Evaluating Latent Generative Paradigms for High-Fidelity 3D Shape Completion from a Single Depth Image

Matthias Humt[1,2] Ulrich Hillenbrand[1] Rudolph Triebel[1,3]
[1]German Aerospace Center [2]TU Munich [3]Karlsruhe Institute of Technology
{matthias.humt,ulrich.hillenbrand,rudolph.triebel}@dlr.de

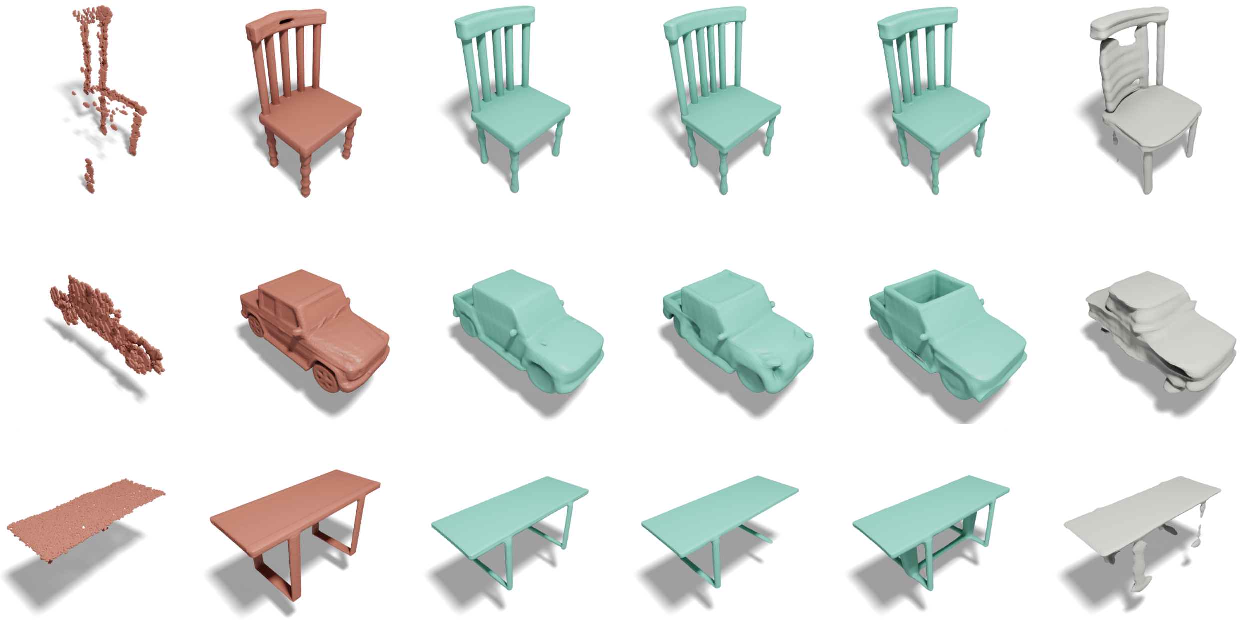

Figure 1. Predicting complete shapes from **partial, noisy inputs** (1) that closely resemble the **ground truth** (2) object remains challenging when the input is highly ambiguous. We explore models that fit generative priors to latent distributions, enabling multi-modal shape completion. The generative models produce **multiple plausible predictions** (3-5) covering the range of possibilities (in descending similarity to ground truth), with some completions surpassing the quality of the **single prediction** (6) from discriminative models.

Abstract

While generative models have seen significant adoption across a wide range of data modalities, including 3D data, a consensus on which model is best suited for which task has yet to be reached. Further, conditional information such as text and images to steer the generation process are frequently employed, whereas others, like partial 3D data, have not been thoroughly evaluated. In this work, we compare two of the most promising generative models–Denoising Diffusion Probabilistic Models and Autoregressive Causal Transformers–which we adapt for the tasks of generative shape modeling and completion. We conduct a thorough quantitative evaluation and comparison of both tasks, including a baseline discriminative model and an extensive ablation study. Our results show that (1) the diffusion model with continuous latents outperforms both the discriminative model and the autoregressive approach and delivers state-of-the-art performance on multi-modal shape completion from a single, noisy depth image under realistic conditions and (2) when compared on the same discrete latent space, the autoregressive model can match or exceed diffusion performance on these tasks.

1. Introduction

In the domain of 3D computer vision, generating complete object shapes from partial and often degraded observations is an enduring challenge, particularly for applications requiring high-fidelity, visually appealing object meshes, such as computer graphics, or accurate geometry for downstream tasks like robotics or augmented reality.

In this work, we focus on the task of single-view 3D shape completion, aiming to infer the complete 3D shape of an object from partial observations, such as a single depth image.

Many previous works have tried to tackle this problem using discriminative models [9, 23, 40, 55, 58, 67, 73], but the inherent ambiguity of the task forces these models to predict the average over all plausible completions [22, 54], often resulting in unrealistic, low-fidelity outcomes.

Meanwhile, generative models have shown impressive results across modalities like text [3, 14, 43, 44] and audio [74] (1D), 2D images [21, 25] and recently also 3D data [66, 72, 75, 76, 79]. The latter have also been conditioned on varying modalities like text or images [76, 79] and, in some cases, limited qualitative results on partial 3D data are presented [63, 66, 76]. Few works on generative 3D shape completion additionally provide limited quantitative evaluation [6, 52, 61, 72, 77], while none include the direct comparison to discriminative models.

The exact quantitative evaluation of generative models in general, and in the context of shape completion in particular, is still an active area of research, and a consensus on the best modeling paradigm and evaluation metrics has yet to be reached. This is notable in the variety of employed metrics and their exact definition and evaluation protocols. The situation gets aggravated by the fact that details and code for evaluation are often not provided, making it hard to reproduce and compare results.

To address this gap, we investigate two of the most promising generative models, Denoising Diffusion Probabilistic Models (DDPM) [21, 25] and Autoregressive (AR) Causal Transformers [59], on the tasks of generative shape modeling and completion. We conduct a thorough quantitative evaluation of both tasks, including a fair comparison between the two models through training on the exact same latent space and between the discriminative versus generative modeling paradigms. An extensive ablation study is also provided. All code, weights, and data used in this work will be made publicly available upon publication.

Our main findings are: (1) Diffusion models outperform autoregressive models on both generative shape modeling and completion, which we are able to clearly attribute to the more expressive latent space of Variational Auto-Encoders [28] (VAE) used by the diffusion models compared to their vector-quantized variants [57] (VQ-VAE) required for latent autoregressive training. Indeed, the advantage of diffusion vanishes, and the outcome is reversed when both models are trained on the VQ-VAE latent space. (2) Our best generative model outperforms the discriminative model in shape completion across all metrics by a large margin under correct evaluation.

We summarize the main contributions of this work as follows:

1. State-of-the-art (SOTA) multi-modal shape completion from a single, noisy depth image under realistic conditions.
2. Rigorous, *quantitative* evaluation of both generative shape modeling and completion.
3. Detailed, quantitative comparison of generative and discriminative models for shape completion.
4. Fair, quantitative comparison of DDPMs and AR Causal Transformers for shape modeling and completion.
5. A runtime-optimized reference implementation of the evaluation protocol, including a large number of commonly used metrics.

2. Related Work

Discriminative shape modeling. Early works, enabled by the advent of large 3D object datasets [4], predicted shapes using 3D convolutional networks on coarse voxel grids [13, 18, 49, 53, 58, 64, 67] and later expanded to point clouds [55, 69, 73] and triangle meshes [17, 60]. More recently, implicit function representations using signed-distance [40, 65] or binary occupancy [9, 22, 23, 36, 41] fields have gained traction due to their simple training objective and strong representation power.

Generative shape modeling. Learning to fit distributions to shapes has followed a similar trajectory, from voxel [11, 52, 63], point [1, 68, 72] and mesh [33] to implicit [6–8, 10, 12, 16, 34, 38, 51, 61, 66, 75–79] representations. These methods can be further demarcated along data [1, 6, 7, 11, 16, 33, 34, 52, 61, 63, 68, 77–79] or latent [8, 10, 12, 38, 51, 66, 72, 75, 76] space generative modeling and into diffusion [10, 12, 34, 51, 56, 76, 79] or autoregressive [38, 66, 75] training paradigms.

Single-view 3D reconstruction. While closely related to shape completion, 3D reconstruction involves the additional challenge of transferring information from 2D to 3D. Due to its relevance and despite its complexity, it has attracted great attention among both discriminative [7, 26, 49, 53, 60, 64, 65] and generative [11, 12, 16, 33, 38] methods.

Shape Completion. Obtaining the full 3D geometry from a partial, potentially degraded observation remains a significant challenge but has advanced significantly through both discriminative [13, 18, 22, 23, 55, 58, 67, 73] and generative [6, 8, 10, 12, 38, 52, 61, 63, 72, 77] modeling paradigms. Some additional works mention but do not focus on shape completion [1, 40, 41, 66, 76]. Most works that focus on shape completion provide some quantitative evaluation [8, 10, 12, 13, 18, 22, 23, 52, 55, 58, 67, 72, 73], but rely either exclusively on global dataset statistics [8, 10, 12, 38, 61] or instance-level reconstruction quality [13, 18, 22, 23, 52, 55, 58, 67, 72, 73]. A direct comparison between generative and discriminative models for shape completion is still missing.

The shape completion task is not clearly defined and is therefore used to mean different things in different works. Most works simply remove parts of the input using a cutting plane or volume. Some works render depth images [6, 13, 18, 22, 23, 52, 58, 67, 77] few of which additionally add some noise to the projected point cloud [22, 23]. Except for [22, 23, 49, 54], the vast majority of works train in an object-centered coordinate system [54] instead of in camera–i.e. *view-centered*–coordinates which significantly simplifies the task.

3. Method

Preliminaries. Given a 3D object shape represented by a point cloud $x = \{x_i \in \mathbb{R}^3\}_{i=1}^N \in \mathcal{X}$ we train a VAE f_θ to predict the binary occupancy probability for any point $p \in \mathbb{R}^3$ as $f_\theta : \mathbb{R}^3 \times \mathcal{X} \to [0, 1]$ which is equivalent to the *discriminative* training objective

$$\hat{y} = p_\theta(y = 1 \mid p, x) \tag{1}$$

where $y \in \{0, 1\}$ is the occupancy label for point p and $\hat{y} \in [0, 1]$ is the predicted occupancy probability. The VAE consists of an encoder E that maps the input point cloud to a latent code $z = E(x)$ and a decoder D that tries to map the latent code back to the input space, giving $\hat{x} = D(E(x))$.

Once the VAE is trained, we fit a generative prior G on its latent distribution $p(z)$ to increase its expressiveness. We can further condition G on signal c during training to control the generation process. We train both a diffusion [21] model,

$$p_\phi(z_{0:T} \mid c) = p(z_T) \prod_{t=1}^{T} p_\phi(z_{t-1}, c) \tag{2}$$

and an autoregressive [59] model,

$$p_\phi(z \mid c) = \prod_{i=1}^{L} p_\phi(z_{<i}, c) \tag{3}$$

, where T is the number of (de)noising steps and L the number of autoregressive steps.

Model architecture. We build on Zhang et al. [76] for the VAE and diffusion model architectures. As shown in Fig. 2, the VAE encoder ingests positional encoded, sampled surface points and cross-attends [59] (also sometimes referred to as *encoder-decoder attention*) to farthest-point-sampled (FPS) *queries* to encode the surface points into a fixed-length latent set. From this latent set, a diagonal Gaussian parameterization is predicted for the VAE while being quantized into fixed codebook entries in the VQ-VAE [57] case, as required for autoregressive training. The sampled (VAE) or quantized (VQ-VAE) latent code is then processed by multiple Transformer [59] encoder layers with layer norm, self-attention, and feed-forward components. Finally, the occupancy probability for p is predicted through cross-attention between positional encoded point coordinates and the latent code. We refer to Zhang et al. [76] for further details.

We make the following changes to the VAE architecture of Zhang et al. [76]: (1) We use the original NeRF [37] positional encoding for both the surface and occupancy points. (2) We add a layer-normalization and feed-forward component to the input encoding stage. (3) We use multi-headed attention [59] throughout the entire model. (4) We half the input dimension of all GeGLU [48] activations. These changes allow us to train a VAE of one-third the size of the original while achieving the same performance (Tab. 10).

Despite a large codebook as suggested in Rombach et al. [46] and various improvements to VQ-VAE training from the literature like K-means initialization [74] and compression of the codebook dimension [70] which indeed increase reconstruction quality, we are unable to match the performance of the continuous VAE (Tab. 1). We found codebook sampling [31] and regularization [70], expiring of stale codes [74] and Finite Scalar Quantization [35] as well as Lookup Free Quantization [71] to be ineffective (ablations can be found in the supplementary material).

For unconditional generative training, both the diffusion and autoregressive models share the same Transformer encoder design. In their conditional configuration, all layers are replaced by Transformer decoder blocks, which add a cross-attention component. The autoregressive model uses *causal* self-attention. As an alternative to conditioning via cross-attention, we can prepend the conditioning vector to the latent code (Tab. 12). The diffusion model uses adaptive layer normalization [42] for time-step conditioning.

Training. We train all models in mixed precision using the Adam [27] optimizer with a linear warmup, cosine annealing learning rate schedule peaking at 0.0001 and effective batch size of 256 on 4-8 NVIDIA A100 80GB GPUs for 800-2000 epochs. We use weight decay of 0.005, exponential moving average over weights and gradient clipping. We found the former to benefit diffusion model performance and the latter being crucial for stable (VQ-)VAE training. During the auto-encoding stage, we augment the inputs by adding Gaussian noise to the surface points and independently randomly scale all axes by up to 20%. Contrary to Zhang et al. [76], we do not use this type of augmentation during training of the latent generative model to prevent the generation of distorted shapes during inference.

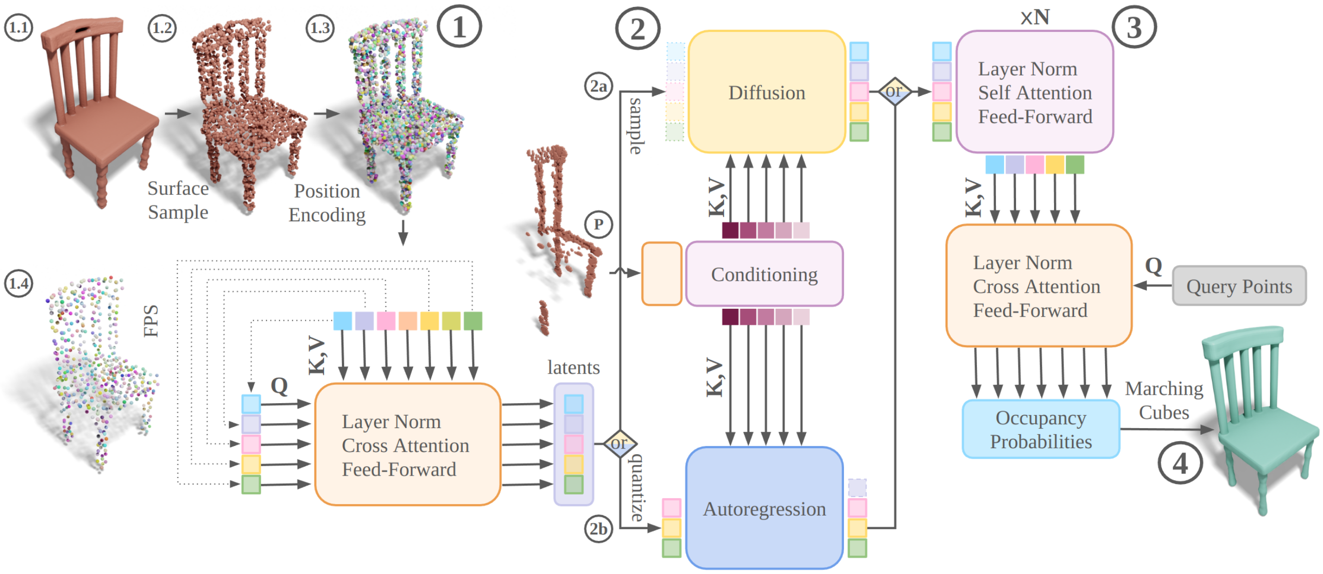

Figure 2. **Generative shape completion.** (1) Given an **input point cloud** (1.2) sampled from the **surface of an object** (1.1), we apply a positional encoding (1.3) and aggregate the entire point cloud into a farthest-point-sampled (FPS) set (1.4) as in Zhang et al. [76], which we additionally passed through a feed-forward network to form a latent code. (2) We then model these latents *either* as a diagonal, multivariate Gaussian (2a) *or* quantize them into a fixed-sized codebook (2b) forming our (VQ-)VAE encoder and train a diffusion *or* autoregressive model on top, respectively. For shape completion, we condition the generative model on the encoding of a **partial view** (P) using a pre-trained feature extractor, which shares the overall architecture of the VAE. (3) We then predict occupancy probabilities through cross-attention between query points and latents sampled from the latent generative model, processed by N Transformer encoder layers, forming the VAE decoder. (4) Optionally, a **mesh** can be extracted using the *Marching Cubes* algorithm. During inference, we discard the VAE encoder and sample latent codes *either* autoregressively *or* via denoising of samples drawn from a standard normal distribution.

4. Experiments

This section comprises four parts: We begin by discussing evaluation metrics, then validate our models' reconstruction and generative modeling performance. Next, we assess shape completion capabilities under increasing complexity and realism. Finally, we conduct ablation studies examining how various design choices affect overall model performance.

All experiments utilize the ShapeNet (v1) dataset [4], unless otherwise specified, with training data generated following the approach of Humt et al. [23].

4.1. Metrics

As alluded to in the introduction, many evaluation metrics for reconstruction and generative modeling have been proposed, and no consensus on their relative importance has been reached. We, therefore, evaluate our models across a wide range of metrics to provide a comprehensive view of their performance.

Instance-level. To evaluate the reconstruction quality, we rely on (volumetric) *Intersection-over-Union* (IoU) (if applicable) and bidirectional L1 *Chamfer Distance* (CD), scaled following Mescheder et al. [36]. We further make use of *F1-score* as well as *Precision* and *Recall*, also referred to as *accuracy* and *completeness* in Tatarchenko et al. [54]. IoU can only be evaluated for watertight meshes, but we opt to evaluate against the original meshes from ShapeNet to facilitate reproduction and comparison and for consistency with the generative modeling evaluation.

Set-level. All of these metrics measure *instance-level* performance, as opposed to the following metrics most commonly used for evaluating the generative quality, which measure *set-level* or *global* performance.

The earliest metrics for evaluating generative models are *Minimum-Matching-Distance* (MMD) and *Coverage* (COV) [1] which we retire in favor of *Leave-One-Out 1-Nearest-Neighbor Accuracy* (1-NNA), proposed to alleviate the shortcomings of MMD and COV [68]. Some works also use *Edge Count Difference* (ECD) [24] as well as *Total Mutual Distance* (TMD) and *Unidirectional Hausdorff Distance* (UHD) [61] which we found to be less informative and refer to the appendix.

More recent additions are Fréchet [50] and Kernel [76] Pointcloud Distance (KPD, FPD), which compute the Fréchet and Kernel distance between point features extracted from the generated and ground truth surface points. These are highly informative but rely on a pre-trained feature extractor, which each work redefines and retrains, making comparison impossible. We reuse the features of our VAE trained on the reconstruction task, which we find to be more informative than the commonly used features from models trained on point cloud classification. Furthermore, this not only frees us from training yet another model but also provides a close to normally distributed feature space, an implicit assumption of the Fréchet distance. Following prior work [1, 68, 76, 78], these set-level metrics are computed on the test split.

Finally, we also evaluate the perceptual quality of the generated shapes using the (shading-image-based) *Fréchet Inception Distance* (FID) [19] and *Kernel Inception Distance* (KID) [2] which measure the distance between the feature distributions of images of generated and real objects, rendered from uniformly sampled viewpoints. Here, we additionally employ CLIP [45] features from a Vision Transformer [15] as proposed in Kynkäänniemi et al. [30] and shown to align better with human perception than Inception features (results for this metric can be found in the appendix).

Sajjadi et al. [47] show how FID can be decomposed into *Precision* and *Recall* of which we use the improved version by Kynkäänniemi et al. [29] relying on k-NN instead of k-means clustering. We propose to employ the same decomposition for KPD. Naeem et al. [39] claim an even better decomposition into *Density* and *Coverage* exists, but due to lack of adoption, these are provided in the appendix.

Again, following prior convention, we evaluate FID and its decompositions as well as KID on the train split.

4.2. Results

Reconstruction. The reconstruction quality of the VAE determines the upper bound for the latent generative model performance. As shown in Tab. 1, first row, our VAE achieves comparable performance to the current SOTA [76] on watertight meshes, but due to differences in training data, no direct comparison can be made. The VQ-VAE (second row) falls short of the VAE but performs reasonably well for a large and diverse dataset such as ShapeNet. We also include the performance on the original ShapeNet meshes in the table's lower half (3rd and 4th row) for reference and future comparison. Interestingly, while the models are well calibrated on the watertight meshes, achieving similar precision and recall, on the original meshes, recall is lacking behind significantly, which we attribute to loss of (interior) detail during the watertightening process.

Generative Modeling. To validate the performance of the latent generative training, we compare our class-conditional LDM against two recent SOTA baselines, LAS-Diffusion [79] (LAS-Dif.) and 3DShape2VecSet [76] (3DS2VS) on the same subset of classes. We use the provided model checkpoints, as retraining these models incurs a significant computational overhead. The results are shown in Tab. 2.

	Chamfer ↓	F1 ↑	Precision ↑	Recall ↑
VAE	**0.032**	**98.33**	**98.62**	**98.13**
VQ-VAE	0.069	89.33	89.34	89.83
VAE	**0.091**	**77.19**	**82.60**	**74.53**
VQ-VAE	0.116	70.53	74.47	69.08

Table 1. Reconstruction quality; class average. Upper half shows performance on watertight meshes, lower half on original meshes.

Due to differences in training data and procedures, we are able to outperform the superior 3DShape2VecSet baseline across all metrics while sharing the overall model architecture. All models show much higher precision than recall, indicating paths toward future improvement.

		1-NNA↓	FPD↓	KPD↓	Prec.↑	Rec.↑
Chair	LAS-Diff.	59.08	99.17	9.31	**96.90**	63.37
	3DS2VS	58.94	94.01	7.16	85.67	**77.10**
	Ours	**58.49**	**89.59**	**6.97**	95.57	60.71
Plane	LAS-Diff.	82.67	257.66	34.79	75.00	11.39
	3DS2VS	69.68	165.01	22.35	68.32	**34.16**
	Ours	**69.06**	**139.68**	**17.05**	**84.16**	30.94
Car	LAS-Diff.	86.32	99.02	**16.27**	**62.62**	**55.67**
	3DS2VS	91.05	170.99	27.71	60.88	39.65
	Ours	**82.18**	**84.74**	16.62	59.41	48.20
Table	LAS-Diff.	55.35	158.87	19.95	94.59	**72.12**
	3DS2VS	56.76	148.10	15.08	92.47	71.53
	Ours	**53.71**	**128.35**	**9.53**	**96.71**	67.65
Rifle	LAS-Diff.	77.43	693.84	115.63	**96.62**	34.60
	3DS2VS	**66.03**	418.01	57.75	91.98	**50.63**
	Ours	70.46	**347.78**	**52.89**	95.78	30.80
Mean	LAS-Diff.	72.17	261.71	39.19	85.15	47.43
	3DS2VS	68.49	199.22	26.01	79.86	**54.62**
	Ours	**66.78**	**158.03**	**20.61**	**86.33**	47.66

Table 2. Comparison of *class-conditional* generative models.

We then proceed to compare our unconditional LDM and AR models. According to Tab. 3, the AR model is outperformed by the LDM, which we attribute to the superior reconstruction quality of the VAE, as established in the previous section.

To test this hypothesis, we train both a class-conditional LDM and AR model on the *same* discrete VQ-VAE latent space. This setup uses an embedding of the class labels as conditioning information c. As evident from Tab. 4, the AR model is able to outperform the LDM in this setting. For reference, we also include the results of the class-conditional

	Diffusion (VAE)	AR (VQ-VAE)
FID ↓	**32.62**	35.76
KID $\times 10^3$ ↓	**13.00**	13.17
Precision ↑	**50.27**	50.26
Recall ↑	**48.08**	42.08

Table 3. Comparison of diffusion and autoregressive *unconditional* generative shape modeling on continuous (VAE) and discrete (VQ-VAE) latents.

LDM trained on the continuous VAE latents.

	VQ-VAE		VAE
	Diffusion	Autoregressive	Diffusion
FID ↓	42.98	**33.58**	30.02
KID$\times 10^3$ ↓	18.03	**12.05**	11.16
Precision ↑	38.59	**51.61**	53.94
Recall ↑	37.98	**43.51**	46.88
1-NNA ↓	67.93	**66.54**	65.01
FPD ↓	80.38	**77.51**	73.03
KPD ↓	**4.30**	5.14	4.43
Precision ↑	**92.17**	91.62	91.51
Recall ↑	54.19	**60.87**	60.00

Table 4. Comparison of diffusion and autoregressive *class-conditional* generative shape modeling on the same latent space.

Shape Completion. We now come to the main results of this work, comparing the discriminative and generative approach on the shape completion task. Our discriminative model architecture is identical to the VAE used for shape auto-encoding, except for the variational part and the fact that the input is now a partial view of the object. The encoder of the trained discriminative model is repurposed as feature extractor to the latent generative model to provide highly informative conditioning information. We tried training a dedicated feature extractor on the classification task but found this to result in worse performance (Tab. 12).

The simplest task we consider is shape completion from a rendered depth image in object-centric coordinates. Due to self-occlusions, this is still significantly more challenging than random removal of parts of the object, as is common practice. Contrary to unconditional and class-conditional generation, shape completion is only evaluated on the test split, as we are interested in generalization to novel instances instead of faithfully capturing the underlying data distribution. In this simplified setup, the discriminative model is able to slightly outperform the generative model on both set-level (upper part) and instance-level (lower part) metrics (Tab. 5). Following Tatarchenko et al. [54], this is to be expected, as the discriminative model can bypass the complex shape completion task and learn the more straightforward retrieval task instead.

	Discriminative	**Generative**
1-NNA ↓	**30.451**	30.806
FPD ↓	**71.126**	71.782
KPD ↓	**5.622**	6.115
Precision ↑	93.928	**94.851**
Recall ↑	**77.184**	76.385
Chamfer ↓	**0.118**	0.122
F1 ↑	**70.681**	69.098
Precision ↑	**74.795**	72.485
Recall ↑	**69.226**	68.149

Table 5. **Generative** vs. **discriminative** shape completion from a single depth image in object-centric coordinates.

Moving on to shape completion in camera coordinates (Tab. 6), the results from the previous tasks are reversed for the set-level metrics. Now, the generative model appears slightly better than the discriminative model, which can no longer entirely rely on the retrieval shortcut. Still, the discriminative model has a slight edge in instance-level performance.

To understand why, recall that discriminative models are forced to predict the best *average* result when faced with ambiguous inputs, whereas generative models when only queried once, can and will predict a single, plausible result, which is not necessarily as close to the ground truth as the average. To test this hypothesis, we move on to the final, most complex shape completion task: the completion of noisy depth images (in camera coordinates) as captured by widely available RGB-D sensors like the *Microsoft Kinect*.

	Discriminative	**Generative**
1-NNA ↓	31.481	**31.099**
FPD ↓	74.817	**70.269**
KPD ↓	6.540	**5.893**
Precision ↑	92.134	**92.276**
Recall ↑	77.557	**77.610**
Density ↑	0.845	**0.920**
Coverage ↑	0.782	**0.795**
Chamfer ↓	**0.125**	0.128
F1 ↑	**69.070**	68.081
Precision ↑	**74.174**	72.669
Recall ↑	**66.475**	65.912

Table 6. **Generative** vs. **discriminative** shape completion from a single depth image in camera coordinates

Instead of generating a single completion, which goes against a generative model's actual benefit and strength, we now instead produce 10 completions per input and pick the one with the highest F1-score. We argue that this is the correct way to assess the generative model's upper-bound performance, as we are interested in its ability to produce not just plausible but also more accurate results than a discriminative model when faced with ambiguous inputs. Tab. 7 confirms our hypothesis in which the generative model (G) now consistently outperforms the discriminative model (D) by a large margin across all metrics. This effect can also be observed in Fig. 1 and 3 where the generative model produces always plausible and, in the best case, also more accurate completions than the discriminative model.

In a final experiment, we investigate the model performance under domain shift and evaluate both the discriminative and generative model on the *Automatica/YCB* dataset by Humt et al. [23]. The results in Tab. 8 show both the superior performance of our discriminative model over the equivalent model of [23] which are still further improved upon by the generative model, as illustrated in Fig. 3. While Chamfer distance remains unchanged for the discriminative model and slightly increases for the generative model, this metric is strongly effected by outliers and poor at distinguishing visual quality [1, 32, 62]. Qualitative results on real Kinect depth data can be found in the appendix.

Figure 3. Examples from the *Automatica/YCB* dataset. Left to right: **input**, **ground truth**, **generative** (best), **discriminative**.

Ablations. To justify and inform our design choices, we perform an extensive ablation study on model size (Tab. 10), number of diffusion steps (Tab. 11), type of conditioning information (Tab. 12), and the conditioning approach (Tab. 13). We also provide an ablation on the number of completions accompanying Tab. 7.

A single completion achieves results comparable to the discriminative model, while as few as two completions already outperform it. Including the results for ten completions from Tab. 7, there is some indication of diminishing returns for larger numbers.

To obtain the small models with approximately one-third of the parameters of the large variants, we simply halve the number of layers and the input dimension of all GeGLU activations. We find that the size of the model has a strong influence on latent generative modeling but not on autoencoding (Tab. 10).

Recent diffusion models [25] require only a fraction of the number of denoising steps during inference as the original DDPMs [21]. We follow Zhang et al. [76] and use as little as 18 steps during inference. Nonetheless, we ablate this choice and find that doubling the number has a discernible impact while further increases show diminishing returns (Tab. 11).

We also investigate the impact of different feature types used for conditioning the generative models on the shape completion task. We find that the features from a model trained on classification are worse than those from a model trained for auto-encoding and, contrary to findings in Chen et al. [5], the features of the final layer are superior to those from the middle of the model for this task. Fine-tuning the feature extractor alongside the training of the generative model provides further improvement (Tab. 12).

Finally, while not competitive against the diffusion models on the shape completion from Kinect depth task, the beginning-of-sequence conditioning where we prepend the conditioning features to the latent code of the VQ-VAE consistently outperforms conditioning via cross-attention (Tab. 13). This has the advantage that no additional cross-attention components must be added to the model, but it doubles the sequence length.

5. Conclusion

This work highlights the potential of generative modeling as an effective approach for high-fidelity 3D shape completion from single-view depth images. Through rigorous quantitative comparison to a discriminative method, we establish the advantage of generative models in effectively handling partial, noisy, and ambiguous input data for shape completion under realistic conditions, both regarding coverage of plausible alternatives and also accuracy in relation to a single ground truth complete shape. While this particular strength of generative models can be partially explained by their ba-

	1-NNA↓		FPD↓		KPD↓		CD↓		F1↑		Prec.↑		Rec.↑	
	D	**G**	**D**	**G**	**D**	**G**	**D**	**G**	**D**	**G**	**D**	**G**	**D**	**G**
Chair	38.26	**33.68**	300	**128**	48.55	**16.27**	0.396	**0.327**	43.27	**52.47**	47.12	**56.66**	41.09	**50.50**
Plane	60.27	**50.99**	426	**214**	62.18	**28.97**	0.312	**0.290**	47.09	**55.70**	50.70	**62.50**	45.32	**51.86**
Car	88.18	**74.70**	172	**114**	33.88	**17.85**	0.283	**0.260**	38.12	**47.64**	50.44	**58.97**	31.19	**40.92**
Table	38.94	**35.94**	274	**134**	31.13	**13.57**	0.447	**0.264**	48.04	**57.15**	49.98	**60.37**	48.28	**56.11**
Rifle	57.38	**54.22**	572	**480**	85.24	**69.68**	0.551	**0.542**	37.39	**46.00**	41.22	**55.53**	35.84	**41.52**
Mean	56.61	**49.91**	349	**214**	52.20	**29.27**	0.398	**0.337**	42.78	**51.79**	47.89	**58.81**	40.35	**48.18**
All	53.60	**48.53**	204	**103**	24.35	**9.26**	/	/	/	/	/	/	/	/

Table 7. **Generative** (**G**) vs. **discriminative** (**D**) shape completion from a single Kinect depth image. Instance-level metrics (CD, F1, Prec., Rec.) are 'best-of-10' for the generative model, which is already competitive with the discriminative model at N=1 (See Tab. 9).

	CD ↓	F1 ↑	Prec. ↑	Rec. ↑
Kinect [23]	0.305	43.37	44.69	42.85
Discriminative	**0.297**	45.99	47.02	45.83
Generative	0.346	**52.92**	**54.23**	**52.43**

Table 8. **Generative** vs. **discriminative** shape completion on the *Automatica/YCB* dataset.

	$N = 1$	$N = 2$	$N = 3$	$N = 5$
Chamfer ↓	0.38	0.36	0.35	**0.35**
F1 ↑	41.04	45.12	47.03	**49.37**
Precision ↑	46.50	51.08	53.26	**55.90**
Recall ↑	38.75	42.42	44.10	**46.14**

Table 9. Ablation on the number of generative completions N.

	VAE			Diffusion	
	Small	Large		Small	Large
CD↓	**0.09**	**0.09**	FID↓	39.46	**32.62**
F1↑	**77.19**	76.52	KID↓	17.48	**13.00**
Prec.↑	82.60	**83.36**	Prec.↑	41.72	**50.27**
Rec.↑	**74.53**	72.60	Rec.↑	**48.64**	48.08

Table 10. Ablation on model size.

	$T = 18$	$T = 35$	$T = 50$	$T = 100$
FID ↓	34.10	31.61	31.14	**30.93**
KID $\times 10^3$ ↓	13.63	12.31	12.05	**11.89**
Precision ↑	47.57	51.64	52.23	**52.70**
Recall ↑	46.49	**46.71**	46.64	46.53

Table 11. Ablation on the number of diffusion steps T.

	Class.	Recon.	Middle	Final	Final FT
FPD↓	84.5	**70.3**	129.1	113.0	**103.4**
KPD↓	7.4	**5.9**	12.2	10.4	**9.3**
Prec.↑	91.2	**92.3**	87.4	88.4	**88.5**
Rec.↑	73.1	**77.6**	60.9	68.1	**70.7**

Table 12. Ablation on conditioning type: classification (class.) vs. reconstruction (recon.) and middle vs. final layer as well as final, fine-tuning (FT) features.

	BOS	Cross-Attn
1-NNA ↓	**55.26**	55.75
FPD ↓	**167.57**	181.96
KPD ↓	**17.57**	19.42
Precision ↑	84.62	**85.74**
Recall ↑	**43.66**	39.68
Chamfer ↓	**0.43**	0.43
F1 ↑	**36.47**	34.61
Precision ↑	**39.18**	36.64
Recall ↑	**35.62**	34.32

Table 13. Ablation on Beginning-of-sequence (BOS) vs. cross-attention (Cross-Attn) conditioning during autoregressive training.

sic design, our empirical analysis uncovers details of their performance characteristics and also highlights key differences between latent diffusion-based and autoregressive approaches.

Limitations include the need to generate, and automatically select from, multiple completions to achieve optimal performance and a focus on specific model architectures, which may limit generalizability.

Future work will explore possible improvements in generative conditioning techniques such as Classifier-Free Guidance [20] and in quantized feature extraction from Residual VQ-VAEs [74] to unlock the full potential of autoregressive models in this domain.

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] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[3] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[4] 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. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[5] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020.

[6] Xuelin Chen, Baoquan Chen, and Niloy J Mitra. Unpaired point cloud completion on real scans using adversarial training. *arXiv preprint arXiv:1904.00069*, 2019.

[7] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5932–5941, 2018.

[8] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023.

[9] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6970–6981, 2020.

[10] Gene Chou, Yuval Bahat, and Felix Heide. Diffusion-sdf: Conditional generative modeling of signed distance functions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2262–2272, 2023.

[11] 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 *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016.

[12] Ruikai Cui, Weizhe Liu, Weixuan Sun, Senbo Wang, Taizhang Shang, Yang Li, Xibin Song, Han Yan, Zhennan Wu, Shenzhou Chen, et al. Neusdfusion: A spatial-aware generative model for 3d shape completion, reconstruction, and generation. *arXiv preprint arXiv:2403.18241*, 2024.

[13] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5868–5877, 2017.

[14] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022.

[17] 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.

[18] Xiaoguang Han, Zhen Li, Haibin Huang, Evangelos Kalogerakis, and Yizhou Yu. High-resolution shape completion using deep neural networks for global structure and local geometry inference. In *Proceedings of the IEEE international conference on computer vision*, pages 85–93, 2017.

[19] 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. *Advances in neural information processing systems*, 30, 2017.

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[22] Matthias Humt, Dominik Winkelbauer, and Ulrich Hillenbrand. Shape completion with prediction of uncertain regions. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1215–1221. IEEE, 2023.

[23] Matthias Humt, Dominik Winkelbauer, Ulrich Hillenbrand, and Berthold Bäuml. Combining shape completion and grasp prediction for fast and versatile grasping with a multifingered hand. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pages 1–8. IEEE, 2023.

[24] Moritz Ibing, Isaak Lim, and Leif Kobbelt. 3d shape generation with grid-based implicit functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13559–13568, 2021.

[25] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

[26] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018.

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[28] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[29] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Neural Information Processing Systems*, 2019.

[30] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022.

[31] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022.

[32] Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu. Morphing and sampling network for dense point cloud completion. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11596–11603, 2020.

[33] Zhen Liu, Yao Feng, Michael J Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. *arXiv preprint arXiv:2303.08133*, 2023.

[34] 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.

[35] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023.

[36] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.

[37] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[38] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 306–315, 2022.

[39] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning (ICML)*, 2020.

[40] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

[41] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020.

[42] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[43] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. OpenAI Technical Report, 2018.

[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[45] 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.

[46] 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.

[47] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lučić, Olivier Bousquet, and Sylvain Gelly. Assessing Generative Models via Precision and Recall. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[48] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[49] Daeyun Shin, Charless C Fowlkes, and Derek Hoiem. Pixels, voxels, and views: A study of shape representations for single view 3d object shape prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3069, 2018.

[50] 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.

[51] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20875–20886, 2023.

[52] David Stutz and Andreas Geiger. Learning 3d shape completion under weak supervision. *International Journal of Computer Vision (IJCV), 2018*, 2018.

[53] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE international conference on computer vision*, pages 2088–2096, 2017.

[54] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3405–3414, 2019.

[55] Lyne P Tchapmi, Vineet Kosaraju, Hamid Rezatofighi, Ian Reid, and Silvio Savarese. Topnet: Structural point cloud de-

coder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 383–392, 2019.

[56] Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022.

[57] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[58] Jacob Varley, Chad DeChant, Adam Richardson, Joaquín Ruales, and Peter Allen. Shape completion enabled robotic grasping. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 2442–2447. IEEE, 2017.

[59] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[60] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018.

[61] Rundi Wu, Xuelin Chen, Yixin Zhuang, and Baoquan Chen. Multimodal shape completion via conditional generative adversarial networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 281–296. Springer, 2020.

[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. *arXiv preprint arXiv:2111.12702*, 2021.

[63] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[64] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2690–2698, 2019.

[65] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems*, 32, 2019.

[66] Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6239–6249, 2022.

[67] Bo Yang, Stefano Rosa, Andrew Markham, Niki Trigoni, and Hongkai Wen. Dense 3d object reconstruction from a single depth view. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2820–2834, 2018.

[68] 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.

[69] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018.

[70] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

[71] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

[72] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12498–12507, 2021.

[73] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 international conference on 3D vision (3DV)*, pages 728–737. IEEE, 2018.

[74] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

[75] Biao Zhang, Matthias Nießner, and Peter Wonka. 3dilg: Irregular latent grids for 3d generative modeling. *Advances in Neural Information Processing Systems*, 35:21871–21885, 2022.

[76] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023.

[77] Junzhe Zhang, Xinyi Chen, Zhongang Cai, Liang Pan, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, Bo Dai, and Chen Change Loy. Unsupervised 3d shape completion through gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1768–1777, 2021.

[78] Xinyang Zheng, Yang Liu, Pengshuai Wang, and Xin Tong. Sdf-stylegan: Implicit sdf-based stylegan for 3d shape generation. In *Computer Graphics Forum*, pages 52–63. Wiley Online Library, 2022.

[79] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *ACM Transactions on Graphics (ToG)*, 42(4):1–13, 2023.