

Generative VoxelNet: Learning Energy-Based Models for 3D Shape Synthesis and Analysis

Jianwen Xie ^{*}, Zilong Zheng ^{*}, Ruiqi Gao, Wenguan Wang,
Song-Chun Zhu, *Fellow, IEEE*, and Ying Nian Wu

Abstract—3D data that contains rich geometry information of objects and scenes is valuable for understanding 3D physical world. With the recent emergence of large-scale 3D datasets, it becomes increasingly crucial to have a powerful 3D generative model for 3D shape synthesis and analysis. This paper proposes a deep 3D energy-based model to represent volumetric shapes. The maximum likelihood training of the model follows an “analysis by synthesis” scheme. The benefits of the proposed model are six-fold: first, unlike GANs and VAEs, the model training does not rely on any auxiliary models; second, the model can synthesize realistic 3D shapes by Markov chain Monte Carlo (MCMC); third, the conditional model can be applied to 3D object recovery and super resolution; fourth, the model can serve as a building block in a multi-grid modeling and sampling framework for high resolution 3D shape synthesis; fifth, the model can be used to train a 3D generator via MCMC teaching; sixth, the unsupervisedly trained model provides a powerful feature extractor for 3D data, which is useful for 3D object classification. Experiments demonstrate that the proposed model can generate high-quality 3D shape patterns and can be useful for a wide variety of 3D shape analysis.

Index Terms—Deep generative models; Energy-based models; Langevin dynamics; Volumetric shape synthesis; Generative VoxelNet; Cooperative learning; Multi-grid sampling.

1 INTRODUCTION

1.1 Background and motivation

UNDERSTANDING 3D world is a core mission of computer vision as many vision-based perception systems in a wide range of real-world application scenarios, such as autonomous navigation [1], home robotics [2], and augmented/virtual reality [3], [4], immensely rely on the ability to accurately parse, reconstruct and interact with the physical 3D environment, as well as the ability to simultaneously detect, recognize, reason, and predict the behaviors of agents within the 3D environment.

The rapid innovations and advancements in 3D sensing technologies, especially 3D data acquisition devices (e.g., structured-light 3D scanners [5] and time-of-flight cameras [6]), as well as the recent emergence of large-scale 3D datasets (e.g., NYU depth [7], SUN3D [8], SUN RGB-D [9], ModelNet [10], and ShapeNet [11]), have tremendously facilitated the accessibility of 3D data. With full geometry information of the 3D sensed objects and scenes, some computer vision tasks, such as classification, localization, segmentation, and alignment, will become much easier than in 2D image space, therefore 3D data is a very useful and valuable asset in the field of computer vision. The amazing success of deep learning

models along with the increased availability of 3D data have been encouraging the computer vision community to investigate the generalization of deep learning technologies to 3D data.

There exist various types of representations of 3D data, such as multi-view data, RGB-D data, volumetric data, 3D point clouds, and 3D meshes. Different types of 3D representations may have different data structures and geometric properties [12]. In this paper, we only focus on volumetric representation of 3D shape, where each 3D object is characterized as a regular voxel grid in the three-dimensional space. The value of a voxel represents opacity on a scale of [0, 1], with 0 indicating “fully transparent” and 1 indicating “fully opaque”.

1.2 Statistical models of 3D volumetric shapes

Recently, some interesting attempts [13], [14], [15] have been made on object recognition and synthesis based on volumetric 3D shape data. From the perspective of statistical modeling, the existing 3D models can be grouped into two main categories: (1) 3D discriminators, such as Voxnet [16], which aim to learn a mapping from 3D voxel input to semantic labels for the purpose of 3D object classification and recognition, and (2) 3D generators, such as 3D-GAN [17], which are in the form of latent variable models that assume that the 3D voxel signals are generated by some latent variables. The training of discriminators usually relies on big data with annotations and is accomplished by a direct minimization of the prediction errors, while the training of the generators learns a mapping from the latent space to 3D voxel data space.

The generator model, while useful for synthesizing 3D shape patterns, involves a challenging inference step (i.e., sampling from the posterior distribution) in maximum likelihood learning, therefore variational inference [18] and adversarial learning [17], [19], [20] methods are commonly used, where an extra network is incorporated into the learning algorithm to get around the difficulty of the posterior inference.

- J. Xie is with the Cognitive Computing Lab, Baidu Research, Bellevue, WA 98004, USA. E-mail: jianwen@ucla.edu
- Z. Zheng is with the Department of Computer Science, University of California, Los Angeles, CA 90095, USA. E-mail: z.zheng@ucla.edu
- R. Gao is with the Department of Statistics, University of California, Los Angeles, CA 90095, USA. E-mail: ruiqigao@ucla.edu
- W. Wang is with ETH Zürich, 8092 Zürich, Switzerland. E-mail: wenguan-wang.ai@gmail.com
- S.-C. Zhu is with Tsinghua University and Peking University, Beijing, China. E-mail: sczhu@stat.ucla.edu
- Y. N. Wu is with the Department of Statistics, University of California, Los Angeles, CA 90095, USA. E-mail: ywu@stat.ucla.edu
- * indicates equal contributions.

Manuscript received Oct 03, 2019; revised Oct 29, 2020; accepted Nov 27, 2020.

The past few years have witnessed impressive progress on developing discriminator models and generator models for 3D shape data, however, there has not been much work in the literature on modeling 3D shape data based on energy-based models. [21] calls this type of models the descriptor models, because the models describe the data based on bottom-up descriptive features learned from the data. The focus of the present paper is to develop a volumetric 3D energy-based model for voxelized shape data. It can be considered an alternative to 3D-GAN [17] for 3D shape generation.

1.3 Energy-based models for 3D volumetric shapes

Specifically, we present a novel framework for probabilistic modeling of volumetric shape patterns by combining the merits of energy-based model [22] and 3D voxel convolutional neural network [16]. The model is a probability density function directly defined on voxelized shape signal, and the model is in the form of an energy-based model, where the feature statistics or the energy function is defined by a bottom-up 3D voxel ConvNet (VoxelNet) that maps the 3D voxel signal to the features. We call the proposed model the generative VoxelNet, because the model learns the VoxelNet in a generative manner.

The training of the proposed model via maximum likelihood estimation follows an “analysis by synthesis” scheme [23], where, within each iteration of the learning algorithm, we synthesize examples by Markov chain Monte Carlo (MCMC) sampling method, and then update the model parameters based on the difference between the observed training examples and the synthesized examples. This process can be interpreted as an alternating mode seeking and mode shifting process.

Different from the variational inference or adversarial learning, the proposed model does not need to incorporate an extra inference network or an adversarial discriminator in the learning process. The learning and sampling processes are guided by the same set of parameters of a single model, which makes it a particularly natural and statistically rigorous framework for probabilistic 3D shape modeling.

Modeling 3D shape data by a probability density function provides distinctive advantages: First, it is able to synthesize realistic 3D shape patterns by sampling examples from the distribution via MCMC, such as Langevin dynamics [24]. Second, the model can be modified to be a conditional model, which is useful for 3D object recovery and 3D object super resolution. Specifically, a conditional probability density function that maps the corrupted (or low resolution) 3D object to the recovered (or high resolution) 3D object is trained, and then the 3D recovery (or 3D super resolution) can be achieved by sampling from the learned conditional distribution given the corrupted or low resolution 3D object as the conditional input. Third, the model can be used in a cooperative training scheme [25], as opposed to adversarial training and variational inference, to train a 3D generator model via MCMC teaching. The training of 3D generator in such a scheme is stable and does not encounter mode collapse issue. Fourth, the model can be used in a multi-grid modeling and sampling framework [26] for high resolution 3D synthesis. Fifth, the model is useful for semi-supervised learning. After learning the model from unlabeled data, the learned features can be used to train a classifier on the labeled data.

We show that the proposed energy-based generative VoxelNet can be used to synthesize realistic 3D shape patterns, and its

conditional version is useful for 3D object recovery and 3D object super resolution. The multi-grid modeling and sampling of the generative VoxelNet can generate high resolution 3D patterns. The 3D generator trained by the generative VoxelNet in a cooperative scheme carries semantic information about 3D objects. The feature maps trained by the generative VoxelNet in an unsupervised manner are useful for 3D object classification.

1.4 Related work

3D object synthesis. Researchers in the fields of graphics and vision have studied the 3D object synthesis problems [27], [28], [29], [30], [31]. For example, component-based 3D object synthesis methods, e.g., [29] and [30], propose to use a graphical model for individual 3D objects that encodes semantic and geometric relationships among shape components. These models can be trained on a given collection of 3D shapes and then used to generate new objects by recombining the training object components. [31] proposes a set evolution method to evolve a set of diverse objects to fit user’s preference for generating sets of 3D novel shapes. However, most of these object synthesis methods are nonparametric and based on meshes or skeletons, and they generate new patterns by retrieving and merging parts from an existing CAD database. The rapid development of generative models has also enabled researchers to utilize parametric model-based approaches, such as [10], [17], [32], [33], [34], [35], [36], [37], for 3D object synthesis. Our framework represents voxelized 3D objects by proposing an energy-based model, which is a parametric probability distribution defined on the 3D shape domain. The model can learn from a repository of 3D shapes. 3D object synthesis can be achieved by running MCMC, such as Langevin dynamics, to draw samples from the learned model, without borrowing components from the training shape repository. Since the space of the 3D shapes is of high dimension, the MCMC-based synthesis is very challenging, but conceptually novel.

Deep learning for volumetric data. Recently, the vision community has witnessed the success of deep learning, and researchers have used the models in the field of deep learning for the sake of voxel-based object synthesis and analysis. For example, 3D ShapeNets [10] represents a 3D volumetric shape by a probability distribution of binary variables on a 3D voxel grid, using a convolutional deep belief network. [16] and [15] adopt a supervised volumetric ConvNet for 3D object recognition. [17] proposes the 3D deep convolutional generative adversarial net (3D-GAN) to synthesize 3D objects by leveraging 3D voxel ConvNets (VoxelNets) and generative adversarial nets (GANs). [32] proposes a deep convolutional decoder that can generate high resolution 3D volumetric shapes in a computationally efficient way by using an octree representation. Building on the early work of [38], recently [39], [40], [41] have developed an introspective learning method to learn the energy-based model, where the energy function is discriminatively learned. A 3D version of Introspective Neural Networks for 3D volumetric shape modeling is studied in [36]. Our 3D model is also powered by the VoxelNets. It incorporates a bottom-up VoxelNet that captures volumetric feature statistics of the 3D shapes for defining the probability density of shape in the 3D voxel grid, and learns the parameters of the VoxelNet by an “analysis by synthesis” scheme.

Energy-based models for synthesis. The energy-based model specifies a probability distribution of the signal, which is based on an energy function that extracts some descriptive feature

statistics from the signal. Our model is related to the following energy-based models. The FRAME (Filters, Random field, And Maximum Entropy) [42] model was developed for modeling stochastic textures. It is a Markov random field model or an energy-based model of stationary spatial processes. The energy function of the distribution is the sum of translation-invariant potential functions that are one-dimensional non-linear transformations of linear Gabor filter responses. The sparse FRAME model [43], [44] generalizes the original FRAME model to modeling object patterns that are non-stationary in the spatial domain. The model is a sparse and inhomogeneous version of original FRAME, where the potential functions are location specific, and they are non-zero only at selected locations. Inspired by the successes of deep convolutional neural networks (CNNs or ConvNets), [45] proposes a deep FRAME model, where the linear filters used in the original FRAME model are replaced by the non-linear filters at a certain convolutional layer of a pre-trained deep ConvNet. Such filters can capture more complicated patterns than linear Gabor filters. We refer readers to [46] for a comprehensive tutorial on FRAME models. Instead of using filters from a pre-trained ConvNet, [47] learns the ConvNet filters from the observed data by maximum likelihood estimation. The resulting model is an energy-based generative ConvNet, which can be considered a recursive multi-layer generalization of the original FRAME model. [48] and [49] further study learning the energy-based generative ConvNet with a non-convergent non-persistent short-run MCMC. [50] adopts a residual network [51] to parameterize the energy function. The energy-based spatial-temporal generative ConvNet [52], [53] is a generalization of [47] by adding the temporal dimension for the purpose of modeling and synthesizing dynamic patterns, such as dynamic textures and action patterns. While previous models focus on modeling 2D images or image sequences, our paper studies an energy-based model with the energy function parameterized by the VoxelNet for 3D volumetric shape patterns. We call the model the generative VoxelNet. We then further generalize the model to a conditional version for 3D shape recovery and super resolution, as well as a multi-grid version for high resolution synthesis. This paper is an expanded version of our conference paper [54].

3D object classification. Even though our paper studies energy-based generative modeling of 3D volumetric shapes, the features learned from the model can be used for classification. We review methods of 3D shape classification below. The computer vision and graphics community has developed various 3D shape descriptors, such as Light Field descriptor [55] and Spherical Harmonic descriptor [56]. Standard classifiers can be easily trained on those descriptors for 3D classification. Another direction is to apply established 2D discriminative ConvNet to the 2D information extracted from the 3D shapes. For example, [57] proposes to convert 3D shapes into geometry images, on which standard 2D ConvNet is directly adopted for classification. [58], [59] propose to classify 3D shapes by applying standard 2D ConvNet to their panoramic [60] view images. Besides, researchers also directly design end-to-end 3D discriminative neural networks that are suitable for 3D raw data. For example, [61] applies graph convolutional neural network to 3D point cloud classification. [16] proposes a 3D volumetric ConvNet for 3D classification. The following approaches learn generative 3D representations in unsupervised manners, and trained classifiers on the features extracted from the learned models along with class labels. 3D ShapeNets [10] trains a convolutional deep belief network from 3D volumetric shapes. The features learned from 3D ShapeNets are discriminatively fine-tuned with class labels for

classification. [62] learns a convolutional volumetric auto-encoder from 3D shapes and uses the learned embedding for classification. [17], [34] train 3D generative adversarial nets, and then train classifiers on the representations learned by the discriminators. [36] learns a 3D Wasserstein introspective neural network [41], and trains a linear SVM on top of the features extracted from the model. Our paper also learns generative features for 3D classification. The major difference between ours and those mentioned above is that our features are extracted by the bottom-up energy function of the energy-based model, which is parameterized by a VoxelNet, and trained via MCMC-based maximum likelihood estimation.

1.5 Contributions

(1) We propose a 3D deep energy-based model that we call generative VoxelNet to model 3D voxel patterns by combining the 3D voxel ConvNets [16] and the energy-based generative ConvNets [47]. The proposed model provides an explicit probability distribution defined on the 3D voxel grid, and its training does not rely on any auxiliary models. (2) For theoretical understanding, we present a mode seeking and mode shifting interpretation of the “analysis by synthesis” learning process of the model, and an adversarial interpretation of the zero temperature limit of the learning process. (4) We propose a conditional learning framework based on the proposed 3D energy-based model for recovery tasks, such as 3D shape super resolution and 3D object recovery. (5) We propose metrics that can be useful for evaluating 3D generative models. (6) We propose a multi-grid energy-based modeling and sampling framework for high resolution 3D volumetric shape synthesis. (7) A 3D cooperative training scheme is provided as an alternative to the adversarial learning method or the variational inference approach to train 3D generators. (8) The proposed model outperforms the other unsupervised baseline 3D models for volumetric data (e.g., 3D-VAE, 3D-GAN, and 3D-WINN), and obtains state-of-the-art performance in terms of both synthesis quality and classification accuracy.

1.6 Organization

The rest of this paper is structured as follows. In section 2, we first propose a deep energy-based model for 3D volumetric object patterns that we call the generative VoxelNet, by leveraging previous advances on 3D voxel convolutional networks [16] and the energy-based generative ConvNets [47]. We then present an interpretation of the learning process of the model by explaining it as a mode seeking and mode shifting dynamics. Section 3 presents a conditional version of the generative VoxelNet model for 3D shape recovery. Section 4 describes a multi-grid energy-based modeling and sampling strategy for high resolution 3D volumetric shape synthesis. In section 5, we show how to learn a top-down ConvNet (3D generator) as a sampler simultaneously with the generative VoxelNet, so that (1) the MCMC simulation of the generative VoxelNet can be initialized by the learned sampler, and (2) the 3D generator can be trained by the generative VoxelNet via MCMC teaching. Section 6 presents extensive qualitative and quantitative experiments to evaluate the proposed models.

2 ENERGY-BASED GENERATIVE VOXELNET

2.1 Probability density for 3D volumetric shapes

Let Y be a 3D volumetric shape defined on the 3D voxel grid. The generative VoxelNet model is a 3D deep convolutional energy-

based model defined on Y , which is in the form of exponential tilting of a reference distribution [47]:

$$p(Y; \theta) = \frac{1}{Z(\theta)} \exp[f(Y; \theta)] p_0(Y), \quad (1)$$

where $p_0(Y)$ is the reference distribution such as Gaussian white noise model, i.e., $p_0(Y) \propto \exp(-\|Y\|^2/2s^2)$ (Standard deviation s is a hyperparameter. We shall set $s = 0.5$). $f(Y; \theta)$ is the scoring function mapping Y to a scalar by a bottom-up 3D voxel convolutional neural network (VoxelNet) that is a composition of L_f layers of 3D volumetric convolutions, subsamplings, and non-linear rectifications (e.g., Rectified Linear Unit (ReLU) [63]), as illustrated by the following diagram:

$$Y \rightarrow h^{(1)} \rightarrow \dots h^{(l-1)} \rightarrow h^{(l)} \rightarrow \dots h^{(L_f)} \rightarrow f(Y; \theta),$$

where $h^{(l)}$ is the hidden output computed recursively by performing 3D convolution on $h^{(l-1)}$ followed by ReLUs. All the weight and bias parameters in this 3D ConvNet f are denoted by θ . The analytically intractable normalizing constant $Z(\theta) = \int \exp[f(Y; \theta)] p_0(Y) dY$ is used to reduce the unnormalized density function to a probability density function with total probability of one. The generative VoxelNet model $p(Y; \theta)$ defines explicit log-likelihood via $f(Y; \theta)$.

$p(Y; \theta)$ can be written in the form of an energy-based model, i.e., $p(Y; \theta) = \frac{1}{Z(\theta)} \exp[-\mathcal{E}(Y; \theta)]$. The energy function is

$$\mathcal{E}(Y; \theta) = \frac{\|Y\|^2}{2s^2} - f(Y; \theta). \quad (2)$$

We may also take $p_0(Y)$ as uniform distribution within a bounded range, then $\mathcal{E}(Y; \theta) = -f(Y; \theta)$. In this paper, we only consider using Gaussian distribution for the reference distribution $p_0(Y)$. The full potential of the energy-based model lies in the structure of the energy landscape defined by $\mathcal{E}(Y; \theta)$, which maps each of the observed examples to an energy value. This energy value serves as a measure of the “goodness” of a configuration of the input variable Y . A well trained model should assign lower energy values (i.e., high probabilities) to the observed examples than the unobserved examples. If the training data are highly varied, the learned energy landscape is likely to be multimodal. The local energy minima [64] of equation (2) satisfy the Hopfield auto-encoder [47]

$$\frac{Y}{s^2} = \frac{\partial}{\partial Y} f(Y; \theta), \quad (3)$$

where $\frac{\partial}{\partial Y} f(Y; \theta)$ can be computed by feed forward computation (bottom-up encoding process) and back-propagation (top-down decoding process). Such a model can be used for associative memory [64], where given an incomplete, occluded, or corrupted memory (i.e., data in the high energy regions), diffusion along the energy manifold starting from the high energy memory can gradually recover the missing memory until it arrives at the low energy regions.

2.2 Analysis by synthesis

The maximum likelihood estimation (MLE) of the energy-based generative VoxelNet follows an “analysis by synthesis” scheme. Suppose we observe 3D training examples $\{Y_i, i = 1, \dots, n\}$ from an unknown data distribution $P_{\text{data}}(Y)$. The MLE seeks to find θ to maximize the log-likelihood function

$$L(\theta) = \frac{1}{n} \sum_{i=1}^n \log p(Y_i; \theta). \quad (4)$$

The gradient of the $L(\theta)$ with respect to θ is

$$L'(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} f(Y_i; \theta) - \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} f(Y; \theta) \right], \quad (5)$$

where \mathbb{E}_θ denotes the expectation with respect to $p(Y; \theta)$. The expectation term in equation (5) is due to $\frac{\partial}{\partial \theta} \log Z(\theta) = \mathbb{E}_\theta[\frac{\partial}{\partial \theta} f(Y; \theta)]$, which is analytically intractable and has to be approximated by MCMC, such as Langevin dynamics, which iterates the following step:

$$\begin{aligned} Y_{\tau+\Delta\tau} &= Y_\tau - \frac{\Delta\tau}{2} \frac{\partial}{\partial Y} \mathcal{E}(Y_\tau; \theta) + \sqrt{\Delta\tau} \epsilon_\tau \\ &= Y_\tau - \frac{\Delta\tau}{2} \left[\frac{Y_\tau}{s^2} - \frac{\partial}{\partial Y} f(Y_\tau; \theta) \right] + \sqrt{\Delta\tau} \epsilon_\tau, \end{aligned} \quad (6)$$

where τ indexes the time steps of the Langevin dynamics, $\Delta\tau$ is the discretized step size, and $\epsilon_\tau \sim \mathcal{N}(0, I)$ is the Gaussian white noise term. The Langevin dynamics consists of a deterministic part, which is a gradient descent on the landscape defined by $\mathcal{E}(Y; \theta)$, and a stochastic part, which is a Brownian motion that helps the chain to escape spurious local minima of the energy $\mathcal{E}(Y; \theta)$. A Metropolis-Hastings step can be added to correct for the finiteness of $\Delta\tau$. Comparing with Gibbs sampling which does not take into account the landscape geometry, the gradient-based Langevin sampling converges quickly and is computationally feasible.

Suppose we draw \tilde{n} samples $\{\tilde{Y}_i, i = 1, \dots, \tilde{n}\}$ from the distribution $p(Y; \theta)$ by running \tilde{n} parallel chains of Langevin dynamics according to (6). The gradient of the log-likelihood $L(\theta)$ can be approximated by

$$L'(\theta) \approx \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} f(Y_i; \theta) - \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \frac{\partial}{\partial \theta} f(\tilde{Y}_i; \theta), \quad (7)$$

which is the difference between the observed examples and the synthesized examples. The above equation leads to the “analysis by synthesis” learning scheme. At t -th iteration, we sample $Y_i \sim p(Y; \theta^{(t)})$ for $i = 1, \dots, \tilde{n}$. Then we update the model parameters via $\theta^{(t+1)} = \theta^{(t)} + \gamma_t L'(\theta^{(t)})$, where γ_t is the learning rate.

If the sample size n is large, the maximum likelihood estimator is equivalent to minimizing $\text{KL}(P_{\text{data}}(Y) \parallel p(Y; \theta))$, the Kullback-Leibler (KL) divergence from the data distribution $P_{\text{data}}(Y)$ to the model distribution $p(Y; \theta)$. That is,

$$\begin{aligned} &\min_{\theta} \text{KL}(P_{\text{data}}(Y) \parallel p(Y; \theta)) \\ &= \min_{\theta} \{ \mathbb{E}_{P_{\text{data}}} [\log P_{\text{data}}(Y)] - \mathbb{E}_{P_{\text{data}}} [\log p(Y; \theta)] \} \\ &= \max_{\theta} \mathbb{E}_{P_{\text{data}}} [\log p(Y; \theta)] \\ &\approx \max_{\theta} \frac{1}{n} \sum_{i=1}^n \log p(Y_i; \theta) \end{aligned} \quad (8)$$

2.3 Mode seeking and mode shifting

The $f(Y; \theta)$ or equivalently the energy function $\mathcal{E}(Y; \theta)$ parameterized by the VoxelNet can be flexible enough to create many local modes to fit P_{data} that is likely to be highly multimodal. The energy function is reshaped by the “analysis by synthesis” learning algorithm in order to put lower energy values on the observed examples than the unobserved examples. This process can be interpreted as density shifting or mode shifting. We can rewrite equation (7) in the form of

$$L'(\theta) \approx \frac{\partial}{\partial \theta} \left[\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \mathcal{E}(\tilde{Y}_i; \theta) - \frac{1}{n} \sum_{i=1}^n \mathcal{E}(Y_i; \theta) \right]. \quad (9)$$

We define a value function

$$V(\{\tilde{Y}_i\}; \theta) = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \mathcal{E}(\tilde{Y}_i; \theta) - \frac{1}{n} \sum_{i=1}^n \mathcal{E}(Y_i; \theta). \quad (10)$$

The equation (9) reveals that the gradient of the log-likelihood $L(\theta)$ coincides with the gradient of V .

The sampling step in (6) can be interpreted as mode seeking, by finding low energy modes or high probability modes in the landscape defined by $\mathcal{E}(Y; \theta)$ via stochastic gradient descent (Langevin dynamics) and placing the synthesized examples around the modes, so that $\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \mathcal{E}(\tilde{Y}_i; \theta)$ tends to be low. It seeks to decrease V .

The learning step can be interpreted as mode shifting (as well as mode creating and mode sharpening) by shifting the low energy modes from the synthesized examples $\{\tilde{Y}_i\}$ toward the observed examples $\{Y_i\}$ by modifying the energy function $\mathcal{E}(Y; \theta)$ in order to increase the difference between the synthesized statistics and the observed statistics, i.e., $\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \mathcal{E}(\tilde{Y}_i; \theta) - \frac{1}{n} \sum_{i=1}^n \mathcal{E}(Y_i; \theta)$. It seeks to increase V .

Even though the learning and sampling algorithm creates and shifts the modes of the energy landscape to the observed examples, there might be still numerous major modes that are not occupied by the observed examples, and these modes represent new examples that are considered similar to the observed examples. Accessing these modes corresponds to synthesizing new examples. While the maximum likelihood learning matches the average statistical properties between model and observed data, the ConvNet $f(Y; \theta)$ is sufficiently flexible and expressive to create modes to encode the highly varied 3D patterns. An illustration of mode seeking and mode shifting is presented in Figure 1. The training algorithm of the generative VoxelNet is presented in Algorithm 1.

Algorithm 1 Generative VoxelNet

Input:

- (1) 3D training data $\{Y_i, i = 1, \dots, n\}$;
- (2) the number of Langevin steps K ;
- (3) the number of learning iterations T .

Output:

- (1) estimated parameters θ ;
- (2) synthesized examples $\{\tilde{Y}_i, i = 1, \dots, \tilde{n}\}$.

- 1: Let $t \leftarrow 0$, randomly initialize $\theta^{(t)}$ and $\{\tilde{Y}_i, i = 1, \dots, \tilde{n}\}$ with Gaussian noise.
 - 2: **repeat**
 - 3: **Mode seeking:** For each i , run K steps of Langevin updates to revise \tilde{Y}_i , i.e., starting from the current \tilde{Y}_i , each step follows equation (6).
 - 4: **Mode shifting:** Update $\theta^{(t+1)} = \theta^{(t)} + \gamma_t L'(\theta^{(t)})$, with learning rate γ_t , where $L'(\theta^{(t)})$ is computed according to equation (7).
 - 5: Let $t \leftarrow t + 1$
 - 6: **until** $t = T$
-

2.4 Alternating back-propagation

Both mode seeking (sampling) and mode shifting (learning) steps involve the derivatives of $f(Y; \theta)$ with respect to Y and θ respectively. Both derivatives can be computed efficiently by back-propagation, and share most of their steps. Specifically, for the bottom-up ConvNet structure $f(Y; \theta)$, the chain rule computation

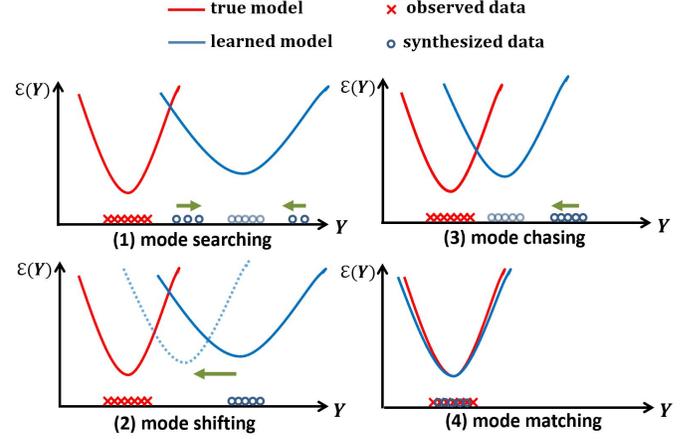


Fig. 1: An illustration of mode seeking and mode shifting. The red curves and blue curves represent the energy landscapes of the true model and the learned model, respectively. The red crosses indicate observed data, and the blue circles indicate the synthesized data. (1) mode searching by Langevin dynamics: finding low energy modes in the energy landscape of the model and placing the synthesized examples around the found modes. (2) model shifting (and sharpening) by model update: shifting the low energy modes toward the observed examples. (3) mode chasing: researching low energy modes in the updated energy landscape of the model. (4) mode matching: the modes in the learned energy landscape eventually match the ones in the data distribution after a few iterations of mode searching and shifting.

of $\partial h^{(l)} / \partial h^{(l-1)}$ for $l = 1, \dots, L$ is shared in calculating $\partial f(Y; \theta) / \partial Y$ and $\partial f(Y; \theta) / \partial \theta$ in terms of coding. The algorithm is thus in the form of alternating back-propagation that iterates the following two steps: (1) Sampling back-propagation: Revise the synthesized examples by Langevin dynamics or gradient descent. (2) Learning back-propagation: Update the model parameters given the synthesized and the observed examples by gradient ascent.

2.5 Zero temperature limit

For notation simplicity, let p_θ be the model distribution, and let p_τ be the distribution of Y_τ of the Langevin dynamics, then $\text{KL}(p_\tau \parallel p_\theta)$ decreases to zero monotonically over τ according to the second law of thermodynamics [65]. In other words, the Langevin dynamics implements a variational approximation to p_θ with p_τ . $\text{KL}(p_\tau \parallel p_\theta) = -\text{entropy}(p_\tau) + \mathbb{E}_{p_\tau}[\mathcal{E}(Y_\tau; \theta)] + \log Z(\theta)$. The gradient descent part of the Langevin dynamics reduces the energy $\mathbb{E}_{p_\tau}[\mathcal{E}(Y_\tau; \theta)]$, while the noise term of the Langevin dynamics increases the entropy of p_τ .

We can add a temperature term to the model $p_T(Y; \theta) = \exp(-\mathcal{E}(Y; \theta)/T) / Z_T(\theta)$, where the original model corresponds to $T = 1$. At zero temperature limit as $T \rightarrow 0$, the Langevin sampling will become gradient descent where the noise term diminishes in comparison to the gradient descent term. The resulting algorithm approximately solves the minimax problem

$$\max_{\theta} \min_{\{\tilde{Y}_i\}} V(\{\tilde{Y}_i\}; \theta), \quad (11)$$

with \tilde{Y}_i initialized from an initial distribution and approaching local modes of V . We can regularize either the diversity of $\{\tilde{Y}_i\}$ or the smoothness of $\mathcal{E}(Y; \theta)$. This is an adversarial interpretation of the

learning algorithm. It is also a generalized version of herding [66]. It is also related to Wasserstein GAN [67], but the critic and the actor are the same model, i.e., the model itself is its own generator and critic.

In our experiments, we find that disabling the noise term of the Langevin dynamics in the later stage of the learning process often leads to better synthesis. Ideally the learning algorithm should create a large number of local modes with similar low energies to capture the diverse observed examples as well as unseen examples.

3 CONDITIONAL LEARNING FOR RECOVERY

The conditional distribution $p(Y|C(Y) = c; \theta)$ can be derived from $p(Y; \theta)$. This conditional form of the generative VoxelNet can be used for recovery tasks such as inpainting and super resolution. In inpainting, $C(Y)$ consists of the visible part of Y . In super resolution, $C(Y)$ is the low resolution version of Y . For such tasks, we can learn the model from the fully observed training data $\{Y_i, i = 1, \dots, n\}$ by maximizing the conditional log-likelihood

$$L(\theta) = \frac{1}{n} \sum_{i=1}^n \log p(Y_i | C(Y_i) = c_i; \theta), \quad (12)$$

where c_i is the observed value of $C(Y_i)$. The learning and sampling algorithm is essentially the same as maximizing the original log-likelihood, except that in the Langevin sampling step, we need to sample from the conditional distribution, which amounts to fixing $C(Y_\tau)$ in the sampling process. The zero temperature limit (with the noise term in the Langevin dynamics disabled) approximately solves the following minimax problem

$$\max_{\theta} \min_{\{\tilde{Y}_i: C(\tilde{Y}_i) = c_i\}} V(\{\tilde{Y}_i\}; \theta). \quad (13)$$

4 3D MULTI-GRID MODELING AND SAMPLING

The maximum likelihood learning of the generative VoxelNet requires MCMC sampling, such as Langevin dynamics, of the 3D objects, which can be expensive and time consuming, because it takes a long time to converge, if the learned $p(Y; \theta)$ is multimodal. This often happens because P_{data} is usually highly varied and multimodal. Therefore, it becomes the bottleneck and the difficulty of learning the energy-based model in the ‘‘analysis by synthesis’’ scheme. In this section, inspired by [26], we will propose a 3D multi-grid modeling and sampling framework as a modified contrastive divergence to train generative VoxelNet models.

4.1 Contrastive divergence

Contrastive divergence (CD) method proposed in [68] generates synthesized examples by finite steps of MCMC initialized from the observed examples, and updates the model parameters with the gradient computed by equation (7). The CD learning approximately minimizes the difference between two KL divergences:

$$\text{KL}(P_{\text{data}} \parallel p_{\theta}) - \text{KL}(M_{\theta} P_{\text{data}} \parallel p_{\theta}), \quad (14)$$

where M_{θ} is the transition kernel of the finite-step MCMC that samples from $p_{\theta}(Y)$, and $M_{\theta} P_{\text{data}}(Y') = \int P_{\text{data}}(Y) M_{\theta}(Y'|Y) dY$ is the distribution after running M_{θ} starting from P_{data} .

The contrastive divergence (CD) learning method is related to score matching estimator [69], [70] and auto-encoder [71], [72], [73]. This method enables the model to learn from large

training datasets via mini-batch training. However, the synthesized examples obtained by CD learning may be distant from fair samples of the current model, thus bringing in bias into the learned model parameters.

An improvement of CD is persistent CD [74], where MCMC chains are still initialized from the observed examples at the initial epoch, but contrary to normal CD, in each subsequent epoch, the finite-step MCMC chains are initialized from the corresponding synthesized examples obtained at the previous epoch instead of starting over. However, it is still difficult for the persistent CD learning method to traverse different modes of the learned model.

4.2 Modified contrastive divergence by multi-grid learning and sampling

For a 3D volumetric shape Y , let $(Y^{(s)}, s = 0, \dots, S)$ denote the multi-grid versions of Y , with $Y^{(0)}$ representing the minimal $1 \times 1 \times 1$ version of Y , and $Y^{(S)}$ representing Y . For each version $Y^{(s)}$, we can divide the 3D voxel grid into cubic blocks of $d \times d \times d$ voxels, and reduce each $d \times d \times d$ block into a single voxel by a down-scaling operation that averages the values of the $d \times d \times d$ voxels of $Y^{(s)}$ to obtain $Y^{(s-1)}$. Conversely, we can also define an up-scaling operation which expands each voxel of $Y^{(s-1)}$ into a $d \times d \times d$ block of constant value to obtain an up-scaled version $\tilde{Y}^{(s)}$ of $Y^{(s-1)}$. Note that the up-scaled $\tilde{Y}^{(s)}$ is not the same as the original $Y^{(s)}$ because the high resolution details are lost after down-scaling and up-scaling. The down-scaling mapping from $Y^{(s)}$ to $Y^{(s-1)}$ is a linear projection onto a set of orthogonal basis vectors, each of which is specified by a $d \times d \times d$ block, while the up-scaling operation is a pseudo-inverse of this linear mapping.

Our multi-grid method learns a separate generative VoxelNet model at each grid. Let $p^{(s)}(Y^{(s)}; \theta^{(s)})$ denote the generative VoxelNet at grid s . We can simply model $p^{(0)}$ by a one dimensional histogram of $Y^{(0)}$, which is pooled from the $1 \times 1 \times 1$ versions of Y . Within each iteration of the learning algorithm, for each observed 3D training example Y_i , we generate the corresponding synthesized examples at multiple grids, i.e., $(\tilde{Y}_i^{(s)}, s = 0, \dots, S)$. Specifically, we initialize the finite-step MCMC sampling from the minimal $1 \times 1 \times 1$ version $Y_i^{(0)}$, and the up-scaled version of $\tilde{Y}_i^{(s-1)}$ sampled from the model $p^{(s-1)}(Y^{(s-1)}; \theta^{(s-1)})$ at the previous coarser grid serves to initialize the finite-step MCMC that samples from the model $p^{(s)}(Y^{(s)}; \theta^{(s)})$ at the subsequent finer grid. The update of the model parameter $\theta^{(s)}$ for each grid follows equation (7), which is based on the difference between the synthesized $\tilde{Y}^{(s)}$ and the observed $Y^{(s)}$. Algorithm 2 presents a full description of the 3D multi-grid learning and sampling algorithm.

The learning gradient of model at grid s approximately minimizes the difference between two KL divergences:

$$\text{KL}(P_{\text{data}}^{(s)} \parallel p_{\theta^{(s)}}^{(s)}) - \text{KL}(M_{\theta^{(s)}}^{(s)} P_{\theta^{(s-1)}}^{(s)} \parallel p_{\theta^{(s)}}^{(s)}), \quad (15)$$

where $M_{\theta^{(s)}}^{(s)}$ is transition kernel of the finite-step Langevin dynamics that samples from $p_{\theta^{(s)}}^{(s)}$. $P_{\theta^{(s-1)}}^{(s)}$ is the up-scaled version of the model $p_{\theta^{(s-1)}}^{(s-1)}$. That is, $P_{\theta^{(s-1)}}^{(s)}$ is the distribution of $\tilde{Y}^{(s)}$, which is the up-scaled version of $Y^{(s-1)}$ that follows distribution $p_{\theta^{(s-1)}}^{(s-1)}$. $P_{\theta^{(s-1)}}^{(s)}$ is smoother than $p_{\theta^{(s)}}^{(s)}$, and the finite-step MCMC specified by $M_{\theta^{(s)}}^{(s)}$ will change $P_{\theta^{(s-1)}}^{(s)}$ into a distribution close to the target distribution $p_{\theta^{(s)}}^{(s)}$ by synthesizing details at the current resolution. Such a coarse-to-fine manner is related to super resolution discussed in Experiment 6.3. Equation (15) can also be regarded as a modified version of CD learning in equation (14).

The learned models are able to synthesize new examples from scratch, because we only need to initialize the MCMC by sampling from the one-dimensional histogram pooled from the $1 \times 1 \times 1$ versions of the training examples.

Algorithm 2 3D multi-grid learning and sampling

Input:

- (1) 3D training data $\{Y_i^{(s)}, s = 1, \dots, S; i = 1, \dots, n\}$;
- (2) the number of Langevin steps K ;
- (3) the number of learning iterations T .

Output:

- (1) estimated parameters $(\theta^{(s)}, s = 1, \dots, S)$;
- (2) synthesized examples $\{\tilde{Y}_i^{(s)}, s = 1, \dots, S; i = 1, \dots, n\}$.

- 1: Let $t \leftarrow 0$. For $s = 1, \dots, S$, initialize $\theta_{(t)}^{(s)}$ with Gaussian noise.
 - 2: **repeat**
 - 3: For $i = 1, \dots, n$, initialize $\tilde{Y}_i^{(0)} = Y_i^{(0)}$
 - 4: For $s = 1, \dots, S$, initialize $\tilde{Y}_i^{(s)}$ as the up-scaled version of $\tilde{Y}_i^{(s-1)}$, and run K steps of the Langevin dynamics to evolve $\tilde{Y}_i^{(s)}$, each step following equation (6).
 - 5: For $s = 1, \dots, S$, update $\theta_{(t+1)}^{(s)} = \theta_{(t)}^{(s)} + \gamma_t L'(\theta_{(t)}^{(s)})$, with learning rate γ_t , where $L'(\theta_{(t)}^{(s)})$ is computed according to equation (7).
 - 6: Let $t \leftarrow t + 1$
 - 7: **until** $t = T$
-

5 TEACHING 3D GENERATOR NET

We can let a 3D generator network [19] learn from the MCMC sampling of the generative VoxelNet, so that the 3D generator network can be used as a much more efficient approximate non-iterative direct sampler of the generative VoxelNet.

5.1 3D generator model

The 3D generator model is a 3D non-linear multi-layer generalization of the traditional factor analysis model. The generator model has the following form

$$\begin{aligned} Z &\sim \mathcal{N}(0, I_d); \\ Y &= g(Z; \alpha) + \epsilon; \epsilon \sim \mathcal{N}(0, \sigma^2 I_D), \end{aligned} \quad (16)$$

where Z is a d -dimensional vector of latent factors that follow $\mathcal{N}(0, I_d)$ independently, and the 3D object Y is generated by first sampling Z from its known prior distribution $\mathcal{N}(0, I_d)$ and then transforming Z to the D -dimensional Y by a top-down 3D deconvolutional network $g(Z; \alpha)$ plus the white noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I_D)$, where σ is the standard deviation and $d < D$. $g(Z; \alpha)$ is a composition of L_g layers of 3D volumetric deconvolutions, upsamplings, and non-linear rectifications (e.g., ReLU), as illustrated by the following diagram:

$$Z \rightarrow Z^{(L_g)} \rightarrow \dots \rightarrow Z^{(l+1)} \rightarrow Z^{(l)} \rightarrow \dots \rightarrow Z^{(1)} \rightarrow g(Z; \alpha),$$

where $Z^{(l)}$ is the hidden output that is computed recursively by performing 3D deconvolution on $Z^{(l+1)}$ followed by ReLUs. α denotes all weight and bias parameters of the 3D top-down ConvNet g . The generator model is a directed graphical model such that it can readily generate a new example Y by ancestral sampling.

The 3D generator model can be trained by a maximum likelihood algorithm studied in [75], without resorting to any assisting networks. The training algorithm alternates two steps: (1) For each observed example, inferring the latent factors by sampling from the current posterior distribution via MCMC, and (2) Updating the model parameters by a non-linear regression of the observed examples on their inferred latent factors, so that the learned parameters enable better reconstructions of the observed examples by the inferred latent factors. The algorithm is also an alternating back-propagation (ABP) algorithm, because both two steps can be powered by back-propagation.

The ABP algorithm requires MCMC sampling to infer latent factors in training 3D generator model. It is possible to avoid MCMC sampling by using either variational inference [18] or adversarial learning [19] where an auxiliary network is recruited to relieve the burden of MCMC-based iterative inference.

As to variational learning, an inference model is simultaneously learned to replace the MCMC sampling of the latent factors. The inference model is an analytically tractable approximation of the posterior distribution. The objective of variational learning is an upper bound of the MLE objective. Consequently, the accuracy of the variational learning as an approximation to the MLE depends on the accuracy of the inference model as an approximation to the posterior distribution.

Adversarial learning provides an alternative to maximum likelihood techniques. A discriminative model is recruited and simultaneously trained with the generator in a minimax two-player game, where the generator seeks to generate “fake” examples that can fool the discriminator, while the discriminator seeks to distinguish the “fake” examples and the real training examples. However, adversarial learning may commonly encounter mode collapse issue, where the generator only outputs samples from a single mode.

5.2 MCMC teaching of 3D generator net

The 3D generator can be trained simultaneously with the generative VoxelNet in a cooperative training scheme [25], [76]. The basic idea is to use the 3D generator to generate examples to initialize a finite-step Langevin dynamics for training the generative VoxelNet. In return, the 3D generator learns from how the Langevin dynamics changes the initial examples it generates. In this scheme, the generative VoxelNet is the teacher network, and the 3D generator model is the student network. The generative VoxelNet teaches the 3D generator via a finite-step MCMC. We call it MCMC teaching. In this cooperative learning process, the generative VoxelNet learns from the real data, while the 3D generator learns from the MCMC sampling of the generative VoxelNet.

Specifically, in each iteration, (1) We generate Z_i from its known prior distribution $\mathcal{N}(0, I_d)$, and then generate the initial synthesized examples by $\hat{Y}_i = g(Z_i; \alpha) + \epsilon_i$ for $i = 1, \dots, \tilde{n}$. (2) Starting from the initial examples $\{\hat{Y}_i\}$, we sample from the generative VoxelNet by running a finite number of steps of Langevin dynamics to obtain the revised synthesized examples $\{\tilde{Y}_i\}$. (3) We then update the parameters θ of the generative VoxelNet based on $\{\tilde{Y}_i\}$ according to equation (7), and update the parameters α of the 3D generator by gradient descent

$$\Delta \alpha \propto - \frac{\partial}{\partial \alpha} \left[\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \|\tilde{Y}_i - g(Z_i; \alpha)\|^2 \right]. \quad (17)$$

We call it MCMC teaching because the revised examples $\{\tilde{Y}_i\}$ generated by the finite-step MCMC are used to teach $g(Z; \alpha)$. For each \tilde{Y}_i , the vector of latent factors Z_i is known to the 3D generator, so that there is no need to infer Z_i , and the learning becomes a much simpler supervised learning problem. Initially the 3D generator maps $\{Z_i\}$ to $\{\hat{Y}_i\}$. After MCMC teaching, the 3D generator seeks to map $\{Z_i\}$ to $\{\tilde{Y}_i\}$. That is, the MCMC teaching enables the 3D generator to absorb and accumulate the MCMC transitions for sampling the generative VoxelNet, and shift its density from $\{\hat{Y}_i\}$ to $\{\tilde{Y}_i\}$ for the sake of reproducing the effect of all the past MCMC transitions by one-step direct ancestral sampling.

Algorithm 3 presents a full description of the learning of the generative VoxelNet with a 3D generator as a sampler.

5.3 Convergence analysis

We briefly discuss the convergence of the cooperative learning algorithm. In such a cooperative learning scheme, the training of the generative VoxelNet follows the modified contrastive divergence:

$$\theta_{t+1} = \arg \min_{\theta} [\text{KL}(P_{\text{data}} \parallel p_{\theta}) - \text{KL}(M_{\theta_t} q_{\alpha} \parallel p_{\theta})], \quad (18)$$

where M_{θ_t} is the transition kernel of the finite-step MCMC samples from p_{θ_t} at t -th iteration, q_{α} is the distribution of the generator, and $M_{\theta_t} q_{\alpha}$ denotes the marginal distribution obtained after running M_{θ_t} starting from q_{α} . The training of the 3D generator follows

$$\alpha_{t+1} = \arg \min_{\alpha} \text{KL}(M_{\theta} q_{\alpha_t} \parallel q_{\alpha}), \quad (19)$$

so that the learned q_{α} eventually is the stationary distribution of M_{θ} , i.e., $q_{\alpha} = M_{\theta} q_{\alpha}$. Since the stationary distribution of M_{θ} is nothing else than p_{θ} , the learned q_{α} is equivalent to p_{θ} . Consequently, the second KL-divergence in (18) will become zero, and the learned θ is to minimize the first KL-divergence, i.e., $\text{KL}(P_{\text{data}} \parallel p_{\theta})$. In other words, θ is a maximum likelihood estimate. The learned α in this way is also a maximum likelihood estimate, because q_{α} traces the p_{θ} that runs towards the data distribution, i.e., $q_{\alpha} \rightarrow p_{\theta} \rightarrow P_{\text{data}}$.

Algorithm 3 MCMC teaching of 3D generator net

Input:

- (1) 3D training examples $\{Y_i, i = 1, \dots, n\}$;
- (2) the numbers of Langevin steps K ;
- (3) the number of learning iterations T .

Output:

- (1) estimated parameters θ and α ;
- (2) synthesized examples $\{\tilde{Y}_i, \tilde{Y}_i, i = 1, \dots, \tilde{n}\}$.

- 1: Let $t \leftarrow 0$, randomly initialize $\theta^{(t)}$ and $\alpha^{(t)}$ with Gaussian distribution.
 - 2: **repeat**
 - 3: **Initializing mode seeking:** For $i = 1, \dots, \tilde{n}$, generate $Z_i \sim \mathcal{N}(0, I_d)$, and generate $\tilde{Y}_i = g(Z_i; \alpha^{(t)}) + \epsilon_i$.
 - 4: **Mode seeking:** For $i = 1, \dots, \tilde{n}$, starting from \hat{Y}_i , run K steps of Langevin dynamics to obtain \tilde{Y}_i , each step following equation (6).
 - 5: **Mode shifting:** Update $\theta^{(t+1)} = \theta^{(t)} + \gamma_t L'(\theta^{(t)})$, where $L'(\theta^{(t)})$ is computed according to equation (7).
 - 6: **Learning from mode seeking:** Update $\alpha^{(t+1)}$ according to equation (17).
 - 7: Let $t \leftarrow t + 1$
 - 8: **until** $t = T$
-

6 EXPERIMENTS

Project page: The code and more results can be found at <http://www.stat.ucla.edu/~jxie/3DEBM/>

In this section, we conduct experiments to evaluate the energy-based generative VoxelNet from different aspects. We first show qualitative results of the synthesized 3D objects, and then we quantitatively evaluate the synthesis quality by Inception score, softmax class probability, and classification error. Further, we show that the conditional generative VoxelNet can be useful for 3D object recovery and 3D super resolution. Additionally, the 3D generator taught by the generative VoxelNet in a cooperative scheme is analyzed from the perspectives of shape synthesis, 3D shape interpolation and arithmetic in the latent space. We also show that the generative VoxelNet trained in an unsupervised manner can be used as a feature extractor for training a classifier with annotated data. Lastly, we show experimental results regarding high resolution 3D shape synthesis by the 3D multi-grid energy-based sampling and modeling strategy.

6.1 3D object synthesis

We conduct experiments on learning generative VoxelNet via algorithm 1 to synthesize 3D objects of categories from ModelNet [10], which is a large-scale 3D CAD model dataset. Specifically, we use ModelNet10, a 10-category subset of ModelNet which is commonly used as a benchmark for 3D object analysis. The categories are chair, sofa, bathtub, toilet, bed, desk, table, nightstand, dresser, and monitor. The size of the training set for each category ranges from 100 to 700. To obtain voxel-based training data, we voxelize each CAD model in the ModelNet dataset by the following method: Each 3D mesh is represented as a binary 3D voxel grid, with 1 indicating the voxel is inside the mesh surface, and 0 indicating the voxel is outside the mesh.

For qualitative experiment, we learn one 3-layer generative VoxelNet for each object category in ModelNet10. The first layer has 200 $16 \times 16 \times 16$ filters with sub-sampling of 3, the second layer has 100 $6 \times 6 \times 6$ filters with sub-sampling of 2, and the final layer is a fully connected layer with a single filter that covers the whole voxel grid. We add ReLU layers between adjacent convolutional layers. We fix the standard deviation of the reference distribution of the model to be $s = 0.5$. The number of Langevin dynamics steps in each learning iteration is $K=20$ and the step size $\Delta\tau = 0.01$. We use Adam [77] for optimization with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is 0.001. The number of learning iterations is 3,000. We disable the noise term in the Langevin step after 100 iterations. The training data are of size $32 \times 32 \times 32$ voxels, whose values are 0 or 1. We pre-process the training data by subtracting the mean value of the dataset. We post-process the synthesized data by adding back the mean value and discretizing each voxel value into 0 or 1 by comparing it with a threshold 0.5. The mini-batch size is 20. The number of parallel chains for each batch is 25.

Figure 2 displays the observed 3D objects randomly sampled from the training set, and the synthesized 3D objects generated by our models for categories chair, bed, sofa, table, dresser, night stand, toilet, and monitor. We visualize volumetric data via isosurfaces in our paper. To show that our model can synthesize new 3D objects beyond the training set, we compare the synthesized patterns with their nearest neighbors in the training set. The retrieved nearest neighbors are based on ℓ_2 distance in the voxel space. As shown in Figure 2, our model can synthesize realistic 3D shape patterns,

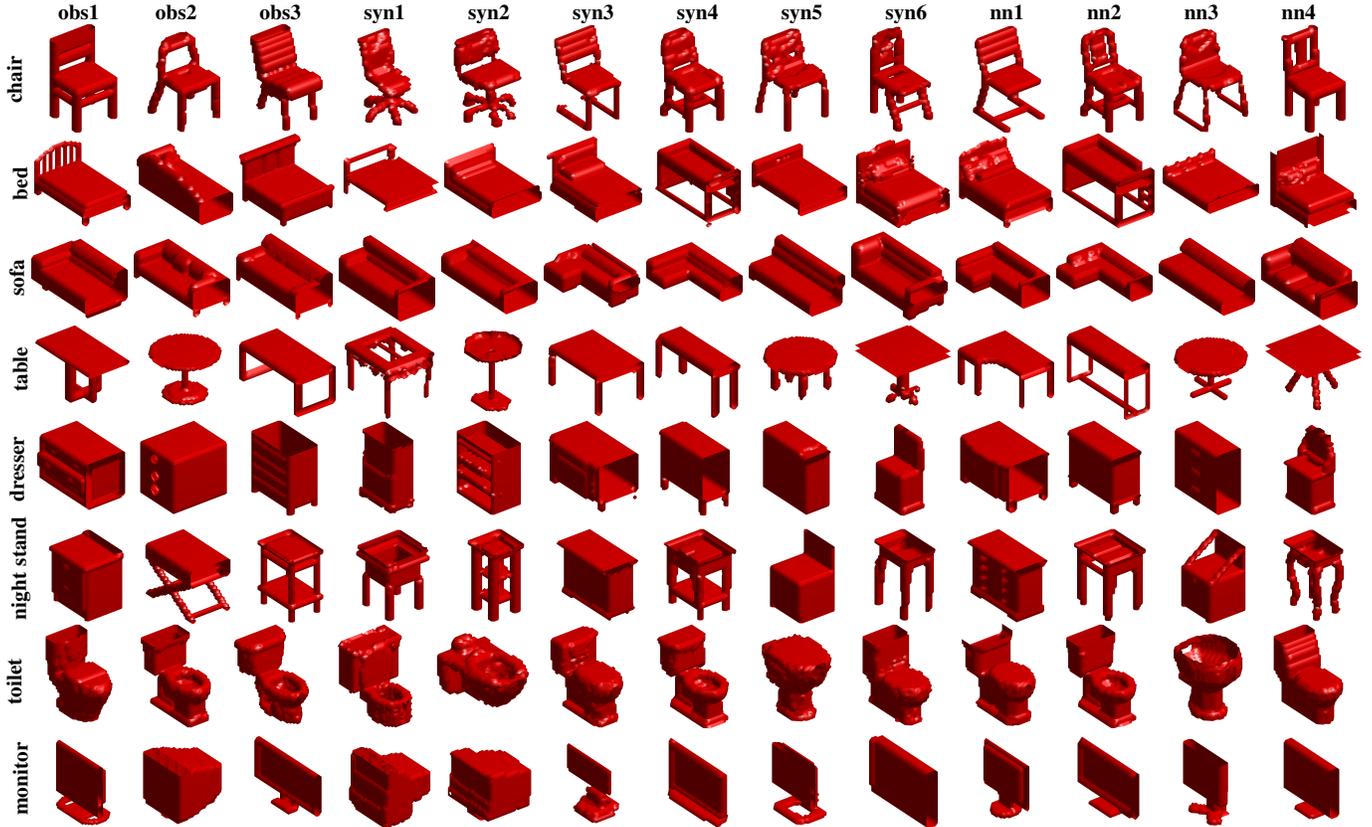


Fig. 2: Generating 3D objects. Each row displays one experiment, where the first three 3D objects are some observed examples, columns 4, 5, 6, 7, 8, and 9 are 6 of the synthesized 3D objects sampled from the learned model by Langevin dynamics. For the last four synthesized objects (shown in columns 6, 7, 8, and 9), their nearest neighbors retrieved from the training set are shown in columns 10, 11, 12, and 13.

and the generated 3D objects are similar, but not identical, to the training set.

To quantitatively evaluate our model, we adopt the Inception score proposed by [78], which uses a reference convolutional neural network to compute

$$I(\{\tilde{Y}_i, i = 1, \dots, \tilde{n}\}) = \exp\left(\mathbb{E}_{\tilde{Y}}\left[\text{KL}(p(c|\tilde{Y}) \parallel p(c))\right]\right),$$

where c denotes category, $\{\tilde{Y}_i, i = 1, \dots, \tilde{n}\}$ are synthesized examples sampled from the model, $p(c|\tilde{Y})$ is obtained from the output of the reference network, and $p(c) \approx \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} p(c|\tilde{Y}_i)$. Both a low entropy conditional category distribution $p(c|\tilde{Y})$ (i.e., the network classifies a given sample with high certainty) and a high entropy category distribution $p(c)$ (i.e., the network identifies a wide variety of categories among the generated samples) can lead to a high inception score. In our experiment, we use a state-of-the-art 3D multi-view ConvNet [15] trained on ModelNet dataset for 3D object classification as the reference network.

We firstly evaluate the generative VoxelNet learned from the training sets of 10-category mixed 3D objects. Table 1 reports the Inception scores of our model and some baseline models, including 3D ShapeNets [10], 3D-GAN [17], 3D-VAE [79], 3D-WINN [36], and Primitive GAN [34].

We then evaluate the quality of the synthesized 3D shapes by the model learned from a single category by using two criteria: (a) average softmax class probability that the reference network assigns to the synthesized examples for the underlying category, and (b) classification error by the reference network, i.e., the probability

TABLE 1: Inception scores of different models learned from 10 3D object categories.

Model	Inception score
3D ShapeNets [10]	4.126±0.193
3D GAN [17]	8.658±0.450
3D VAE [79]	11.015±0.420
3D WINN [36]	8.810±0.180
Primitive GAN [34]	11.520±0.330
generative VoxelNet (ours)	11.772±0.418

that the underlying category does not belong to the categories with the top softmax probability. Table 2 and 3 display the results for all 10 categories. It can be seen that our model generates 3D shape patterns with higher softmax class probabilities and lower classification errors than other baseline models. In other words, our model can generate more realistic examples than other methods, in the sense that the examples synthesized by our method are very vivid such that higher probabilities a state-of-the-art or well-trained classifier would assign for their underlying categories.

6.2 3D object recovery

We then test the conditional generative VoxelNet on the 3D object recovery task. For each testing 3D object, we randomly corrupt some voxels of the 3D object. We then seek to recover the corrupted voxels by sampling from the conditional distribution $p(Y_M|Y_{\bar{M}}; \theta)$ according to the learned model $p(Y; \theta)$, where M and \bar{M} denote the corrupted and uncorrupted voxels, and Y_M

TABLE 2: Softmax class probability

category	ours	[17]	[79]	[10]
bathhtub	0.8348	0.7017	0.7190	0.1644
bed	0.9202	0.7775	0.3963	0.3239
chair	0.9920	0.9700	0.9892	0.8482
desk	0.8203	0.7936	0.8145	0.1068
dresser	0.7678	0.6314	0.7010	0.2166
monitor	0.9473	0.2493	0.8559	0.2767
night stand	0.7195	0.6853	0.6592	0.4969
sofa	0.9480	0.9276	0.3017	0.4888
table	0.8910	0.8377	0.8751	0.7902
toilet	0.9701	0.8569	0.6943	0.8832
Avg.	0.8811	0.7431	0.7006	0.4596

TABLE 3: Classification errors on the synthesized examples

	ours	[17]	[79]	[10]
bathhtub	0.1226	0.2642	0.2642	0.8208
bed	0.0583	0.1961	0.5981	0.6583
chair	0.0050	0.0125	0.0100	0.1475
desk	0.0165	0.1900	0.1700	0.8900
dresser	0.0165	0.3000	0.2400	0.7950
monitor	0.0452	0.7484	0.1312	0.7204
night stand	0.2700	0.3000	0.3200	0.4800
sofa	0.0397	0.0471	0.6868	0.4912
table	0.0800	0.1333	0.1067	0.1833
toilet	0.0174	0.1233	0.2994	0.1047
Avg.	0.0968	0.2315	0.2826	0.5291

and $Y_{\bar{M}}$ are the corrupted part and the uncorrupted part of the 3D object Y respectively. The sampling of $p(Y_M|Y_{\bar{M}}; \theta)$ is again accomplished by the Langevin dynamics, which is the same as the Langevin dynamics that samples from the full distribution $p(Y; \theta)$, except that we fix the uncorrupted part $Y_{\bar{M}}$ and only update the corrupted part Y_M throughout the Langevin dynamics. In the learning stage, we learn the model from the fully observed training 3D objects. To specialize the learned model to this recovery task, we learn the conditional distribution $p(Y_M|Y_{\bar{M}}; \theta)$ directly. That is, in the learning stage, we also randomly corrupt each fully observed training 3D object Y , and run Langevin dynamics by fixing $Y_{\bar{M}}$ to obtain the synthesized 3D object. The parameters θ are then updated by gradient ascent according to equation (7). Also see Algorithm 4 for a description. We try two different network architectures in this experiment. The first one is a 3-layer network, which is the same as the one used in Section 6.1 for synthesis, and the other one is a 2-layer network consisting of 100 $16 \times 16 \times 16$ filters with sub-sampling of 3 in its first layer, followed by a fully connected layer with a single filter. The number of Langevin dynamics steps for recovery in each iteration is set to be $K = 90$ and the step size is $\Delta\tau = 0.0049$. The number of learning iterations is 1,000. The size of the mini-batch is 50. The 3D training data are of size $32 \times 32 \times 32$ voxels.

After learning the model, we recover the corrupted voxels in each testing data Y by sampling from $p(Y_M|Y_{\bar{M}}; \theta)$ by running 90 Langevin dynamics steps. In the training stage, we randomly corrupt 70% of each training 3D shape. In the testing stage, we experiment with the same percentage of corruption. We compare our methods with 3D-GAN and 3D ShapeNets.

The original 3D-GAN does not naturally support shape recovery. To use the 3D-GAN algorithm to recover corrupted voxels, we firstly train the 3D-GAN model from the complete training data, and then use the learned 3D generator to predict the corrupted voxels from the observed ones by adding an extra inference step.

Specifically, following [75], we derive the posterior distribution $p(Z|Y; \alpha)$ from the generator model, and perform inference by sampling from the posterior distribution via Langevin dynamic to infer the vector of latent factors for each corrupted example. The unobserved voxels can then be recovered from their inferred factors by a direct top-down generation. Only those corrupted voxels get updated by the generated voxels. We adopt the same 3D-GAN architecture as in [17].

We measure the recovery error by the average of per-voxel differences between the original testing data and the corresponding recovered data on the corrupted voxels. Table 4 displays the numerical comparison results for the 10 categories, where we use our-2 and our-3 to represent the 2-layer and the 3-layer models respectively. Our methods outperform the other baseline methods in terms of recovery error. Figure 3 displays some examples of 3D object recovery. For each experiment, the first row displays the original 3D objects, the second row displays the corrupted 3D objects, and the third row displays the recovered 3D objects that are sampled from the learned conditional distributions given the corrupted 3D objects as inputs.

TABLE 4: Recovery errors in occlusion experiments

category	ours-3	ours-2	[17]	[10]
bathhtub	0.0152	0.0316	0.0266	0.0621
bed	0.0068	0.0071	0.0240	0.0617
chair	0.0118	0.0126	0.0238	0.0444
desk	0.0122	0.019	0.0298	0.0731
dresser	0.0038	0.0076	0.0384	0.1558
monitor	0.0103	0.0133	0.0220	0.0783
night stand	0.0080	0.0091	0.0248	0.2925
sofa	0.0068	0.0071	0.0186	0.0563
table	0.0051	0.0064	0.0326	0.0340
toilet	0.0119	0.0129	0.0180	0.0977
Avg.	0.0092	0.0127	0.0259	0.0956

Algorithm 4 Conditional generative VoxelNet for 3D Recovery

Input:

- (1) 3D training data $\{Y_i, i = 1, \dots, n\}$;
- (2) percentage points of corruption ϱ ;
- (3) the number of Langevin steps K ;
- (4) the number of learning iterations T .

Output:

- (1) estimated parameters θ .

- 1: Let $t \leftarrow 0$.
 - 2: Randomly initialize $\theta^{(t)}$ with Gaussian distribution.
 - 3: Randomly corrupt ϱ percent of voxels of each training example and obtain a corrupted dataset $\{Y'_i, i = 1, \dots, n\}$ with corresponding uncorrupted voxels $\{\tilde{M}_i, i = 1, \dots, n\}$.
 - 4: **repeat**
 - 5: For each i , starting from the corrupted data Y'_i , run K steps of Langevin dynamics to obtain \tilde{Y}_i , where each step follows equation (6).
 - 6: Fixing the uncorrupted parts of voxels $\tilde{Y}_i(\tilde{M}_i) \leftarrow Y_i(\tilde{M}_i)$
 - 7: Update $\theta^{(t+1)} = \theta^{(t)} + \gamma_t L'(\theta^{(t)})$, with learning rate γ_t , where $L'(\theta^{(t)})$ is computed according to equation (7).
 - 8: Let $t \leftarrow t + 1$
 - 9: **until** $t = T$
-

6.3 3D object super resolution

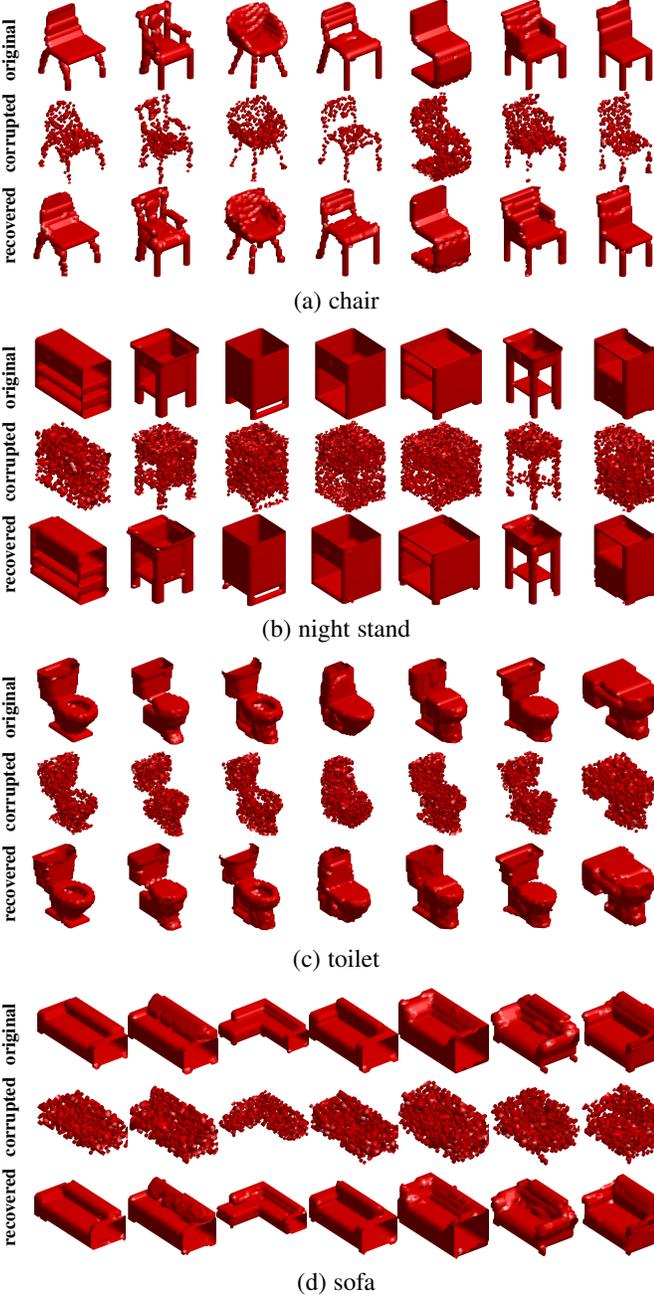


Fig. 3: 3D object recovery by sampling from the conditional generative VoxelNet models. In each category, the first row displays the original 3D objects, the second row shows the corrupted 3D objects, and the third row displays the recovered 3D objects by running Langevin dynamics starting from the corrupted objects. (a) chair (b) night stand (c) toilet (d) sofa.

We test the conditional generative VoxelNet on the 3D object super resolution task. Similar to Experiment 6.2, we can perform super resolution on a low resolution 3D object by sampling from a conditional generative VoxelNet $p(Y_{\text{high}}|Y_{\text{low}}; \theta)$, where Y_{high} denotes a high resolution version of Y_{low} . The sampling of the conditional model $p(Y_{\text{high}}|Y_{\text{low}}; \theta)$ is accomplished by the Langevin dynamics initialized with the given low resolution 3D object that needs to be super-resolved. In the learning stage, we learn the conditional model from the fully observed training 3D objects as well as their low resolution versions. To specialize the

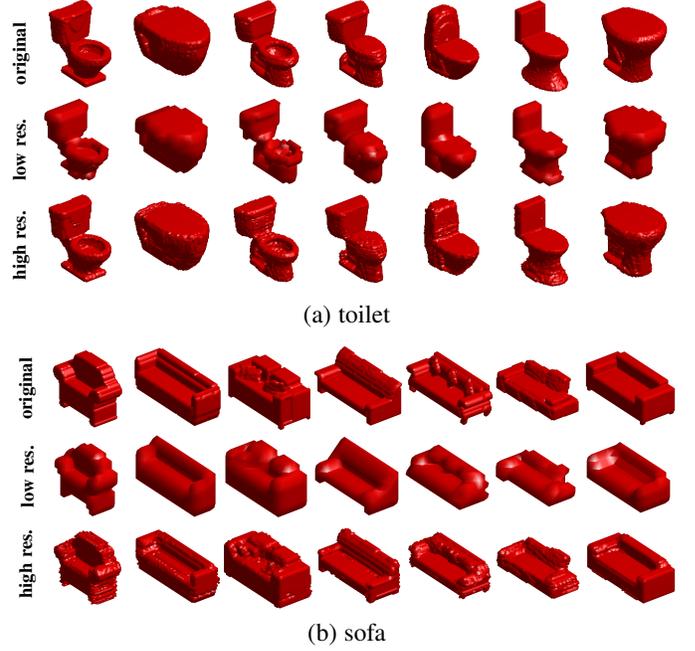


Fig. 4: 3D object super resolution by conditional generative VoxelNet. The first row displays some original 3D objects ($64 \times 64 \times 64$ voxels). The second row shows the corresponding low resolution 3D objects ($16 \times 16 \times 16$ voxels). The last row displays the corresponding super resolution results which are obtained by sampling from the conditional generative VoxelNet by running 10 steps of Langevin dynamics initialized with the objects shown in the second row.

learned model to this super resolution task, in the training process, we down-scale each fully observed training 3D object Y_{high} into a low resolution version Y_{low} , which leads to information loss. In each iteration, we first up-scale Y_{low} by expanding each voxel of Y_{low} into a $d \times d \times d$ block (where d is the ratio between the sizes of Y_{high} and Y_{low}) of constant values to obtain an up-scaled version Y'_{high} of Y_{low} (The up-scaled Y'_{high} is not identical to the original high resolution Y_{high} since the high resolution details are lost), and then run Langevin dynamics starting from Y'_{high} . The parameters θ are then updated by gradient ascent according to equation (7). Figure 4 shows some qualitative results of 3D super resolution, where we use a 2-layer conditional generative VoxelNet. The first layer has 200 $16 \times 16 \times 16$ filters with sub-sampling of 3. The second layer is a fully-connected layer with one single filter. The Langevin step size is $\Delta\tau = 0.0001$, and the number of Langevin steps is $K = 90$. To be more specific, let $Y_{\text{low}} = CY_{\text{high}}$, where C is the down-scaling matrix, e.g., each voxel of Y_{low} is the average of the corresponding $d \times d \times d$ block of Y_{high} . Let C^- be the pseudo-inverse of C , e.g., C^-Y_{low} gives us a high resolution shape by expanding each voxel of Y_{low} into a $d \times d \times d$ block of constant values. Then the sampling of $p(Y_{\text{high}}|Y_{\text{low}}; \theta)$ is similar to sampling the unconditioned model $p(Y_{\text{high}}; \theta)$, except that for each step of the Langevin dynamics, let ΔY be the change of Y , we update $Y \leftarrow Y + (I - C^-C)\Delta Y$, i.e., we project ΔY to the null space of C , so that the low resolution version of Y , i.e., CY , remains fixed. From this perspective, super resolution is similar to inpainting, except that the visible voxels are replaced by low resolution voxels. See Algorithm 5 for a detailed description.

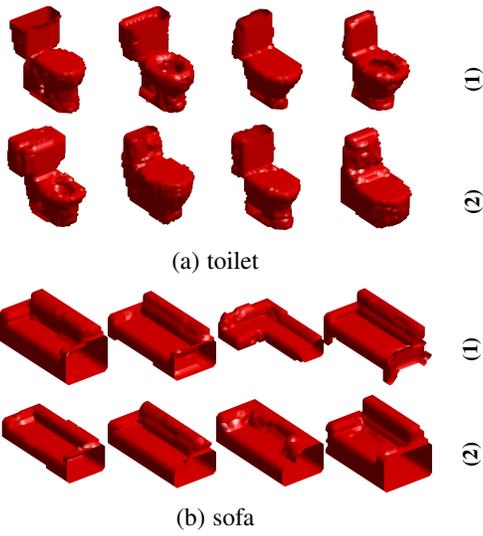


Fig. 5: Synthesis by 3D generators. (a) toilet (b) sofa. The 3D generators are trained by the generative VoxelNet via MCMC teaching. The 3D shapes are generated by first sampling Z from $\mathcal{N}(0, I_d)$ and then mapping Z to 3D shape space via $g(Z; \alpha)$.

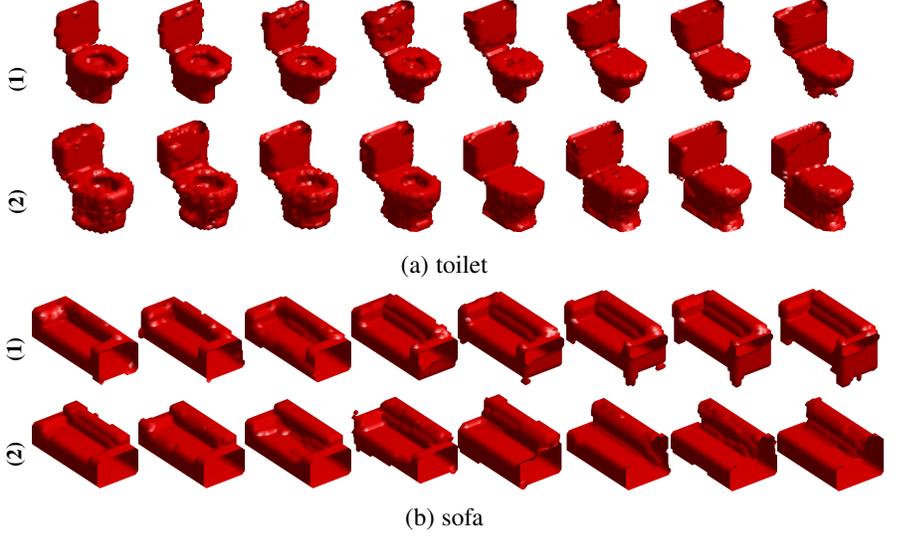


Fig. 6: Interpolation between latent vectors of the 3D objects on the two ends. (a) toilet (b) sofa. The 3D generators are trained via MCMC teaching on each category. For each row in each category, the 3D shapes are generated from the learned generator model $g(Z; \alpha)$ with linear interpolation in latent space (i.e., $(1 - \rho) \cdot Z_0 + \rho \cdot Z_1$), where Z_0 and Z_1 are latent vectors of the 3D objects on the two ends, ρ is discretized into 8 values within $[0, 1]$.

Algorithm 5 Conditional generative VoxelNet for 3D Super resolution

Input:

- (1) 3D training data $\{Y_i, i = 1, \dots, n\}$;
- (2) the number of Langevin steps K ;
- (3) the number of learning iterations T .

Output:

- (1) estimated parameters θ .

- 1: Let $t \leftarrow 0$.
 - 2: Randomly initialize $\theta^{(t)}$ with Gaussian distribution.
 - 3: Down-scale each training example via $Y_{\text{low}i} = CY_i$ and obtain a low resolution dataset $\{Y_{\text{low}i}, i = 1, \dots, n\}$.
 - 4: Compute up-scaled data $Y_{\text{high}i} \leftarrow C^- Y_{\text{low}i}$, for $i = 1, \dots, n$.
 - 5: **repeat**
 - 6: For $i = 1, \dots, n$, initialize via $\tilde{Y}_i \leftarrow Y_{\text{high}i}'$, run K steps of modified Langevin dynamics, where each step follows $\tilde{Y}_i \leftarrow \tilde{Y}_i + (I - C^-C)\Delta Y_i$, where ΔY_i is the change of \tilde{Y}_i by running one step of equation (6).
 - 7: Update $\theta^{(t+1)} = \theta^{(t)} + \gamma_t L'(\theta^{(t)})$, with learning rate γ_t , where $L'(\theta^{(t)})$ is computed according to equation (7).
 - 8: Let $t \leftarrow t + 1$
 - 9: **until** $t = T$
-

6.4 Analyzing the learned 3D generator trained by MCMC teaching

We evaluate a 3D generator trained by a generative VoxelNet via MCMC teaching. The generator network $g(Z; \alpha)$ has 4 layers of volumetric deconvolution with $4 \times 4 \times 4$ kernels, with up-sampling factors $\{1, 2, 2, 2\}$ at different layers respectively. The numbers of channels at different layers are 256, 128, 64, and 1. There is a fully connected layer under the 100 dimensional vector of latent factors Z . The output size is $32 \times 32 \times 32$. Batch normalization

and ReLU layers are used between adjacent deconvolution layers and tanh non-linearity is added at the bottom layer. We train a generative VoxelNet with the above 3D generator as a sampler in a cooperative training scheme presented in Algorithm 3 for the categories of toilet, sofa, and nightstand in ModelNet10 dataset independently. The generative VoxelNet has a 4-layer network, where the first layer has $64 \times 9 \times 9$ filters, the second layer has $128 \times 7 \times 7 \times 7$ filters, the third layer has $256 \times 4 \times 4 \times 4$ filters, and the fourth layer is a fully connected layer with a single filter. The sub-sampling factors are $\{2, 2, 2, 1\}$. ReLU layers are added between adjacent convolutional layers.

We use Adam for optimization of the generative VoxelNet model with $\beta_1 = 0.4$ and $\beta_2 = 0.999$, and 3D generator with $\beta_1 = 0.6$ and $\beta_2 = 0.999$. The learning rates for generative VoxelNet and 3D generator are 0.001 and 0.0003 respectively. The number of parallel chains is 50, and the mini-batch size is 50. The training data are scaled into the range of $[-1, 1]$. The synthesized data are re-scaled back into $[0, 1]$ for visualization. We run $K = 20$ steps of Langevin dynamics with a step size $\Delta\tau = 0.09$. Figure 5 shows some examples of 3D objects generated by the 3D generators trained by the generative VoxelNet via MCMC teaching on categories toilet and sofa.

We show results of interpolating between two latent vectors of Z in Figure 6. For each row in each category, the 3D objects at the two ends are generated from Z vectors that are randomly sampled from $\mathcal{N}(0, I_d)$. Each object in the middle is obtained by first interpolating the Z vectors of the two end objects, and then generating the objects using the 3D generator. We observe smooth transitions in 3D shape structures and that most intermediate objects are also physically plausible. This experiment demonstrates that the learned 3D generator embeds the 3D object distribution into a smooth low dimensional manifold. Another way to investigate the learned 3D generator is to show shape arithmetic in the latent space. As shown in Figure 7, the 3D generator is able to encode semantic knowledge of 3D shapes in its latent space such that arithmetic can

be performed on Z vectors for visual concept manipulation of 3D shapes. For example, the second row of images show the obtained “toilet lid” vector can be added to another toilet.

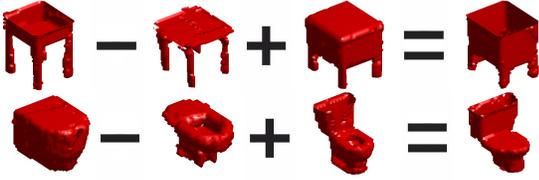


Fig. 7: 3D shape arithmetic in the latent space. Each row of images show that the obtained “part” vector can be added to a new object. The latent space is of 100 dimensions. The size of 3D shape is $32 \times 32 \times 32$ voxels.

6.5 3D object classification

We evaluate the feature maps learned by our generative VoxelNet model. We perform a classification experiment on ModelNet10 dataset. We first train a single model on all categories of the training set in an unsupervised manner. The network architecture and learning configuration are the same as the one used for synthesis in Section 6.1. Then we use the model as a feature extractor. Specifically, for each input 3D object, we use the model to extract its first and second layers of feature maps, apply max pooling of kernel sizes $4 \times 4 \times 4$ and $2 \times 2 \times 2$ respectively, and concatenate the outputs as a feature vector of length 8,100. We train a multinomial logistic regression classifier using labeled data based on the extracted feature vectors. We evaluate the classification accuracy of the classifier on the testing data using the one-versus-all rule. For the purpose of comparison, Table 5 lists 12 published results on this dataset obtained by other baseline methods. Our method outperforms the other methods in terms of classification accuracy on this dataset.

TABLE 5: 3D object classification on ModelNet10 dataset

Method	Accuracy
Geometry Image [57]	88.4%
PANORAMA-NN [59]	91.1%
ECC [61]	90.0%
3D ShapeNets [10]	83.5%
DeepPano [58]	85.5%
SPH [56]	79.8%
LFD [55]	79.9%
VConv-DAE [62]	80.5%
VoxNet [16]	92.0%
3D-GAN [17]	91.0%
3D-WINN [36]	91.9%
Primitive GAN [34]	92.2%
generative VoxelNet (ours)	92.4%

6.6 3D multi-grid modeling and sampling

We conduct qualitative experiments to evaluate the proposed 3D multi-grid sampling algorithm on the toilet and sofa categories in ModelNet10 dataset. We learn the models at 5 grids: $4 \times 4 \times 4$, $16 \times 16 \times 16$, $32 \times 32 \times 32$, $64 \times 64 \times 64$ and $128 \times 128 \times 128$, which we refer to as grid1, grid2, grid3, grid4 and grid5, respectively. The training 3D shapes are of size $128 \times 128 \times 128$ voxels. Since the models of the 5 grids act on 3D shapes of different resolutions, we design a specific network structure per grid: grid1 has a 2-layer

network, where the first layer has $128 \ 4 \times 4 \times 4$ filters with sub-sampling of 1 and the second layer is a fully-connected layer with 1 channel output; grid2 has a 2-layer network, where the first layer has $256 \ 8 \times 8 \times 8$ filters with a sub-sampling factor 2 and the second layer is a fully-connected layer with one single filter; grid3 has a 3-layer network, where the first layer has $256 \ 16 \times 16 \times 16$ filters with a sub-sampling factor 3, the second layer has $128 \ 6 \times 6 \times 6$ filters with a sub-sampling factor 2, and the third layer is a fully-connected layer with one single filter; grid4 has the same network structure as that of grid3; grid5 has a similar 3-layer network as that of grid4, except that the sub-sampling factor in the first layer is 4 and the filters in the second layer are of size $8 \times 8 \times 8$. For all grids, ReLU layers are added between adjacent layers. At each iteration, we run $K = 20$ steps of Langevin dynamics for each grid with a step size $\Delta\tau = 0.01$. All networks are trained simultaneously with mini-batches of size 40 and an initial learning rate of 0.001. Figure 8 displays the synthesized 3D shapes at different grids for toilet and sofa categories in ModelNet10 dataset. For each category, we show 5 examples of the 3D multi-grid sampling results. Each column corresponds to one example, in which synthesized results shown from top to bottom are generated by models from low resolution grid (i.e., $16 \times 16 \times 16$ voxels) to high resolution grid (i.e., $128 \times 128 \times 128$ voxels), respectively.

For comparison purpose, we also train a generative VoxelNet with a high resolution of $128 \times 128 \times 128$ voxels as a single-grid baseline. The model has a 5-layer network, where the first four layers have filter sizes $\{16, 8, 4, 4\}$, sub-sampling factors $\{4, 2, 2, 2\}$, and numbers of output channels $\{128, 128, 256, 512\}$, and the last layer is a fully-connected layer with one filter. The batch size is 20. We run $K = 20$ Langevin steps with a step size $\Delta\tau = 0.0025$. Figure 9 displays some synthesized examples with a resolution of $128 \times 128 \times 128$ voxels from toilet and sofa categories. To quantitatively compare the quality of the 3D shapes generated by the multi-grid sampling strategy and the original one, we calculate the “Fréchet Inception Distance” [80] (FID) score based on a reference network, which is a state-of-the-art discriminative neural network for 3D object classification [15]. FID is a metric that computes the distance between feature vectors extracted from observed and synthesized 3D objects by the reference network,

$$\text{FID} = \|\tilde{\mu} - \mu\|^2 + \text{Tr} \left(\tilde{\Sigma} + \Sigma - 2(\tilde{\Sigma}\Sigma)^{1/2} \right),$$

where $V \sim \mathcal{N}(\mu, \Sigma)$ and $\tilde{V} \sim \mathcal{N}(\tilde{\mu}, \tilde{\Sigma})$ are the feature vectors of the reference network for observed and generated samples respectively. A lower FID score means better quality of the synthesized 3D objects. We compare the multi-grid sampling strategy with the original one using single-grid sampling on synthesizing 3D shapes with a resolution of $128 \times 128 \times 128$ voxels in Table 6, where FID scores and computation times (in seconds) are reported. We generate 300 examples to compute the FID score. Although the computation time for each training epoch of the multi-grid model is longer than that of the single-grid one, the former model usually needs much less epochs to converge and is more stable than the latter one. In general, the multi-grid method outperforms the single-grid one in terms of sampling stability, computational efficiency, and synthesis quality.

The experiment demonstrates that the proposed multi-grid sampling method can learn realistic models of 3D objects in a coarse-to-fine scheme, and we hope that this experiment will push forward the development of efficient MCMC algorithms for learning high resolution deep 3D volumetric energy-based models.

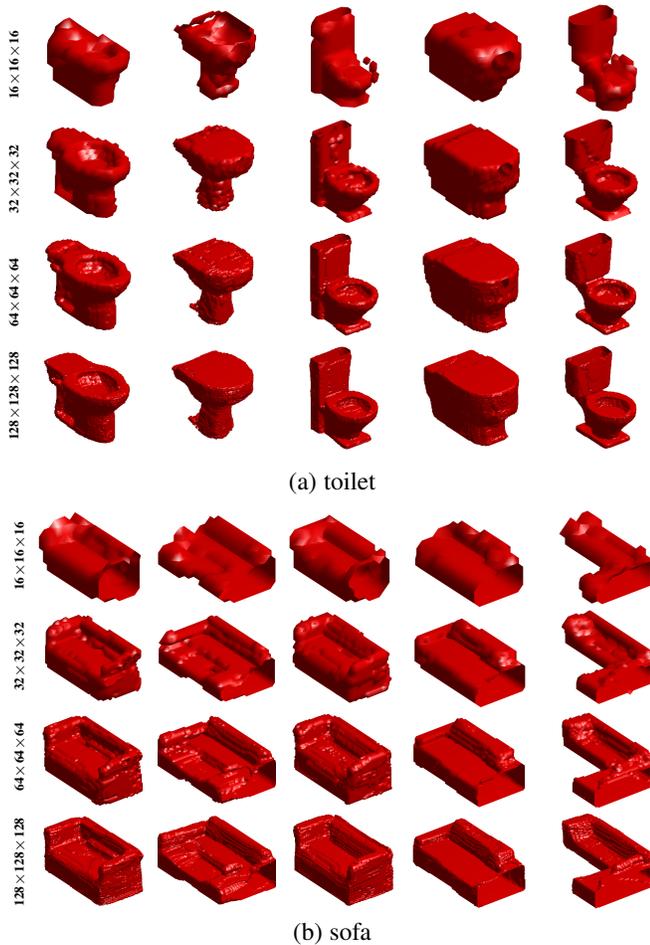


Fig. 8: Synthesized 3D shapes at multi-grids by 3D multi-grid sampling. (a) toilet (b) sofa. From top to bottom: $16 \times 16 \times 16$ grid, $32 \times 32 \times 32$ grid, $64 \times 64 \times 64$ grid and $128 \times 128 \times 128$ grid. Synthesized example at each grid is obtained by 20 steps Langevin sampling initialized from the synthesized examples at the previous coarser grid, starting from the $1 \times 1 \times 1$ grid.

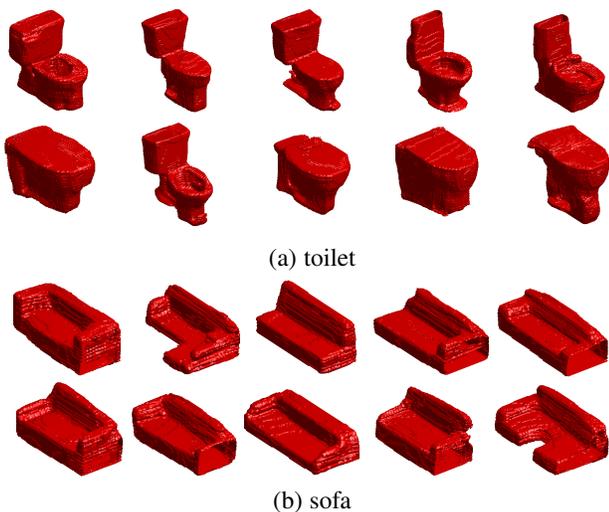


Fig. 9: Generating high resolution 3D objects with $128 \times 128 \times 128$ voxels by the generative VoxelNet. (a) toilet (b) sofa.

TABLE 6: A comparison of Fréchet Inception Distance (FID) scores and computation times (seconds per epoch \times number of epochs) between single-grid and multi-grid models on generating $128 \times 128 \times 128$ 3D shapes. The running times were recorded in a PC with an Intel i7-6700k CPU and a Titan Xp GPU.

Method	FID \downarrow		Time \downarrow (sec / epoch \times # of epochs)	
	toilet	sofa	toilet	sofa
single-grid	18.477	15.551	348.96×1240	632.04×680
multi-grid	10.850	7.857	792.90×280	1650.12×70

7 CONCLUSION

We propose the generative VoxelNet model for volumetric object synthesis, and the conditional generative VoxelNet for 3D object recovery and 3D object super resolution. The proposed model is a deep energy-based model parameterized by a VoxelNet, which can be trained by an “analysis by synthesis” scheme. The training of the model can be interpreted as a mode seeking and mode shifting process, and the zero temperature limit has an adversarial interpretation. We also propose a 3D multi-grid energy-based modeling and sampling algorithm to train multi-scaled generative VoxelNets for high resolution 3D shape synthesis. The generative VoxelNet can also be useful in training a 3D generator network via MCMC teaching. The 3D generator trained in such a cooperative scheme is capable of synthesizing meaningful 3D shapes and 3D shape embedding. Experiments demonstrate that our models are able to generate realistic 3D shape patterns and are useful for 3D shape analysis, such as 3D shape recovery, 3D shape super resolution, and 3D shape classification.

ACKNOWLEDGMENTS

The work is supported by NSF DMS-2015577, DARPA SIMPLEX N66001-15-C-4035, ONR MURI N00014-16-1-2007, DARPA ARO W911NF-16-1-0579, DARPA N66001-17-2-4029, and XSEDE grant ASC180018. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. We thank Erik Nijkamp for his help with coding. We thank Siyuan Huang for helpful discussions.

REFERENCES

- [1] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the KITTI vision benchmark suite,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.
- [2] Y.-J. Oh and Y. Watanabe, “Development of small robot for home floor cleaning,” in *Proceedings of the 41st SICE Annual Conference.*, vol. 5, 2002, pp. 3222–3223.
- [3] R. T. Azuma, “A survey of augmented reality,” *Presence: Teleoperators & Virtual Environments*, vol. 6, no. 4, pp. 355–385, 1997.
- [4] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre, “Recent advances in augmented reality,” *IEEE Computer Graphics and Applications*, vol. 21, no. 6, pp. 34–47, 2001.
- [5] J. Geng, “Structured-light 3d surface imaging: a tutorial,” *Advances in Optics and Photonics*, vol. 3, no. 2, pp. 128–160, 2011.
- [6] M. Hansard, S. Lee, O. Choi, and R. P. Horaud, *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012.
- [7] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from RGBD images,” in *European Conference on Computer Vision*, 2012, pp. 746–760.
- [8] J. Xiao, A. Owens, and A. Torralba, “SUN3D: A database of big spaces reconstructed using sfm and object labels,” in *International Conference on Computer Vision*, 2013, pp. 1625–1632.
- [9] S. Song, S. P. Lichtenberg, and J. Xiao, “Sun rgb-d: A rgb-d scene understanding benchmark suite,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 567–576.

- [10] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D shapenets: A deep representation for volumetric shapes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1912–1920.
- [11] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3D model repository," *arXiv preprint arXiv:1512.03012*, 2015.
- [12] E. Ahmed, A. Saint, A. E. R. Shabayek, K. Cherenkova, R. Das, G. Gusev, D. Aouada, and B. Ottersten, "Deep learning advances on different 3D data representations: A survey," *arXiv preprint arXiv:1808.01462*, 2018.
- [13] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta, "Learning a predictable and generative vector representation for objects," in *European Conference on Computer Vision*, 2016, pp. 484–499.
- [14] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *International Conference on Computer Vision*, 2015, pp. 945–953.
- [15] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view CNNs for object classification on 3D data," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5648–5656.
- [16] D. Maturana and S. Scherer, "Voxnet: A 3D convolutional neural network for real-time object recognition," in *International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 922–928.
- [17] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling," in *Advances in Neural Information Processing Systems*, 2016, pp. 82–90.
- [18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.
- [19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.
- [20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
- [21] S.-C. Zhu, "Statistical modeling and conceptualization of visual patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 691–712, 2003.
- [22] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. J. Huang, "A tutorial on energy-based learning," in *Predicting Structured Data*. MIT Press, 2006.
- [23] U. Grenander and M. I. Miller, *Pattern theory: from representation to inference*. Oxford University Press, 2007.
- [24] R. M. Neal *et al.*, "Mcmc using hamiltonian dynamics," *Handbook of markov chain monte carlo*, vol. 2, no. 11, p. 2, 2011.
- [25] J. Xie, Y. Lu, R. Gao, S.-C. Zhu, and Y. N. Wu, "Cooperative training of descriptor and generator networks," *arXiv preprint arXiv:1609.09408*, 2016.
- [26] R. Gao, Y. Lu, J. Zhou, S.-C. Zhu, and Y. Nian Wu, "Learning generative convnets via multi-grid modeling and sampling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9155–9164.
- [27] W. E. Carlson, "An algorithm and data structure for 3D object synthesis using surface patch intersections," *ACM SIGGRAPH Computer Graphics*, vol. 16, no. 3, pp. 255–263, 1982.
- [28] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.
- [29] S. Chaudhuri, E. Kalogerakis, L. Guibas, and V. Koltun, "Probabilistic reasoning for assembly-based 3D modeling," in *ACM Transactions on Graphics*, vol. 30, no. 4, 2011, p. 35.
- [30] E. Kalogerakis, S. Chaudhuri, D. Koller, and V. Koltun, "A probabilistic model for component-based shape synthesis," *ACM Transactions on Graphics*, vol. 31, no. 4, p. 55, 2012.
- [31] K. Xu, H. Zhang, D. Cohen-Or, and B. Chen, "Fit and diverse: set evolution for inspiring 3D shape galleries," *ACM Transactions on Graphics*, vol. 31, no. 4, p. 57, 2012.
- [32] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs," in *International Conference on Computer Vision*, 2017, pp. 2088–2096.
- [33] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, "Learning representations and generative models for 3D point clouds," *arXiv preprint arXiv:1707.02392*, 2017.
- [34] S. H. Khan, Y. Guo, M. Hayat, and N. Barnes, "Unsupervised primitive discovery for improved 3D generative modeling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9739–9748.
- [35] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5939–5948.
- [36] W. Huang, B. Lai, W. Xu, and Z. Tu, "3d volumetric modeling with introspective neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8481–8488.
- [37] L. Gao, J. Yang, T. Wu, Y.-J. Yuan, H. Fu, Y.-K. Lai, and H. Zhang, "Sdmnet: Deep generative network for structured deformable mesh," *ACM Transactions on Graphics*, vol. 38, no. 6, pp. 1–15, 2019.
- [38] Z. Tu, "Learning generative models via discriminative approaches," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
- [39] L. Jin, J. Lazarow, and Z. Tu, "Introspective classification with convolutional nets," in *Advances in Neural Information Processing Systems*, 2017, pp. 823–833.
- [40] J. Lazarow, L. Jin, and Z. Tu, "Introspective neural networks for generative modeling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2774–2783.
- [41] K. Lee, W. Xu, F. Fan, and Z. Tu, "Wasserstein introspective neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3702–3711.
- [42] S. C. Zhu, Y. Wu, and D. Mumford, "Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling," *International Journal of Computer Vision*, vol. 27, no. 2, pp. 107–126, 1998.
- [43] J. Xie, W. Hu, S.-C. Zhu, and Y. N. Wu, "Learning sparse FRAME models for natural image patterns," *International Journal of Computer Vision*, vol. 114, no. 2-3, pp. 91–112, 2015.
- [44] J. Xie, Y. Lu, S.-C. Zhu, and Y. N. Wu, "Inducing wavelets into random fields via generative boosting," *Applied and Computational Harmonic Analysis*, vol. 41, no. 1, pp. 4–25, 2016.
- [45] Y. Lu, S.-C. Zhu, and Y. N. Wu, "Learning FRAME models using CNN filters," *arXiv preprint arXiv:1509.08379*, 2015.
- [46] Y. N. Wu, J. Xie, Y. Lu, and S.-C. Zhu, "Sparse and deep generalizations of the frame model," *Annals of Mathematical Sciences and Applications*, vol. 3, no. 1, pp. 211–254, 2018.
- [47] J. Xie, Y. Lu, S.-C. Zhu, and Y. N. Wu, "A theory of generative convnet," in *International Conference on Machine Learning*, 2016.
- [48] E. Nijkamp, M. Hill, S.-C. Zhu, and Y. N. Wu, "Learning non-convergent non-persistent short-run mcmc toward energy-based model," in *Advances in Neural Information Processing Systems*, 2019, pp. 5232–5242.
- [49] E. Nijkamp, M. Hill, T. Han, S.-C. Zhu, and Y. N. Wu, "On the anatomy of MCMC-based maximum likelihood learning of energy-based models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5272–5280.
- [50] Y. Du and I. Mordatch, "Implicit generation and generalization in energy-based models," *arXiv preprint arXiv:1903.08689*, 2019.
- [51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [52] J. Xie, S.-C. Zhu, and Y. Nian Wu, "Synthesizing dynamic patterns by spatial-temporal generative convnet," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7093–7101.
- [53] J. Xie, S.-C. Zhu, and Y. N. Wu, "Learning energy-based spatial-temporal generative convnets for dynamic patterns," *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [54] J. Xie, Z. Zheng, R. Gao, W. Wang, S.-C. Zhu, and Y. Nian Wu, "Learning descriptor networks for 3D shape synthesis and analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8629–8638.
- [55] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung, "On visual similarity based 3D model retrieval," in *Computer Graphics Forum*, vol. 22, no. 3, 2003, pp. 223–232.
- [56] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3D shape descriptors," in *Symposium on geometry processing*, vol. 6, 2003, pp. 156–164.
- [57] A. Sinha, J. Bai, and K. Ramani, "Deep learning 3D shape surfaces using geometry images," in *European Conference on Computer Vision*, 2016, pp. 223–240.
- [58] B. Shi, S. Bai, Z. Zhou, and X. Bai, "DeepPano: Deep panoramic representation for 3-D shape recognition," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2339–2343, 2015.
- [59] K. Sfikas, T. Theoharis, and I. Pratikakis, "Exploiting the PANORAMA representation for convolutional neural network classification and retrieval," in *Eurographics Workshop on 3D Object Retrieval*, 2017.

- [60] P. Papadakis, I. Pratikakis, T. Theoharis, and S. Perantonis, "PANORAMA: A 3D shape descriptor based on panoramic views for unsupervised 3D object retrieval," *International Journal of Computer Vision*, vol. 89, no. 2-3, pp. 177–192, 2010.
- [61] M. Simonovsky and N. Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [62] A. Sharma, O. Grau, and M. Fritz, "VConv-DAE: Deep volumetric shape learning without object labels," in *European Conference on Computer Vision*, 2016, pp. 236–250.
- [63] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [64] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.
- [65] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
- [66] M. Welling, "Herding dynamical weights to learn," in *International Conference on Machine Learning*, 2009, pp. 1121–1128.
- [67] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875*, 2017.
- [68] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.
- [69] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, pp. 695–709, 2005.
- [70] A. Hyvärinen, "Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables," *IEEE Transactions on Neural Networks*, vol. 18, no. 5, pp. 1529–1531, 2007.
- [71] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.
- [72] K. Swersky, M. Ranzato, D. Buchman, B. M. Marlin, and N. de Freitas, "On autoencoders and score matching for energy based models," in *International Conference on Machine Learning*, 2011, pp. 1201–1208.
- [73] G. Alain and Y. Bengio, "What regularized auto-encoders learn from the data-generating distribution," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3563–3593, 2014.
- [74] T. Tieleman, "Training restricted boltzmann machines using approximations to the likelihood gradient," in *International Conference on Machine Learning*, 2008, pp. 1064–1071.
- [75] T. Han, Y. Lu, S.-C. Zhu, and Y. N. Wu, "Alternating back-propagation for generator network," in *31st AAAI Conference on Artificial Intelligence*, 2017.
- [76] J. Xie, Y. Lu, R. Gao, and Y. N. Wu, "Cooperative learning of energy-based model and latent variable model via mcmc teaching," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [77] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.
- [78] D. Warde-Farley and Y. Bengio, "Improving generative adversarial networks with denoising feature matching," in *International Conference on Learning Representations*, 2017.
- [79] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Generative and discriminative voxel modeling with convolutional neural networks," *arXiv preprint arXiv:1608.04236*, 2016.
- [80] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.



Jianwen Xie received his Ph.D. degree in statistics from University of California, Los Angeles (UCLA) in 2016. He is currently a senior research scientist at Baidu Research USA. Before joining Baidu, he was a senior research scientist at Hikvision Research Institute USA from 2017 to 2020, and a staff research associate and postdoctoral researcher in the Center for Vision, Cognition, Learning, and Autonomy (VCLA) at UCLA from 2016 to 2017. His research interests focus on generative modeling and learning with

applications in computer vision.



Zilong Zheng is currently a Ph.D. candidate in the Center for Vision, Cognition, Learning and Autonomy at the University of California, Los Angeles (UCLA). He received his B.S. degree in computer science at University of Minnesota, and B.E. degree in micro-electronic technology from University of Electronic Science and Technology of China (UESTC). His research interests lie in multimodal representation learning on computer vision and natural language.



Ruiqi Gao is currently a Ph.D. candidate in the Center for Vision, Cognition, Learning and Autonomy at the University of California, Los Angeles (UCLA). She received her B.S. degree in Statistics from Peking University, China in 2016. Her research interests are in statistical modeling and learning, with a focus on deep generative models and representation learning.



Wenguan Wang received his Ph.D. degree from Beijing Institute of Technology in 2018. He is currently a postdoc researcher at ETH Zurich, Switzerland. From 2016 to 2018, he was a joint Ph.D. candidate in University of California, Los Angeles. From 2018 to 2019, he was a senior scientist at Inception Institute of Artificial Intelligence, UAE. His current research interests include computer vision, image processing and deep learning.



Song-Chun Zhu received Ph.D. degree from Harvard University in 1996, and is Chair Professor jointly with Tsinghua University and Peking University, director of Institute for Artificial Intelligence, Peking University. He worked at Brown, Stanford, Ohio State, and UCLA before returning to China in 2020 to launch a non-profit organization – Beijing Institute for General Artificial Intelligence. He has published over 300 papers in computer vision, statistical modeling and learning, cognition, Language, robotics, and AI. He received the Marr Prize in 2003, the Aggarwal prize from the Intl Association of Pattern Recognition in 2008, the Helmholtz Test-of-Time prize in 2013, twice Marr Prize honorary nominations in 1999 and 2007, a Sloan Fellowship, the US NSF Career Award, and ONR Young Investigator Award in 2001. He is a Fellow of IEEE since 2011. He serves as General co-Chair for CVPR 2012 and CVPR 2019.

received the Marr Prize in 2003, the Aggarwal prize from the Intl Association of Pattern Recognition in 2008, the Helmholtz Test-of-Time prize in 2013, twice Marr Prize honorary nominations in 1999 and 2007, a Sloan Fellowship, the US NSF Career Award, and ONR Young Investigator Award in 2001. He is a Fellow of IEEE since 2011. He serves as General co-Chair for CVPR 2012 and CVPR 2019.



Ying Nian Wu received his Ph.D. degree in statistics from Harvard University in 1996. He was an assistant professor in the Department of Statistics, University of Michigan from 1997 to 1999. He joined University of California, Los Angeles (UCLA) in 1999, and is currently a professor in UCLA Department of Statistics. His research interests include generative modeling, representation learning, and computer vision. He received Honorable Mention for the David Marr Prize with S. C. Zhu et al. in 1999 and 2007 for

generative modeling in computer vision.