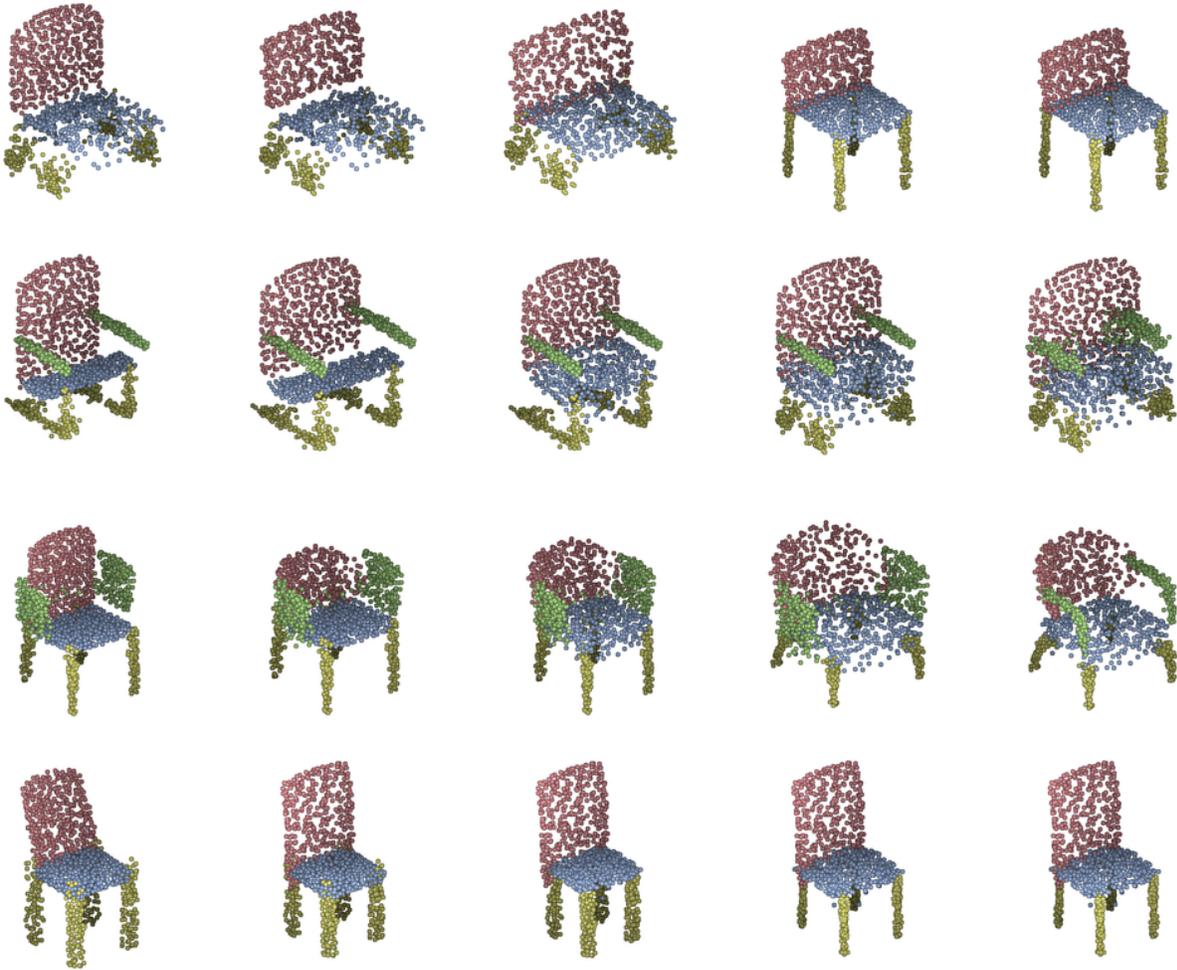


Figure 13: Chairs part-by-part interpolation. In each row we present part-by-part interpolation between two chairs (see left and right objects). The order of replacement is back-seat-legs-armrests. Please note, for certain pairs the arm-rests part does not exist, thus the interpolation of the arm-rest part is identity (no change).

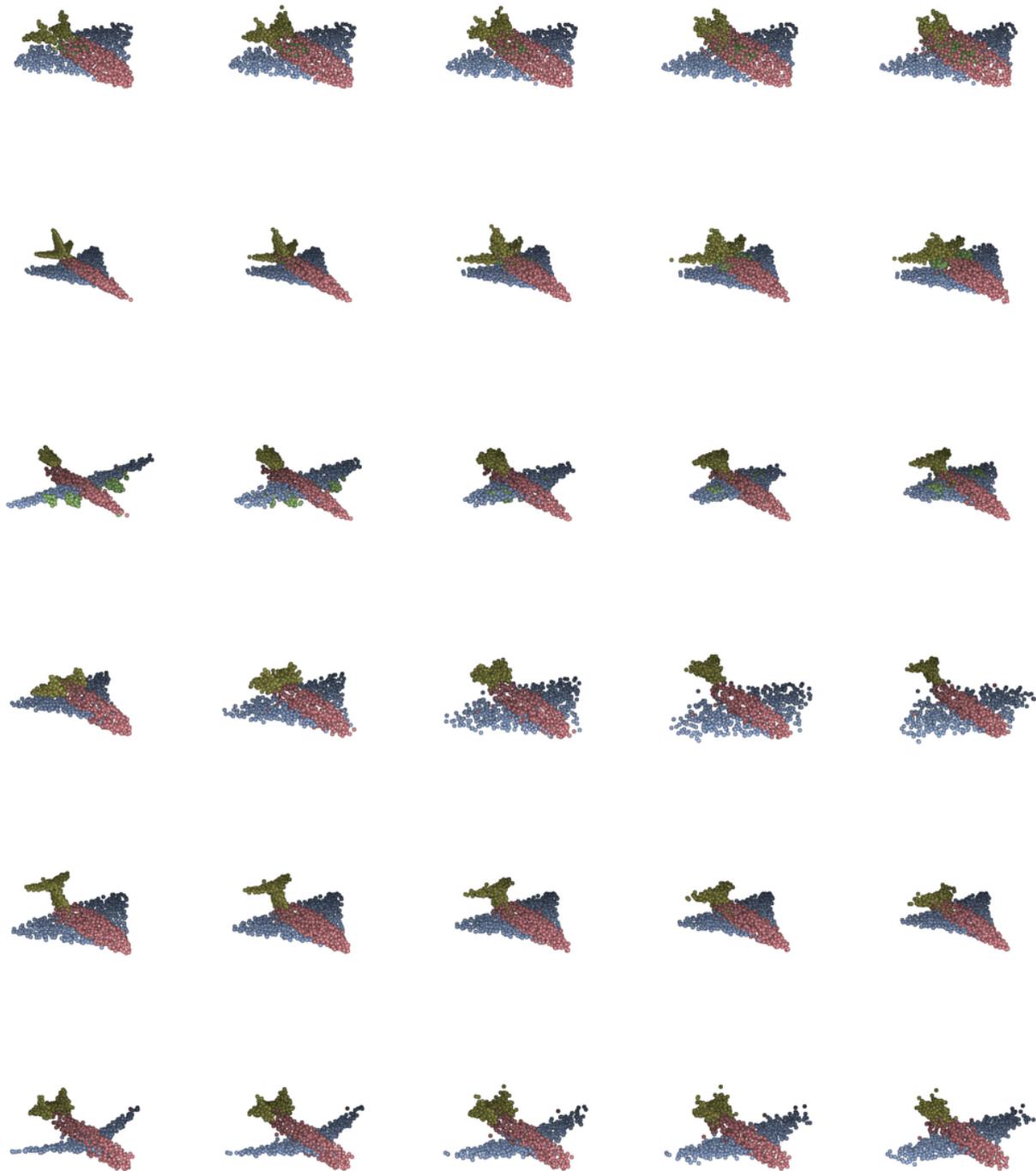


Figure 14: Complete latent space interpolation (parts and composition) for the airplane category. The shape's code on the left is linearly interpolated to the shape's code on the right.

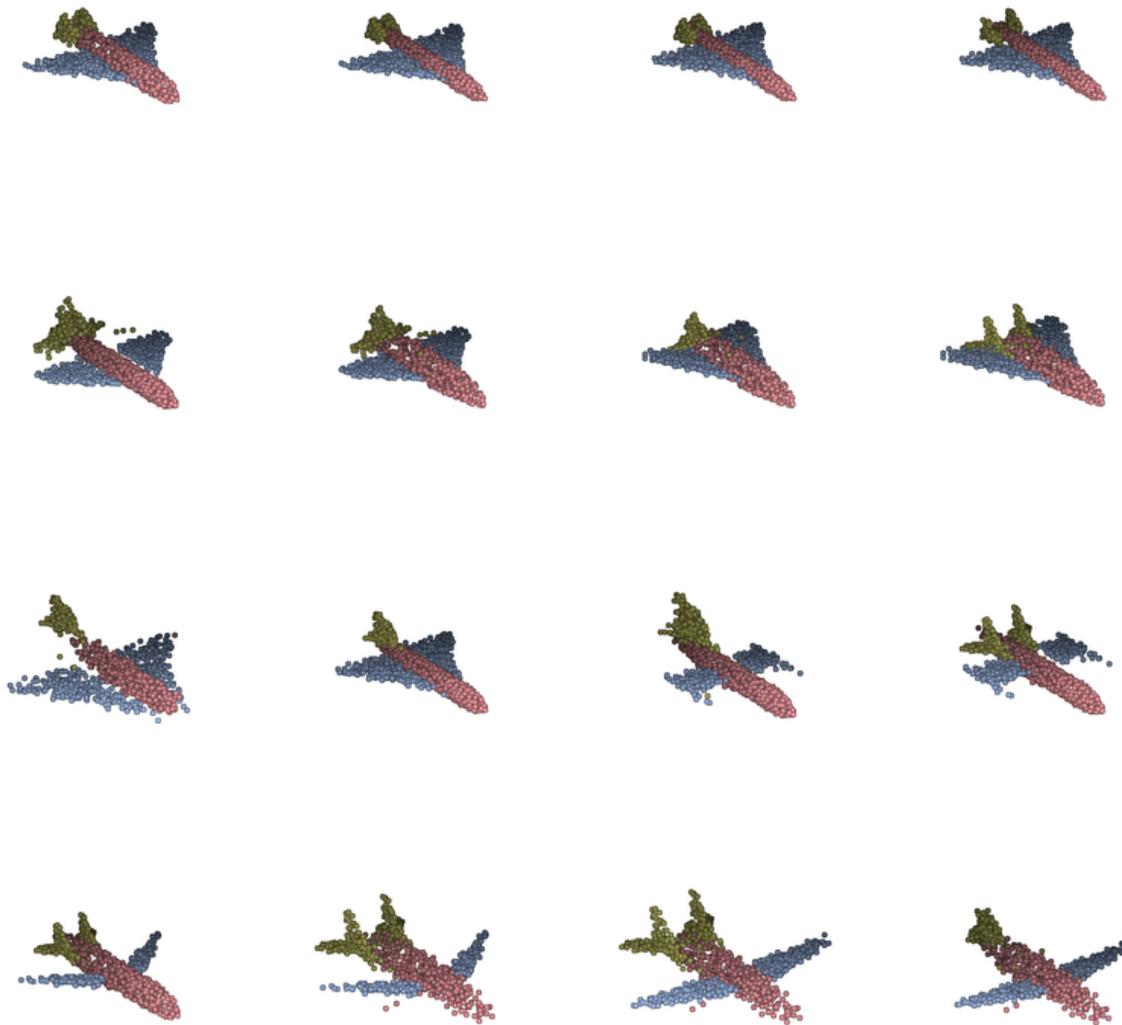


Figure 15: Airplanes part-by-part interpolation. In each row we present part-by-part interpolation between two airplanes (see left and right objects). The order of replacement is body-wings-tail.



Figure 16: Chairs nearest-neighbors. We present random generated results from CompoNet and the baseline, for each result, we present its three nearest-neighbors based on the Chamfer distance. In the left half we show results from CompoNet (in colors), where the left column is generated and the next three columns to the right are its three nearest-neighbors. In the right half we present results from the baseline (gray), where the left column is generated and the next three columns to the right are its three nearest-neighbors.

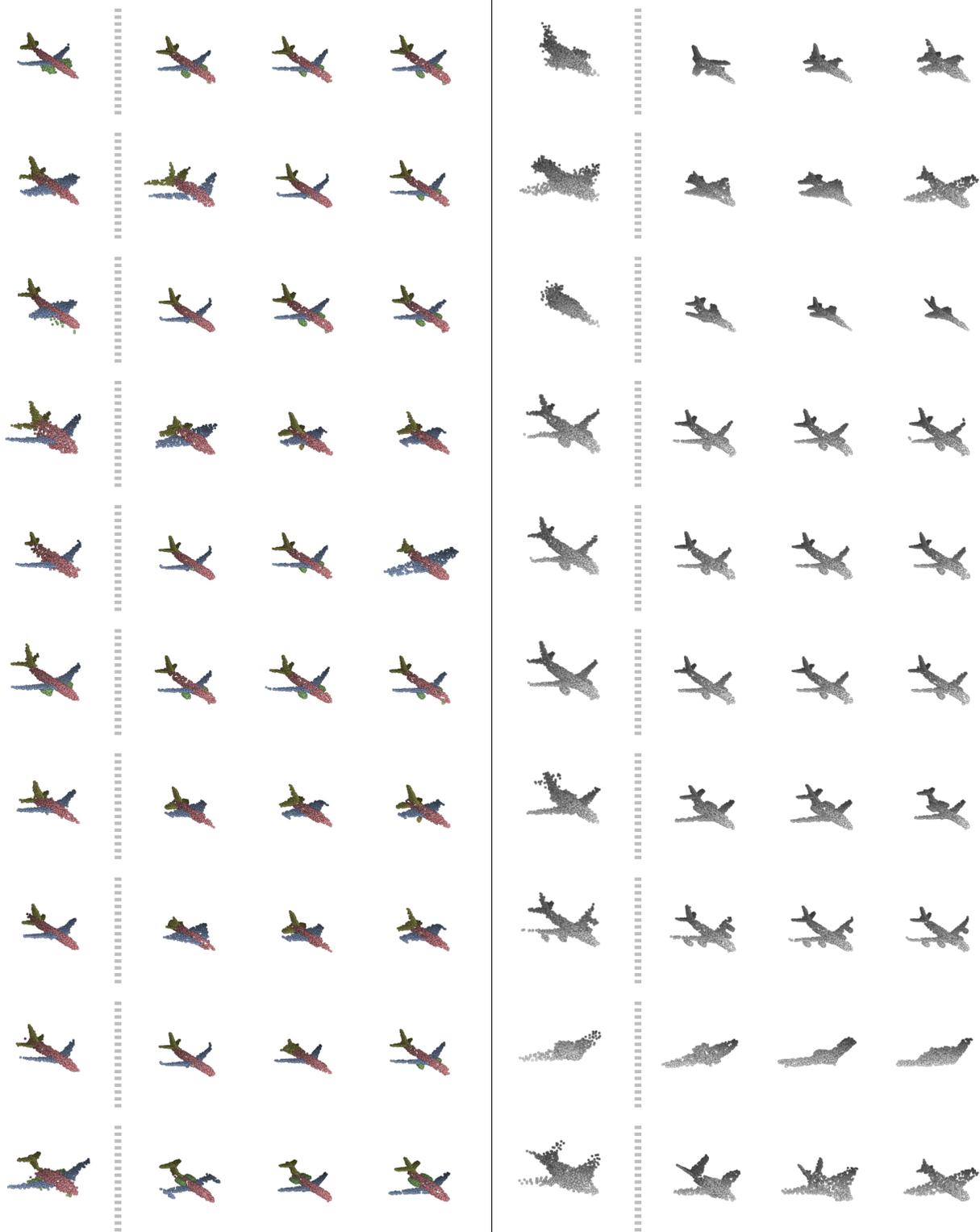


Figure 17: Airplanes nearest-neighbors. We present random generated results from CompoNet and the baseline, for each result, we present its three nearest-neighbors based on the Chamfer distance. In the left half we show results from CompoNet (in colors), where the left column is generated and the next three columns to the right are its three nearest-neighbors. In the right half we present results from the baseline (gray), where the left column is generated and the next three columns to the right are its three nearest-neighbors.