

# Neural Wavelet-domain Diffusion for 3D Shape Generation

Ka-Hei Hui  
The Chinese University of Hong Kong  
HK SAR, China  
khhui@cse.cuhk.edu.hk

Ruihui Li  
Hunan University  
China  
liruihui@hnu.edu.cn

Jingyu Hu, Chi-Wing Fu  
The Chinese University of Hong Kong  
HK SAR, China  
{jyhu,cwfu}@cse.cuhk.edu.hk



Figure 1: Our method is able to generate diverse shapes with complex structures and topology, fine details, and clean surfaces.

## ABSTRACT

This paper presents a new approach for 3D shape generation, enabling direct generative modeling on a continuous implicit representation in wavelet domain. Specifically, we propose a *compact wavelet representation* with a pair of coarse and detail coefficient volumes to implicitly represent 3D shapes via truncated signed distance functions and multi-scale biorthogonal wavelets, and formulate a pair of neural networks: a *generator* based on the diffusion model to produce diverse shapes in the form of coarse coefficient volumes; and a *detail predictor* to further produce compatible detail coefficient volumes for enriching the generated shapes with fine structures and details. Both quantitative and qualitative experimental results manifest the superiority of our approach in generating diverse and high-quality shapes with complex topology and structures, clean surfaces, and fine details, exceeding the 3D generation capabilities of the state-of-the-art models.

## CCS CONCEPTS

• Computing methodologies → Shape analysis; Neural networks; Mesh models.

## KEYWORDS

3D shape generation, diffusion model, wavelet representation

## 1 INTRODUCTION

Generative modeling of 3D shapes enables rapid creation of 3D contents, enriching extensive applications across graphics, vision, and VR/AR. With the emerging large-scale 3D datasets [Chang et al. 2015], data-driven shape generation has gained increasing attention from the research community recently. In general, a good 3D generative model should be able to produce diverse, realistic, and novel shapes, not necessarily the same as the existing ones.

Existing shape generation models are developed mainly for voxels [Girdhar et al. 2016; Yang et al. 2018; Zhu et al. 2017], point

clouds [Achlioptas et al. 2018; Fan et al. 2017; Jiang et al. 2018], and meshes [Groueix et al. 2018; Smith et al. 2019; Tang et al. 2019; Wang et al. 2018]. Typically, these representations cannot handle high resolutions or irregular topology, thus unlikely producing high-fidelity results. In contrast, implicit functions [Chen and Zhang 2019; Mescheder et al. 2019; Park et al. 2019] show improved performance in surface reconstructions. By representing a 3D shape as a level set of discrete volume or a continuous field, we can flexibly extract a mesh object of arbitrary topology at desired resolution.

Existing generative models such as GANs and normalizing flows have shown great success in generating point clouds and voxels. Yet, they cannot effectively generate implicit functions. To represent a surface in 3D, a large number of point samples are required, even though many nearby samples are redundant. Taking the occupancy field for instance, only regions near the surface have varying data values, yet we need huge efforts to encode samples in constant and smoothly-varying regions. Such representation non-compactness and redundancy demands a huge computational cost and hinders the efficiency of direct generative learning on implicit surfaces.

To address these challenges, some methods attempt to sample in a pre-trained latent space built on the reconstruction task [Chen and Zhang 2019; Mescheder et al. 2019] or convert the generated implicit functions into point clouds or voxels for adversarial learning [Kleineberg et al. 2020; Luo et al. 2021]. However, these regularizations can only be indirectly applied to the generated implicit functions, so they are not able to ensure the generation of realistic objects. Hence, the visual quality of the generated shapes often shows a significant gap, as compared with the 3D reconstruction results, and the diversity of their generated shapes is also quite limited.

This work introduces a new approach for 3D shape generation, enabling direct generative modeling on a continuous implicit representation in the wavelet frequency domain. Overall, we have three key contributions: (i) a compact wavelet representation (*i.e.*, a pair of coarse and detail coefficient volumes) based on biorthogonal

wavelets and truncated signed distance field to implicitly encode 3D shapes, facilitating effective learning of 3D shape distribution for shape generation; (ii) a generator network formulated based on the diffusion probabilistic model [Sohl-Dickstein et al. 2015] to produce coarse coefficient volumes from random noise samples, promoting the generation of diverse and novel shapes; and (iii) a detail predictor network, formulated to produce compatible detail coefficients to enhance the fine details in the generated shapes.

With the two trained networks, we can start from random noise volumes and flexibly generate diverse and realistic shapes that are not necessarily the same as the training shapes. Both quantitative and qualitative experimental results manifest the 3D generation capabilities of our method, showing its superiority over the state-of-the-art approaches. As Figure 1 shows, our generated shapes exhibit diverse topology, clean surfaces, sharp boundaries, and fine details, without obvious artifacts. Fine details such as curved/thin beams, small pulley, and complex cabinets are very challenging for the existing 3D generation approaches to synthesize.

## 2 RELATED WORK

*3D reconstruction via implicit function.* Recently, many methods leverage the flexibility of implicit surface for 3D reconstructions from voxels [Chen and Zhang 2019; Mescheder et al. 2019], complete/partial point clouds [Liu et al. 2021; Park et al. 2019; Yan et al. 2022], and RGB images [Li and Zhang 2021; Tang et al. 2021; Xu et al. 2019, 2020]. On the other hand, besides ground-truth field values, various supervisions have been explored to train the generation of implicit surfaces, *e.g.*, multi-view images [Liu et al. 2019; Niemeyer et al. 2020] and unoriented point clouds [Atzmon and Lipman 2020; Gropp et al. 2020; Zhao et al. 2021]. Yet, the task of 3D reconstruction focuses mainly on synthesizing a high-quality 3D shape that best matches the input. So, it is fundamentally very different from the task of 3D shape generation, which aims to learn the shape distribution of a given set of shapes for generating diverse, high-quality, and possibly novel shapes accordingly.

*3D shape generation via implicit function.* Unlike 3D reconstruction, the 3D shape generation task has no fixed ground truth to supervise the generation of each shape sample. Exploring efficient guidance for implicit surface generation is still an open problem. Some works attempt to use the reconstruction task to first learn a latent embedding [Chen and Zhang 2019; Hao et al. 2020; Ibing et al. 2021; Mescheder et al. 2019] then generate new shapes by decoding codes sampled from the learned latent space. Recently, [Hertz et al. 2022] learn a latent space with a Gaussian-mixture-based autoencoder for shape generation and manipulation. Though these approaches ensure a simple training process, the generated shapes have limited diversity restricted by the pre-trained shape space. Some other works attempt to convert implicit surfaces to some other representations, *e.g.*, voxels [Kleineberg et al. 2020; Zheng et al. 2022], point cloud [Kleineberg et al. 2020], and mesh [Luo et al. 2021], for applying adversarial training. Yet, the conversion inevitably leads to information loss in the generated implicit surfaces, thus reducing the training efficiency and generation quality.

In this work, we propose to adopt a compact wavelet representation for modeling the implicit surface and learn to synthesize it with a diffusion model. By this means, we can effectively learn to

generate the implicit representation without a pre-trained latent space and a representation conversion. The results also show that our new approach is capable of producing diversified shapes of high visual quality, exceeding the state-of-the-art methods.

*Other representations for 3D shape generation.* [Smith and Meger 2017; Wu et al. 2016] explore voxels, a natural grid-based extension of 2D image. Yet, the methods learn mainly coarse structures and fail to produce fine details due to memory restriction. Some other methods explore point clouds via GAN [Gal et al. 2020; Hui et al. 2020; Li et al. 2021], flow-based models [Cai et al. 2020; Kim et al. 2020], and diffusion models [Zhou et al. 2021]. Due to the discrete nature of point clouds, 3D meshes reconstructed from them often contain artifacts. This work focuses on implicit surface generation, aiming at generating high-quality and diverse meshes with fine details and overcoming the limitations of the existing representations.

*Multi-scale neural implicit representation.* This work also relates to multi-scale representations, so we discuss some 3D deep learning works in this area. [Chen et al. 2021; Chibane et al. 2020; Liu et al. 2020; Martel et al. 2021; Takikawa et al. 2021] predict multi-scale latent codes in an adaptive octree to improve the reconstruction quality and inference efficiency. [Fathony et al. 2020] propose a band-limited network to obtain a multi-scale representation by restricting the frequency magnitude of the basis functions. Recently, [Saragadam et al. 2022] adopt the Laplacian pyramid to extract multi-scale coefficients for multiple neural networks. Unlike our work, this work overfits each input object with an individual representation for efficient storage and rendering. In contrast to our work on shape generation, the above methods focus on improving 3D reconstruction performance by separately handling features at different levels. In our work, we adopt a multi-scale implicit representation based on wavelets (motivated by [Velho et al. 1994]) to build a compact representation for high-quality shape generation.

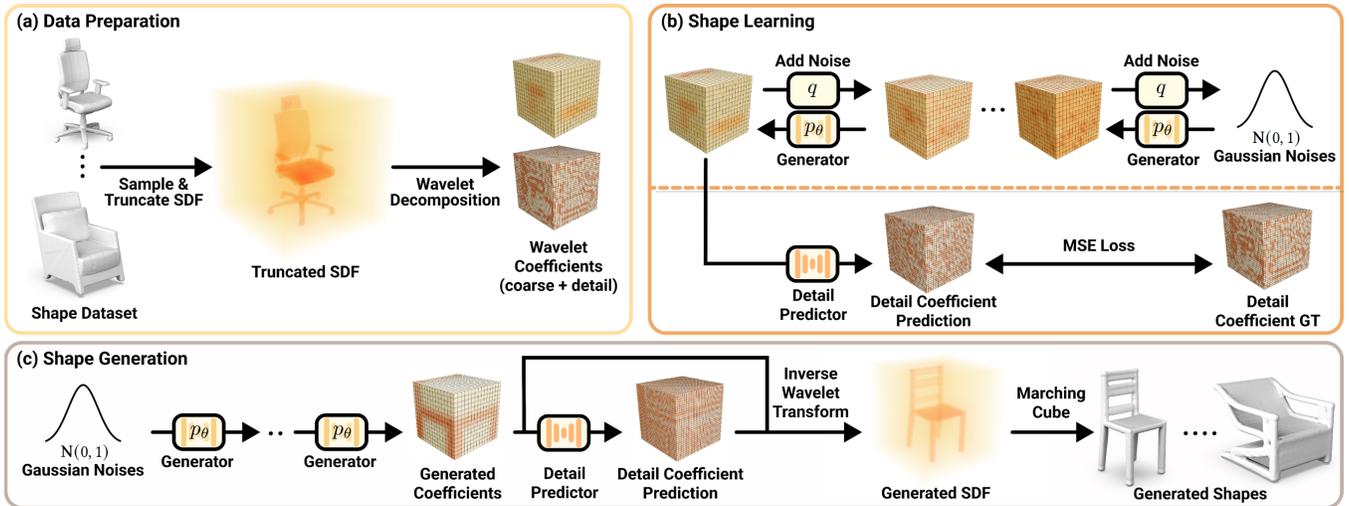
*Denoising diffusion models.* These models [Ho et al. 2020; Nichol and Dhariwal 2021; Sohl-Dickstein et al. 2015; Song et al. 2020] recently show top performance in image generation, surpassing GAN-based models [Dhariwal and Nichol 2021]. Very recently, [Luo and Hu 2021; Zhou et al. 2021] adopt diffusion models for point cloud generation. Yet, they fail to generate smooth surfaces and complex structures, as point clouds contain only discrete samples. Distinctively, we adopt diffusion model with a compact wavelet representation to model a continuous signed distance field, promoting shape generation with diverse structures and fine details.

## 3 OVERVIEW

Our approach consists of the following three major procedures:

(i) *Data preparation* is a one-time process for preparing a compact wavelet representation from each input shape; see Figure 2(a). For each shape, we sample a signed distance field (SDF) and truncate its distance values to avoid redundant information. Then, we transform the truncated SDF to the wavelet domain to produce a series of multi-scale coefficient volumes. Importantly, we take a *pair of coarse and detail coefficient volumes* at the same scale as our compact wavelet representation for implicitly encoding the input shape.

(ii) *Shape learning* aims to train a pair of neural networks to learn the 3D shape distribution from the coarse and detail coefficient volumes; see Figure 2(b). First, we adopt the denoising diffusion



**Figure 2: Overview of our approach.** (a) *Data preparation* builds a compact wavelet representation (a pair of coarse and detail coefficient volumes) for each input shape using a truncated signed distance field (TSDF) and a multi-scale wavelet decomposition. (b) *Shape learning* trains the generator network to produce coarse coefficient volumes from random noise samples and trains the detail predictor network to produce detail coefficient volumes from coarse coefficient volumes. (c) *Shape generation* employs the trained generator to produce a coarse coefficient volume and then the trained detail predictor to further predict a compatible detail coefficient volume, followed by an inverse wavelet transform and marching cube, to generate the output 3D shape.

probabilistic model [Sohl-Dickstein et al. 2015] to formulate and train the *generator network* to learn to iteratively refine a random noise sample for generating diverse 3D shapes in the form of the coarse coefficient volume. Second, we design and train the *detail predictor network* to learn to produce the detail coefficient volume from the coarse coefficient volume for introducing further details in our generated shapes. Using our compact wavelet representation, it becomes feasible to train both the generator and detail predictor to successfully produce coarse coefficient volumes with plausible 3D structures and detail coefficient volumes with fine details.

(iii) *Shape generation* employs the two trained networks to generate 3D shapes; see Figure 2(c). Starting from a random Gaussian noise sample, we first use the trained generator to produce the coarse coefficient volume then the detail predictor to produce an associated detail coefficient volume. After that, we can perform an inverse wavelet transform, followed by the marching cube operator [Lorensen and Cline 1987], to generate the output 3D shape.

## 4 METHOD

### 4.1 Compact Wavelet Representation

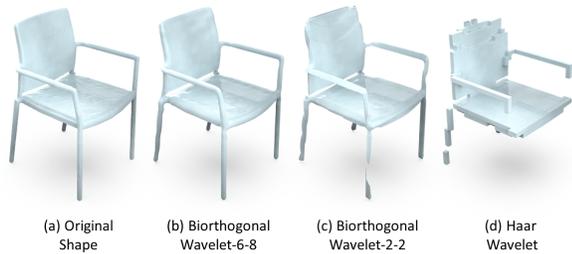
Preparing a compact wavelet representation from an input shape (see Figure 2(a)) involves the following two steps: (i) implicitly represent the shape using a signed distance field (SDF); and (ii) decompose the implicit representation via wavelet transform into coefficient volumes, each encoding a specific scale of the shape.

In the first step, we scale each shape to fit  $[-0.45, +0.45]^3$  and sample an SDF of resolution  $256^3$  to implicitly represent the shape. Importantly, we truncate the distance values in the SDF to  $[-0.1, +0.1]$ , so regions not close to object surface are clipped to a constant. We

denote the truncated signed distance field (TSDF) for the  $i$ -th shape in training set as  $S_i$ . By using  $S_i$ , we can significantly reduce the shape representation redundancy and enable the shape learning process to better focus on the shape’s structures and fine details.

The second step is a multi-scale wavelet decomposition [Daubechies 1990; Mallat 1989; Velho et al. 1994] on the TSDF. Here, we decompose  $S_i$  into a high-frequency detail coefficient volume and a low-frequency coarse coefficient volume, which is roughly a compressed version of  $S_i$ . We repeat this process  $J$  times on the coarse coefficient volume of each scale, decomposing  $S_i$  into a series of multi-level coefficient volumes. We denote the coarse and detail coefficient volumes at the  $j$ -th step (scale) as  $C_i^j$  and  $D_i^j$ , respectively, where  $j = \{1, \dots, J\}$ . The representation is lossless, meaning that the extracted coefficient volumes together can faithfully reconstruct the original TSDF via a series of inverse wavelet transforms.

There are three important considerations in the data preparation. First, multi-scale decomposition can effectively separate rich structures, fine details, and noise in the TSDF. Empirically, we evaluate the reconstruction error on the TSDF by masking out all higher-scale detail coefficients and reconstructing  $S_i$  only from the coefficients at scale  $J = 3$ , i.e.,  $C_i^3$  and  $D_i^3$ . We found that the reconstructed TSDF values have relatively small changes from the originals (only 2.8% in magnitude), even without 97% of the coefficients for the Chair category in ShapeNet [Chang et al. 2015]. Comparing Figures 3 (a) vs. (b), we can see that reconstructing only from the coarse scale of  $J = 3$  already well retains the chair’s structure. Motivated by this observation, we propose to construct the compact wavelet representation at a coarse scale ( $J = 3$ ) and drop other detail coefficient volumes, i.e.,  $D_i^1$  and  $D_i^2$ , for efficient shape



**Figure 3: Reconstructions with different wavelet filters. (a) An input shape from ShapeNet. (b,c) Reconstructions from the  $J=3$  coefficient volumes with biorthogonal wavelets. The two numbers mean the vanishing moment of the synthesis and analysis wavelets. (d) Reconstruction with the Haar wavelet.**

learning. More details on the wavelet decomposition are given in the supplementary material.

Second, we need a suitable wavelet filter. While Haar wavelet is a popular choice due to its simplicity, using it to encode smooth and continuous signals such as the SDF may introduce serious voxelization artifacts, see, e.g., Figure 3 (d). In this work, we propose to adopt the biorthogonal wavelets [Cohen 1992], since it enables a more smooth decomposition of the TSDF. Specifically, we tried different settings in the biorthogonal wavelets and chose to use high vanishing moments with six for the synthesis filter and eight for the analysis filter; see Figures 3 (b) vs. (c). Also, instead of storing the detail coefficient volumes with seven channels, as in traditional wavelet decomposition, we follow [Velho et al. 1994] to efficiently compute it as the difference between the inverse transformed  $C_i^j$  and  $C_i^{j-1}$  in a Laplacian pyramid style. Hence, the detail coefficient volume has a higher resolution than the coarser one, but both have much lower resolution than the original TSDF volume ( $256^3$ ).

Last, it is important to truncate the SDF before constructing the wavelet representation for shape learning. By truncating the SDF, regions not close to the shape surface would be cast to a constant function to make efficient the wavelet decomposition and shape learning. Otherwise, we found that the shape learning process will collapse and the training loss cannot be reduced.

## 4.2 Shape Learning

Next, to learn the 3D shape distribution in the given shape set, we collect coefficient volumes  $\{C_i^j, D_i^j\}$  from different input shapes for training (i) the *generator network* to learn to iteratively remove noise from a random Gaussian noise sample to generate  $C_i^j$ ; and (ii) the *detail predictor network* to learn to predict  $D_i^j$  from  $C_i^j$  to enhance the details in the generated shapes.

*Network structure.* To start, we formulate a simple but efficient neural network structure for both the generator and detail predictor networks. The two networks have the same structure, since they both take a 3D volume as input and then output a 3D volume of same resolution as the input. Specifically, we adopt a modified 3D version of the U-Net architecture [Nichol and Dhariwal 2021]. We first apply 3D convolution to progressively compose and down-sample the input into a set of multi-scale features and a bottleneck

feature volume. Then, we apply a single self-attention layer to aggregate features in the bottleneck volume, so that we can efficiently incorporate non-local information into the features. Further, we upsample and concatenate features in the same scale and progressively perform an inverse convolution to produce an output of same size as the input. Note also that for all convolution layers in the network structure, we use a filter size of three with a stride of one.

*Modeling the generator network.* We formulate the 3D shape generation process based on the denoising diffusion probabilistic model [Sohl-Dickstein et al. 2015]. For simplicity, we drop the subscript and superscript in  $C_i^j$ , and denote  $\{C_0, \dots, C_T\}$  as the shape generation sequence, where  $C_0$  is the target, which is  $C_i^j$ ;  $C_T$  is a random noise volume from the Gaussian prior; and  $T$  is the total number of time steps. As shown on top of Figure 2(b), we have (i) a forward process (denoted as  $q(C_{0:T})$ ) that progressively adds noises based on a Gaussian distribution to corrupt  $C_0$  into a random noise volume; and (ii) a backward process (denoted as  $p_\theta(C_{0:T})$ ) that employs the generator network (with network parameter  $\theta$ ) to iteratively remove noise from  $C_T$  to generate the target. Note that all 3D shapes  $\{C_0, \dots, C_T\}$  are represented as 3D volumes and each voxel value is a wavelet coefficient at its spatial location.

Both the forward and backward processes are modeled as Markov processes. The generator network is optimized to maximize the generation probability of the target, i.e.,  $p_\theta(C_0)$ . Also, as suggested in [Ho et al. 2020], this training procedure can be further simplified to use the generator network to predict noise volume  $\epsilon_\theta$ . Hence, we adopt a mean-squares loss to train our framework:

$$L_2 = E_{t, C_0, \epsilon} [\|\epsilon - \epsilon_\theta(C_t, t)\|^2], \epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad (1)$$

where  $t$  is a time step;  $\epsilon$  is a noise volume; and  $\mathcal{N}(0, \mathbf{I})$  denotes a unit Gaussian distribution. In particular, we first sample noise volume  $\epsilon$  from a unit Gaussian distribution  $\mathcal{N}(0, \mathbf{I})$  and a time step  $t \in [1, \dots, T]$  to corrupt  $C_0$  into  $C_t$ . Then, our generator network learns to predict noise  $\epsilon$  based on the corrupted coefficient volume  $C_t$ . Further, as the network takes time step  $t$  as input, we convert value  $t$  into an embedding via two MLP layers. Using this embedding, we can condition all the convolution modules in the prediction and enable the generator to be more aware of the amount of noise contaminated in  $C_t$ . For more details on the derivation of the training objectives, please refer to the supplementary material.

*Detail predictor network.* With the trained generator, we can obtain diverse and good-quality coarse coefficient volumes, i.e.,  $C_0$ . Next, we train the detail predictor network to produce detail coefficient volume  $D_0$  from  $C_0$  (see the bottom part of Figure 2(b)), so that we can further enhance the details in our generated shapes.

To train the detail predictor network, we leverage the paired coefficient volumes  $\{C_i^j, D_i^j\}$  from the data preparation. Importantly, detail coefficient volume  $D_0$  should be highly correlated to coarse coefficient volume  $C_0$ . Hence, we pose detail prediction as a conditional regression on the detail coefficient volume, aiming at learning neural network function  $f: C_0 \rightarrow D_0$ ; hence, we optimize  $f$  via a mean squared error loss. Overall, the detail predictor has the same network structure as the generator, but we include more convolution layers to accommodate the cubic growth in the number of nonzero values in the detail coefficient volume.



**Figure 4: Gallery of our generated shapes: Table, Chair, Cabinet, and Airplane (top to bottom). Our shapes exhibit complex structures, fine details, and clean surfaces, without obvious artifacts, compared with those generated by others; see Figure 5.**

### 4.3 Shape Generation

Now, we are ready to generate 3D shapes. First, we can randomize a 3D noise volume as  $C_T$  from the standard Gaussian distribution. Then, we can employ the trained generator for  $T$  iterations to produce  $C_0$  from  $C_T$ . This process is iterative and inter-dependent. We cannot parallelize the operations in different iterations, so leading to a very long computing time. To speed up the inference process, we adopt an approach in [Song et al. 2020] to sub-sample a set of time steps from  $[1, \dots, T]$  during the inference; in practice, we evenly sample  $1/10$  of the total time steps in all our experiments.

After we obtain the coarse coefficient volume  $C_0$ , we then use the detail predictor network to predict detail coefficient volume  $D_0$  from  $C_0$ . After that, we perform a series of inverse wavelet transforms from  $\{C_0, D_0\}$  at scale  $J=3$  to reconstruct the original TSDF. Hence, we can further extract an explicit 3D mesh from the reconstructed TSDF using the marching cube algorithm [Lorensen and Cline 1987]. Figure 2(c) illustrates the shape generation procedure.

### 4.4 Implementation Details

We employed ShapeNet [Chang et al. 2015] to prepare the training dataset used in all our experiments. Following the data split in [Chen and Zhang 2019], we use only the training split to supervise our network training. Also, similar to [Hertz et al. 2022; Li et al. 2021; Luo and Hu 2021], we train a single model for generating shapes of each category in the ShapeNet dataset [Chang et al. 2015].

We implement our networks using PyTorch and run all experiments on a GPU cluster with four RTX3090 GPUs. We follow [Ho et al. 2020] to set  $\{\beta_t\}$  to increase linearly from  $1e^{-4}$  to 0.02 for 1,000 time steps and set  $\sigma_t = \frac{1-\alpha_t-1}{1-\alpha_t} \beta_t$ . We train the generator for 800,000 iterations and the detail predictor for 60,000 iterations, both using the Adam optimizer [Kingma and Ba 2014] with a learning rate of  $1e^{-4}$ . Training the generator and detail predictor takes around

three days and 12 hours, respectively. The inference takes around six seconds per shape on an RTX 3090 GPU. We adapt [Cotter 2020] to implement the 3D wavelet decomposition and *will release our code and training data upon the publication of this work*.

## 5 RESULTS AND EXPERIMENTS

### 5.1 Galleries of our generated shapes

Besides Figure 1, we present Figure 4 to showcase the compelling capability of our method on generating shapes of various categories. Our generated shapes exhibit *diverse topologies, fine details*, and also *clean surfaces without obvious artifacts*, covering a rich variety of small, thin, and complex structures that are typically very challenging for the existing approaches to produce. More 3D shape generation results are provided in the supplementary material.

### 5.2 Comparison with Other Methods

Next, we compare the shape generation capability of our method with four state-of-the-art methods: IM-GAN [Chen and Zhang 2019], Voxel-GAN [Kleineberg et al. 2020], Point-Diff [Luo and Hu 2021], and SPAGHETTI [Hertz et al. 2022]. To our best knowledge, ours is the first work that generates implicit shape representations in frequency domain and considers coarse and detail coefficients to enhance the generation of structures and fine details.

Our experiments follow the same setting as the above works. Specifically, we leverage our trained model on the Chair and Airplane categories in ShapeNet [Chang et al. 2015] to randomly generate 2,000 shapes for each category. Then, we uniformly sample 2,048 points on each generated shape and evaluate the shapes using the same set of metrics as in the previous methods (details to be presented later). As for the four state-of-the-art methods, we employ publicly-released trained network models to generate shapes.

**Table 1: Quantitative comparison between the generated shapes produced by our method and four state-of-the-art methods. We follow the same setting to conduct this experiment as in the state-of-the-art methods. From the table, we can see that our generated shapes have the best quality for almost all cases (lowest MMD, largest COV, and lowest 1-NNA) for both the Chair and Airplane categories. The units of CD and EMD are  $10^{-3}$  and  $10^{-2}$ , respectively.**

Method	Chair						Airplane					
	COV		MMD		1-NNA		COV		MMD		1-NNA	
	CD	EMD										
IM-GAN [Chen and Zhang 2019]	56.49	54.50	11.79	14.52	61.98	63.45	61.55	62.79	3.320	8.371	76.21	76.08
Voxel-GAN [Kleineberg et al. 2020]	43.95	39.45	15.18	17.32	80.27	81.16	38.44	39.18	5.937	11.69	93.14	92.77
Point-Diff [Luo and Hu 2021]	51.47	<b>55.97</b>	12.79	16.12	61.76	63.72	60.19	62.30	3.543	9.519	74.60	72.31
SPAGHETTI [Hertz et al. 2022]	49.19	51.92	14.90	15.90	70.72	68.95	58.34	58.38	4.062	8.887	78.24	77.01
<b>Ours</b>	<b>58.19</b>	55.46	<b>11.70</b>	<b>14.31</b>	<b>61.47</b>	<b>61.62</b>	<b>64.78</b>	<b>64.40</b>	<b>3.230</b>	<b>7.756</b>	<b>71.69</b>	<b>66.74</b>



**Figure 5: Visual comparisons with state-of-the-art methods. Our generated shapes exhibit finer details and cleaner surfaces, without obvious artifacts.**

*Evaluation metrics.* Following [Hertz et al. 2022; Luo and Hu 2021], we evaluate the generation quality using (i) minimum matching distance (MMD) measures the fidelity of the generated shapes; (ii) coverage (COV) indicates how well the generated shapes cover the shapes in the given 3D repository; and (iii) 1-NN classifier accuracy (1-NNA) measures how well a classifier differentiates the generated shapes from those in the repository. Overall, a low MMD, a high COV, and an 1-NNA close to 50% indicate good generation quality. More details are provided in the supplementary material.

*Quantitative evaluation.* Table 1 reports the quantitative comparison results, showing that our method surpasses all others for almost all the evaluation cases over the three metrics for both the Chair and Airplane categories. We employ the Chair category, due to its large variations in structure and topology, and the Airplane category, due to the fine details in its shapes. As discussed in [Luo

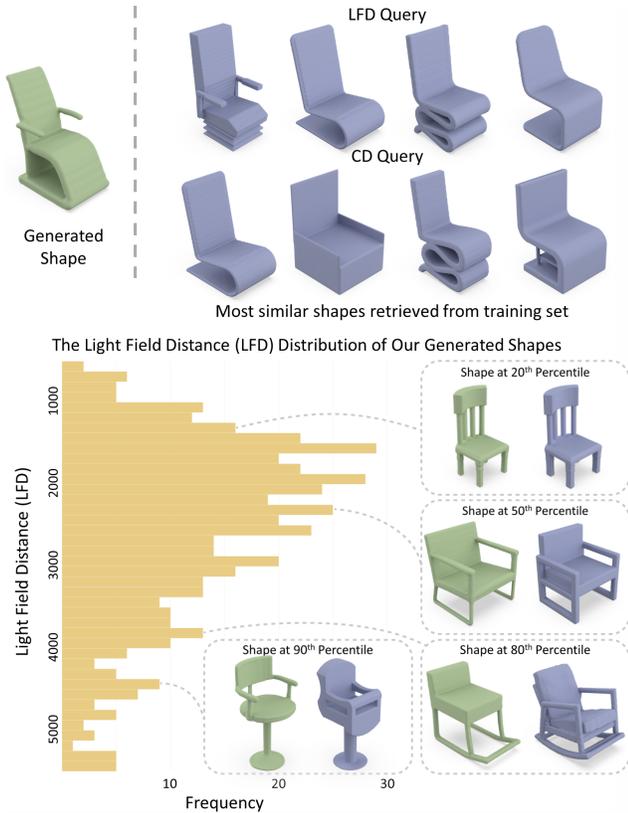
and Hu 2021; Yang et al. 2019], the COV and MMD metrics have limited capabilities to account for details, so they are not suitable for measuring the fine quality of the generation results, e.g., the generated shapes sometimes show a better performance even when compared with the ground-truth training shapes on these metrics. In contrast, 1-NNA is more robust and can better correlate with the generation quality. In this metric, our approach outperforms all others, while having a significant margin in the Airplane category, manifesting the diversity and fidelity of our generated results.

*Qualitative Evaluation.* Figure 5 show some visual comparisons. For each random shape generated by our method, we find a similar shape (with similar structures and topology) generated by each of the other methods to make the visual comparison easier. See supplementary material Sections B and D for more visual comparisons. Further, as different methods likely have different statistical modes in the shape generation distribution, we also take random shapes generated by IM-GAN and find similar shapes generated by our method for comparison; see supplementary material Section C for the results. From all these results, we can see that the 3D shapes generated by our method clearly exhibit finer details, higher fidelity structures, and cleaner surfaces, without obvious artifacts.

### 5.3 Model Analysis

*Shape novelty analysis.* Next, we analyze whether our method can generate shapes that are not necessarily the same as the training-set shapes, meaning that it does not simply memorize the training data. To do so, we use our method to generate 500 random shapes and retrieve top-four most similar shapes in the training set separately via two different metrics, i.e., Chamfer Distance (CD) and Light Field Distance (LFD) [Chen et al. 2003]. It is noted that LFD is computed based on rendered images from multiple views on each shape, so it focuses more on the visual similarity between shapes and is considered to be more robust for shape retrieval. For the details on the metrics, please see the supplementary material.

Figure 6 (top) shows a shape generated by our method, together with top-four most similar shapes retrieved from the training set by CD and LFD; due to the page limit, another ten examples are shown in the supplementary material. Comparing our shapes with the retrieved ones, we can see that the shapes share similar structures, showing that our method is able to generate realistic-looking



**Figure 6: Shape novelty analysis.** Top: From our generated shape (in green), we retrieve top-four most similar shapes (in blue) in training set by CD and LFD. Bottom: We generate 500 chairs using our method; for each chair, we retrieve the most similar shape in the training set by LFD; then, we plot the distribution of LFDs for all retrievals, showing that our method is able to generate shapes that are more similar (low LFDs) or more novel (high LFDs) compared to the training set. Note that the generated shape at 50<sup>th</sup> percentile is already not that similar to the associated training-set shape.

structures like those in the training set. Beyond that, our shapes exhibit noticeable differences in various local structures.

As mentioned earlier, a good generator should produce diverse shapes that are not necessarily the same as the training shapes. So, we further statistically analyze the novelty of our generated shapes relative to the training set. To do so, we use our method to generate 500 random chairs; for each generated chair shape, we use LFD to retrieve the most similar shape in the training set. Figure 6 (bottom) plots the distribution of LFDs between our generated shapes (in green) and retrieved shapes (in blue). Also, we show four shape pairs at various percentiles, revealing that shapes with larger LFDs are more different from the most similar shapes in the training set. From the LFD distribution, we can see that our method can learn a generation distribution that covers shapes in the training set (low LFD) and also generates novel and realistic-looking shapes that are more different (high LFD) from the training-set shapes.

**Table 2: Comparing our full pipeline with various ablated cases on the Chair category. The units of CD and EMD are  $10^{-3}$  and  $10^{-2}$ , respectively.**

Method	COV $\uparrow$		MMD $\downarrow$		1-NNA $\downarrow$	
	CD	EMD	CD	EMD	CD	EMD
Full Model	<b>58.19</b>	<b>55.46</b>	<b>11.70</b>	<b>14.31</b>	<b>61.47</b>	<b>61.62</b>
W/o detail predictor	54.20	50.96	12.32	14.54	62.46	62.57
VAD Generator	21.83	26.77	21.83	26.77	95.20	93.62
Direct predict TSDF	50.51	50.67	12.83	15.24	68.69	68.29

*Ablation Study.* To evaluate the major components in our method, we conducted an ablation study by successively changing our full pipeline. First, we evaluate the generation performance with/without the detail predictor. Next, we study the importance of the diffusion model and the wavelet representation in the generator network.

The results in Table 2 demonstrate the capability of the detail predictor, which introduces a substantial improvement on all metrics (first vs. second rows). Further, replacing our generator with the VAD model or directly predicting TSDF leads to a performance degrade (second & last two rows). Due to the page limit, please refer to the supplementary material for the details on how the ablation cases are implemented and the visual comparison results.

*Limitations.* Due to the page limit, please refer to Section K of the supplementary material for the discussion on limitations.

## 6 CONCLUSION

This paper presents a new generative approach for learning 3D shape distribution and generating diverse, high-quality, and possibly novel 3D shapes. Unlike prior works, we operate on the frequency domain. By decomposing the implicit function in the form of TSDF using biorthogonal wavelets, we build a compact wavelet representation with a pair of coarse and detail coefficient volumes, as an encoding of 3D shape. Then, we formulate our generator upon a probabilistic diffusion model to learn to generate diverse shapes in the form of coarse coefficient volumes from noise samples, and a detail predictor to further learn to generate compatible detail coefficient volumes for reconstructing fine details. Both quantitative and qualitative experimental results demonstrate the superiority of our method in generating diverse and realistic shapes that exhibit fine details, complex and thin structures, and clean surfaces, beyond the generation capability of the state-of-the-art methods.

To our best knowledge, this is the first work that successfully adopts a compact wavelet representation for an unconditional generative modeling on 3D shape generation, enabling many directions for future research. At first glance, our benefits can be extended to other downstream tasks with extra conditions, e.g., shape reconstruction from images or point clouds, and shape editing with user inputs. Another promising future direction is to adopt wavelet-based 3D generation to animation production, e.g., generating sequences of character motion with spatio-temporal wavelet representations. Also, we would like to explore more challenging cases, e.g., objects with extremely fine details and generation of 3D scenes.

## ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments. We also acknowledge help from Tianyu Wang for various visualizations in the paper. This work is supported by Research Grants Council of the Hong Kong Special Administrative Region (Project No. CUHK 14206320 & 14201921) and National Natural Science Foundation of China (No. 62202151).

## REFERENCES

- Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas J. Guibas. 2018. Learning representations and generative models for 3D point clouds. In *Proceedings of International Conference on Machine Learning (ICML)*. 40–49.
- Matan Atzmon and Yaron Lipman. 2020. SAL: Sign agnostic learning of shapes from raw data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2565–2574.
- Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snively, and Bharath Hariharan. 2020. Learning gradient fields for shape generation. In *European Conference on Computer Vision (ECCV)*. 364–381.
- Angel X. Chang, Thomas Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012* (2015).
- Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. 2003. On visual similarity based 3D model retrieval. In *Computer Graphics Forum*, Vol. 22. 223–232.
- Zhiqin Chen and Hao Zhang. 2019. Learning implicit fields for generative shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5939–5948.
- Zhang Chen, Yinda Zhang, Kyle Genova, Sean Fanello, Sofien Bouaziz, Christian Häne, Ruofei Du, Cem Keskin, Thomas Funkhouser, and Danhang Tang. 2021. Multiresolution Deep Implicit Functions for 3D Shape Representation. In *IEEE International Conference on Computer Vision (ICCV)*. 13087–13096.
- Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. 2020. Implicit functions in feature space for 3D shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6970–6981.
- Albert Cohen. 1992. Biorthogonal wavelets. *Wavelets: A Tutorial in Theory and Applications 2* (1992), 123–152.
- Fergal Cotter. 2020. *Uses of Complex Wavelets in Deep Convolutional Neural Networks*. Ph.D. Dissertation. University of Cambridge.
- Ingrid Daubechies. 1990. The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory* 36, 5 (1990), 961–1005.
- Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat GANS on image synthesis. *Conference on Neural Information Processing Systems (NeurIPS)* (2021), 8780–8794.
- Haoqiang Fan, Hao Su, and Leonidas J. Guibas. 2017. A point set generation network for 3D object reconstruction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 605–613.
- Rizal Fathony, Anit Kumar Sahu, Devin Willmott, and J. Zico Kolter. 2020. Multiplicative filter networks. In *International Conference on Learning Representations (ICLR)*.
- Rinon Gal, Amit Bermano, Hao Zhang, and Daniel Cohen-Or. 2020. MRGAN: Multi-Rooted 3D Shape Generation with Unsupervised Part Disentanglement. In *IEEE International Conference on Computer Vision (ICCV)*. 2039–2048.
- Rohit Girdhar, David F. Fouhey, Mikel Rodriguez, and Abhinav Gupta. 2016. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision (ECCV)*. 484–499.
- Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. 2020. Implicit geometric regularization for learning shapes. In *Proceedings of International Conference on Machine Learning (ICML)*. 3569–3579.
- Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. 2018. A papier-mâché approach to learning 3D surface generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 216–224.
- Zekun Hao, Hadar Averbuch-Elor, Noah Snively, and Serge Belongie. 2020. DualSDF: Semantic shape manipulation using a two-level representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7631–7641.
- Amir Hertz, Or Perel, Raja Giryes, Olga Sorkine-Hornung, and Daniel Cohen-Or. 2022. SPAGHETTI: Editing Implicit Shapes Through Part Aware Generation. *arXiv preprint arXiv:2201.13168* (2022).
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Conference on Neural Information Processing Systems (NeurIPS)* (2020), 6840–6851.
- Le Hui, Rui Xu, Jin Xie, Jianjun Qian, and Jian Yang. 2020. Progressive point cloud deconvolution generation network. In *European Conference on Computer Vision (ECCV)*. 397–413.
- Moritz Ibing, Isaak Lim, and Leif Kobbelt. 2021. 3D Shape Generation With Grid-Based Implicit Functions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 13559–13568.
- Li Jiang, Shaoshuai Shi, Xiaojuan Qi, and Jiaya Jia. 2018. GAL: Geometric adversarial loss for single-view 3D-object reconstruction. In *European Conference on Computer Vision (ECCV)*. 802–816.
- Hyeonju Kim, Hyeonseung Lee, Woo Hyun Kang, Joun Yeop Lee, and Nam Soo Kim. 2020. SoftFlow: Probabilistic framework for normalizing flow on manifolds. In *Conference on Neural Information Processing Systems (NeurIPS)*. 16388–16397.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- Marian Kleineberg, Matthias Fey, and Frank Weichert. 2020. Adversarial generation of continuous implicit shape representations. *arXiv preprint arXiv:2002.00349* (2020).
- Manyi Li and Hao Zhang. 2021. D<sup>2</sup>IM-Net: Learning detail disentangled implicit fields from single images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 10246–10255.
- Ruihui Li, Xianzhi Li, Ke-Hei Hui, and Chi-Wing Fu. 2021. SP-GAN: Sphere-Guided 3D Shape Generation and Manipulation. *ACM Transactions on Graphics (SIGGRAPH)* 40, 4 (2021).
- Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. 2020. Neural sparse voxel fields. *Conference on Neural Information Processing Systems (NeurIPS)* (2020), 15651–15663.
- Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. 2019. Learning to infer implicit surfaces without 3D supervision. *Conference on Neural Information Processing Systems (NeurIPS)* (2019).
- Shi-Lin Liu, Hao-Xiang Guo, Hao Pan, Pengshuai Wang, Xin Tong, and Yang Liu. 2021. Deep Implicit Moving Least-Squares Functions for 3D Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1788–1797.
- William E. Lorensen and Harvey E. Cline. 1987. Marching Cubes: A high resolution 3D surface construction algorithm. In *Proceedings of SIGGRAPH*, Vol. 21. 163–169.
- Andrew Luo, Tianqin Li, Wen-Hao Zhang, and Tai Sing Lee. 2021. SurfGen: Adversarial 3D Shape Synthesis with Explicit Surface Discriminators. In *IEEE International Conference on Computer Vision (ICCV)*. 16238–16248.
- Shitong Luo and Wei Hu. 2021. Diffusion probabilistic models for 3D point cloud generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2837–2845.
- Stephane G. Mallat. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence* 11, 7 (1989), 674–693.
- Julien N. P. Martel, David B. Lindell, Connor Z. Lin, Eric R. Chan, Marco Monteiro, and Gordon Wetzstein. 2021. ACORN: Adaptive coordinate networks for neural scene representation. *ACM Transactions on Graphics (SIGGRAPH)* 40, 4 (2021), 13.
- Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3D reconstruction in function space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4460–4470.
- Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *Proceedings of International Conference on Machine Learning (ICML)*. 8162–8171.
- Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. 2020. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3504–3515.
- Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. DeepSDF: Learning continuous signed distance functions for shape representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 165–174.
- Vishwanath Saragadam, Jasper Tan, Guha Balakrishnan, Richard G. Baraniuk, and Ashok Veeraraghavan. 2022. MINER: Multiscale Implicit Neural Representations. *arXiv preprint arXiv:2202.03532* (2022).
- Edward J. Smith, Scott Fujimoto, Adriana Romero, and David Meger. 2019. GEOMETrics: Exploiting geometric structure for graph-encoded objects. In *Proceedings of International Conference on Machine Learning (ICML)*. 5866–5876.
- Edward J. Smith and David Meger. 2017. Improved adversarial systems for 3D object generation and reconstruction. In *Conference on Robot Learning*. PMLR, 87–96.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of International Conference on Machine Learning (ICML)*. 2256–2265.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).
- Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. 2021. Neural geometric level of detail: Real-time rendering with implicit 3D shapes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 11358–11367.
- Jiapeng Tang, Xiaoguang Han, Junyi Pan, Kui Jia, and Xin Tong. 2019. A skeleton-bridged deep learning approach for generating meshes of complex topologies from single RGB images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4541–4550.
- Jiapeng Tang, Xiaoguang Han, Minghui Tan, Xin Tong, and Kui Jia. 2021. SkeletonNet: A topology-preserving solution for learning mesh reconstruction of object surfaces from RGB images. *IEEE Transactions Pattern Analysis & Machine Intelligence* (2021).

- to appear.
- Luiz Velho, Demetri Terzopoulos, and Jonas Gomes. 1994. Multiscale implicit models. In *Proceedings of SIBGRAPI*, Vol. 94. 93–100.
- Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. 2018. Pixel2Mesh: Generating 3D mesh models from single RGB images. In *European Conference on Computer Vision (ECCV)*. 52–67.
- Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. 2016. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *Conference on Neural Information Processing Systems (NeurIPS)*. 82–90.
- Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 2019. DISN: Deep implicit surface network for high-quality single-view 3D reconstruction. In *Conference on Neural Information Processing Systems (NeurIPS)*. 490–500.
- Yifan Xu, Tianqi Fan, Yi Yuan, and Gurprit Singh. 2020. Ladybird: Quasi-Monte Carlo sampling for deep implicit field based 3D reconstruction with symmetry. In *European Conference on Computer Vision (ECCV)*. 248–263.
- Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2022. Shapeformer: Transformer-based shape completion via sparse representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6239–6249.
- Guandao Yang, Yin Cui, Serge Belongie, and Bharath Hariharan. 2018. Learning single-view 3D reconstruction with limited pose supervision. In *European Conference on Computer Vision (ECCV)*. 86–101.
- Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. 2019. PointFlow: 3D point cloud generation with continuous normalizing flows. In *IEEE International Conference on Computer Vision (ICCV)*. 4541–4550.
- Wenbin Zhao, Jiabao Lei, Yuxin Wen, Jianguo Zhang, and Kui Jia. 2021. Sign-Agnostic Implicit Learning of Surface Self-Similarities for Shape Modeling and Reconstruction from Raw Point Clouds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 10256–10265.
- Xin-Yang Zheng, Yang Liu, Peng-Shuai Wang, and Xin Tong. 2022. SDF-StyleGAN: Implicit SDF-Based StyleGAN for 3D Shape Generation. In *Eurographics Symposium on Geometry Processing (SGP)*.
- Linqi Zhou, Yilun Du, and Jiajun Wu. 2021. 3D shape generation and completion through point-voxel diffusion. In *IEEE International Conference on Computer Vision (ICCV)*. 5826–5835.
- Rui Zhu, Hamed Kiani Galoogahi, Chaoyang Wang, and Simon Lucey. 2017. Rethinking Reprojection: Closing the loop for pose-aware shape reconstruction from a single image. In *IEEE International Conference on Computer Vision (ICCV)*. 57–65.