VPP: Efficient Conditional 3D Generation via Voxel-Point Progressive Representation Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

A Implementation Details

A.1 Experimental Details

Training Details We use ShapeNetCore from ShapeNet [1] as the training dataset. ShapeNet is a clean set of 3D CAD object models with rich annotations, including 51K unique 3D models from 55 common object categories. For the acquisition of multi-modal data, we follow ReCon [8] for multi-view rendering and utilize BLIP [4] based on the rendered images to obtain textual data. In Table 1, we show the training hyperparameters and model architecture information of each part of our VPP.

Config	3D VQGAN	Voxel Generator	Grid Smoother	Point Upsampler
Training Parameters				
Optimizer	Adam	AdamW	AdamW	AdamW
Learning rate	1e-4	1e-3	1e-3	1e-3
Weight decay	1e-4	5e-2	5e-2	5e-2
Training epochs	300	100	100	300
Learning rate scheduler	cosine	cosine	cosine	cosine
Batch size	128	128	128	128
Drop path rate	-	0.1	0.1	0.1
Input point size	8192	8192	8192	1024
Model Architecture				
Backbone	CNN	Transformer	Transformer	Transformer
Layers	6	12	4	6
Hidden size	384	384	64	384
Heads	-	6	4	6
Voxel resolution	24	24	24	24
Point patch size	-	-	-	32
GPU device	NVIDIA A100	NVIDIA A100	NVIDIA A100	NVIDIA A100

Table 1: Training recipes for 3D VQGAN, Voxel Generator, Grid Smoother and Point Upsampler.

Downstream Tasks Details Following Point-E [6], we use pre-trained PointNet++ as a classifier in all ACC, FID, and IS evaluations to extract the features and calculate the accuracy of generated point clouds. In point cloud generation and editing, we employ 8 or 4 steps for a parallel generation. The generation task utilizes initial voxel codebooks composed entirely of `[MASK]` tokens. In editing, we

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

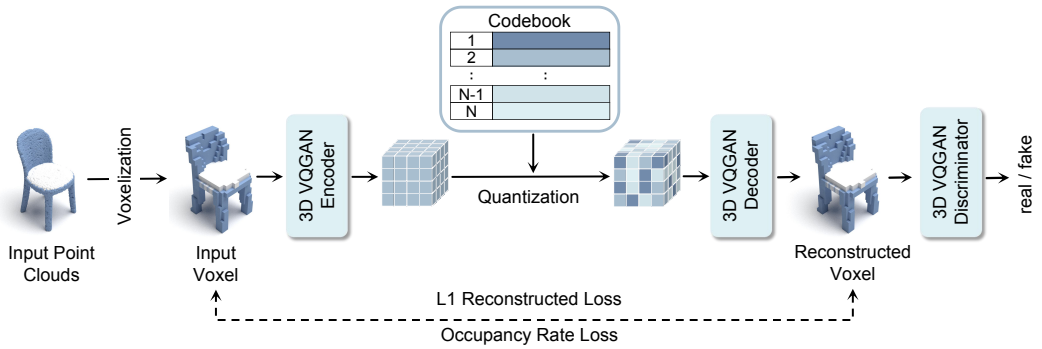

Figure 1: The training overview of 3D VQGAN. We introduce the occupancy rate loss to have a better reconstruction of the voxel.

extract VQGAN features from the original point cloud to initialize the voxel codebooks. As for the transfer classification task on ScanObjectNN [10] and ModelNet40 [14], we fully follow the previous work [2, 7] configuration and trained 300 epochs with the AdamW optimizer, and used the voting strategy in testing.

A.2 Implementation Details of 3D VQGAN

We show the detailed training overview of 3D VQGAN in Figure 1. Fllowing the tranditional training of VQGAN [3], we use L1 loss to supervise the reconstruction of voxel, and feed the reconstructed voxel into the discriminator to ensure the generated authenticity by GAN loss. Besides the L1 loss and GAN loss, we also introduce the occupancy rate loss to make the occupancy rate of the reconstructed voxel grid similar to the ground truth, so as to obtain a better reconstruction of the voxel.

B Additional Experiments

We conduct more experiments to further demonstrate the generation quality and universality of VPP. Including diversity & specific text-conditional generation, point-based mesh reconstruction, and few-shot transfer classification.

B.1 Diversity & Specific Text-Conditional Generation

We show the diversity qualitative results of VPP on text-conditional 3D generation in Figure 2 (a). It can observe that VPP can generate a broad category of shapes with rich diversity while remain faithful to the provided text descriptions. Meanwhile, we present the qualitative results of VPP on more specific text-conditional 3D generation in Figure 2 (b). Notably, VPP can generate high-fidelity shapes that well react to very specific text descriptions, like `'a round chair with arms'`, `'a round table with four legs'` and `'an old-style propeller aircraft'`, etc.

B.2 Point-Based Mesh Reconstruction

Besides the representation of point clouds, we also show the common mesh representation of generated shapes to demonstrate the high generation quality of VPP. We use DMTet [9] to reconstruct the mesh from the generated point clouds, where the reconstructed examples are presented in Figure 3. It is observed that the reconstructed mesh is still high-quality with fine geometric details, which further proves that VPP not only has high fidelity generation but also supports the output of mesh representation.

B.3 Few-Shot Transfer Classification

We conduct few-shot experiments on the ModelNet40 [14] dataset, and the results are shown in Table 2. In the self-supervised benchmark without the use of additional modality data, VPP achieves excellent performance compared to previous works.

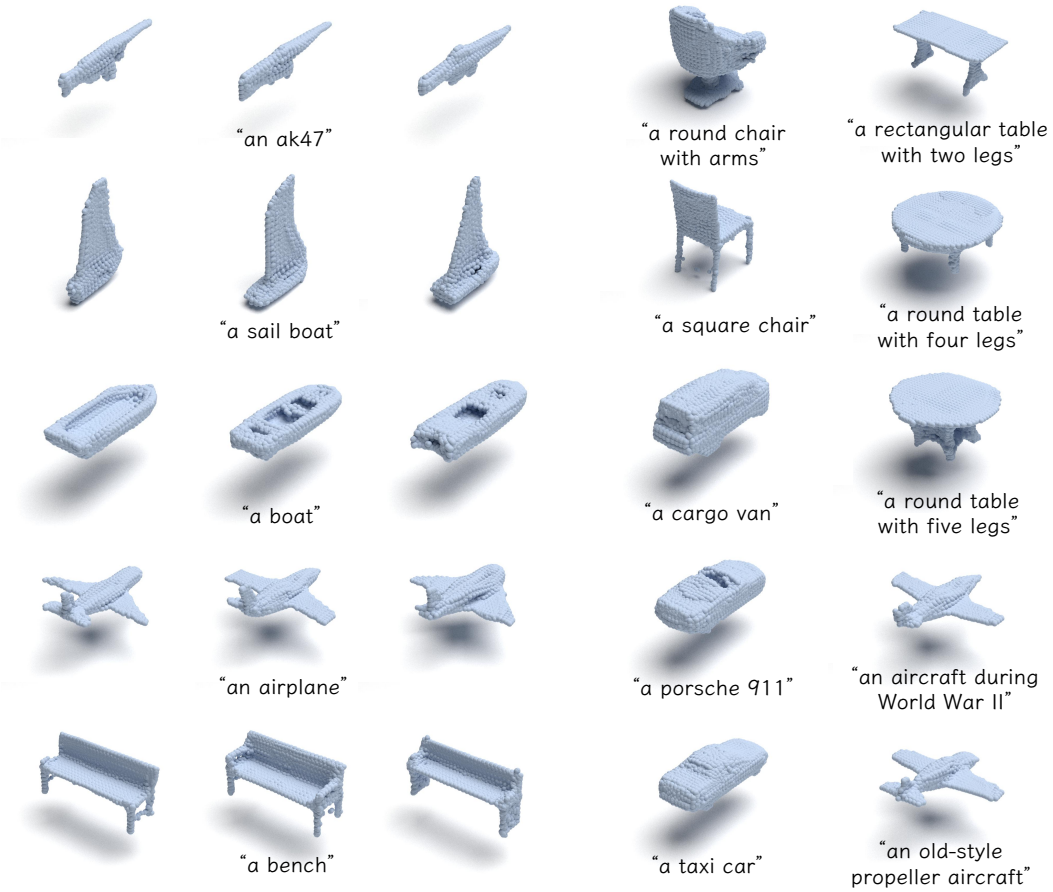

Figure 2: (a) **Diversity** qualitative results of VPP on text-conditional 3D generation. (b) Qualitative results of VPP on more **specific** text-conditional 3D generation. VPP can generate a broad category of shapes with rich diversity and high fidelity, while remain faithful to the provided text descriptions. Besides, VPP can react to very specific text descriptions, like `"a round chair with arms"`.

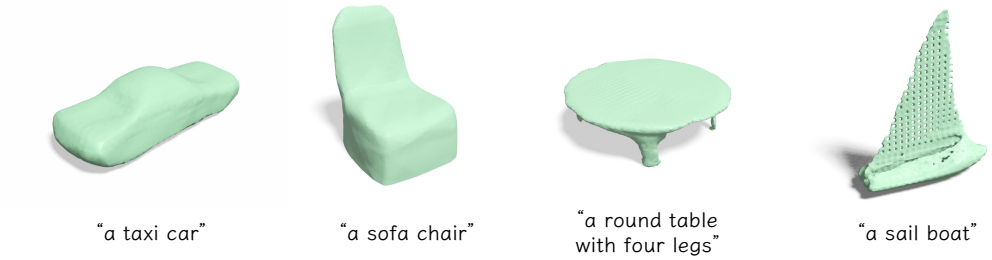

Figure 3: Examples of reconstructed mesh from the generated point clouds. VPP not only has high-fidelity generation but also supports the output of mesh representation.

Method	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
Supervised Learning Only				
DGCNN [13]	31.6 ± 2.8	40.8 ± 4.6	19.9 ± 2.1	16.9 ± 1.5
OcCo [12]	90.6 ± 2.8	92.5 ± 1.9	82.9 ± 1.3	86.5 ± 2.2
with Self-Supervised Representation Learning				
Transformer [11]	87.8 ± 5.2	93.3 ± 4.3	84.6 ± 5.5	89.4 ± 6.3
OcCo [12]	94.0 ± 3.6	95.9 ± 2.3	89.4 ± 5.1	92.4 ± 4.6
Point-BERT [15]	94.6 ± 3.1	96.3 ± 2.7	91.0 ± 5.4	92.7 ± 5.1
MaskPoint [5]	95.0 ± 3.7	97.2 ± 1.7	91.4 ± 4.0	93.4 ± 3.5
Point-MAE [7]	96.3 ± 2.5	97.8 ± 1.8	92.6 ± 4.1	95.0 ± 3.0
Point-M2AE [16]	96.8 ± 1.8	98.3 ± 1.4	92.3 ± 4.5	95.0 ± 3.0
VPP (ours)	**96.9 ± 1.9**	**98.3 ± 1.5**	**93.0 ± 4.0**	**95.4 ± 3.1**

Table 2: Few-shot transfer classification results on ModelNet40. Overall accuracy (%) without voting is reported.

C Limitation and Future Work

Although, as a generative model, the proposed VPP has achieved the unified model on multiple downstream 3D tasks, the current work focuses on some relatively simple tasks. For instance, we perform 3D completion task on point cloud upsampling, and the pre-training transfer learning on 3D classification. It would be more valuable to extend it to more complex 3D applications, e.g., shape completion and 3D pre-training on detection & segmentation. Furthermore, the current model is trained with ShapeNet [1], which is still a limited-scale dataset across different categories. It would be promising to explore large-scale 3D dataset for training the model to further improve the generation quality and capability.

D Broader Impact

Conditional 3D generation has the potential to improve the 3D design and assist practitioners efficiently to work on content creation. Similarly, we propose VPP with the hope of enhancing better creativity and work as an alternative tool for artists to more efficiently design. Besides, conditional 3D generation methods allow the public to have more convenient access to 3D craft, which can bring many benefits to human-AI collaborative application.

References

[1] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015. 1, 4

[2] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 2

[3] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[4] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022. 1

[5] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 4

[6] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 1

[7] Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 2, 4

[8] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023. 1

[9] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *NeurIPS*, pages 6087–6101, 2021. 2

[10] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1588–1597, 2019. 2

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst. (NIPS)*, pages 5998–6008, 2017. 4

[12] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9782–9792, 2021. 4

[13] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph CNN for learning on point clouds. *ACM Trans. Graph.*, 38(5):146:1–146:12, 2019. 4

[14] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1912–1920, 2015. 2

[15] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 4

[16] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2AE: Multi-scale masked autoencoders for hierarchical point cloud pre-training. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 4