Rethinking Metrics and Diffusion Architecture for 3D Point Cloud Generation

Matteo Bastico[1,*], David Ryckelynck[3] Laurent Corté[1], Yannick Tillier[3], Etienne Decencière[2]
Mines Paris, PSL University
[1]Centre for material sciences (MAT), UMR7633 CNRS, 91003 Evry, France
[2]Centre for mathematical morphology (CMM), 77300 Fontainebleau, France
[3]Centre for material forming (CEMEF), UMR7635 CNRS, 06904 Sophia Antipolis, France

Abstract

As 3D point clouds become a cornerstone of modern technology, the need for sophisticated generative models and reliable evaluation metrics has grown exponentially. In this work, we first expose that some commonly used metrics for evaluating generated point clouds, particularly those based on Chamfer Distance (CD), lack robustness against defects and fail to capture geometric fidelity and local shape consistency when used as quality indicators. We further show that introducing samples alignment prior to distance calculation and replacing CD with Density-Aware Chamfer Distance (DCD) are simple yet essential steps to ensure the consistency and robustness of point cloud generative model evaluation metrics. While existing metrics primarily focus on directly comparing 3D Euclidean coordinates, we present a novel metric, named Surface Normal Concordance (SNC), which approximates surface similarity by comparing estimated point normals. This new metric, when combined with traditional ones, provides a more comprehensive evaluation of the quality of generated samples. Finally, leveraging recent advancements in transformer-based models for point cloud analysis, such as serialized patch attention , we propose a new architecture for generating high-fidelity 3D structures, the Diffusion Point Transformer (DiPT). We perform extensive experiments and comparisons on the ShapeNet dataset, showing that our model outperforms previous solutions, particularly in terms of quality of generated point clouds, achieving new state-of-the-art. Code available at `https://github.com/matteo-bastico/DiffusionPointTransformer`.

1. Introduction

The analysis of 3D point clouds, critical for applications ranging from autonomous vehicles [7] and robotics [12] to the medical domain [34, 69], faces persistent challenges in collecting and annotating large-scale data. With recent advancements in deep generative models [52], point cloud generation and synthesis have attracted growing interest from the research community, aiming to produce high-fidelity, realistic samples [1, 13, 26, 28, 40, 42, 54, 57, 67, 68, 74]. Like other generative tasks, this field presents two major challenges: (1) designing effective deep learning architectures and (2) developing robust evaluation methods to ensure fair model comparisons.

Generative AI has achieved significant success across various domains, producing high-quality 2D images [47, 50], among others, mainly leveraging transformer-based architectures [58]. These models are built upon attention mechanisms to capture relationships between input tokens. This makes them inherently suited to point cloud analysis, where understanding spatial relationships between points is essential. As a result, deep learning algorithms for point cloud processing have recently received a significant boost [15, 30, 37, 59, 60, 63, 64, 70, 72]. Classification and segmentation tasks, in particular, have achieved impressive performance thanks to recent developments, such as Point Cloud Transformer (PCT) [15], Point Transformer (PT) and its successors [63, 64, 72]. Meanwhile, Denoising Diffusion Probabilistic Models (DDPMs) [18] have demonstrated immense potential in generative tasks [10, 19, 21, 24, 40, 47, 65] by employ a forward noising process and learning a reverse process that restores the original data. Several efforts have been made to apply DDPMs to 3D shapes [23, 40, 42, 48, 71, 74]. However, many of these approaches rely on partitioning input data into voxels [42], using downsampled encoded tokens [23], or leveraging skeletons [48], often leading to the loss of local structure details. Despite these advancements, point cloud generation and evaluation remain challenging due to the complexity of 3D data and the difficulty of assessing spatial relationships. As we will show, some traditional point cloud generative model evaluation metrics [1, 67] frequently fail to capture geometric fidelity and structural consistency, especially in the presence of noise and translations on generated samples,

*Corresponding author: matteo.bastico@minesparis.psl.eu

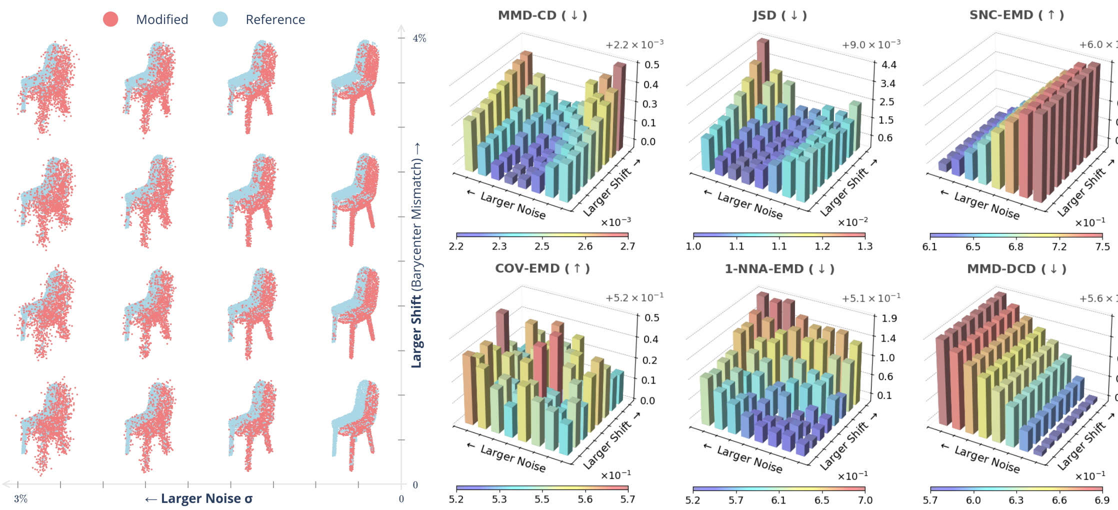

Figure 1. Response of several metrics to random noise and barycenter shift on generated samples. (**Left**) An example comparing a reference sample (blue) and its modified version (red) as noise and barycenter translations are added in proportion to its diameters. (**Right**) An overview of the robustness of some traditional metrics (MMD-CD, COV-EMD, JSD and 1-NNA-EMD) and some proposed metrics (SNC-EMD, MMD-DCD) for evaluating point cloud generation.

slowing progress in developing more robust and reliable solutions. Thus, new guidelines for assessment are needed to better meet the demands of real-world applications. In this work, we propose enhancements to existing metrics to improve their stability and better reflect the true quality of generated shapes. Our approach involves performing rigid alignment of synthesized shapes to ensure consistent matching with reference samples, along with incorporating recent improvements of Chamfer Distance (CD) to account for point density rather than relying solely on Euclidean distance, i.e. the Density-aware Chamfer Distance (DCD) [62]. Additionally, we introduce a new metric, the Surface Normal Concordance (SNC), which facilitates the evaluation of point cloud structures by incorporating point normals, particularly in contexts where surface regularity and local geometry are critical for generating realistic synthetic data [22, 51]. Through a small scale user study, we show that SNC better reflects human visual perception than current quality indicators.

Furthermore, to enhance the quality of generated point clouds, we introduce a novel plain transformer-based architecture for DDPM, inspired by recent advancements on point cloud processing [63, 64, 72], called Diffusion Point Transformer (DiPT). Unlike existing methods, our model preserves the raw input size (in number of points) throughout its layers, avoiding voxelization or downsampling, which often compromise output surface quality. Experiments on the ShapeNet benchmark [6] show that our point-wise diffusion approach consistently produces higher-fidelity generated samples, demonstrating a clear improvement over previous methods.

Our contributions are summarized as follows:

- We propose new guidelines to improve the evaluation metrics for point cloud generative models.
- We introduce a new metric, Surface Normal Concordance (SNC), to assess the samples quality by also considering point normals rather than only Euclidean distances.
- We present Diffusion Point Transformer (DiPT), a novel model for point-wise diffusion that avoids voxelization or downsampling, boosting final quality.
- We provide extensive evaluation and comparison of DiPT on the ShapeNet dataset [6] on various object categories.

2. Related Works

Metrics. Several metrics have been defined to assess the quality of point cloud generative models [1, 42, 56, 67, 74]. These metrics always compare a set of generated samples, S_g, with a reference set, S_r. The Fréchet Point Cloud Distance (FPD) [54], inspired by the Fréchet Inception Distance (FID) [16], defined to evaluate 2D image generation, was initially used to measure the distance between real and generated samples in the feature spaces extracted by PointNet [49]. In recent studies, FPD has been replaced by newer metrics that leverage Euclidean distances to quantify point clouds similarity [45]. Two widely used distance measures for point clouds are the CD and the Earth Mover's Distance (EMD). CD calculates the sum of the squared Euclidean distances from each point in one point cloud to the nearest point in the other point cloud, while EMD, also known as Wasserstein distance, computes the minimal cost required to transform one point cloud into another. Metrics built on such measures aim to effectively capture both the quality, i.e. realism, of generated samples and/or their diversity or representativeness. Based on these two principles, *Achlioptas et al.* [1] introduced three key evaluation metrics:

- Coverage (**COV**): Evaluates the diversity of generated samples relative to the reference set.
- Minimum Matching Distance (**MMD**): Measures the average distance to the nearest (i.e., most similar) reference, aiming to capture the quality of generated samples.
- Jensen-Shannon Divergence (**JSD**): Quantifies the similarity between the marginal point distributions of voxelized reference and generated shapes.

Recently, to overcome some limitations of these metrics, *Yang et al.* [67] introduced a new metric, the 1-Nearest Neighbour Accuracy (**1-NNA**) [38, 66]. It essentially measures to what extent the distributions of S_g and S_r are similar, focusing primarily on the diversity of generated point clouds, with a marginal consideration of quality. Furthermore, *Triess et al.* [56] proposed a learning-based metric to quantify the realism of local regions in LiDAR point clouds. However, this approach requires a proxy classification task trained on both real-world and synthetic point clouds. Following previous works, we refer to a given metric computed with a specific distance measure as METRIC-MEASURE (e.g., MMD-CD refers to MMD calculated using CD). As shown in Fig. 1 and discussed in the next section, certain traditional metrics can lead to misleading evaluations. To address this, we introduce metric enhancements, together with SNC, to provide a more reliable and comprehensive assessment of generative models.

Formal definitions of the distance measures and traditional metrics are provided in Sec. 8 of the Supplementary.

3D Point Cloud Generation. Different techniques have been exploited for 3D point cloud generation, mostly deep-learning methods such as Variational AutoEncoders (VAE) [13, 26, 68], Generative Adversarial Networks (GANs) [1, 54, 57], normalized flows [28, 67], and diffusion models [40, 42, 74]. Among these, FoldingNet [68] was an early attempt, built upon PointNet [49] to address unsupervised learning challenges on point clouds using a VAE. SetVAE [26] approached point cloud generation as a set generation task using a hierarchical VAE based on a set transformer [31]. ShapeGF [5] proposed to learn distributions over gradient fields that model shape surfaces. PointFlow [67] introduced a novel approach using continuous normalizing flows to simultaneously model the distribution of latent variables and the distribution of points for a given shape. SoftFlow [25] extended this idea by estimating the conditional distribution of noisy input point clouds perturbed by random noise sampled from various distributions.

More recently, the advent of DDPMs has led to substantial improvements in 3D point cloud generation. Early diffusion-based methods for point clouds, such as DPM [40], employed PointNet [49] backbone. Others, including Point-Voxel Diffusion (PVD) [74] and LION [71], implemented instead the Point-Voxel Convolutional (PVConv) architecture [35]. PVD combines a low-resolution voxel-based branch to encode coarse-grained information with a high-resolution point-based branch to capture fine-grained features. LION [71] introduced the diffusion in two different latent spaces combining global shape representation with point-structured features. More recently, plain transformer-based diffusion models have gained popularity also for 3D point cloud generation, achieving outperforming results. In particular, DiT-3D [42] adapted the Diffusion Transformer (DiT) architecture [47] to voxelized point clouds, enabling multi-class training with learnable class embeddings. Similarly, Latent Diffusion Transformer (LDT) [23] proposed an AE latent compressor to convert raw point clouds into latent tokens, which are then processed by diffusion models.

As a result, many previous works rely on point encoding techniques such as voxelization, downsampling, or compression, which can degrade the final quality. In contrast, our approach, DiPT, performs diffusion directly on raw point clouds without reducing their resolution, enabling the generation of fine-grained, high-quality samples.

3. Metrics Rethinking

We identify three key properties for a generative point cloud evaluation metric: (1) invariance to rigid translations of generated samples, (2) consistent behavior across different point distributions, and (3) an inverse monotonic response to noise. The latter property should strictly hold for quality metrics (e.g., MMD), whereas for variability metrics (e.g., COV) we expect invariance at low noise levels and an inverse response only when noise is high enough to alter the underlying shape structure. As shown in Fig. 1, and in more detail in Sec. 10 of the Supplementary, one or more of these properties does not always hold for some traditional metrics. For example, MMD-CD and JSD do not exhibit a monotonic inverse response to noise, and none of the traditional metrics are invariant to barycenter shifts.

The proposed enhancements are jointly formalized and validated below, using traditional calculations as a baseline. Analyses are conducted on a set of 4573 training samples, considered as ideal generations S_g, and compared against a reference set S_r of 753 samples. We progressively introduce random Gaussian noise and/or rigid barycenter shifts proportional to sample diameters (i.e., the maximum inter-point distance), as in Fig. 1 (Left). Shapes contain 2,048 points, following the literature [5, 23, 25, 26, 40, 42, 71, 74], sampled from the original point clouds either uniformly or randomly in separate trials to simulate uniform and inhomogeneous point distributions.

Barycenter Alignment. Prior works on point cloud gen-

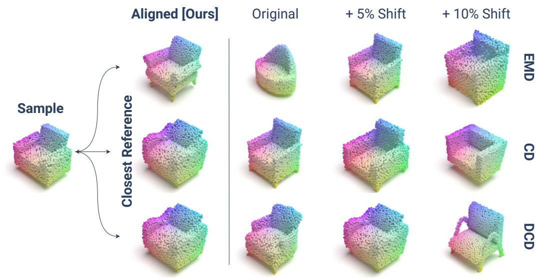

Figure 2. Closest references to a sample under different distance measures with alignment and in response to small shifts.

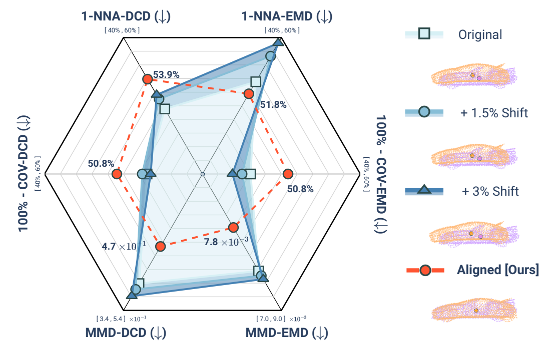

Figure 3. Comparison of 1-NNA, MMD, and COV computed with (red) and without (blue) barycenter alignment. Each metric is evaluated using both DCD and EMD for three levels of shifting.

eration typically apply global rather than per-sample normalization, using the training set mean and standard deviation [25, 26, 42, 67, 71, 74]. This ensures the model learns a distribution in normalized space (e.g., varying scales) instead of adapting to each sample specific characteristics. Generated point clouds are eventually de-normalized before evaluating model performances. As a result, barycenters can vary within the same set and between S_g and S_r. Furthermore, even with sample-wise normalization, generative models have no theoretical guarantee of producing centered objects, and current evaluation distance measures [1, 67] do not inherently account for such displacements, compromising metric invariance to sample positioning. For example, the same generated point cloud with different small shifts may be matched as closest to different reference samples when alignment is not applied, as in Fig. 2, affecting COV and 1-NNA values. To overcome this issue and obtain the desired invariance, we propose a barycenter alignment of point clouds before computing their distances. That is, instead of computing directly a distance measure $D(X, Y)$ between two point clouds, $X = \{\boldsymbol{x}_i\}_{i=1}^N$ and $Y = \{\boldsymbol{y}_j\}_{j=1}^M$, we compute $D(X - \boldsymbol{x}_b, Y - \boldsymbol{y}_b)$, where $\boldsymbol{x}_b = \frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i$ and $\boldsymbol{y}_b = \frac{1}{M}\sum_{j=1}^M \boldsymbol{y}_j$. In this way, a generated point cloud with a given structure will always be associated with the same reference regardless of its position in the Euclidean space. A comparison of several metrics computed with and without alignment is shown in Fig. 3. Specifically, the stable metric value achieved using the proposed barycenter alignment is compared to traditional computation, which exhibits undesired variability under small barycenter shifts.

Replacing CD with DCD. The CD, traditionally used to evaluate generative point cloud models, is well known for its limitations. Among these, it is insensitive to mismatched local density [62], weakly rotation-aware [33], and vulnerable to outliers [32]. As a result, CD-based metrics do not always respond inversely to noise. In fact, metrics such as MMD-CD can exhibit improvements when low to mid levels of noise are added to the samples in S_g, as shown in Fig. 4, making them unsuitable as quality indicators. Barycenter alignment mitigates but does not eliminate this issue. To address these limitations, we propose replacing CD in the metrics calculation with the recently introduced DCD [62], detailed in Eq. (5) of the Supplementary. DCD is inherited from CD but benefits from a higher sensitivity to distribution quality and has been proven to be a more robust measure of point clouds similarity. These properties make DCD more suitable than CD for evaluating generative models. To validate this intuition, in Fig. 4 we compare the robustness of the MMD metric against the amount of noise added to S_g when computed using different distance measures: CD, EMD, and DCD. Additionally, to cover all scenarios, we compare the metrics computed with and without barycenter alignment for both uniformly and randomly sampled point clouds. In contrast to CD, EMD and DCD demonstrate a monotonically increasing behavior in response to noise. However, MMD-DCD without alignment shows a slight improvement at low noise levels, which disappears once barycenter alignment is applied before distance calculation (see zoom in Fig. 4). Interestingly, for uniform samples, MMD-DCD increases more rapidly, as perturbations cause stronger density variations that amplify the effect of DCD. This analysis shows that DCD outperforms CD in MMD calculation and underscores the importance of alignment for reliable evaluation of generative models. Intuitively, improving distance calculation with DCD also benefits other metrics, such as 1-NNA and COV. A more detailed comparison between DCD- and CD-based metrics is available in Sec. 10 of the Supplementary.

Surface Normal Concordance. Several methods have been proposed in the literature for estimating point cloud normals [27], ranging from the Principal Component Analysis (PCA) of a neighborhood region [4, 20], which is detailed in Sec. 9 of the Supplementary, to more recent deep learning-based approaches [3, 14], as well as other techniques [41, 46, 61, 73]. SNC measures the average similar-

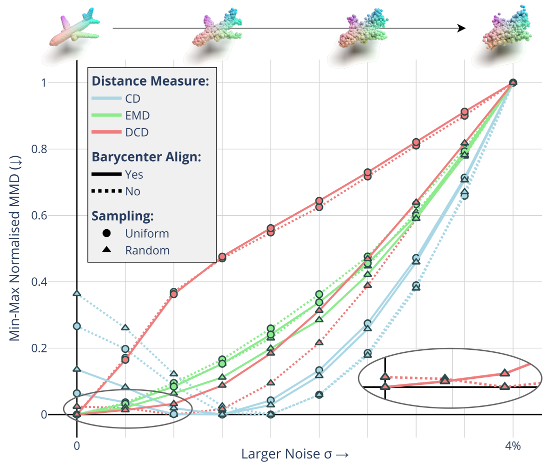

Figure 4. Evolution of the normalized MMD with respect to noise added to the samples of S_g, comparing distance measures (CD, EMD, DCD) under different conditions: with or without barycenter alignment and using uniformly or randomly sampled points.

ity of these normals, calculated using any chosen method, between generated samples and their closest references. Specifically, let $M_X \in S_r$ represent the closest reference sample, i.e. the best match, after barycenter alignment, to $X \in S_g$, such that

$$M_X = \underset{Y \in S_r}{\arg\min}\, D(X - \boldsymbol{x}_b, Y - \boldsymbol{y}_b) \tag{1}$$

where $D(\cdot,\cdot)$ is any point clouds distance function, e.g. EMD or DCD. Additionally, let $\hat{n}(\cdot)$ denote any method for computing point normals. The SNC is then defined as

$$\begin{aligned}\mathrm{SNC}(S_r, S_g) = \frac{1}{|S_g|} \sum_{X \in S_g} \frac{1}{|X|} \sum_{\boldsymbol{x} \in X} \\ \left| \hat{n}(\boldsymbol{x}) \cdot \hat{n}\big(\underset{\boldsymbol{y} \in M_X}{\arg\min}\, ||\boldsymbol{x} - \boldsymbol{y}||_2 \big) \right|. \end{aligned} \tag{2}$$

Namely, for each point in a generated point cloud, SNC computes the similarity between its normal direction with the normal direction of the closest point from the best-matching shape in the set of references. The proposed metric is highly flexible and can be computed independently of the specific distance measure $D(\cdot,\cdot)$ or normal estimation method $\hat{n}(\cdot)$, as it uses only the absolute value of the cosine to address sign disambiguity, e.g. in PCA-based methods. SNC demonstrates a very strong inverse response to noise, as shown in Fig. 1. This is because small perturbations in point positions cause significant variations in their normals. Thus, SNC is highly sensitive to fine-grained details, making it an ideal complement to traditional metrics for evaluating the quality of generated point clouds. Additionally, as discussed in the experiments, normals are independent of global scaling and normalization, enabling fair model comparisons. When computed with a robust method, they are also less sensitive to point distribution than pure Euclidean distances, ensuring consistent behavior.

4. Diffusion Point Transformer

Inspired by DiT-3D [42] for its diffusion structure and PTv3 [64] for its backbone architecture, we propose the Diffusion Point Transformer (DiPT) model, in Fig. 5, for 3D point cloud generation. Motivated by some recent advancements [37, 59, 60, 64], we transition from the traditional unordered paradigm of point clouds to a serialized structured format. To achieve this, we employ space-filling curves to reorganize point clouds into a one-dimensional sequence by using the Z-order curve [43], Hilbert curve [17], and their variants Trans-Hilbert and Trans-Z [64]. Importantly, this serialization does not require voxelization nor downsampling. Sparse points are placed into a grid of a given resolution to define the serialized order, as on the right of Fig. 5, allowing the input data to retain its original dimensionality. To enhance generalization capabilities, we incorporate random shuffling of the serialized orders, following the approach of [60, 64]. This ensures that each DiPT block can learn diverse patterns rather than focusing on a single space-filling curve. Moreover, the serialization enables input points to be grouped into non-overlapping patches, with attention performed independently within each patch, inspired by window attention [36]. This approach, named Serialized Patch Attention [64], reduces the computational cost compared to traditional local structure creation methods such as K-Nearest Neighbors (KNN) [63, 72]. Moreover, we replace the absolute sine-cosine embeddings of DiT-3D [42] or Relative Positional Embeddings (RPE) with enhanced Conditional Positional Encoding (xCPE) [59, 64]. It consists of a sparse convolution layer with a skip connection before the attention layer of each block, offering more flexibility than traditional positional embeddings for point clouds. As xCPE operates outside the attention mechanism, unlike RPE, it enables optimizations such as flash attention [8, 9, 53], significantly reducing computational time.

Following DiT [42, 47], we adapt the model for diffusion by incorporating Adaptive Layer Normalization (AdaLN) for feature modulation and scaling based on the input condition. The latter includes time embedding, representing the forward diffusion step, and a learnable class embedding, encoding the category to generate. This design enables multi-class training since in each DiPT block the features scale and shift parameters γ and β are regressed from the input condition. Additionally, a scaling parameter α is applied after each operation and before residual connections within a block, ensuring condition-dependent feature scaling.

DiPT is designed for scalability, performing point-wise

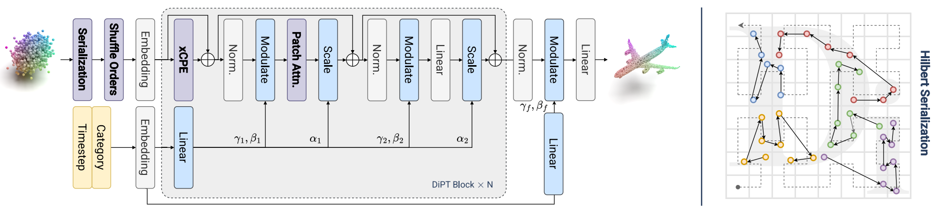

Figure 5. Proposed Diffusion Point Transformer (DiPT) for 3D point cloud generation. (**Left**) The model serializes the raw input and shuffles the serialization orders before processing it through N DiPT blocks, each performing xCPE, Serialized Patch Attention, and a linear layer. Features are modulated and scaled based on the input condition, composed of the sample category and the diffusion noising timestamp. (**Right**) Example of Hilbert serialization, where each color represents a patch of maximum size 8.

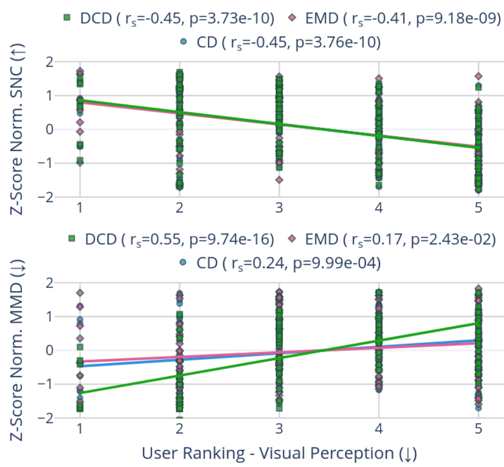

Figure 6. Quality metrics user study. User point clouds perceptual quality rankings are compared to SNC (top) and MMD (bottom).

rather than voxel-wise diffusion, and can adapt to different window sizes and model configurations by tuning the patch size and number of blocks. As shown below, it achieves superior point cloud generation quality, producing high-fidelity outputs compared to state-of-the-art methods.

5. Experiments

5.1. Metrics User Study

We conducted a small scale user study to validate the proposed SNC based on human perception of point clouds quality. 15 participants from a mixed audience were asked to sort 5 point clouds, comprising a random reference and its DCD-closest generation from 4 different models, from most to least realistic. Samples were presented in random order via an interactive 3D GUI. Each user ranked samples from the 3 different categories in separate trails, resulting in a total of 45 trials. Correlations between user rankings and quality metrics are shown in Fig. 6, with average Spearman scores of -0.44 for SNC and 0.32 for MMD. These results strengthens the proposed SNC by suggesting that it better reflects human perception than MMD. Furthermore, this study confirms the earlier intuition from the MMD analysis in Fig. 4, with DCD achieving the highest correlation to visual perception among all metrics, and improving CD of 0.31. In contrast, SNC maintains a similar correlation regardless of the base distance measure, indicating its desired weaker dependence on Euclidean distances.

5.2. Experiments Settings

Dataset. Following previous works [5, 23, 25, 26, 40, 42, 71, 74], we used the chair, airplane, and car categories from ShapeNet [6] to train the DiPT model. For each training shape, we sampled 2048 points using Furthest Point Sampling (FPS). We adopted the same dataset splits and preprocessing steps introduced in PointFlow [67], which are widely adopted in the community [25, 26, 42, 71, 74], including global sample normalization. Additional DiPT experiments on 10 mixed ShapeNet categories, as well as ablations on model size and components (e.g., positional embeddings), are provided in Sec. 11 of the Supplementary.

Implementation Details. For comparison with other methods, we trained the proposed DiPT model following the Small (S) ViT and DiT architecture [11, 42, 47]. Namely, we used 12 blocks with feature size 384 and 6 attention heads. Inspired by Swin-Transformer [36], we alternate small and large patch sizes for the serialized attention, repeating the pattern 256 - 512 - 1024 - 1024 and aiming to capture both local and global information relevant for generation variability and quality, respectively. The models were trained on 32 NVIDIA H100 GPUs for 10000 epochs using the AdamW optimizer [39] and one-cycle learning rate policy [55] with a maximum learning rate of $2e{-}4$. Finally, we used a DDPM scheduler with 1000 noising steps with linearly increasing forward process variances from $1e{-}4$ to 0.02, as in [18]. SNC was calculated using PCA-based normals [4, 20] extracted from neighborhoods of 20 points. We found this value to be a good trade-off between local and global normal information for the ShapeNet samples (see Fig. 8 of the Supplementary). We chose this method for its simplicity and flexibility in handling varying distributions, as it focuses on local geometric structures.

Table 1. Comparison of metrics across different models for 3D point cloud generation. All models are evaluated using the same uniformly sampled reference set and their public generated samples or weights. MMD is omitted for models trained with different input normalization, as it does not provide a fair comparison. The best scores are highlighted in bold. MMD-DCD is scaled by 10, and MMD-EMD by 10^3.

	Model	Variability				Quality			
		1-NNA (%,↓)		COV (%,↑)		MMD (↓)		SNC (%, ↑)	
		DCD	EMD	DCD	EMD	DCD	EMD	DCD	EMD
Chair	PointFlow [67]	60.72	60.18	43.64	**52.07**	6.49	8.91	70.93	69.42
	SoftFlow [25]	61.64	67.76	37.67	43.34	6.47	9.07	73.11	71.48
	ShapeGF [5]	55.28	64.47	**49.16**	45.48	-	-	73.99	73.08
	SetVAE [26]	62.33	66.54	43.19	41.04	6.44	8.73	**77.41**	74.61
	DPM [40]	70.21	91.65	40.12	33.84	-	-	70.14	68.21
	PVD [74]	52.60	**54.13**	45.33	48.24	6.46	**8.46**	75.94	73.72
	LION [71]	**51.61**	54.98	44.72	49.46	6.44	8.54	75.47	73.19
	DiT-3D [42]	99.00	91.35	17.00	19.14	6.68	9.85	76.12	73.52
	DiPT [Ours]	68.68	64.47	41.81	43.95	**6.08**	8.47	77.29	**75.10**
Airplane	PointFlow [67]	66.67	86.30	40.99	38.27	4.30	2.35	83.25	81.12
	SoftFlow [25]	66.79	90.37	40.00	38.52	4.26	2.40	84.05	81.98
	ShapeGF [5]	64.94	92.10	**47.41**	30.86	-	-	83.27	81.32
	SetVAE [26]	64.69	88.52	38.52	36.79	4.26	2.24	87.39	85.51
	DPM [40]	68.40	92.96	40.99	28.15	-	-	82.86	81.20
	PVD [74]	**60.62**	82.35	43.46	40.00	4.36	2.17	84.42	82.60
	LION [71]	65.68	84.94	44.44	39.01	4.24	2.30	83.01	80.90
	LDT [23]	90.25	90.86	44.20	34.32	-	-	86.35	83.95
	DiPT [Ours]	63.70	**74.32**	44.20	**46.42**	**3.29**	**1.65**	**87.50**	**86.00**
Car	PointFlow [67]	50.85	61.97	43.02	49.00	5.41	3.69	76.84	74.67
	SoftFlow [25]	**50.57**	67.38	37.89	45.01	5.39	3.75	78.31	75.84
	ShapeGF [5]	52.71	68.23	46.72	45.01	-	-	77.62	75.69
	SetVAE [26]	53.42	72.65	36.47	49.29	5.38	3.55	82.54	79.82
	PVD [74]	50.71	64.25	42.74	51.28	5.55	4.54	78.99	76.45
	LION [71]	50.85	64.39	41.88	**53.28**	5.48	3.70	78.01	75.73
	LDT [23]	75.93	73.08	**47.86**	50.43	-	-	82.26	78.89
	DiPT [Ours]	61.11	**60.26**	36.47	44.44	**4.65**	**3.28**	**82.69**	**80.64**

5.3. Experiments Results

We present, in Tab. 1, a quantitative comparison of generative models using the proposed enhanced evaluation metrics. Note that JSD is excluded from the analysis, as it remains the only metric that lacks robustness and stability, even after the refinements (see Fig. 10 of the Supplementary). Our DiPT model demonstrates its superiority over the others, achieving the best performance on qualitative metrics MMD and SNC across all categories. Compared to DiT-3D with the same Small (S) model size [42], our model demonstrates significantly better generalization (greater variability) while simultaneously producing higher-quality samples. Furthermore, the introduced SNC metric complements MMD by providing a deeper understanding of the quality of the generated samples. When MMD cannot be compared fairly due to different normalization, and consequently different generated point cloud sizes, SNC can be used as the only reliable quality indicator, as it is not affected by scale. Additionally, when MMD values are very close or discordant when computed with different distance measures, SNC helps in better interpreting the results. For example, in airplane generation, SetVAE [26] and PVD [74] exhibit discordant MMD-DCD and MMD-EMD values, with one model outperforming the other on only one measure. The SNC metric, however, reveals that SetVAE produces higher-quality samples, as its value is consistently higher for both DCD and EMD. In fact, as shown in the graphical comparison in Fig. 7, the airplane sample generated by SetVAE seems less noisy than the one generated by PVD, in concordance with MMD-DCD and SNCs. The figure also illustrates the superiority of our DiPT model, which generates high-fidelity samples with sharper contour definitions and smoother normals compared to those of other models. In terms of variability, measured by the 1-NNA and COV metrics, the proposed DiPT outperforms the other methods in the airplane category. However, for the other categories, no single method clearly outperforms the others. This is expected, as variability metrics depend solely on the diversity within each category and can fluctuate in the presence of noise. Nevertheless, DiPT outperforms in 4 out of 12 scores (3 on airplane, 1 on car). LION leads in 2, while others top only 1. As the overall best in quality, DiPT thus also offers the best variety-quality tradeoff among models.

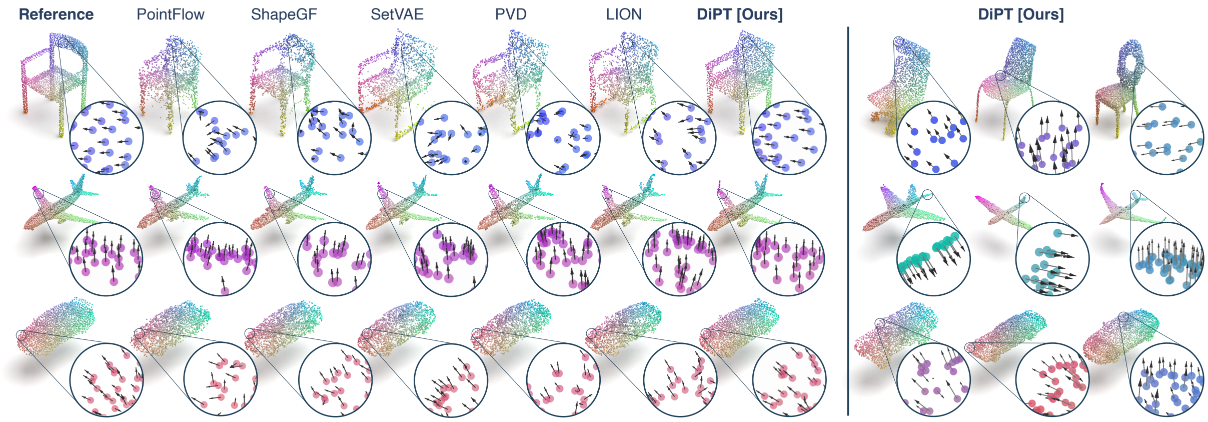

Figure 7. (**Left**) Qualitative comparison of the closest generated 3D point clouds to the reference based on DCD, across different models for the chair, airplane, and car categories. Additionally, point normals in zoomed regions are shown for samples smoothness comparison. (**Right**) Additional samples generated by DiPT.

5.4. Discussion

In this work, since the ShapeNet data share the same orientation, we introduced only barycenter alignment for simplicity. Nevertheless, as all traditional metrics, SNC is also sensitive to rotation and geometry, therefore a rigid registration method like ICP or CPD [44] might be required in more general scenarios, before computing point cloud distances, to handle rotation mismatches, e.g. on DiPT-S, SNC improves in mean 0.06% with ICP but runs $3.35\times$ slower. Moreover, Tab. 2 in the Supplementary presents the same comparison as in Tab. 1, but with inhomogeneous references. The results show that SNCs, along with MMD-DCD, are the most consistent metrics for preserving relative model rankings across different reference distributions, achieving the highest rank correlations. This supports SNC's reliability despite geometry mismatches, provided a consistent reference set is used. Furthermore, the proposed SNC is designed for single objects where the evaluation of surface smoothness is of interest. Consequently, it may struggle with irregular objects, such as trees, and may require tuning of the normal estimation techniques, e.g. by changing the neighbors region size for PCA or dynamically adapting them based on object complexity. SNC is analyzed in details under mismatched point densities and different normal estimation settings in Sec. 9 of the Supplementary. Additionally, for generating scenes, such as in LiDAR sequences [2], SNC can still be used, along with other metrics [56], by decomposing the scene into smaller objects, such as cars, pedestrians, and buildings, and evaluating the surface quality of each compared to a set of references.

6. Conclusions

We introduced new guidelines to ensure a more reliable assessment of 3D point cloud generative models by enhancing the fidelity of evaluation metrics in reflecting the true quality of generated samples, making them robust to shifts and more sensitive to defects such as noise. Additionally, we introduced the SNC metric to evaluate the surface quality of generated samples by comparing their estimated point normals with those of the references. We believe that the proposed SNC can help assess, and consequently improve, the quality of synthesized shapes by complementing MMD in cases where it struggles and particularly when surface regularity is of primary interest. When normals are less relevant, our work encourages future metrics to target other meaningful properties as needed. Furthermore, the proposed DiPT model combines innovations from point cloud processing and diffusion models, outperforming previous methods in generative quality, as shown on the ShapeNet dataset. Our framework strengthens evaluation methods and opens avenues for further research in 3D generation. Advancing these techniques could lead to more accurate, realistic, and consistent 3D models. A promising direction for future work is to adapt the proposed model and metrics to other fields, such as LiDAR scans or domain-specific datasets, while also dynamically adjusting metrics like SNC based on shape complexity, irregularities, and requirements, leading to more generalizable assessments of 3D generative models.

7. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945304-Cofund AI4theSciences hosted by PSL University. This work was granted access to the HPC/AI resources of IDRIS under the allocation 2022-AD011013902 made by GENCI.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning Representations and Generative Models for 3D Point Clouds. In *Proceedings of the 35th International Conference on Machine Learning*, pages 40–49. PMLR, 2018. ISSN: 2640-3498. 1, 2, 3, 4

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9296–9306, Seoul, Korea (South), 2019. IEEE. 8

[3] Yizhak Ben-Shabat, Michael Lindenbaum, and Anath Fischer. Nesti-Net: Normal Estimation for Unstructured 3D Point Clouds Using Convolutional Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10104–10112, Long Beach, CA, USA, 2019. IEEE. 4

[4] J. Berkmann and T. Caelli. Computation of surface geometry and segmentation using covariance techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(11):1114–1116, 1994. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 4, 6, 2

[5] Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snavely, and Bharath Hariharan. Learning Gradient Fields for Shape Generation. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III*, pages 364–381, Berlin, Heidelberg, 2020. Springer-Verlag. 3, 6, 7, 2

[6] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository, 2015. arXiv:1512.03012 [cs]. 2, 6, 8, 9

[7] Siheng Chen, Baoan Liu, Chen Feng, Carlos Vallespi-Gonzalez, and Carl Wellington. 3D Point Cloud Processing and Learning for Autonomous Driving: Impacting Map Creation, Localization, and Perception. *IEEE Signal Processing Magazine*, 38(1):68–86, 2021. Conference Name: IEEE Signal Processing Magazine. 1

[8] Tri Dao. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 5

[9] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022. 5

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. 1

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 6, 8, 10

[12] Haonan Duan, Peng Wang, Yayu Huang, Guangyun Xu, Wei Wei, and Xiaofei Shen. Robotics Dexterous Grasping: The Methods Based on Point Cloud and Deep Learning. *Frontiers in Neurorobotics*, 15, 2021. Publisher: Frontiers. 1

[13] Matheus Gadelha, Rui Wang, and Subhransu Maji. Multiresolution Tree Networks for 3D Point Cloud Processing. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, pages 105–122, Berlin, Heidelberg, 2018. Springer-Verlag. 1, 3

[14] Paul Guerrero, Yanir Kleiman, Maks Ovsjanikov, and Niloy J. Mitra. PCPNET: Learning Local Shape Properties from Raw Point Clouds. *Computer Graphics Forum*, 37(2): 75–85, 2018. arXiv:1710.04954 [cs]. 4

[15] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R. Martin, and Shi-Min Hu. PCT: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021. arXiv:2012.09688 [cs]. 1

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 2

[17] David Hilbert. *Dritter Band: Analysis · Grundlagen der Mathematik · Physik Verschiedenes*. Springer, Berlin, Heidelberg, 1935. 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 6840–6851, Red Hook, NY, USA, 2020. Curran Associates Inc. 1, 6

[19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video Diffusion Models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 1

[20] Hugues Hoppe, Tony DeRose, Tom Duchamp, John McDonald, and Werner Stuetzle. Surface reconstruction from unorganized points. *SIGGRAPH Comput. Graph.*, 26(2):71–78, 1992. 4, 6, 2

[21] Rongjie Huang, Max W. Y. Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 4157–4163, Vienna, Austria, 2022. International Joint Conferences on Artificial Intelligence Organization. 1

[22] ZhangJin Huang, Yuxin Wen, ZiHao Wang, Jinjuan Ren, and Kui Jia. Surface Reconstruction From Point Clouds: A Survey and a Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9727–9748, 2024. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 2

[23] Junzhong Ji, Runfeng Zhao, and Minglong Lei. Latent diffusion transformer for point cloud generation. *The Visual Computer*, 40(6):3903–3917, 2024. 1, 3, 6, 7, 2

[24] Jaehyeong Jo, Dongki Kim, and Sung Ju Hwang. Graph Generation with Diffusion Mixture. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. 1

[25] Hyeongju Kim, Hyeonseung Lee, Woo Hyun Kang, Joun Yeop Lee, and Nam Soo Kim. SoftFlow: Probabilistic Framework for Normalizing Flow on Manifolds. In *Advances in Neural Information Processing Systems*, pages 16388–16397. Curran Associates, Inc., 2020. 3, 4, 6, 7, 2

[26] Jinwoo Kim, Jaehoon Yoo, Juho Lee, and Seunghoon Hong. SetVAE: Learning Hierarchical Composition for Generative Modeling of Set-Structured Data. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15054–15063, Nashville, TN, USA, 2021. IEEE. 1, 3, 4, 6, 7, 2

[27] Klaas Klasing, Daniel Althoff, Dirk Wollherr, and Martin Buss. Comparison of surface normal estimation methods for range sensing applications. In *2009 IEEE International Conference on Robotics and Automation*, pages 3206–3211, 2009. ISSN: 1050-4729. 4, 2

[28] Roman Klokov, Edmond Boyer, and Jakob Verbeek. Discrete Point Flow Networks for Efficient Point Cloud Generation. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*, pages 694–710, Berlin, Heidelberg, 2020. Springer-Verlag. 1, 3

[29] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. Publisher: Institute of Mathematical Statistics. 1

[30] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified Transformer for 3D Point Cloud Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8490–8499, 2022. ISSN: 2575-7075. 1

[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 *Proceedings of the 36th International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019. ISSN: 2640-3498. 3

[32] Fangzhou Lin, Yun Yue, Songlin Hou, Xuechu Yu, Yajun Xu, Kazunori D Yamada, and Ziming Zhang. Hyperbolic Chamfer Distance for Point Cloud Completion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14549–14560, 2023. ISSN: 2380-7504. 4, 1

[33] Dongrui Liu, Chuanchaun Chen, Changqing Xu, Robert C. Qiu, and Lei Chu. Self-Supervised Point Cloud Registration With Deep Versatile Descriptors for Intelligent Driving. *IEEE Transactions on Intelligent Transportation Systems*, 24 (9):9767–9779, 2023. Conference Name: IEEE Transactions on Intelligent Transportation Systems. 4, 1

[34] Yifan Liu, Wuyang Li, Jie Liu, Hui Chen, and Yixuan Yuan. GRAB-Net: Graph-Based Boundary-Aware Network for Medical Point Cloud Segmentation. *IEEE Transactions on Medical Imaging*, 42(9):2776–2786, 2023. Conference Name: IEEE Transactions on Medical Imaging. 1

[35] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-Voxel CNN for Efficient 3D Deep Learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Montreal, QC, Canada, 2021. IEEE. 5, 6

[37] Zhijian Liu, Xinyu Yang, Haotian Tang, Shang Yang, and Song Han. FlatFormer: Flattened Window Attention for Efficient Point Cloud Transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1200–1211, Vancouver, BC, Canada, 2023. IEEE. 1, 5

[38] David Lopez-Paz and Maxime Oquab. Revisiting Classifier Two-Sample Tests. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 3, 1

[39] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 6

[40] Shitong Luo and Wei Hu. Diffusion Probabilistic Models for 3D Point Cloud Generation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2836–2844, Nashville, TN, USA, 2021. IEEE. 1, 3, 6, 7, 2

[41] Niloy J. Mitra and An Nguyen. Estimating surface normals in noisy point cloud data. In *Proceedings of the nineteenth annual symposium on Computational geometry*, pages 322–328, New York, NY, USA, 2003. Association for Computing Machinery. 4

[42] Shentong Mo, Enze Xie, Ruihang Chu, Lewei Yao, Lanqing Hong, Matthias Nießner, and Zhenguo Li. DiT-3D: exploring plain diffusion transformers for 3D shape generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 67960–67971, Red Hook, NY, USA, 2023. Curran Associates Inc. 1, 2, 3, 4, 5, 6, 7

[43] G. M. Morton. *A Computer Oriented Geodetic Data Base and a New Technique in File Sequencing*. International Business Machines Company, 1966. Google-Books-ID: 9FFdHAAACAAJ. 5

[44] Andriy Myronenko, Xubo Song, and Miguel Carreira-Perpiñán. Non-rigid point set registration: Coherent Point Drift. In *Advances in Neural Information Processing Systems*. MIT Press, 2006. 8

[45] Trung Nguyen, Quang-Hieu Pham, Tam Le, Tung Pham, Nhat Ho, and Binh-Son Hua. Point-set Distances for Learning Representations of 3D Point Clouds. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10458–10467, Montreal, QC, Canada, 2021. IEEE. 2

[46] Mark Pauly, Richard Keiser, Leif P. Kobbelt, and Markus

Gross. Shape modeling with point-sampled geometry. *ACM Trans. Graph.*, 22(3):641–650, 2003. 4

[47] William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2023. ISSN: 2380-7504. 1, 3, 5, 6, 8, 10

[48] Dmitry Petrov, Pradyumn Goyal, Vikas Thamizharasan, Vladimir Kim, Matheus Gadelha, Melinos Averkiou, Siddhartha Chaudhuri, and Evangelos Kalogerakis. GEM3D: GEnerative Medial Abstractions for 3D Shape Synthesis. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, New York, NY, USA, 2024. Association for Computing Machinery. 1

[49] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2, 3

[50] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022. arXiv:2204.06125 [cs]. 1

[51] Haoxi Ran, Jun Liu, and Chengjie Wang. Surface Representation for Point Clouds. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18920–18930, 2022. ISSN: 2575-7075. 2

[52] Lars Ruthotto and Eldad Haber. An introduction to deep generative modeling. *GAMM-Mitteilungen*, 44(2):e202100008, 2021. Publisher: John Wiley & Sons, Ltd. 1

[53] Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low-precision. *Advances in Neural Information Processing Systems*, 37: 68658–68685, 2025. 5

[54] Dongwook Shu, Sung Woo Park, and Junseok Kwon. 3D Point Cloud Generative Adversarial Network Based on Tree Structured Graph Convolutions. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3858–3867, Seoul, Korea (South), 2019. IEEE. 1, 2, 3

[55] Leslie N. Smith and Nicholay Topin. Super-convergence: very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, pages 369–386. SPIE, 2019. 6

[56] Larissa T. Triess, Christoph B. Rist, David Peter, and J. Marius Zöllner. A Realism Metric for Generated LiDAR Point Clouds. *International Journal of Computer Vision*, 130(12): 2962–2979, 2022. 2, 3, 8

[57] Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Learning Localized Generative Models for 3D Point Clouds via Graph Convolution. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 1, 3

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1

[59] Peng-Shuai Wang. OctFormer: Octree-based Transformers for 3D Point Clouds. *ACM Transactions on Graphics*, 42(4): 1–11, 2023. arXiv:2305.03045 [cs]. 1, 5

[60] Tao Wang, Wei Wen, Jingzhi Zhai, Kang Xu, and Haoming Luo. Serialized Point Mamba: A Serialized Point Cloud Mamba Segmentation Model, 2024. arXiv:2407.12319 [cs]. 1, 5

[61] Weijia Wang, Xuequan Lu, Di Shao, Xiao Liu, Richard Dazeley, Antonio Robles-Kelly, and Wei Pan. Weighted Point Cloud Normal Estimation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2015–2020, 2023. ISSN: 1945-788X. 4

[62] Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. Density-aware chamfer distance as a comprehensive metric for point cloud completion. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 29088–29100, Red Hook, NY, USA, 2021. Curran Associates Inc. 2, 4, 1

[63] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer V2: grouped vector attention and partition-based pooling. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 33330–33342, Red Hook, NY, USA, 2022. Curran Associates Inc. 1, 2, 5

[64] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point Transformer V3: Simpler, Faster, Stronger. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4840–4851, Seattle, WA, USA, 2024. IEEE. 1, 2, 5, 11

[65] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. DiffIR: Efficient Diffusion Model for Image Restoration. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13049–13059, 2023. ISSN: 2380-7504. 1

[66] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks, 2018. arXiv:1806.07755. 3, 1

[67] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. PointFlow: 3D Point Cloud Generation With Continuous Normalizing Flows. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4540–4549, Seoul, Korea (South), 2019. IEEE. 1, 2, 3, 4, 6, 7

[68] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. FoldingNet: Point Cloud Auto-Encoder via Deep Grid Deformation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 206–215, Salt Lake City, UT, 2018. IEEE. 1, 3

[69] Jianhui Yu, Chaoyi Zhang, Heng Wang, Dingxin Zhang, Yang Song, Tiange Xiang, Dongnan Liu, and Weidong Cai. 3D Medical Point Transformer: Introducing Convolution to Attention Networks for Medical Point Cloud Analysis, 2021. arXiv:2112.04863 [eess]. 1

[70] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-BERT: Pre-training 3D Point

Cloud Transformers with Masked Point Modeling. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19291–19300, New Orleans, LA, USA, 2022. IEEE. 1

[71] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. LION: latent point diffusion models for 3D shape generation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 10021–10039, Red Hook, NY, USA, 2022. Curran Associates Inc. 1, 3, 4, 6, 7, 2

[72] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point Transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16239–16248, 2021. ISSN: 2380-7504. 1, 2, 5

[73] Jun Zhou, Wei Jin, Mingjie Wang, Xiuping Liu, Zhiyang Li, and Zhaobin Liu. Fast and Accurate Normal Estimation for Point Clouds Via Patch Stitching. *Comput. Aided Des.*, 142 (C), 2022. 4

[74] Linqi Zhou, Yilun Du, and Jiajun Wu. 3D Shape Generation and Completion through Point-Voxel Diffusion. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5806–5815, Montreal, QC, Canada, 2021. IEEE. 1, 2, 3, 4, 6, 7