

GECCO: Geometrically-Conditioned Point Diffusion Models

Michał J. Tyszkiewicz¹

Pascal Fua¹

Eduard Trulls²

¹École Polytechnique Fédérale de Lausanne (EPFL)

²Google Research, Zurich

michal.tyszkiewicz@epfl.ch

pascal.fua@epfl.ch

trulls@google.com

Abstract

Diffusion models generating images conditionally on text, such as Dall-E 2 [51] and Stable Diffusion[53], have recently made a splash far beyond the computer vision community. Here, we tackle the related problem of generating point clouds, both unconditionally, and conditionally with images. For the latter, we introduce a novel geometrically-motivated conditioning scheme based on projecting sparse image features into the point cloud and attaching them to each individual point, at every step in the denoising process. This approach improves geometric consistency and yields greater fidelity than current methods relying on unstructured, global latent codes. Additionally, we show how to apply recent continuous-time diffusion schemes [59, 21]. Our method performs on par or above the state of art on conditional and unconditional experiments on synthetic data, while being faster, lighter, and delivering tractable likelihoods. We show it can also scale to diverse indoors scenes.

1. Introduction

Given the popularity of depth sensors and laser scanners, point clouds have become ubiquitous, with applications to robotics, autonomous driving, and augmented reality. Furthermore, they do not suffer from the precision/complexity trade-off inherent to voxel grids, and scale better and more generally than graph-based representations such as meshes. As a result, they have been extensively used for analytical tasks such as classification [46, 47, 60, 70, 48] and segmentation [46, 47, 16, 60, 12, 70, 72]. Recent work has turned to point cloud synthesis and its many applications to 3D content creation. However, this remains an emerging field and state of the art methods [66, 4, 25, 38, 71, 68] operate on small datasets that feature only a handful of object types [5]. More importantly, the generated shapes are typically not anchored in any prior and are thus difficult to control.

We present a novel approach that can mitigate these issues, taking our inspiration from generative methods that perturb samples with a diffusion process [58, 21] and de-

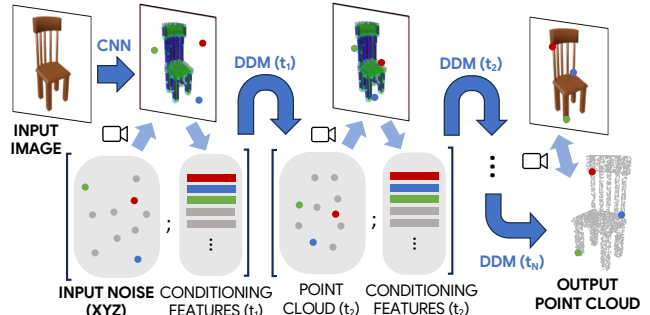


Figure 1: Our generative approach is based on denoising diffusion models (DDMs) and can be conditioned on images. At each denoising step we project the point cloud to the image, sample sparse features, and concatenate them to the locations, thus guiding the denoising process and yielding point clouds consistent with the images.

noise them with a deep network, which can later be used to synthesize new samples by iteratively denoising a signal. Specifically, most denoising-based approaches generate novel samples from pure noise. But to mirror the success of text-based image synthesis [51, 53, 54], a generative approach must be able to not only produce samples of sufficient quality and diversity, but also ground them in contextual information. Applying this generic idea to point clouds is, however, not straightforward. We show how to achieve this by conditioning the network with sparse image features.

Unlike previous works relying on unstructured, global embeddings, we do so in a geometrically-principled way, by projecting the point cloud into an image, sampling sparse features at those locations, and feeding them to the network along with the point location, at each denoising step, as illustrated in Fig. 1. This allows us to render 3D objects geometrically and semantically consistent with the image content, while controlling the viewpoint. Unlike regression models, such as monocular depth, our method can generate plausible hypotheses for occluded regions. This work is thus a first step towards unlocking the applicability of denoising diffusion models to practical scenarios such as 3D content creation, generating priors for automotive or

robotics applications, and single-view 3D reconstruction.

In short, we propose a novel generative point cloud model and show how to condition it on images. Our main contributions are: (i) We propose a framework composed of a permutation-equivariant Set Transformer [31] trained with a continuous-time diffusion scheme, which performs on par with the state of the art on *unconditional synthesis* while running 10x faster and delivering exact probabilities. (ii) We augment it with *geometrically-principled conditioning* to generate point clouds from images, yielding better reconstructions than with unstructured global embeddings, with state of the art performance. (iii) We bring denoising diffusion models for point cloud synthesis to the real world by applying our method to the Taskonomy dataset [67].

2. Related work

Denoising diffusion models. Denoising diffusion models [58, 15] are trained to denoise data perturbed by Gaussian noise. This process is applied iteratively during inference, and models are able to generate high-quality samples mirroring the distribution of the training data from random noise, optionally with a conditioning signal. They have shown great success synthesizing images from text [51, 53, 54], speech [7, 29, 43, 8], 3D objects [18, 42], and recently point clouds [38, 71, 68]. Diffusion models can be applied discretely, with a Markov chain [58, 15], or continuously with stochastic differential equations [27, 59]. We use a continuous formulation first proposed in [59], specifically an extension proposed in [21].

Generative point clouds models. Point cloud synthesis has been tackled with a wide array of techniques, including Variational Auto-Encoders (VAEs) [26], Generative Adversarial Networks (GANs) [63, 1, 33, 9, 57, 17, 26, 34], and autoregressive models [61]. Set-VAE [26] proposed an attention-based hierarchical VAE applicable to sets, such as point clouds. Achlioptas *et al.* [1] introduced l-GAN, operating over latents encoding shape, and r-GAN, directly on point clouds. SP-GAN [34] guides the generator with a global, uniformly distributed spherical prior and a local, random latent code to disentangle global and local shape. PointGrow [61] relies on an autoregressive model that samples each point conditionally on previously-generated points. ShapeGF [4] learns distributions over gradient fields, moving randomly sampled points to high-density areas such as surfaces. Xie *et al.* [65] formulate a permutation-invariant energy-based model with a PointNet. Of more direct importance to us are two other families, discussed separately: those based on normalizing flows (NF) and denoising diffusion models (DDM).

NFs for point cloud synthesis. Normalizing flows make powerful generative models and have been applied to point



Figure 2: **Image-conditioned generation.** We condition denoising diffusion models with images (a) in a geometrically-principled manner. Our model can reconstruct the shapes accurately from that view (b), and also generate plausible completions for regions not visible in the image (c, d), even under drastic occlusions (rows 3-5).

cloud synthesis [66, 25, 45, 28, 44]. PointFlow [66] broke ground by proposing a framework dividing the problem into two stages. First, a latent code responsible for the shape of the object, $s \sim P_{\theta_s}$, is sampled. Second, individual points p_i are sampled i.i.d. conditionally on s , meaning that a cloud $\{p_i\}$ is sampled with probability

$$P(\{p_i\}) = \int_{s \in \mathcal{S}} P_{\theta_s}(s) \sum_{i=0}^n P_{\theta_p}(p_i | s). \quad (1)$$

A downside of this formulation is the intractability of the probability computation, requiring an integral over all s : they thus train with an ELBO loss. In contrast, C-flow [45] models the point cloud jointly with a NF. This avoids the intractable probability, but they encounter difficulty in defining invertible layers which respect the permutation equivariance of point clouds. Their solution is to canonicalize the order of points with space-filling curves, which makes the approach complex. An alternative would be to use continuous-time NFs [13]. Unfortunately, powerful continuous-time models are known to be slow and expen-

sive to train [22]: as training progresses and the dynamics of the ODE become more complex, the cost of solving it with the precision necessary for stable backpropagation becomes impractically large.

DDMs for point cloud synthesis. This family, which our approach belongs to, has seen significant developments over the past year [38, 71, 68]. DPM [38] revisited the PointFlow [66] formulation (Eq. 2), replacing the normalizing flow $P(p_i|s)$ with a diffusion model. It relies on shape latents to parameterize the prior distribution with NFs, and uses them to condition a discrete DDM. It splits the loss into two additive terms, which requires tuning hyperparameters – our approach is simpler. PVD [71] trains a discrete DDM directly on point clouds (without shape latents) using a point-voxel network (PVCNN) that enables 3D convolutions [36]. It can optionally take in depth images as input to perform shape completion on occluded regions. This is achieved by freezing a set of points, extracted from the depth map, and optimizing over a set of ‘free’ points – we do geometrically-principled conditioning with RGB images instead, which is more widely applicable. Moreover, the authors argue that conventional permutation-equivariant architectures operating on pure point representations such as PointNet++ [47] are difficult to apply to diffusion models – we show we can achieve similar performance with a very simple architecture [31]. In a different direction, PDR [39] proposes a dual-network approach to shape completion based on DDMs.

More recently, LION [68] proposed another two-stage approach. First, VAEs are used to obtain latent representations of both global shape, and points. Second, a DDM is trained to model those latent spaces. They use a continuous diffusion model, like we do, but their reliance on shape latents means that computing exact probabilities is not tractable. Like PVD, LION uses PVCNN [36] for the encoder, decoder, and diffusion models. Finally, it may condition samples with different signals, such as images or text embeddings, by conditioning the shape latent with adaptive Group Normalization in the PVCNN layers. In contrast, we use a convolutional backbone to extract image features at the locations 3D points project to, concatenate them to the positional features, and feed them to the network.

Other single-view reconstruction approaches. Most of the generative point cloud methods our approach belongs to tackle only the unconditional problem. There is a wide array of relevant works on shape synthesis from single images, including regression and generative models, and using different representations such as voxels or meshes. 3D-R²N² [10] maps multiview images to occupancy grids with recurrent networks, but its resolution is limited due to using voxel grids. Pixel2Mesh [62] produces meshes from images by progressively deforming an ellipsoid using intermediate features extracted from the image. AtlasNet [14],

also mesh-based, generates 3D shapes by mapping multiple squares to shape surfaces. Chen *et al.* [6] learn point clouds from images by supervising their projections onto the image plane with samples from ground-truth silhouettes. Pix2Point [32] uses a 2D-3D hybrid network with an optimal transport loss to reconstruct point clouds from outdoor images. PSGN [11] combines an image and a random vector with a feedforward network to turn them into a point cloud, supervising with a permutation-equivariant loss. OcNet [40] tackles 3D reconstruction as the decision boundary of a learned occupancy classifier, which unlike voxel-based methods can be evaluated at arbitrary resolutions. Finally, there is of course a large body of work on monocular depth estimation: we refer the reader to [41].

3. Method

In order to follow the framework of [21, 59], we need to design a network s_θ to approximate the score $s_\theta(\mathbf{p}, t, c) \approx \nabla_{\mathbf{p}} \log p_t(\mathbf{p}|c)$, where c is an optional conditioning signal. We treat the point set $\{p_i\}, p_i \in \mathbb{R}^D, i = 1, \dots, N$ as a vector $\mathbf{p} \in \mathbb{R}^{N \times D}$. Since our network s_θ is permutation-equivariant, working with \mathbf{p} is sound. We present experiments with $D = 3$, but the approach is general.

3.1. Score network

Our score network $s_\theta(\mathbf{p}, t, c)$ (where dependency on c is optional) is inspired by the Set Transformer [31], which we choose as a powerful permutation-invariant architecture specifically designed for unordered inputs, such as point clouds. It treats each point as a token, but due to the quadratic scaling of self-attention, it instead uses cross-attention with a number of *inducers*, whose initial values are learned. Compared to the original Set Transformer, in each layer we use an extra shallow MLP on the inducers and not just on the tokens. Specifically to the task of diffusion, we “inject” the noise parameter t through the bias and scale of the Group Normalization [64] layers in the network, similarly to [68]. We also follow the recent approach of [50] and use Gaussian activations, allowing us to use a simple linear projection of input coordinates $\mathbb{R}^3 \rightarrow \mathbb{R}^{d_{\text{nn}}}$, where d_{nn} is the dimensionality of the input to the Set Transformer. We found this approach substantially better than the more common Fourier feature embedding.

3.2. Image-based conditioning

We undertake the goal of geometrically-principled point cloud generation conditioned on images. Specifically, we wish to: (a) generate point clouds in the reference frame of the camera; (b) accurately reconstruct the visible part of the object; and (c) build plausible hypotheses for occluded or ambiguous regions in a generative fashion. Note that the last property sets our approach apart from most supervised

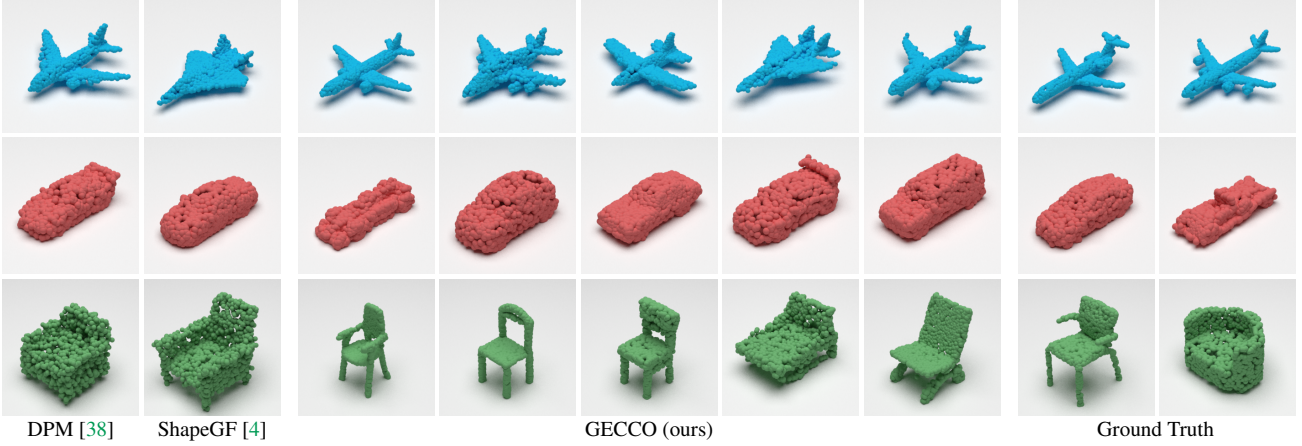


Figure 3: **Unconditional point cloud synthesis on ShapeNet.** All examples contain 2048 points.

approaches, such as monocular depth, which treat the problem point/pixel-wise and largely ignore the different modes of the posterior. Property (a) differs from most prior work on conditional generative models for point clouds, which usually condense the image into an unstructured, global embedding, losing much of the geometric detail. We achieve our goals through *camera frustum reparameterization* of the point cloud and *point-wise projective conditioning*.

Projective conditioning. In order to provide the network with a highly accurate conditioning signal we need to maintain a geometrical interpretation of the model. We do so by using a ConvNeXt backbone [35] to extract multi-scale image features. At any time t in the diffusion process, we take the point cloud $\{p_i\}_t$, project the points onto the image plane, look up the image features corresponding to each projection using bilinear interpolation, and concatenate them to the point’s location before feeding them to the transformer. This way each point knows the individual image properties *at its location*. We use this approach, which has a small computational overhead, to condition the reverse diffusion process *at every step*, as shown in Fig. 1. We concatenate the point location and its associated feature prior to the MLP that projects the features to d_{nn} dimensions, and use the exact same transformer architecture for conditional and unconditional settings, making this the only difference between $s_\theta(\mathbf{p}, t)$ and $s_\theta(\mathbf{p}, t, c)$. For points outside of image bounds the bilinear look-up simply returns zeros.

Camera frustum reparameterization. Current diffusion models are constrained to well-centered small objects and thus work directly in \mathbb{R}^n , but in Sec. 4.3 we wish to model points contained in the camera’s viewing frustum only. Our solution is to *reparameterize* the coordinates. Since the projection of each point on the image plane $(p_h, p_w) \in [0, 1]^2$, and its depth $p_d \in \mathbb{R}^+$, we can biject it to

$$(u, v, l) = (S^{-1}(p_h), S^{-1}(p_w), \log(p_d)) \in \mathbb{R}^3, \quad (2)$$

where S^{-1} is the inverse sigmoid function. We use standard diffusion in (u, v, l) and map points back to (x, y, z) at the end. We do not use any reparameterization for ShapeNet.

Implementation details. Our Set Transformer-based network has 6 layers and takes inputs of dimensionality $d_{nn} = 384$. For unconditional models, we simply project the 3D point location to d_{nn} . For image-conditioned models, we extract multi-scale ConvNeXt features of sizes 96, 192, 384 (total: 672), concatenate them to the point’s location, and project them to d_{nn} . We train our models with 1024 or 2048 points, subsampling datasets which contain more points per example, for data augmentation. We follow [20] in initializing transformer skip-connections with small weights and optimise the network (including the ConvNeXt) with AdaBelief [73] with learning rate $2 \cdot 10^{-4}$, for a number of steps depending on the dataset (see appendix). For each dataset we scale the data globally to zero mean, unit variance (which we undo at inference time) and pick σ_{\max} by estimating the maximum pairwise distance between training examples. We train using the preconditioning and loss formulation of [21], with the main departure in that we sample σ values log-uniformly over $(10^{-4}, \sigma_{\max})$ instead of log-normally. As in [59], we found it beneficial to apply an exponential moving average to the weights of our model: we use a rate of 0.999. We implement our software with JAX [3] and Equinox [24] and use DiffraX [23] for ODE solving. For inference we use the 2nd-order stochastic sampler of [21] (later referred to as ‘SDE’) or the probability flow of [59] (later: ‘ODE’), with 128 steps. Please refer to Sec. 4.4 for an ablation study, and to Table 6 for computational details: our approach is both *faster* and *lighter* than comparable methods. Code and models are available.

4. Experiments

We use ShapeNet [5] to evaluate unconditional point cloud synthesis in Sec. 4.1, and with image conditioning

	Model	MMD↓		COV↑ (%)		1-NNA↓ (%)	
		CD	EMD	CD	EMD	CD	EMD
AIRPLANE	Oracle	0.214	0.369	46.17	49.88	64.44	63.58
	r-GAN [1]	0.447	2.309	30.12	14.32	98.40	96.79
	I-GAN-CD [1]	0.340	0.583	38.52	21.23	87.30	93.95
	I-GAN-EMD [1]	0.397	0.417	38.27	38.52	89.49	76.91
	PointFlow [66]	0.224	0.390	47.90	46.41	75.68	70.74
	SoftFlow [25]	0.231	0.375	46.91	47.90	76.05	65.80
	SetVAE [26]	0.200	0.367	43.70	48.40	76.54	67.65
	DPF-Net [28]	0.264	0.409	46.17	48.89	75.18	65.55
	DPM [38]	0.213	0.572	48.64	33.83	76.42	86.91
	PVD [71]	0.224	0.370	48.88	52.09	73.82	64.81
	LION [68]	0.219	0.372	47.16	49.63	67.41	61.23
	GECCO (ours)	0.245	0.368	48.15	48.40	72.10	62.96
CHAIR	Oracle	2.618	1.555	53.02	51.21	51.28	54.76
	r-GAN [1]	5.151	8.312	24.27	15.13	83.69	99.70
	I-GAN-CD [1]	2.589	2.007	41.99	29.31	68.58	83.84
	I-GAN-EMD [1]	2.811	1.619	38.07	44.86	71.90	64.65
	PointFlow [66]	2.409	1.595	42.90	50.00	62.84	60.57
	SoftFlow [25]	2.528	1.682	41.39	47.43	59.21	60.05
	SetVAE [26]	2.545	1.585	46.83	44.26	58.84	60.57
	DPF-Net [28]	2.536	1.632	44.71	48.79	62.00	58.53
	DPM [38]	2.399	2.066	44.86	35.50	60.05	74.77
	PVD [71]	2.622	1.556	49.84	50.60	56.26	53.32
	LION [68]	2.640	1.550	48.94	52.11	53.70	52.34
	GECCO (ours)	2.793	1.601	46.68	49.40	56.57	54.68
CAR	Oracle	0.938	0.791	50.85	55.68	51.70	50.00
	r-GAN [1]	1.446	2.133	19.03	6.539	94.46	99.01
	I-GAN-CD [1]	1.532	1.226	38.92	23.58	66.49	88.78
	I-GAN-EMD [1]	1.408	0.899	37.78	45.17	71.16	66.19
	PointFlow [66]	0.901	0.807	46.88	50.00	58.10	56.25
	SoftFlow [25]	1.187	0.859	42.90	44.60	64.77	60.09
	SetVAE [26]	0.882	0.733	49.15	46.59	59.94	59.94
	DPF-Net [28]	1.129	0.853	45.74	49.43	62.35	54.48
	DPM [38]	0.902	1.140	44.03	34.94	68.89	79.97
	PVD [71]	1.077	0.794	41.19	50.56	54.55	53.83
	LION [68]	0.913	0.752	50.00	56.53	53.41	51.14
	GECCO (ours)	1.044	0.769	50.00	56.82	56.82	49.15

Table 1: **Unconditional generation (global normalization)**. Generation metrics on three ShapeNet categories. MMD-CD is multiplied by 10^3 , and MMD-EMD by 10^2 . The top 3 are highlighted in gray (darker is better).

in Sec. 4.2. We then show that our method can translate to larger-scale, real data on the Taskonomy dataset [67] in Sec. 4.3, and ablate it and showcase some of its properties in Sec. 4.4. We render point clouds with Mitsuba 3 [19].

4.1. Unconditional generation on ShapeNet

Dataset. We evaluate our approach on the dataset most commonly used for generative shape modelling: ShapeNet [5]. We follow the methodology, splits, and metrics introduced by PointFlow [66], which provides point clouds sampled from the original meshes, and train single-class models for three categories: airplane, chair, and car. In order to ensure reproducibility we compare against the results published in [68], the most recent and thorough. While PointFlow normalizes the data globally to zero-mean per axis, and unit variance, others methods use per-shape normalization: we

	Model	MMD↓		COV↑ (%)		1-NNA↓ (%)	
		CD	EMD	CD	EMD	CD	EMD
AIRPLANE	Oracle	0.230	0.539	42.72	45.68	69.26	67.78
	TreeGAN [57]	0.558	1.460	31.85	17.78	97.53	99.88
	ShapeGF [4]	0.313	0.636	45.19	40.25	81.23	80.86
	SP-GAN [34]	0.403	0.766	26.42	24.44	94.69	93.95
	PDGN [17]	0.409	0.701	38.77	36.54	94.94	91.73
	GCA [69]	0.359	0.765	38.02	36.30	88.15	85.93
	LION [68]	0.356	0.593	42.96	47.90	76.30	67.04
	GECCO (ours)	0.354	0.572	44.20	50.12	76.17	68.89
	Oracle	3.864	2.302	49.7	42.11	55.14	54.76
	TreeGAN [57]	4.841	3.505	39.88	26.59	88.37	96.37
CHAIR	ShapeGF [4]	3.724	2.394	48.34	44.26	58.01	61.25
	SP-GAN [34]	4.208	2.620	40.03	32.93	72.58	83.69
	PDGN [17]	4.224	2.577	43.20	36.71	71.83	79.00
	GCA [69]	4.403	2.582	45.92	47.89	64.27	64.50
	LION [68]	3.846	2.309	46.37	50.15	56.50	53.85
	GECCO (ours)	4.119	2.410	48.64	52.42	55.36	56.80
	Oracle	1.05	0.829	47.44	48.01	57.53	56.68
	TreeGAN [57]	1.142	1.063	40.06	31.53	89.77	94.89
	ShapeGF [4]	1.020	0.824	44.03	47.16	61.79	57.24
	SP-GAN [34]	1.168	1.021	34.94	31.82	87.36	85.94
CAR	PDGN [17]	1.184	1.063	31.25	25.00	89.35	87.22
	GCA [69]	1.074	0.867	42.05	48.58	70.45	64.20
	LION [68]	1.064	0.808	42.90	50.85	59.52	49.29
	GECCO (ours)	1.063	0.802	46.31	49.15	60.51	47.87

Table 2: **Unconditional generation (per-shape normalization)**. Same as Table 1, with per-shape normalization.

consider both. We use 2048 points for all methods.

Metrics. We consider two distance metrics between point clouds: the chamfer distance (CD), which measures the average squared distance between each point in one set to its nearest neighbor on the other set; and the earth mover’s distance (EMD), which solves the optimal transport problem. Given point sets $\mathbf{p} = \{p_i\}$ and $\mathbf{q} = \{q_i\}$ with N points each, and ϕ a bijection between them, they are defined as:

$$\text{CD}(\mathbf{p}, \mathbf{q}) = \frac{1}{N} \left[\sum_{p \in \mathbf{p}} \min_{q \in \mathbf{q}} \|p - q\|_2^2 + \sum_{q \in \mathbf{q}} \min_{p \in \mathbf{p}} \|p - q\|_2^2 \right], \quad (3)$$

$$\text{EMD}(\mathbf{p}, \mathbf{q}) = \frac{1}{N} \min_{\phi: \mathbf{p} \rightarrow \mathbf{q}} \sum_{p \in \mathbf{p}} \|p - \phi(p)\|_2. \quad (4)$$

Given these two similarity functions, we sample as many point clouds as there are in the reference set \mathbf{S}_r to obtain a generated set \mathbf{S}_g and compute three metrics between the ground truth and sampled collections. To compute **coverage (COV)** we find the nearest neighbor in the reference set for each point cloud in the generated set, and compute the fraction of shapes in the reference set that are matched to at least one shape in the reference set. It can capture mode collapse, but not the quality of the samples. The **minimum matching distance (MMD)** is a complementary metric that measures the average minimum distance from every sample in the reference set to every sample in the generated set. PointFlow [66] proposes an arguably better metric also used

	airplane	bench	cabinet	car	chair	display	lamp	loudspeaker	rifle	sofa	table	telephone	vessel	average
3D-R ² N ² [10]	0.227	0.194	0.217	0.213	0.270	0.314	0.778	0.318	0.183	0.229	0.239	0.195	0.238	0.278
PSGN [11]	0.137	0.181	0.215	0.169	0.247	0.284	0.314	0.316	0.134	0.224	0.222	0.161	0.188	0.215
Pixel2Mesh [62]	0.187	0.201	0.196	0.180	0.265	0.239	0.308	0.285	0.164	0.212	0.218	0.149	0.212	0.216
AtlasNet [14]	0.104	0.138	0.175	0.141	0.209	0.198	0.305	0.245	0.115	0.177	0.190	0.128	0.151	0.175
OccNet [40]	0.140	0.157	0.156	0.153	0.209	0.260	0.394	0.269	0.142	0.185	0.176	0.129	0.200	0.198
GECCO	0.106	0.097	0.110	0.103	0.142	0.138	0.186	0.158	0.097	0.123	0.107	0.090	0.132	0.122
OccNet [40] w/ ICP	0.151	0.158	0.141	0.139	0.196	0.247	0.380	0.251	0.155	0.188	0.207	0.138	0.203	0.196 (+1.02%)
GECCO w/ ICP	0.081	0.088	0.100	0.093	0.117	0.119	0.164	0.141	0.073	0.114	0.111	0.083	0.128	0.109 (+11.9%)

Table 3: **Image-conditional generation on ShapeNet-Vol.** L1 chamfer distance between samples reconstructed from an image and the ground truth point clouds (lower is better), following [40]. Qualitative results are available in Fig. 2.

for GANs: **1-Nearest Neighbour Accuracy (1-NNA)**. It is defined as the accuracy of a leave-one-out classifier that assigns each sample in $\mathbf{S}_r \cup \mathbf{S}_g$ to the ‘class’ (set) of its closest neighbor, other than itself. Note that a perfect oracle would score $\sim 50\%$. As all three metrics rely on nearest neighbours, they can be computed with CD or EMD: we report both. Details are provided in the supplementary material.

Results. Results are shown in Table 1 for global normalization, and Table 2 for per-shape normalization. 1-NNA is the metric favored by most recent papers. Our method performs on par with LION, the state of the art, on the first benchmark and slightly outperforms it on the second. In addition to the baselines, we consider an oracle that spits out samples from the training set instead of generating novel ones. Our method outperforms this oracle in 1-NNA-EMD for all three categories in Table 1, which suggests that the dataset is at the saturation point, or that the distance metrics fail to fully capture the quality of the samples. We show samples and ground truth examples in Fig. 3.

4.2. Conditional generation on ShapeNet-Vol

Dataset. None of the baselines used in the previous section are able to condition the generative process with images, with the exception of LION, which may optionally train DDMs conditioned with CLIP embeddings [49, 55] extracted from ShapeNet renders. While seemingly effective this approach is not geometrically principled, and the paper offers only qualitative examples. We thus turn to the ShapeNet-Vol benchmark, originally introduced by 3D-R²N² [10], which provides rendered images and voxelized models for 13 ShapeNet categories: each shape is rendered from 24 viewpoints at 137×137 pixels. We align the point clouds to the camera pose for each view and train our models with the conditioning scheme of Sec. 3.2. Note that unlike the unconditional experiments in the previous section, here we train *a single model for all 13 categories*.

Results. We follow the evaluation protocol for single-view reconstruction introduced by OccNet [40]. For mesh-based methods, such as OccNet, the benchmark samples 100k points from the mesh and computes the chamfer distance

Split	Model	Chamfer ↓	Chamfer (ICP) ↓
Test	GECCO	0.661 / 0.502	0.444 / 0.242
	Monocular depth [56]	0.632 / 0.527	0.558 / 0.451
Val	GECCO	0.541 / 0.427	0.361 / 0.222
	Monocular depth [56]	0.567 / 0.490	0.497 / 0.418

Table 4: **Evaluation on Taskonomy** with mean / median values. Note that we do not use the validation set for early stopping, so ‘validation’ and ‘test’ both act as test sets. The difference in performance in ‘test’ is due to out-of-distribution scenes in that subset (see appendix for details).

between the generated and ground truth point clouds as a quality metric – including points occluded in the image. For point-based methods such as PSGN [11], the benchmark simply samples more points (no meshing). In order to reach the required number of points, we simply generate multiple samples for each image and concatenate them. Note that following [40], we use the *L1 chamfer distance*, defined as in Eq. 3 but without squaring the norms. OccNet reconstructs the model in a canonical reference frame, which is also used for evaluation, but our method generates predictions from the point of view of the camera: we move our predictions to this canonical frame with the ground truth pose. We report results in Table 3¹. Our models outperform all mesh-based methods, including OccNet, and also PSGN. Additionally, we noticed that our generated point clouds were slightly misaligned, but not those from OccNet. We hypothesized this was due to OccNet generating samples in a canonical reference frame rather than the camera’s point of view, which while advantageous here does lose generalization. We confirmed this by aligning the point clouds with ICP [2, 52] and re-computing the metric: our method improves by 12% relative, compared to OccNet’s 1%. Note that unlike OccNet, our approach does not need normalized data in a canonical pose and it can deal with non-watertight meshes. On the other hand, it does require known camera intrinsics. We show qualitative examples in Fig. 2.

¹For OccNet we run the latest model available in their repository, improved from the original paper. For others we use the results from [40].

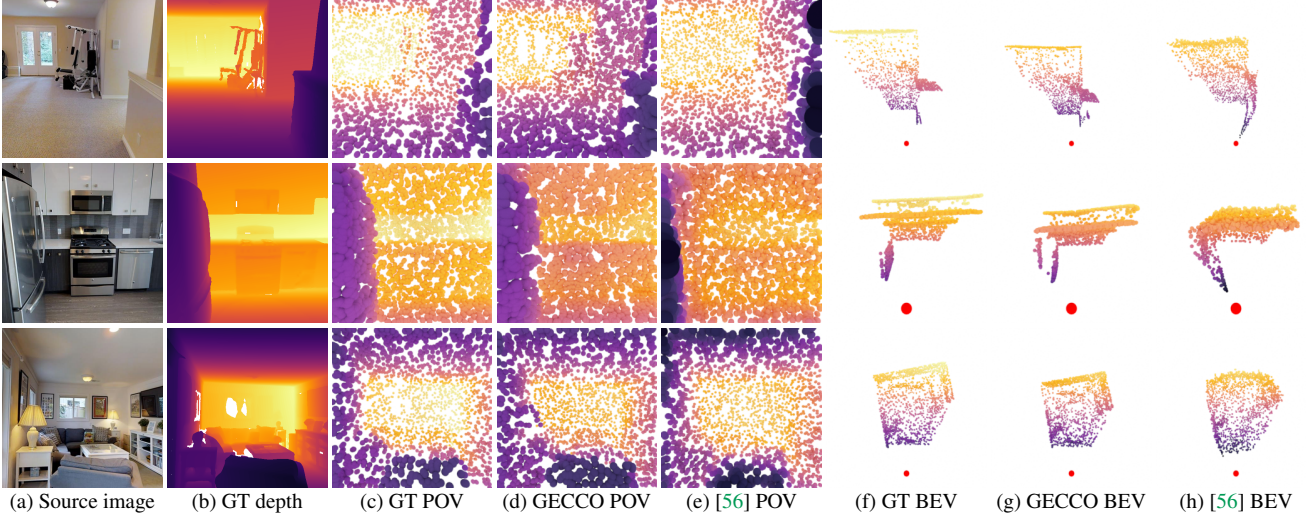


Figure 4: **Qualitative examples on Taskonomy.** Color encodes depth. We show point clouds with 2048 points each, from the camera’s point of view, and from a bird’s eye view. The red dot in BEV marks the location of the camera.

4.3. Conditional generation on Taskonomy

We also wish to showcase how our method scales to real data, and past the object-centric nature of ShapeNet that most generative methods are limited to. For this purpose we turn to Taskonomy [67], which contains a large dataset of scanned indoor scenes with high-quality depth maps, and convert them into point clouds by sampling and unprojecting 8192 points per image. We sample points inversely proportionally to pixel depth, to emulate per-surface-area densities. This yields a rich image-point cloud dataset.

Given the lack of generative methods that can scale up to this data, we compare with a monocular depth method from [56], also trained on Taskonomy. We first adjust the absolute scale and shift of its output by comparing with ground truth depth (as in the loss function of MiDaS [30]) and proceed by unprojecting with the same procedure as when creating the dataset. For GECCO, we directly predict the point clouds in absolute units, using the (u, v, l) reparameterization introduced in Sec. 3.2. We use 2048 points for both training and evaluation. We compare the two approaches qualitatively in Fig. 4, and quantitatively in Table 4, using the same metric as in Sec. 4.2 and Table 3. This experiment confirms our method extends beyond object-centric views to real scenes, and greatly outperforms similarly-sized baselines benefitting from years of research on monocular depth. As in Sec 4.2, we also report results with ICP. For the baseline we disabled scale estimation, as it degrades the results.

4.4. Ablation studies and further experiments

Ablation study: Global vs projective conditioning. We evaluate our approach to conditioning the denoising process through projective geometry with the more standard

Conditioning	Sampler	N_{denoise}	ICP	Subset	L1-CD↓	$\Delta \uparrow$ (%)
Global	ODE	128	✗	–	0.305	–
Projective	ODE	128	✗	–	0.286	+6.6%
Global	SDE	128	✗	–	0.302	–
Projective	SDE	128	✗	–	0.283	+6.7%
Global	ODE	128	✓	–	0.280	–
Projective	ODE	128	✓	–	0.259	+8.1%
Global	SDE	128	✓	–	0.276	–
Projective	SDE	128	✓	–	0.257	+7.4%
Projective	SDE	8	✗	–	0.615	-117.3%
Projective	SDE	16	✗	–	0.309	-9.2%
Projective	SDE	32	✗	–	0.286	-1.1%
Projective	SDE	64	✗	–	0.287	-1.4%
Projective	SDE	128	✗	–	0.283	–
Projective	SDE	8	✓	–	0.353	-37.4%
Projective	SDE	16	✓	–	0.267	-3.9%
Projective	SDE	32	✓	–	0.258	-0.4%
Projective	SDE	64	✓	–	0.258	-0.4%
Projective	SDE	128	✓	–	0.257	–
Global	SDE	128	✗	50%	0.308	-2.0%
Global	SDE	128	✓	50%	0.283	-2.5%
Projective	SDE	128	✗	50%	0.289	-2.1%
Projective	SDE	128	✓	50%	0.261	-1.6%

Table 5: **Ablation study on the OccNet benchmark.** We compare conditioning with global context vs projective lookups (sec. 3.2), probability flow (‘ODE’) [59] vs the stochastic solver (‘SDE’) of [21], the number of solver iterations, and the impact of reducing the training data to 50%. We conduct these experiments with 1024 points, for speed.

approach relying on a global embedding. Instead of bilinear lookup for each point, we mean-pool the CNN features and inject them globally as in [68], alongside t , through the bias and scale of normalization layers. We compare both approaches on the OccNet benchmark of Sec. 4.2. As we do not aim to compare against those baselines, we use 1024

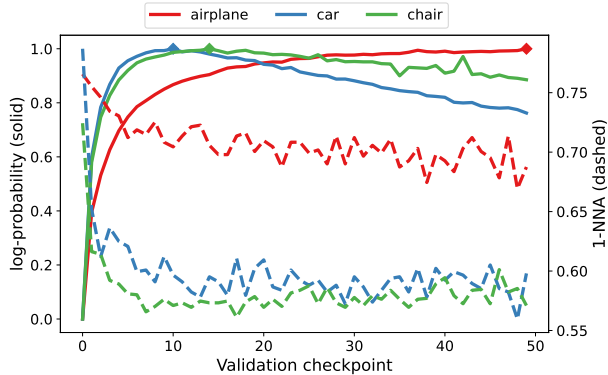


Figure 5: **Log-probabilities while training.** Dashed line: CD 1-NNA (Sec. 4.1) on the validation set. Solid: validation set log-probability (normalized); diamonds mark maxima. Notice how the decreasing likelihood of sampling validation examples does not increase (*i.e.*, degrade) 1-NNA.

points rather than 100k, for simplicity. Results are shown in Table 5. Our projective conditioning is 7-8% more accurate.

Ablation study: Numerical solvers. We consider the OccNet [40] benchmark and ablate the use of the probability flow solver proposed in [59] versus the stochastic solver used in [21]. We find the latter superior and turn to the effect of the number of solver steps. We observe that more solver steps bring larger perceptual improvements which are not always captured by the chamfer distance, up to about 128 steps, which we use in all other experiments in the paper. We report the numbers in Table 5: relative values take $N_{denoise} = 128$ as reference. It should be noted that technically, both solvers use two network evaluations per step.

Ablation study: training with fewer samples. We train conditional models with only 50% of the data and report a surprisingly small drop in performance: $\sim 2\%$ (Table 5). This holds for both projective and global conditioning.

Probabilities as a metric. Our approach allows us to compute exact probabilities over the validation set, following the method in [59]. The maxima in probability corresponds to the optimal state in terms of *novel* generative performance. On ShapeNet, which is quite small, we notice that our models overfit in terms of this metric while maintaining low 1-NN accuracy: we report their evolution in the unconditional setting in Fig. 5. This corroborates our observation that the ShapeNet benchmark is relatively saturated: by some metrics, state-of-the-art methods may even outperform an oracle which simply returns the training set. We argue that tractable likelihoods may prove very useful in datasets and tasks with no other easy means of validation.

Point cloud upsampling by inpainting. Another application of our method is point cloud upsampling. To upsample

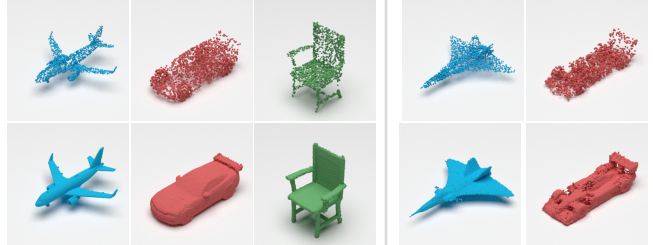


Figure 6: **Point cloud upsampling examples.** Top: original point clouds with 2048 points. Bottom: we upsample them by 50x to 102k points. Left: Our inpainting technique yields high-quality point clouds. Right: Artifacts may appear occasionally, especially on complex, irregular structures. Note: figure is rendered with a smaller point size.

	Num. params	Inference speed
LION (unconditional)	110M	2.51 s/example
PVD (unconditional)	27.7M	2.95 s/example
GECCO (unconditional)	13.7M	0.25 s/example
GECCO (conditional)	47.7M	0.27 s/example

Table 6: **Size and speed comparison.** Measured on an NVIDIA A100 GPU with 40Gb, with 2048 points and batch of 64. GECCO uses the SDE sampler with 128 steps.

from n to m points, following the blueprint of [37], we diffuse input $\{p_{0..n}\}$ to σ_{\max} and concatenate with new points sampled from the prior, $\{p_{n..m}\}$. We then reverse-diffusion, computing the scores on all points $\{p_{0..m}\}$, but use them only to update $\{p_{n..m}\}$, while $\{p_{0..n}\}$ is reversed deterministically to the input. Compared to models following Eq. 2, this approach treats the input $\{p_{0..n}\}$ as the latent code s . We use 4 resampling sub-steps (see [37]) per solver step, and when upsampling by large factors, in order to stay in the range of m the network is trained for, we concatenate multiple conditionally-independent completions of $\{p_{0..n}\}$. We find this procedure to result in coherent, high-resolution point clouds: see Fig. 6 for examples.

5. Conclusions

We propose a novel approach to condition denoising diffusion models in a geometrically-principled manner by projecting generated point clouds to an image and augmenting point locations with sparse features from a convolutional backbone. Our framework relies on a simple permutation-equivariant transformer, trained with a continuous-time diffusion scheme. It yields state-of-the-art results in single-view synthesis, while performing on par in the unconditional setting. It can also deliver exact probabilities, and upsample by inpainting. We believe this is a first step towards controllable diffusion point cloud models on real data. Future work will explore occlusion on large-scale datasets, multi-view inference, and completion via inpainting.

Acknowledgements This research was partially funded by Google’s *Visual Positioning System*.

Supplementary Material

A. Experimental details

For simplicity we measure training in terms of gradient updates rather than dataset epochs.

Unconditional ShapeNet [66]. We use the download links provided by the PointFlow paper [66] directly. Their ShapeNet models contain 100k points: we subsample them to 2048 points while training, as explained in the paper. We train our models for 500k steps, with batch size of 64, and evaluate the model every 10k steps. For this evaluation we run a simplified variant of the benchmark described in section 4.1, which uses only the chamfer distance to pick the best checkpoint over the validation set – a variant of early stopping.

ShapeNet-vol [40]. We use the download links provided by the authors of OccNet [40], which contain prerendered images in addition to point clouds. Their point clouds are already subsampled to 2048 points. We train our models for 500k steps, with batch size of 48 and evaluate the model every 10k steps with a simplified version of the benchmark described in 4.2 not using ICP and with a random subset of the validation set. Since we observed no overfitting, we do not do early stopping.

Taskonomy [67]. The dataset comes in 4 sizes: *tiny*, *medium*, *full* and *full-plus*. We noticed that *full-plus* contains certain scenes with incomplete/faulty scans. We train our model for 1M steps on the *full* variant and evaluate it on a randomly sampled 1024-element subset of the ‘validation’ split every 10k steps - the subsampling is to save computation. Note that the monocular depth baseline of [56] was also trained (by the authors) on the Taskonomy dataset.

We use the same evaluation procedure as for ShapeNet-vol, this time with 8192 examples, and observe no overfitting, so we report numbers for the final checkpoint. This means that the results over ‘validation’ and over ‘test’ should be equivalent, up to uniformity in the dataset split, and report both in the paper.

However, we notice that performance is lower on the ‘test’ split, particularly in terms of the mean, and not the median. We trace this discrepancy to two out-of-distribution scenes in the ‘test’ subset: ‘vacherie’, a residential apartment with very high ceilings which GECCO mis-scales by a factor of 2×, and ‘german’, a gym filled with large windows and mirrors, resulting in ambiguous depth (see Fig. 9). Note that the monocular baseline predicts *relative* depth, which we then rescale using the ground truth, which makes it largely immune to such outliers. In fact, the large difference in performance between GECCO without ICP and with ICP is mostly explained by mis-estimating the scale of the scene, which the baseline doesn’t suffer from,

due to its use of ground truth scale. This is also likely the reason why the baseline performs worse if ICP estimates both scale and rotation – as explained in the paper, we disable scale estimation for it.

B. Unconditional evaluation metrics

Here we define the metrics of coverage (COV), minimum matching distance (MMD) and 1-nearest neighbour accuracy (1 – NNA). They are equivalent to those used in [66], but using a different notation. We assume n given reference point clouds from the test set $\mathbf{S}_r = \{p_1, \dots, p_n\}$, n samples generated by the model $\mathbf{S}_g = \{q_1, \dots, q_n\}$, and a distance function $D(\cdot, \cdot)$ which can be either CD or EMD (defined in main text).

Coverage considers the distances between $p \in \mathbf{S}_r$ and $q \in \mathbf{S}_g$ and measures the fraction of p which are the nearest neighbour of some q :

$$\text{COV}(\mathbf{S}_r, \mathbf{S}_g) = \frac{1}{n} \left| \left\{ q \in \mathbf{S}_g : \left(\exists p \in \mathbf{S}_r : q = \arg \min_{\hat{q} \in \mathbf{S}_r} D(\hat{q}, p) \right) \right\} \right|. \quad (5)$$

Mean matching distance measures the average distance between a sample and its nearest reference point

$$\text{MMD}(\mathbf{S}_r, \mathbf{S}_g) = \frac{1}{n} \sum_{p \in \mathbf{S}_r} \min_{q \in \mathbf{S}_g} D(p, q). \quad (6)$$

1-nearest neighbour accuracy measures the accuracy of a classifier distinguishing between elements of \mathbf{S}_r and \mathbf{S}_g by returning the class of the nearest example in $\mathbf{S}_r \cup \mathbf{S}_g - \{u\}$, where u is the element under consideration itself. We define this formally by introducing the asymmetric $H(\mathbf{S}_a, \mathbf{S}_b)$ which counts the $a \in \mathbf{S}_a$ correctly classified w.r.t. the union of \mathbf{S}_a and \mathbf{S}_b and then define $1 - \text{NNA}$ as the average of H applied in both directions. We introduce $\mathbf{S}_a, \mathbf{S}_b, u$ and v to emphasize that in the final equation H is applied in both directions.

$$H(\mathbf{S}_a, \mathbf{S}_b) = \left| \left\{ u \in \mathbf{S}_a : \left[\arg \min_{v \in (\mathbf{S}_a \cup \mathbf{S}_b - \{u\})} D(u, v) \in \mathbf{S}_a \right] \right\} \right| \quad (7)$$

$$1 - \text{NNA}(\mathbf{S}_r, \mathbf{S}_g) = \frac{1}{2n} (H(\mathbf{S}_r, \mathbf{S}_g) + H(\mathbf{S}_g, \mathbf{S}_r)) \quad (8)$$

We note that EMD is usually approximated due to its computational cost, and that CUDA-based implementations may vary drastically; we use [PyTorchEMD](#).

C. Additional visualizations

We invite the reader to view the video included in this supplementary material for conditional and unconditional samples from ShapeNet, as well as for image-conditional

samples from Taskonomy. On the following pages of this document, in Fig. 7 we show additional unconditional samples from models trained on the ShapeNet, compared with baselines [38, 4]. In Fig. 8 we show additional samples from the image-conditional model trained on the Taskonomy dataset, along with the results of their upsampling. Note that we upsample the point clouds with the technique introduced in sec. 4.4 of the main paper, and how showcase it in the image-conditional case. Due to performance issues with the point cloud renderer we render the results upsampled only by 5x, but GECCO itself works equally well with 100k points, as shown in the main paper.

D. Efficient upsampling

During upsampling by inpainting, for optimal performance, it is preferable to keep the network’s input distribution as close as possible to that at training time. This means that ideally, we would want $m = n + 1$, where n and m are the cardinality of the original and upsampled point clouds, respectively. Note that a cloud can be upsampled by a factor of f , to $(1 + f)n$ points, by repeating that procedure $f \times n$ times and concatenating the results. This however is as expensive as generating a cloud of n points $f \times n$ times from scratch.

Fortunately, the Set Transformer [31] architecture enables a workaround which brings the cost down to that of sampling a cloud of $f \times n$ points once. Since the points do not interact directly with each other, but rather via the inducers (see the main paper and [31] for definition), if we make the assumption that in the limit of large n the influence of any single point on the inducers is negligible, we can reverse-diffuse many new (inpainted) points *in parallel*, with shared inducer state. This leads to a simple algorithm starting from the conditioning set $\{p\}_{1..n}^{\sigma=0}$ and the pure noise inputs $\{p\}_{n..m}^{\sigma=\sigma_{\max}}$. At each step t of reverse diffusion:

1. The conditioning input is diffused to σ_t to obtain $\{p\}_{1..n}^{\sigma_t}$.
2. The score network is evaluated on $\{p\}_{1..n}^{\sigma_t}$. The score estimate is discarded and instead we cache the activations of the inducers across the network’s layers.
3. The score network is ran again, this time on $\{p\}_{n..m}^{\sigma_t}$, but with the inducer activations provided by the cache from point 2. We obtain the score for the inpainted points $n..m$.
4. We update $\{p\}_{n..m}$ as usual, using the score from point 3.


The results provided in the main paper and here are obtained by the naive procedure, upsampling multiple times with a context of $\frac{1}{2}n = 1024$ but we also release code implementing the improved scheme described here.

E. Dataset licenses

ShapeNet. The original ShapeNet [5] dataset of 3D meshes is [licensed](#) for non-commercial use, with clauses allowing for redistribution of derived assets. Our use of derived datasets released with [66] and [40] fulfills these provisions.

Taskonomy. Our dataset of real world imagery is derived from the data released with Taskonomy [67], which allows non-commercial use in its [license](#). Bound by the license, we are not able to release our derivatives until authorized by the authors of [67]. Depending on their permission, we intend to release either our dataset or the code to re-generate it from the original files Taskonomy files.

F. Emoji license

In the video material we use an emoji () from the [Noto emoji bank](#), licensed under Apache 2.0 license.

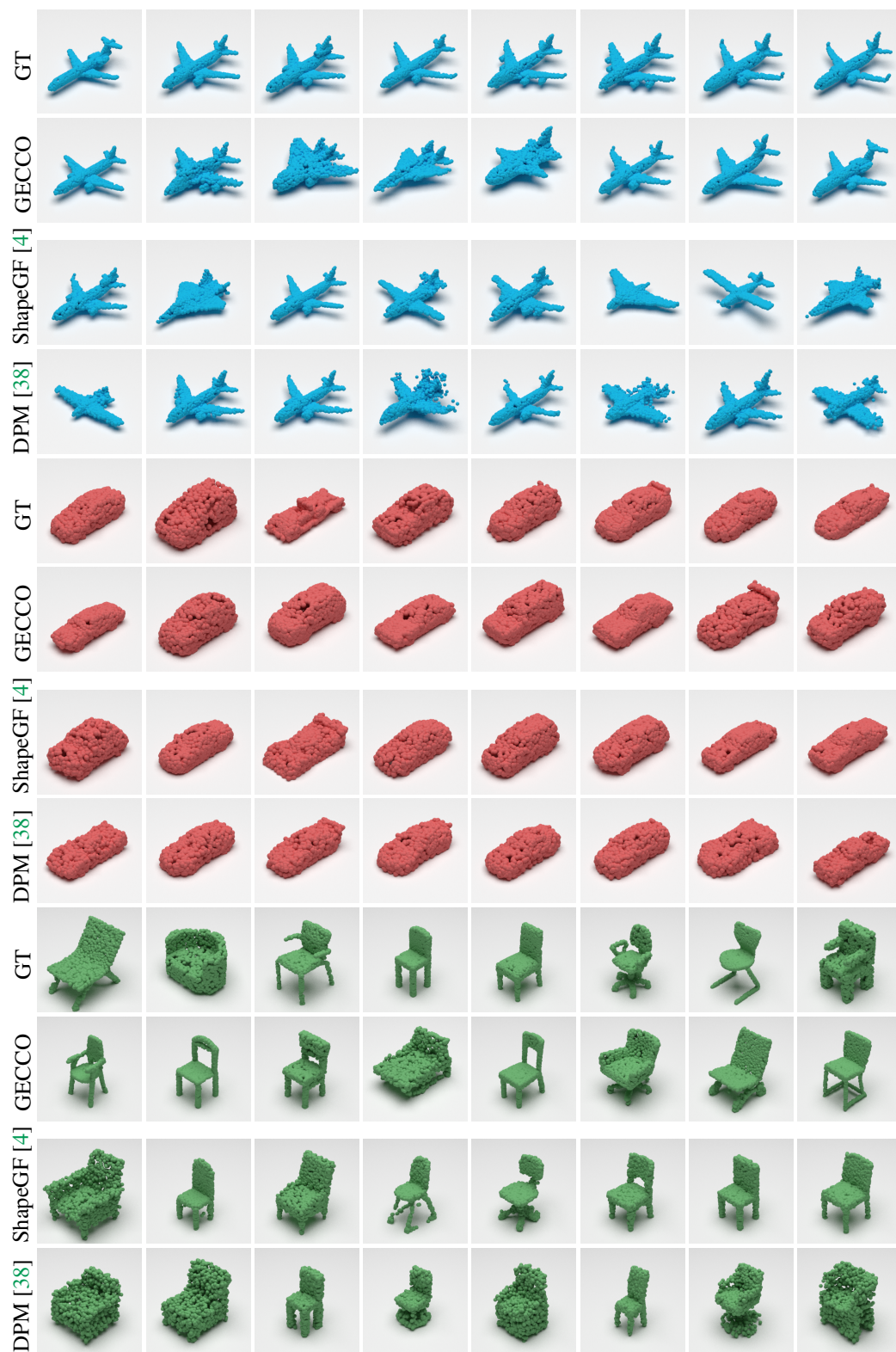


Figure 7: **Unconditional point cloud synthesis on ShapeNet.** All examples contain 2048 points.

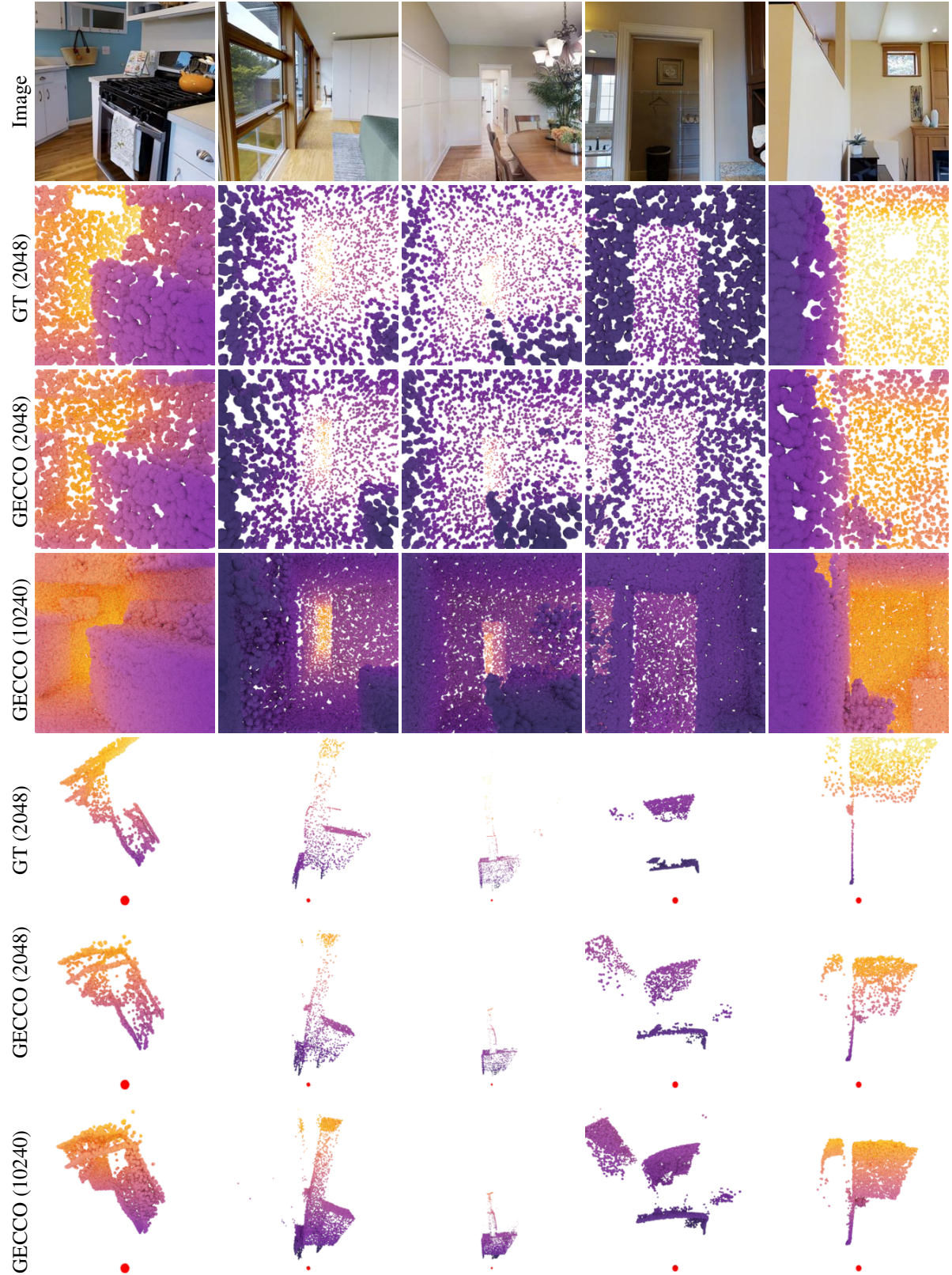


Figure 8: **Image-conditional point cloud synthesis and upsampling on the test set of Taskonomy.** We render the point clouds with smaller point radii, to highlight the difference made by upsampling. Note how in column 4 GECCO is tricked by a mirror, on the left side.

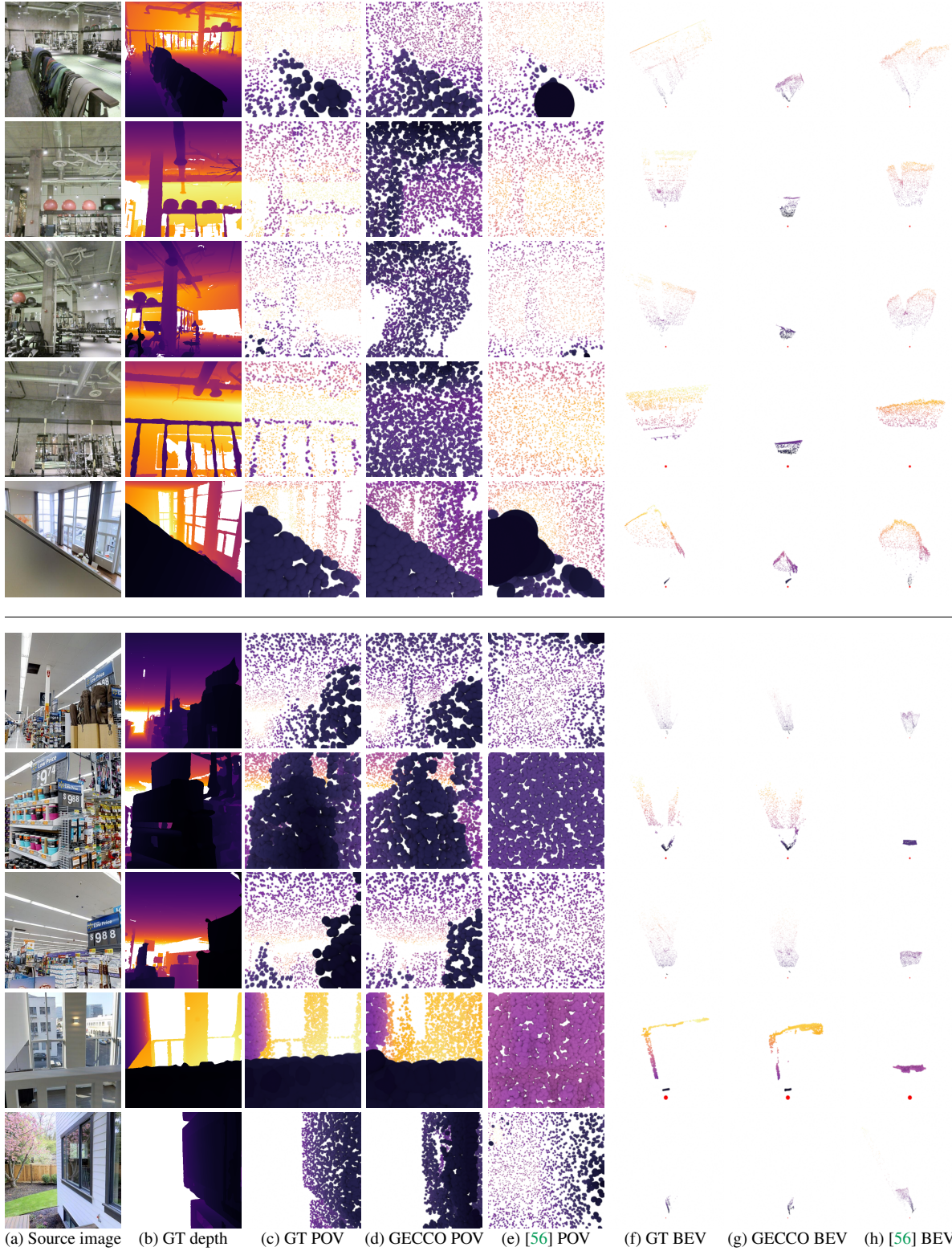


Figure 9: **Outliers in Taskonomy.** We showcase the top 5 test images where the difference in chamfer distance is most in favor of the monocular baseline (top) and GECCO (bottom). GECCO suffers from mirrors (top 4 rows, ‘german’) and unusual room dimensions (last row, ‘vacherie’), while the baseline cannot resolve areas of undefined depth, such as in the two bottom rows.

References

- [1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 2, 5
- [2] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992. 6
- [3] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. 4
- [4] Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snively, and Bharath Hariharan. Learning gradient fields for shape generation. In *European Conference on Computer Vision*, pages 364–381. Springer, 2020. 1, 2, 4, 5, 11, 12
- [5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 4, 5, 11
- [6] Chao Chen, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. Unsupervised learning of fine structure generation for 3d point clouds by 2d projection matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12466–12477, 2021. 3
- [7] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *International Conference on Learning Representations*, 2021. 2
- [8] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. Wavegrad 2: Iterative refinement for text-to-speech synthesis. *Inter-speech*, 2021. 2
- [9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 2
- [10] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016. 3, 6
- [11] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 3, 6
- [12] Kyle Genova, Xiaoqi Yin, Abhijit Kundu, Caroline Pantofaru, Forrester Cole, Avneesh Sud, Brian Brewington, Brian Shucker, and Thomas Funkhouser. Learning 3d semantic segmentation with only 2d image supervision. In *2021 International Conference on 3D Vision (3DV)*, pages 361–372. IEEE, 2021. 1
- [13] Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFFJORD: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018. 2
- [14] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 3, 6
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2
- [16] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11108–11117, 2020. 1
- [17] Le Hui, Rui Xu, Jin Xie, Jianjun Qian, and Jian Yang. Progressive point cloud deconvolution generation network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 397–413. Springer, 2020. 2, 5
- [18] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 2
- [19] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. <https://mitsuba-renderer.org>. 5
- [20] Andrej Karpathy. Mingpt. <https://github.com/karpathy/minGPT/>, 2020. 4
- [21] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 1, 2, 3, 4, 7, 8
- [22] Jacob Kelly, Jesse Bettencourt, Matthew J Johnson, and David K Duvenaud. Learning differential equations that are easy to solve. *Advances in Neural Information Processing Systems*, 33:4370–4380, 2020. 3
- [23] Patrick Kidger. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2021. 4
- [24] Patrick Kidger and Cristian Garcia. Equinox: neural networks in JAX via callable PyTrees and filtered transformations. *Differentiable Programming workshop at Neural Information Processing Systems 2021*, 2021. 4
- [25] Hyeonju Kim, Hyeonseung Lee, Woo Hyun Kang, Joun Yeop Lee, and Nam Soo Kim. Softflow: Probabilistic framework for normalizing flow on manifolds. *Advances in Neural Information Processing Systems*, 33:16388–16397, 2020. 1, 2, 5
- [26] Jinwoo Kim, Jaehoon Yoo, Juho Lee, and Seunghoon Hong. SetVAE: Learning hierarchical composition for generative modeling of set-structured data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15059–15068, 2021. 2, 5

- [27] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 2
- [28] Roman Klokov, Edmond Boyer, and Jakob Verbeek. Discrete point flow networks for efficient point cloud generation. In *European Conference on Computer Vision*, pages 694–710. Springer, 2020. 2, 5
- [29] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *International Conference on Learning Representations*, 2021. 2
- [30] Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019. 7
- [31] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019. 2, 3, 11
- [32] Rémy Leroy, Pauline Trouvé-Peloux, Frédéric Champagnat, Bertrand Le Saux, and Marcela Carvalho. Pix2point: Learning outdoor 3d using sparse point clouds and optimal transport. In *2021 17th International Conference on Machine Vision and Applications (MVA)*, pages 1–5. IEEE, 2021. 3
- [33] Chun-Liang Li, Manzil Zaheer, Yang Zhang, Barnabas Poczos, and Ruslan Salakhutdinov. Point cloud gan. *arXiv preprint arXiv:1810.05795*, 2018. 2
- [34] Ruihui Li, Xianzhi Li, Ka-Hei Hui, and Chi-Wing Fu. SP-GAN: sphere-guided 3d shape generation and manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021. 2, 5
- [35] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 4
- [36] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel CNN for efficient 3d deep learning. *Advances in Neural Information Processing Systems*, 32, 2019. 3
- [37] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 8
- [38] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 1, 2, 3, 4, 5, 11, 12
- [39] Zhaoyang Lyu, Zhifeng Kong, Xudong Xu, Liang Pan, and Dahua Lin. A conditional point diffusion-refinement paradigm for 3d point cloud completion. *arXiv preprint arXiv:2112.03530*, 2021. 3
- [40] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 6, 8, 10, 11
- [41] Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. Deep learning for monocular depth estimation: A review. *Neuro-computing*, 438:14–33, 2021. 3
- [42] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2
- [43] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021. 2
- [44] Janis Postels, Mengya Liu, Riccardo Spezialetti, Luc Van Gool, and Federico Tombari. Go with the flows: Mixtures of normalizing flows for point cloud generation and reconstruction. In *2021 International Conference on 3D Vision (3DV)*, pages 1249–1258. IEEE, 2021. 2
- [45] Albert Pumarola, Stefan Popov, Francesc Moreno-Noguer, and Vittorio Ferrari. C-flow: Conditional generative flow models for images and 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7949–7958, 2020. 2
- [46] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1
- [47] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1, 3
- [48] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Abed Al Kader Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 2022. 1
- [49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6
- [50] Sameera Ramasinghe and Simon Lucey. Beyond periodicity: Towards a unifying framework for activations in coordinate-MLPs. In *European Conference on Computer Vision*, pages 142–158. Springer, 2022. 3
- [51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2
- [52] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 6
- [53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2

- [54] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022. 1, 2
- [55] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshah. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022. 6
- [56] Alexander Sax, Jeffrey O Zhang, Bradley Emi, Amir Zamir, Silvio Savarese, Leonidas Guibas, and Jitendra Malik. Learning to navigate using mid-level visual priors. In *Conference on Robot Learning*, pages 791–812. PMLR, 2020. 6, 7, 10, 14
- [57] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3859–3868, 2019. 2, 5
- [58] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1, 2
- [59] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1, 2, 3, 4, 7, 8
- [60] Weiwei Sun, Wei Jiang, Eduard Trulls, Andrea Tagliasacchi, and Kwang Moo Yi. Acne: Attentive context normalization for robust permutation-equivariant learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11286–11295, 2020. 1
- [61] Yongbin Sun, Yue Wang, Ziwei Liu, Joshua Siegel, and Sanjay Sarma. Pointgrow: Autoregressively learned point cloud generation with self-attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 61–70, 2020. 2
- [62] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2Mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018. 3, 6
- [63] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 2
- [64] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3
- [65] Jianwen Xie, Yifei Xu, Zilong Zheng, Song-Chun Zhu, and Ying Nian Wu. Generative PointNet: Deep energy-based learning on unordered point sets for 3d generation, reconstruction and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14976–14985, 2021. 2
- [66] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4541–4550, 2019. 1, 2, 3, 5, 10, 11
- [67] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 2, 5, 7, 10, 11
- [68] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. LION: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 5, 7
- [69] Dongsu Zhang, Changwoon Choi, Jeonghwan Kim, and Young Min Kim. Learning to generate 3d shapes with generative cellular automata. *arXiv preprint arXiv:2103.04130*, 2021. 5
- [70] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021. 1
- [71] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 1, 2, 3, 5
- [72] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuxin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9939–9948, 2021. 1
- [73] Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *Advances in neural information processing systems*, 33:18795–18806, 2020. 4