RealCity3D: A Large-scale Georeferenced 3D Shape Dataset of Real-world Cities

Yuqiong Li
New York University
yl5090@nyu.edu

Hang Zhao
MIT
hangzhao@mit.edu

Zhiding Yu
NVIDIA
zhidingy@nvidia.com

Chen Feng*
New York University
cfeng@nyu.edu

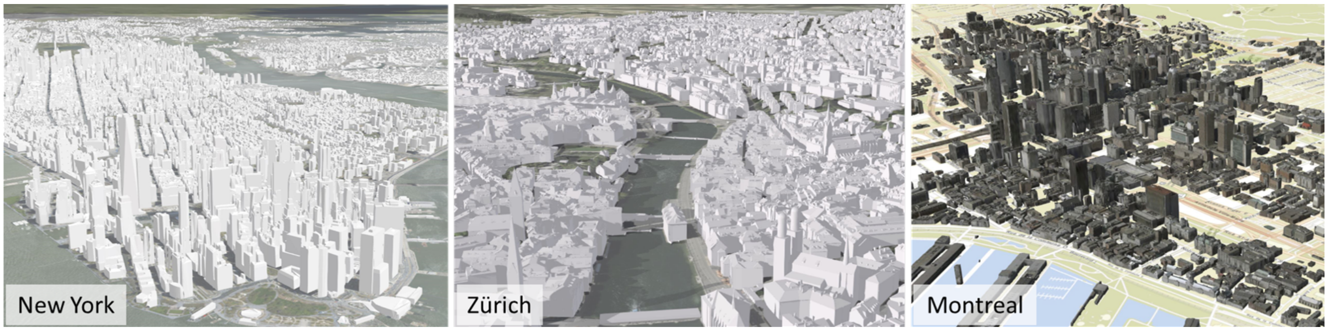

Figure 1. Overview of several cities in our RealCity3D dataset. Each building shape is of LoD2 complexity.

Abstract

Existing 3D shape datasets foster 3D deep learning research in the vision, graphics, and robotics communities by motivating research, specifying challenges, and enabling model comparisons. However, most existing 3D shape datasets are comprised of CAD models or point cloud scans at either object-level or room-level, leaving out a large source of 3D shape data: real-world cities. Cities are important because they contain complex shapes such as skyscrapers, residential buildings, roads, and bridges. These shapes contain rich details that can be significantly different from object-level and room-level 3D shapes. Such inherent domain differences bring challenges to existing deep learning methods on 3D data, especially unsupervised ones, therefore inviting additional research in this area. In this work, we collect and process more than $950,000$ *georeferenced 3D shapes from the city of New York, and demonstrate the performance gap of three unsupervised 3D deep learning methods on our dataset and existing datasets. We are also actively working to include other major world cities and benchmarking more 3D deep learning methods on this dataset. We will release the dataset and tools to the public and invite research collaborations on the topic.*

1. Introduction

The world we live in is 3D, and our eyes have naturally evolved to perceive this 3D world. The area of 3D computer vision has seen great progress in both research and real applications, thanks to the recent advances in deep representation learning of 3D shapes. The increasingly more powerful but also more data-hungry models have created needs for large-scale datasets for 3D deep learning. As a result many large-scale 3D datasets were proposed.

In this paper, we present RealCity3D, a large-scale data repository of real-world city models containing georeferenced 3D building shapes, each represented as a set of semantically labeled polygon meshes. Such data format can then be readily converted to either voxels, point clouds, or multi-view images for potential evaluations of many 3D deep representation learning methods [6, 14, 9, 17, 2]. Despite the enormous progress happening in this field, we see a potential contribution opportunity to enrich this collection of datasets with a benchmark that has: 1) a large number of single 3D objects in the scale of millions and 2) a different semantic scene (outdoor) compared to previous datasets, which are mostly indoor scenes.

2. Related Work

3D deep learning has been an active research area with considerable recent advances. Multiple large-scale 3D datasets have been proposed to facilitate research and benchmarking in this research area. One of the early well-known and frequently used benchmark is the Princeton Shape Benchmark in 2004 [10] consisting of $6,670$ single 3D models, among which $1,814$ are manually classified into 161 categories. The IKEA dataset in 2013 [5] is another popular dataset of images and 3D models rep-

*Corresponding author.

resenting typical indoor scenes and contains 759 images and 219 3D models across around 30 categories. PASCAL 3D+ [16] was proposed in the following year (2014) for 3D object detection and pose estimation. The dataset has 12 categories of rigid objects, with each containing more than $3,000$ instances. Subsequent datasets include the Princeton ModelNet (2015) [15] with $127,915$ 3D CAD models from 662 categories, covering most common object categories in the world. The sizes of datasets continued to explode. ShapeNet [3] contains over three million 3D models with a core dataset of about 51,300 unique 3D models from 55 common object categories. It also has manually verified category and alignment annotations.

Apart from the above datasets focusing on 3D models, some other datasets focus on 3D scenes, taking the form of RGBD images. An example is the NYU-Depth V2 dataset (2012) [7] containing video sequences recorded by depth cameras from a variety of indoor scenes. There are 1449 densely labeled pairs of RGBD images from 464 new scenes in 3 cities, in which the objects are labeled and classified. Similarily, SUNRGB-D (2015) [11] is a popular benchmark consisting of $10,355$ RGB-D scenes in the training set and $2,860$ in the testing set, with even richer annotation information for scene classification, semantic segmentation, object detection and pose estimation. Similar to the trend in 3D model datasets, the sizes of datasets kept increasing. In 2017, ScanNet [4] was released with 2.5 million RGBD views in more than 1500 scans and annotated with 3D camera poses, surface reconstructions, and instance-level semantic segmentation. The second version of ScanNet was released in 2018 with an extra 100 scans.

3. Data Collection and Processing

In our dataset, all 3D objects are extracted from 3D city models in the format of **CityGML**, an open data model and XML-based format for the storage and exchange of virtual 3D city models widely used by the geographic community. It extends XML by adding sets of primitives, including topology, features, and geometry, as well as constraints specific to cities. Example 3D object classes in CityGML include buildings, tunnels, bridges, and water bodies. For every 3D object class, CityGML has a hierarchical model complexity system from LoD1 (Levels of Detail) to LoD4. A LoD1 building, for example, is represented by a horizontal polygon with a height, which essentially defines an elevated footprint. A LoD2 building or building part will have a geometrically simplified outer shell represented by horizontal and vertical outer surfaces as well as simplified roof shapes. Surfaces also have semantic information, including ground, wall, roof, outer ceiling, outer floor, and virtual closure surfaces. LoD3 buildings have more complex outer shells represented by detailed outer surfaces and detailed roof shapes, most notably with windows and doors (i.e. holes in surfaces). Finally, LoD4 buildings have interior semantic objects such as furniture.

Table 1. New York City Building Mesh Statistics

	height	area	volume	**#F**	**#V**
avg	8.4	156.6	2228.0	12.1	19.2
std	6.3	514.8	1.9×10^4	13.4	24.3
min	0.3	0.0	0.0	1	4
25%	5.9	64.4	478.6	7	10
50%	8.0	91.9	732.3	10	14
75%	9.4	125.1	1085.0	14	22
max	377.6	1.1×10^5	3.5×10^6	3093	5148

We collect CityGML city models from free 3D geospatial datasets publicly available on the Internet. Most of these models have LoD2 complexity, although we do notice city models of LoD1 or other 3D formats and skip them. At the moment, we have collected CityGML data of 63 major cities around the world, of which 54 are from Germany and 9 from other countries. Apart from New York and Montreal, all the others cities are in Europe. Data qualities can vary. Only 76% of Zurich buildings have valid CityGML surfaces (the other 24% have non-planar and duplicated surfaces, violating the CityGML format standard, requiring more data cleaning), while over 99% of buildings have valid geometry in New York City (NYC).

We thus choose NYC as the focus of our current release because of its high data quality. Our work is based on a published dataset by NYC [12]. In total, we extract $955,120$ individual polygon mesh building models through parsing the raw CityGML data. We then triangulate polygon meshes to acquire $955,023$ traingular meshes, use Poisson disk sampling to acquire $953,058$ point clouds of size 4096×3 from triangular mesh, and use the open source program binvox [8] to acquire $955,120$ building voxels of size 256^3.

Semantic information of buildings are preserved in polygon meshes through adding class labels to each surfaces as comments in the *.obj* files in our dataset. There are 3 categories for building surfaces: GroundSurface, RoofSurface, and WallSurfaces. Statistics demonstrating variations in the dataset is shown in Table 1. As can be seen from the number of vertices and faces, some building shapes are highly complex with thousands of faces, while others have far fewer, adding learning challenges.

4. Benchmark

How to justify our dataset by evaluating its differences against other popular datasets? Qualitatively, one unique property of our dataset is the geometrically highly constrained shapes. For example, most buildings have vertical walls, piece-wise planar surfaces, but some have intricate details (e.g. the Empire State Building compared to a town house). This would pose a challenge for suitable represen-

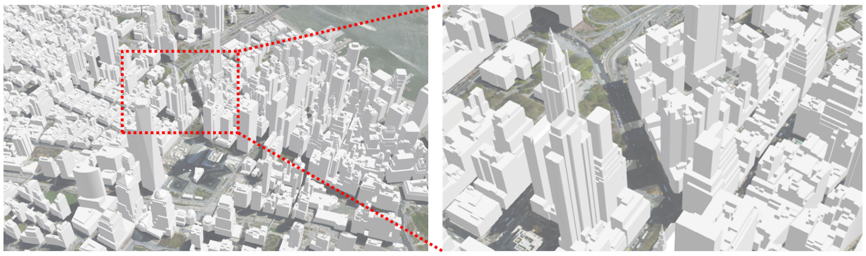

Figure 2. Detailed views of 3D building shapes in New York, U.S.

Table 2. Benchmark of point cloud generation on two datasets.

RealCity3D	JSD	Coverage	MMD
Raw-GAN	0.068	47.6	0.061
Latent-GAN [1]	0.024	57.3	0.088
ShapeNet[3]	JSD	Coverage	MMD
Raw-GAN	0.176	52.3	0.0020
Latent-GAN [1]	0.020	68.9	0.0018

tations of buildings that would facilitate 3D learning.

To quantitatively show this difference, we compare RealCity3D with ShapeNet in terms of properties related to 3D generation by training and testing the same generative model on both datasets independently. The rationale is if the same generative model performs differently on the two datasets, this would indicate the two datasets have different properties related to 3D shape generations, thus highlighting the uniqueness of our datasets. In Table 2, we report results using the same evaluation criteria as in [13]:

Jensen-Shannon Divergence (JSD) is a classic measure of the similarity between two data distributions. A point cloud's distribution is calculated by counting number of points in a voxel grid. Here we calculated the JSD between real and generated point clouds.

Coverage measures the fraction of generated point clouds that are matched to training point clouds.

Minimum Matching Distance (MMD) measures the *fidelity* of a set of point clouds A to another set B by reporting the average distance in minimum distance matching.

In addition, we trained FoldingNet [17], a 3D shape Auto-Encoder, on RealCity3D, and qualitatively demonstrated the challenge it faces in reconstructing the 3D building shape, as shown in Figure 3.

5. Future Work

We are actively working on the following:

- Extend the dataset to include more cities, such as Zürich and Montreal.
- Extend the dataset to include more shape categories than building, such as road, bridges, etc.
- Benchmark more 3D deep learning algorithms (mainly unsupervised ones) on this RealCity3D.

References

[1] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas. Learning representations and generative models for 3d point clouds. *arXiv preprint arXiv:1707.02392*, 2017. 3, 4

[2] A. Arsalan Soltani, H. Huang, J. Wu, T. D. Kulkarni, and J. B. Tenenbaum. Synthesizing 3d shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1511–1519, 2017. 1

[3] 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. 2, 3

[4] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 2

[5] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing IKEA Objects: Fine Pose Estimation. *ICCV*, 2013. 1

[6] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5115–5124, 2017. 1

[7] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[8] F. S. Nooruddin and G. Turk. Simplification and repair of polygonal models using volumetric techniques. *IEEE Transactions on Visualization and Computer Graphics*, 9(2):191–205, 2003. 2

[9] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 1

[10] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The princeton shape benchmark. In *Proceedings Shape Modeling Applications, 2004.*, pages 167–178. IEEE, 2004. 1

[11] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 2

[12] The New York City Department of Information Technology and Telecommunications. NYC 3-D building

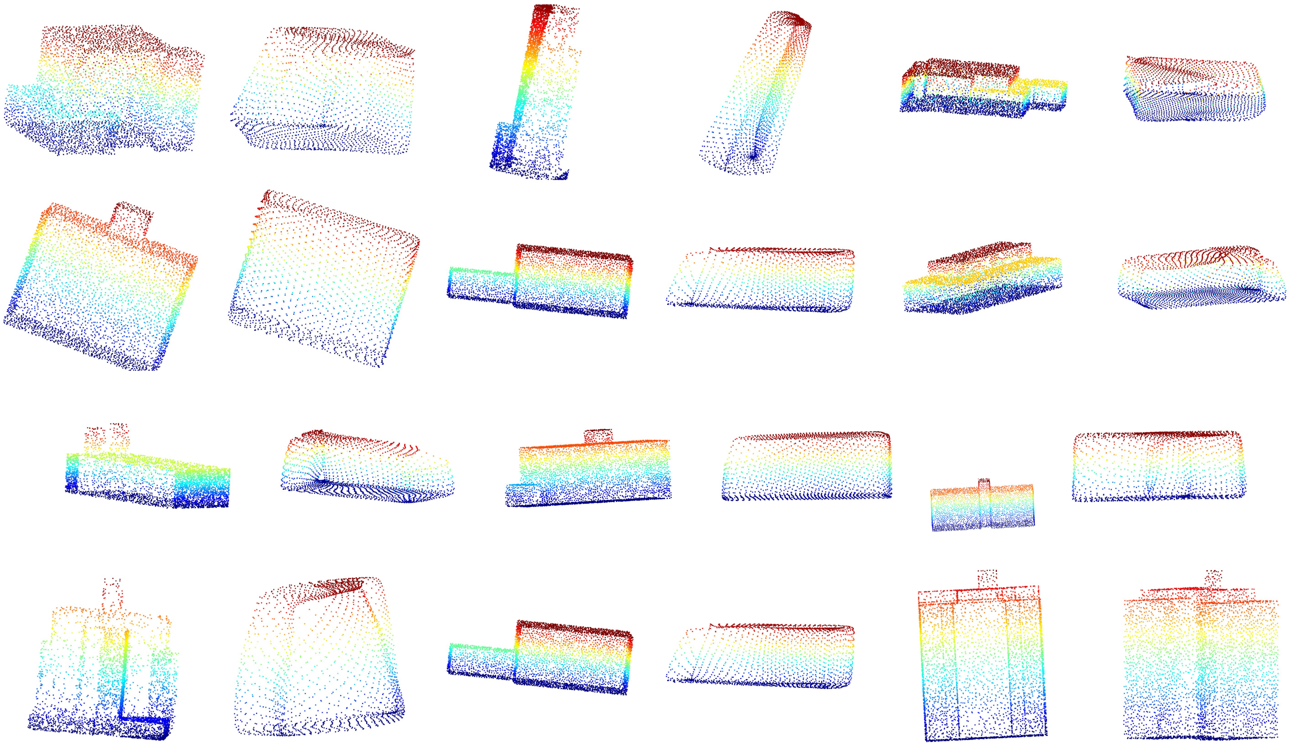

Figure 3. Some testing results of FoldingNet [17] trained on RealCity3D. For each pair of shapes, the first one is the input point cloud and the second one is the auto-encoder's reconstructed point cloud. It can be seen that the reconstructions lost many important geometric details of the 3D building shapes. This suggests that the RealCity3D is a non-trivial 3D shape dataset.

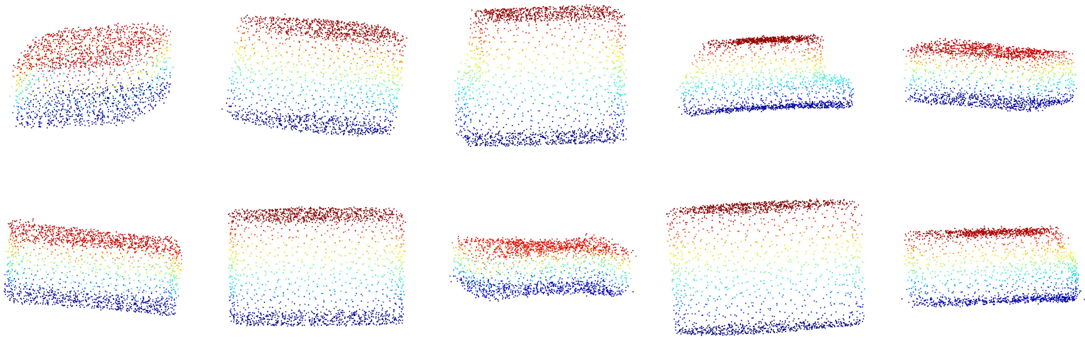

Figure 4. Some shape generation results of Latent-GAN [1] trained on RealCity3D. It can be seen that the reconstructions lost many important geometric details and variations of the 3D building shapes. This suggests again that the RealCity3D dataset is non-trivial.

model, 2016. https://www1.nyc.gov/site/doitt/initiatives/3d-building.page. 2

[13] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in neural information processing systems*, pages 82–90, 2016. 3

[14] 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*, pages 82–90, 2016. 1

[15] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 2

[16] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, pages 75–82. IEEE, 2014. 2

[17] Y. Yang, C. Feng, Y. Shen, and D. Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 206–215, 2018. 1, 3, 4