Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling

Jiajun Wu* MIT CSAIL
Chengkai Zhang* MIT CSAIL
Tianfan Xue MIT CSAIL
William T. Freeman MIT CSAIL, Google Research
Joshua B. Tenenbaum MIT CSAIL

1 Introduction

What makes a 3D generative model of object shapes appealing? We believe a good generative model should be able to synthesize 3D objects that are both highly varied and realistic. Specifically, a generative model should be able to go beyond memorizing and recombining parts from a pre-defined repository, and generate examples with fine details.

In the past decades, researchers have made impressive progress on 3D object modeling and synthesis [7, 3, 1, 20–22], mostly based on meshes or skeletons. Many of these traditional methods synthesize new objects by borrowing parts from objects in an existing CAD model library. Therefore, the synthesized objects look realistic, but not conceptually novel.

Recently, with the advances in deep representation learning and the introduction of large 3D CAD datasets like ShapeNet [2, 23], there have been some inspiring attempts in learning deep object representations based on voxelized objects [13, 5, 19, 14, 18, 17]. Different from part-based models, their generative approach aims to synthesize new objects based on learned hierarchical object representations. This is an encouraging approach, but there is room for improvement on the performance of object synthesis.

In this paper, drawing on recent advances in general-adversarial 2D image modeling [6, 15] and volumetric convolutional networks [13, 23], we demonstrate that modeling volumetric objects in a generative-adversarial manner could be a promising solution which captures the best of both worlds. Different from traditional heuristic criteria, generative-adversarial modeling introduces an additional adversarial discriminator to classify synthesized vs. real objects. This could be a particularly favorable framework for 3D object modeling: as 3D objects are highly structured, a generative-adversarial criterion, but not a voxel-wise independent heuristic one, has the potential to capture the structural difference of two 3D objects. The use of a generative-adversarial loss also arguably avoids possible criterion-dependent overfitting (*e.g.*, generating mean-shape-like, blurred objects when minimizing a mean-squared error).

We show that our generative representation can be used to synthesize high-quality realistic objects, and our discriminative representation can be used for 3D object recognition, achieving comparable performance with recent supervised methods [13, 18], and outperforming other unsupervised methods by a large margin. Our learned generative and discriminative representations also have wide applications. For example, we show that our network can be combined with a variational autoencoder [9, 10] to directly reconstruct a 3D object from a 2D input image.

2 Models

In this section we first discuss how we build our framework, named 3D Generative Adversarial Network (3D-GAN), by leveraging previous advances on volumetric convolutional networks and generative adversarial nets. We then show how we can learn a variational autoencoder simultaneously so that our framework can capture an image to 3D object mapping.

* indicates equal contributions. Emails: {jiajunwu, ckzhang, tfxue, billf, jbt}@mit.edu

3D Deep Learning Workshop at NIPS 2016, Barcelona, Spain.

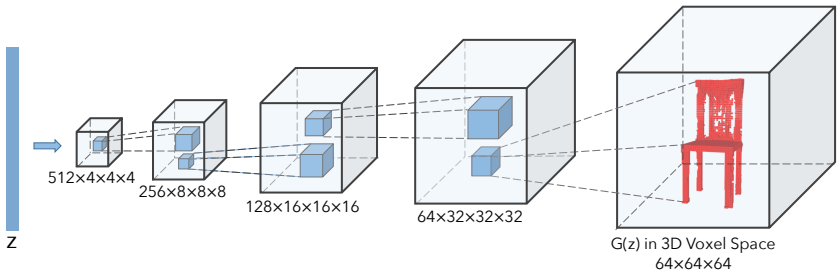

Figure 1: The generator in 3D-GAN. The discriminator mostly mirrors the generator.

2.1 3D Generative Adversarial Network (3D-GAN)

As proposed in [6], the Generative Adversarial Network (GAN) consists of a generator and a discriminator, where the discriminator tries to differentiate real objects and synthesized objects generated by the generator, and the generator attempts to confuse the discriminator. In our 3D Generative Adversarial Network (3D-GAN), the generator G maps a 200 dimensional latent vector z, randomly sampled from a probabilistic latent space, to a $64 \times 64 \times 64$ cube, representing an object $G(z)$ in 3D voxel space. The discriminator D gives its confidence $D(x)$ of whether a 3D object input x is real or synthesized by the generator.

Following [6], we use binary cross entropy as discriminator classification loss, and present our overall adversarial loss function as

$$L_{\text{VAN}} = \log D(x) + \log(1 - D(G(z))), \tag{1}$$

where x are real objects in a $64 \times 64 \times 64$ voxel space, and z are randomly sampled noise vectors from a noise distribution $p(z)$. In this work, $p(z)$ is an i.i.d. uniform distribution over $[0, 1]$.

Network Structure Inspired by [15], we design an all-convolutional neural network to generate 3D objects. As shown in Figure 1, the generator consists of five volumetric fully convolutional layers of kernel sizes $4 \times 4 \times 4$ and strides 2, with batch normalization and ReLU layers added in between and a Sigmoid layer at the end. The discriminator basically mirrors the generator, except that it uses Leaky ReLU [12] instead of ReLU layers. There are no pooling or linear layers in our network.

Training Details Since the discriminator usually learns much faster than the generator, in which case the generator extracts no signals as all examples it generated would be identified as synthetic objects by the discriminator. Therefore, to keep the training of both networks in pace, we employ an adaptive training strategy: for each batch, the discriminator only gets updated if its accuracy in last batch is not higher than 80%. We observe this helps to stabilize the training.

2.2 3D-VAE-GAN

To map images to the latent representation, we introduce VAE-VAN, an extention to VAN that can recover the 3D geometry of an object from a 2D image. We add an additional image encoder E to VAN, which takes a 2D image as input and outputs the latent representation vector z. This is inspired by VAE-GAN proposed by [10], which combines a variational autoencoder and a generative adversarial net by sharing the decoder component of VAE and the generator component of GAN.

The image encoder consists of five spatial convolution layers with kernel size $\{11, 5, 5, 5, 8\}$ and strides $\{4, 2, 2, 2, 1\}$, respectively. There are batch normalization and ReLU layers in between, and a sampler at the end to sample a 200 dimensional vector used by the 3D-GAN. The structure of the generator and the discriminator is the same as those in Section 2.1.

Similar to VAE-GAN [10], our loss function consists of three parts: an object reconstruction loss L_{recon}, a cross entropy loss L_{VAN} for 3D-GAN, and a KL divergence loss L_{KL} to restrict the distribution of the output of the encoder. Formally, these loss functions write as

$$L = L_{\text{VAN}} + \alpha_1 L_{\text{KL}} + \alpha_2 L_{\text{recon}}, \tag{2}$$

where α_1 and α_2 are weights of the KL divergence loss and the reconstruction loss. We have

$$L_{\text{VAN}} = \log D(x) + \log(1 - D(G(z))) + \log(1 - D(G(E(y)))), \tag{3}$$

$$L_{\text{KL}} = D_{\text{KL}}(q(z|y) \,||\, p(z)), \tag{4}$$

$$L_{\text{recon}} = ||G(E(y)) - x||_2, \tag{5}$$

where x is a 3D shape from the training set, y is its corresponding 2D image, and $q(z|y)$ is the variational distribution of the latent representation z. The KL-divergence pushes this variational

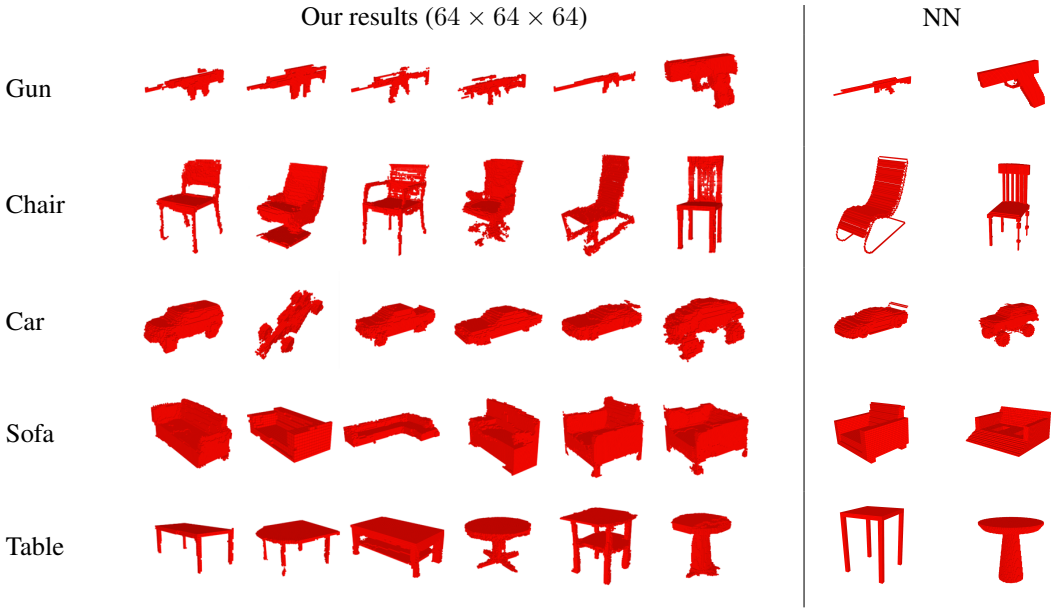

Figure 2: Objects generated by 3D-GAN from vectors, without a reference image/object. We show, for the last two objects in each row, the nearest neighbor retrieved from the training set. We see that the generated objects are similar, but not identical, to examples in the training set.

Supervision	**Pretraining**	**Method**	**Classification (Accuracy)**	
			ModelNet40	ModelNet10
Category labels	ImageNet	MVCNN [19]	90.1%	-
		MVCNN-MultiRes [14]	**91.4%**	-
	None	3D ShapeNets [23]	77.3%	83.5%
		DeepPano [18]	77.6%	85.5%
		VoxNet [13]	83.0%	92.0%
		ORION [16]	-	**93.8%**
Unsupervised	-	SPH [8]	68.2%	79.8%
		LFD [4]	75.5%	79.9%
		T-L Network [5]	74.4%	-
		VConv-DAE [17]	75.5%	80.5%
		3D-GAN (ours)	**83.3%**	**91.0%**

Table 1: Classification results on the ModelNet dataset. Our 3D-GAN outperforms other unsupervised learning methods by a large margin, while being comparable to some recent voxel-based supervised learning frameworks.

distribution towards to the prior distribution $p(z)$, so that the generator can sample the latent representation from the same distribution $p(z)$. In this work, we choose $p(z)$ as i.i.d. Gaussian distribution $N(0, \mathbf{I})$. For more details, please refer to [10].

3 Evaluation

In this section, we evaluate the unsupervisedly learned representation by the discriminator, by using them as features for 3D object classification. We show both qualitative and quantitative results on the popular benchmark ModelNet [23]. Further, we evaluate our 3D-VAE-GAN on 3D object recovery from a single image, and show both qualitative and quantitative results on the IKEA dataset [11].

3.1 3D-GAN Qualitative Results

Figure 2 shows 3D objects generated by our 3D-GAN. For this experiment, we trained a 3D-GAN for each object category. For generation, we sample 200-dimensional vectors following an i.i.d. uniform distribution over $[0, 1]$.

When synthesized results get better, a natural concern is whether the network is simply memorizing objects from training data. We show that the network is generalizing beyond the training set by comparing synthesized objects with their nearest neighbor in the training set. Note that finding nearest

Figure 3: Qualitative results on IKEA [11], from 3D-VAE-GAN separately trained for each class

Method	Bed	Bookcase	Chair	Desk	Sofa	Table	Overall
AlexNet-fc8 [5]	29.5	17.3	20.4	19.7	38.8	16.0	19.8
AlexNet-conv4 [5]	38.2	26.6	31.4	26.6	69.3	19.1	31.1
T-L Network [5]	56.3	30.2	32.9	25.8	71.7	23.3	38.3
3D-VAE-GAN (jointly trained)	49.1	31.9	42.6	34.8	**79.8**	33.1	45.2
3D-VAE-GAN (separately trained)	**63.2**	**46.3**	**47.2**	**40.7**	78.8	**42.3**	**53.1**

Table 2: Average precision for voxel prediction on the IKEA dataset

neighbors for a 3D object is a non-trivial problem, and using L2 distance on voxel level does not produce reasonable results due to possible translation and scale difference. We find that using the output of the last convolutional layer in our discriminator (with a 2x pooling) for retrieval gives good matches. We observe from Figure 2 that generated objects are similar, but not identical, to examples in the training set. There exist many reasonable variations that make the generated objects novel.

3.2 3D Object Classification

We then evaluate the representations learned by our discriminator. A typical way of evaluating representation that are learned without supervision is to use them as features for classification using linear SVM. Here we use the second to last layer of the discriminator as the feature representation for an input 3D object. We train a single 3D-GAN on the seven major object categories (chairs, sofas, tables, boats, airplanes, rifles, and cars) in the ShapeNet [2] training set. This also tests the out-of-category generalization ability of 3D-GAN.

Following [23, 17, 13, 14], we use ModelNet [23] for this task. We compare with the state-of-the-art methods [23, 5, 17, 16] and show per-class accuracy in Table 1. Specifically, our framework outperforms other features learned without supervision [5, 17] by a large margin (83.3% vs. 75.5% on ModelNet 40, and 91.0% vs 80.5% on ModelNet 10). Further, our classification accuracy is also higher than some recent supervised methods [18] and close to the state-of-the-art voxel-based supervised learning approaches [13, 16]. Multi-view CNNs [19, 14] outperforms us, though their method is designed for classification, and requires rendered multi-view images and an ImageNet-pretrained model.

3.3 Single Image 3D Reconstruction

As an application, our show that the 3D-VAE-GAN can perform well single image 3D reconstruction. Following previous work [5], we test it on all six categories of the IKEA dataset [11]: bed, bookcase, chair, desk, sofa, and table. We crop the images so that the objects are centered in the images, and show both qualitative and quantitative results.

We show our results in Figure 3 and Table 2. Following [5], we evaluate results at resolution $20 \times 20 \times 20$, use average precision as our evaluation metric, and attempt to align each prediction with the ground-truth over permutations, flips, and translational alignments (up to 10%), as IKEA ground truth objects are not in a canonical viewpoint. We see that our model consistently outperforms previous state-of-the-art in voxel-level prediction and other baseline methods.

4 Conclusion

In this paper, we proposed 3D-GAN for 3D object generation, as well as 3D-VAE-GAN for learning an image to 3D model mapping. We demonstrated that 3D-GAN and 3D-VAE-GAN are able to generate novel objects and also reconstruct 3D objects from images. We also showed that the discriminator in GAN, learned without supervision, can be used as an informative feature representation for 3D objects, and showed its performance on object classification.

References

[1] Wayne E Carlson. An algorithm and data structure for 3d object synthesis using surface patch intersections. In *SIGGRAPH*, 1982. 1

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 4

[3] Siddhartha Chaudhuri, Evangelos Kalogerakis, Leonidas Guibas, and Vladlen Koltun. Probabilistic reasoning for assembly-based 3d modeling. *ACM TOG*, 30(4):35, 2011. 1

[4] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3d model retrieval. *CGF*, 2003. 3

[5] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 1, 3, 4

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 2

[7] Evangelos Kalogerakis, Siddhartha Chaudhuri, Daphne Koller, and Vladlen Koltun. A probabilistic model for component-based shape synthesis. *ACM TOG*, 31(4):55, 2012. 1

[8] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *SGP*, 2003. 3

[9] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1

[10] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016. 1, 2, 3

[11] Joseph J. Lim, Hamed Pirsiavash, and Antonio Torralba. Parsing ikea objects: Fine pose estimation. In *ICCV*, 2013. 3, 4

[12] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 2

[13] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, 2015. 1, 3, 4

[14] Charles R Qi, Hao Su, Matthias Niessner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *CVPR*, 2016. 1, 3, 4

[15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 1, 2

[16] Nima Sedaghat, Mohammadreza Zolfaghari, and Thomas Brox. Orientation-boosted voxel nets for 3d object recognition. *arXiv preprint arXiv:1604.03351*, 2016. 3, 4

[17] Abhishek Sharma, Oliver Grau, and Mario Fritz. Vconv-dae: Deep volumetric shape learning without object labels. *arXiv preprint arXiv:1604.03755*, 2016. 1, 3, 4

[18] Baoguang Shi, Song Bai, Zhichao Zhou, and Xiang Bai. Deeppano: Deep panoramic representation for 3-d shape recognition. *IEEE SPL*, 22(12):2339–2343, 2015. 1, 3, 4

[19] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *ICCV*, 2015. 1, 3, 4

[20] Johan WH Tangelder and Remco C Veltkamp. A survey of content based 3d shape retrieval methods. *Multimedia tools and applications*, 39(3):441–471, 2008. 1

[21] Oliver Van Kaick, Hao Zhang, Ghassan Hamarneh, and Daniel Cohen-Or. A survey on shape correspondence. *CGF*, 2011.

[22] Andrew KC Wong, Si W Lu, and Marc Rioux. Recognition and shape synthesis of 3-d objects based on attributed hypergraphs. *IEEE TPAMI*, 11(3):279–290, 1989. 1

[23] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, 2015. 1, 3, 4