

ReCo: A Dataset for Residential Community Layout Planning

Xi Chen
Yun Xiong
x_chen21@m.fudan.edu.cn
yunx@fudan.edu.cn
Shanghai Key Laboratory of Data Science,
School of Computer Science, Fudan University
Shanghai, China

Siqi Wang
Haofen Wang*
siqi_wang@tongji.edu.cn
carter.whfcarter@gmail.com
College of Design and Innovation,
Tongji University
Shanghai, China

Tao Sheng
Yao Zhang
tsheng16@fudan.edu.cn
yaozhang@fudan.edu.cn
Shanghai Key Laboratory of Data Science,
School of Computer Science, Fudan University
Shanghai, China

Yu Ye
yye@tongji.edu.cn
College of Architecture and Urban Planning,
Tongji University
Shanghai, China

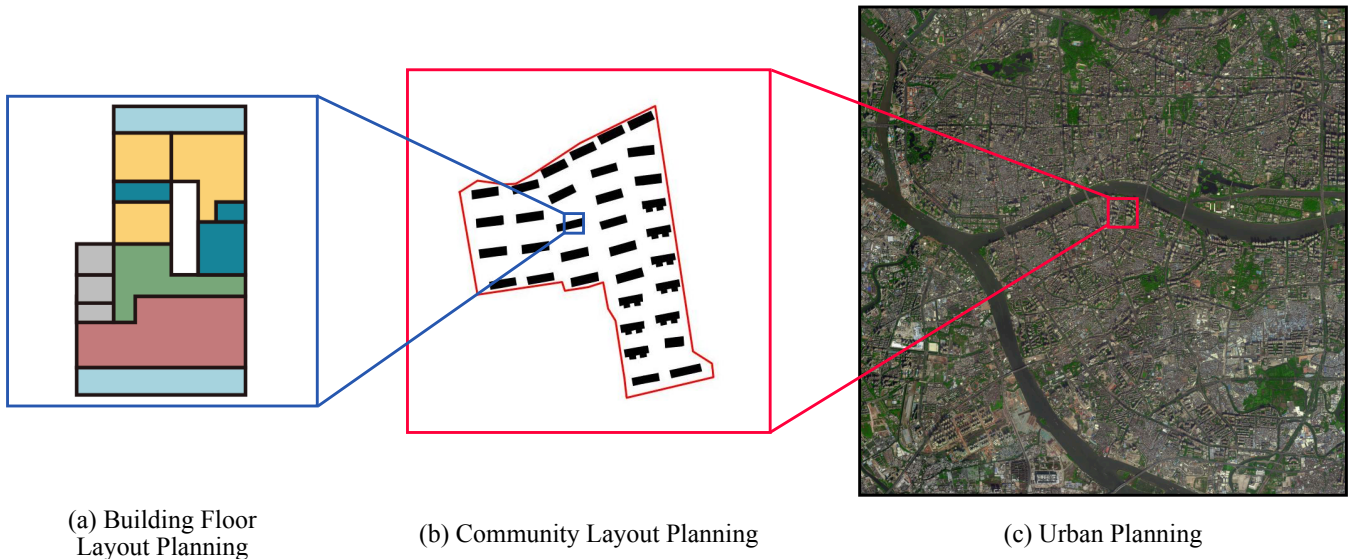


Figure 1: Typical tasks of layout planning from fine- to coarse-grained (the projection relationship in the figure is for illustration only). (a) Building Floor Layout Planning [34]. (b) Community Layout Planning (an example generated from ReCo). (c) Urban Planning (an example screenshot from map.)

*Haofen Wang is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10...\$15.00
<https://doi.org/10.1145/3581783.3612465>

ABSTRACT

Layout planning is centrally important in the field of architecture and urban design. Among the various basic units carrying urban functions, residential community plays a vital part for supporting human life. Therefore, the layout planning of residential community has always been of concern, and has attracted particular attention since the advent of deep learning that facilitates the automated layout generation and spatial pattern recognition. However, the research circles generally suffer from the insufficiency of residential community layout benchmark or high-quality datasets, which hampers the future exploration of data-driven methods for residential community layout planning. The lack of datasets is largely

due to the difficulties of large-scale real-world residential data acquisition and long-term expert screening. In order to address the issues and advance a benchmark dataset for various intelligent spatial design and analysis applications in the development of smart city, we introduce **Residential Community Layout Planning (ReCo) Dataset**, which is the first and largest open-source vector dataset related to real-world community to date. ReCo Dataset is presented in multiple data formats with 37,646 residential community layout plans, covering 598,728 residential buildings with height information. ReCo can be conveniently adapted for residential community layout related urban design tasks, e.g., generative layout design, morphological pattern recognition and spatial evaluation. To validate the utility of ReCo in automated residential community layout planning, two Generative Adversarial Network (GAN) based generative models are further applied to the dataset. We expect ReCo Dataset to inspire more creative and practical work in intelligent design and beyond. The ReCo Dataset is published at: <https://www.kaggle.com/fdudsde/reco-dataset> and related code can be found at: <https://github.com/FDUDSDE/ReCo-Dataset>.

CCS CONCEPTS

• **Applied computing**; • **Arts and humanities**; • **Architecture (buildings)**; • **Computer-aided design**;

KEYWORDS

Dataset, Residential Community Layout, Layout Planning and Design, Layout Generation

ACM Reference Format:

Xi Chen, Yun Xiong, Siqi Wang, Haofen Wang, Tao Sheng, Yao Zhang, and Yu Ye. 2023. ReCo: A Dataset for Residential Community Layout Planning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, October 29–November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3581783.3612465>

1 INTRODUCTION

Layout tasks in architecture and urban planning refer to the physical arrangement of urban spatial components at different scales, in a creative and functionally reasonable way [39], where building floor layout planning, community layout planning and urban planning are the typical tasks from fine- to coarse-grained (shown in Fig. 1), contributing significantly to the quality and sustainability of buildings, neighborhoods or entire cities. The essence of layout planning is an analytical and problem-solving activity that has to meet various requirements and specifications, while traditional expert-empirical-led planning methods are becoming sluggish in the increasingly complex contemporary design context. Recent years have witnessed a rapid growing research interest in intelligent design, e.g., design pattern recognition [5, 18], building volume generation [36], and street network generation [17], using advanced data-driven approaches, which greatly promote the quantitative digital layout planning tasks in a more automated, rational, and efficient way.

Among the three typical layout tasks, the research on building floor layout planning, represented by indoor furniture placement [8, 37, 43, 44, 50] and floor plan generation [3, 4, 34, 35, 41], is the most active, thanks in large part to the relatively high-quality

mature datasets. In contrast, studies on data-driven urban planning and residential community layout planning are still limited [10, 22]. As a key task that directly affects the quality of residence life and urban environmental space [24], residential community layout planning plays a linking role between floor layout and urban planning. However, the effective work cannot be carried out widely due to the lack of large-scale, reliable, and open-source benchmark dataset. Specifically, current work on automatic generative residential layout and design mainly relies on rule-based approaches [10, 32, 36, 49]. Although some efforts have been made to apply Generative Adversarial Networks (GAN) [16] to generative design, the datasets involved cannot be accessed publicly [9]. Besides, the relevant analytic tasks, e.g., community layout pattern recognition [5, 13, 18, 48, 51], also leave the issue of limited performance and inadequate datasets. Despite some algorithms, e.g., online reinforcement learning [42], show promising results to rely less on data or even require no historical data during training processes, sufficient data provided by dataset is still essential when it comes to model effectiveness evaluation [1, 26].

To resolve the data inadequacy issues and pave the way to robust and open data-driven modelling for residential community layout related tasks, in this paper, we introduce the **Residential Community Layout Planning Dataset (ReCo Dataset)**, which is by far the first and largest vector dataset based on real-world residential communities. The ReCo Dataset contains 37,646 residential community layout plans sampled from 60 different cities covering 598,728 residential buildings. The height information of buildings is also included for the extension of two-dimensional (2D) information to 3D space, making ReCo applicable to more 3D modeling and spatial evaluation tasks [13]. Unlike other raster image-based datasets in architectural design fields, e.g., LIFULL HOME's dataset [27] and RPLAN dataset [47], ReCo provides more fine-grained coordinate information, through which commonly used raster data or vector (or polygon) data formats, such as image, Shapefile [15], as well as 3D geometry (with height information) can be exported flexibly. By providing data in this form, the spatial attribute information of buildings, including distance and size, and the fundamental metrics of communities, including Floor Area Ratio (FAR) and Building Coverage Ratio (BCR) can be preserved and calculated, so that the dataset can be adapted to generation and analysis tasks at different granularity.

In addition to ReCo Dataset construction involving data collection and processing, this paper also demonstrates the dataset usability and benchmark in one of the principal downstream data-driven tasks, i.e., automated residential community layout planning, using the Deep Convolutional Generative Adversarial Network (DCGAN) [38] and Conditional Generative Adversarial Nets (cGAN) [33] as backbone. The experiment results confirm the potential of applying our dataset for the tasks in architecture and urban design.

The contributions of our paper can be summarized as follows:

- We release ReCo, the **first large-scale open-source** residential community layout planning dataset. ReCo can be applied to numerous promising applications such as generative layout planning, pattern recognition, classification and spatial evaluation.

- ReCo is a **diverse and extensive** dataset containing layout information of 37,646 residential communities and 598,728 buildings across 60 cities. These sample cities span a large geographical area covering inland cities, coastal cities, etc., with different urban characteristics.
- ReCo is a **fine-grained coordinate information-based** dataset that can be flexibly exported to various common spatial file formats. It provides more spatial attribute information than image-based datasets, so that can be applied to a wider range of tasks at different scales.
- We build two generative models to **validate dataset usability**, which can serve as baselines for benchmarking the task of automated residential community layout planning. We believe ReCo has a great potential to expedite research in the growing field of intelligent layout planning.

2 RELATED WORK

In this section, we firstly review the related methods of the three typical layout tasks, namely community layout planning, building floor layout planning and urban planning. Then we conclude the datasets applied to these tasks. In general, the maturity and scale of existing datasets vary a lot in terms of layout types, resulting in inconsistent development of technical methods in the field. It is particularly time-consuming and challenging to collect high-quality data for community layout research, thus there are only a few small-scale datasets at community level.

Methods. The study of building floor layout planning has been at the forefront of the three, with numerous efforts using advanced algorithms to artificial intelligence task, which has largely replaced the traditional experience-oriented design method. Earlier research is mainly based on computer-aided methods by exploiting explicit rules, i.e., translating domain prior knowledge into computer algorithms [8, 14, 29, 31]. However, methods based on finite rules are bound to be difficult to deal with the complex relational modeling in layout planning, the development of related datasets and models provides solutions to this challenge [2, 6]. As for community layout planning tasks, the existing research mainly focuses on community layout pattern recognition and classification [5, 13]. In contrast with building floor layout planning, there are few studies on community generative design. The development of this field is hindered by the small-scale, closed-source or proprietary datasets [9, 25, 36]. For the coarse-grained tasks of urban planning, most of the current research focuses on design optimization, generative design, and urban environmental evaluation [10, 17, 28, 32]. However, commonly used machine learning methods in the field of Computer Vision (CV), e.g., Variational Auto-Encoder [12] and GAN are under-utilized in design optimization [32]. Only a few methods are developed to generate road networks, but exhibit potential to enhance generative urban design [17]. As for urban environmental evaluation, when numerous objective evaluation metrics are required, only a concept of interactive machine learning integrating clustering, feature extraction, and human subjective-oriented Reinforcement Learning [42] has been proposed [10] where adequate data is particularly important to help establish the objective reward function and evaluation system.

Datasets. Two commonly used open-source datasets for building floor layout planning tasks are *RPLAN dataset* [47] and *LIFULL HOME dataset* [27], which offer 80K annotated house layouts and 5 million ground-truth floorplans, respectively. They have been applied to automatic floorplan generative models [20, 34, 35, 41, 46]. With large-scale datasets provide the sufficient training data for GAN, existing models can automatically formulate floor plans that are indistinguishable from the ground-truth [34, 35]. Community layout planning models are struggling with limited amount of data. For example, *Dong et al.* [13] proposed a Convolutional Auto-Encoder model to embedding plots by applying a dataset with 1,887 samples. The study suggests that larger-scale data covering more attributes can help improve model performance and utility. *Bei et al.* [5] introduced Graph Convolutional Network (GCN) [23] to accomplish different tasks of building state identification, building node clustering and building pattern recognition. Nevertheless, the individual blocks containing building contours in dataset are randomly selected rectangular areas rather than strict community boundaries. In addition, *Yan et al.* [48] presented a GCN to classify building patterns which also remains limited by the small dataset (2,194 samples). For the work on intelligent community layout planning, *Cheng et al.* [9] applied the Convolutional GAN to generate residential layout planning while the training process is still limited by the small sample size of 1,050. The diversity of data is also crucial for pattern recognition and layout generation tasks since it can help provide more sufficient information and wider range of data generation distribution [11, 45]. Furthermore, due to the lack of benchmark datasets, it is difficult to evaluate and compare the performance of different models. In summary, data-driven community layout planning tasks rely heavily on datasets. In terms of urban planning, *Hartmann et al.* [17] applied data extracted from *OpenStreetMap (OSM)*¹ to generate road networks. Although *OSM* contains a large amount of raw geographic data, it cannot be directly applied to model training without complex preprocessing. We summarize and compare the above datasets and ReCo dataset, as shown in Tab. 2.

Comparing three types of layout planning. Generally, the smaller scale building floor layout planning has been more widely studied, especially in generation tasks, benefiting from numerous relatively mature datasets; while the development of the larger-scale of urban planning and community layout planning is subject to the lack of high-quality benchmark datasets. Regarding community layout planning, there is still a lot of room for improvement in the performance of data-driven models, which highlights the urgent need of large-scale open-source benchmark datasets for this research area. We hope to tackle the field development issues through the release of ReCo Dataset, thereby accelerating the maturity of data-driven methods for community layout planning, and even for the smaller- or larger-scale layout tasks in the field of architecture design.

3 RECO DATASET

In this section, we describe the unique characteristics of ReCo, as well as the pipeline to acquire the dataset from the collection of raw

¹<https://www.openstreetmap.org/>

data in real-world residential communities, to the calibration and standardization of community boundaries and building outlines.

3.1 Properties of ReCo Dataset

The ReCo dataset is based on high-precision vector coordinates rather than raster images, which can be easily converted to various data types, such as 2D image, 3D geometry and Shapefile. The properties of ReCo can be summarized in the following three major points:

Diversity. ReCo covers residential community layout data in 60 cities, with different scales of residential areas, residential distribution characteristics, and residential forms, which constitute the diversity of the dataset. See Section 3.4 for specific statistics. ReCo allows researchers to classify datasets for different research purposes based on features that are not limited to location, number of buildings, and plot size.

Flexibility. To our knowledge, ReCo is so far the first and largest open-source vector dataset related to real-world community. The form of the dataset based on coordinates makes it flexible to output various common spatial data formats, and retain the original information of data to the greatest extent.

Uniformity. ReCo can serve as a standard dataset for the residential community layout planning related tasks, providing a benchmark for evaluating the performance of various data-driven models, to facilitate the convergence and progress of advanced algorithms and urban planning.

3.2 Data format and description

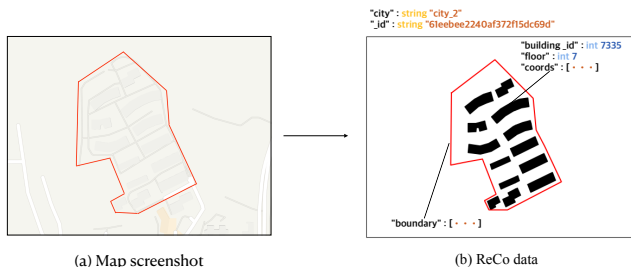


Figure 2: Examples of image data. (a) an example community screenshot in a city from map. (b) the community from ReCo Dataset corresponding to (a).

ReCo is provided with data-interchange formats of JSON and GeoJSON [7] that describes the information of coordinates and spatial attributes. These types of vector data formats can support commonly used spatial format conversions. To make it easier for users to apply ReCo to image-based layout tasks, we provide a way to generate the 2D image from existing dataset formats, with the code published at GitHub².

To explain the content of the ReCo Dataset in detail, we randomly select an example community (shown in Fig. 2 (a)) and convert

the corresponding data in ReCo Dataset into 2D image (shown in Fig. 2 (b)). ReCo consists of three types of instances, namely building, residential community and city. While the basic elements for describing the instances are coordinates. We summarize the basic element and instance types as following (a more detailed example of data instances is shown in Appendix. A):

- **Coordinate:** geographical coordinates have been converted to Mercator coordinates [30] and desensitized for legal and privacy concerns.
- **Building:** residential buildings arranged in the community which consists of a set of coordinates describing the outline, and each building has a unique identifier (“building_id”) within its city limits. The building height attribute is represented by “building floor”, with an assumption that each floor is 3 meters high.
- **Community:** the community where buildings are located which can be recognized by the “_id” (the unique identifier, automatically generated by MongoDB³) with city attribute to explain the location. A set of coordinates (community boundary coordinates) describe community’s outline, and the value of coordinates is constrained to be non-negative.
- **City:** the city where communities are located. In the ReCo, “city” is also considered as one of the attributes of communities. A set of community instances with the same “city” attribute is a sub-set of the dataset.

3.3 Data collection and generating pipeline

In order to capture the morphology of residential community layout plan on a large scale, we prepare two parts of raw data. The first is the information of buildings in the map including buildings’ vector outline and height information (as Building Morphology Data) collected from OpenStreetMap¹ and Google Earth Engine⁴. The second is the residential community information including boundary coordinates information (as Community Morphology Data) acquired through the Baidu Map APIs⁵.

The visualized dataset generation pipeline is presented in Fig. 3. As shown in Fig. 3, after two parts of raw data are collected, the information of corresponding coordinates is extracted. Since the geographic coordinate systems of these multi-source data are different, unification is required to align the two parts of coordinate data. The data from different geographic coordinate systems are projected onto the same 2D plane, i.e., the Transverse Mercator projection [30]. Due to the sensitivity of the Geo information, coordinate correction and desensitization are added after the unification process. Next, we are allowed to determine whether the building belongs to the community by calculating whether the building centroid is within the area enclosed by the community boundary, under the unified spatial environment. Finally, the two parts of data are combined into one as the pipeline output, which completely describes the information of residential community layout plannings. In addition, the information of the building height is also kept. We save the data in JSON and GeoJSON formats for users to export images or Geographic maps.

³<https://www.mongodb.com>

⁴<https://developers.google.com/earth-engine>

⁵<https://lbsyun.baidu.com>

²Related code is at our GitHub repository: <https://github.com/FDUDSDE/ReCo-Dataset>.

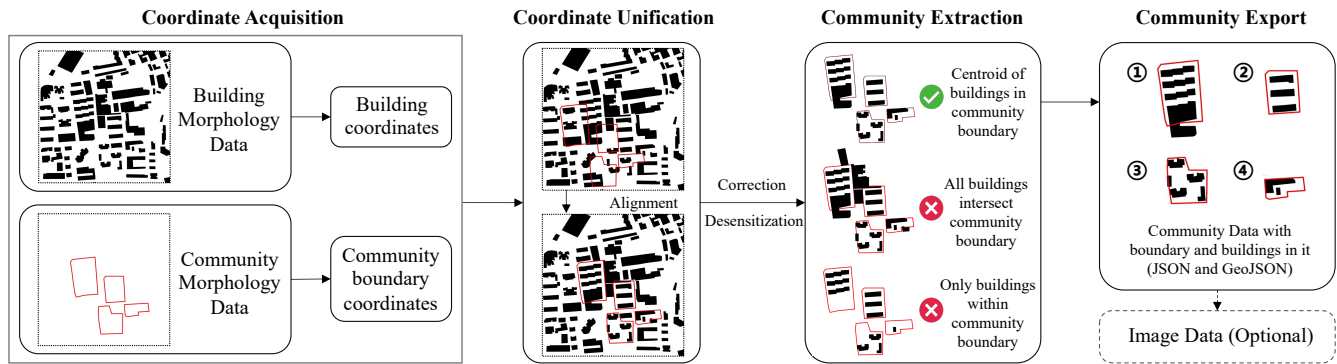


Figure 3: Dataset generating pipeline.

3.4 Dataset statistics

Table 1: Statistics of ReCo Dataset.

Stats	# of Communities	# of Buildings	Average
Max	3,947	70,614	23.40
Min	7	143	10.35
Mean	627.43	9,978.80	/
Total	37,646	598,728	15.90

In order to demonstrate the ReCo more specifically, we provide statistics of the ReCo (shown in Tab. 1). Significantly, all data in ReCo comes from the real world and meets the realistic requirements for the layout of residential buildings. Comparing the datasets for the three typical tasks mentioned in Section 2, the ReCo dataset is by far the largest and the only open-source dataset in the community layout planning field (shown in Tab. 2). Moreover, the data volume of ReCo has increased by more than 17 times compared to previous largest residential layout dataset [48]. In addition, ReCo has the widest data distribution with samples from 60 different cities, which increases the diversity of the dataset. However, compared with the two datasets [27, 47] commonly used in the building floor layout planning, ReCo still has a huge room for improvement in data volume.

4 EXPERIMENTS

The GAN [16] models have achieved a breakthrough in the field of building floor layout planning[3, 34, 35]. We expect that GANs can also be applied to residential community layout planning generation tasks if supported by sufficient data. Based on the ReCo Dataset, therefore, we propose a residential community layout planning generative model, and conduct a baseline experiment. In addition, the residential community layout planning tasks are often subject to various constraints, e.g., community boundary constraints, in the actual design process. Therefore, we propose another generative model for residential community layout planning based on boundary constraints. We aim to answer the following research questions:

- **RQ1:** Can our dataset be applied to residential community layout planning generative tasks, and how does it perform?

- **RQ2:** How does the size of training dataset affect the performance of residential community layout planning generative model?
- **RQ3:** What is the effect of different data distribution (i.e., sampled from different regions) on training of the model?
- **RQ4:** Can our dataset be applied to generative tasks based on community boundary constraints, and how does it perform?
- **RQ5:** Which model performs better? The model with or without the boundary as constraint?

4.1 Residential community layout planning generative model (RQ1,2,3)

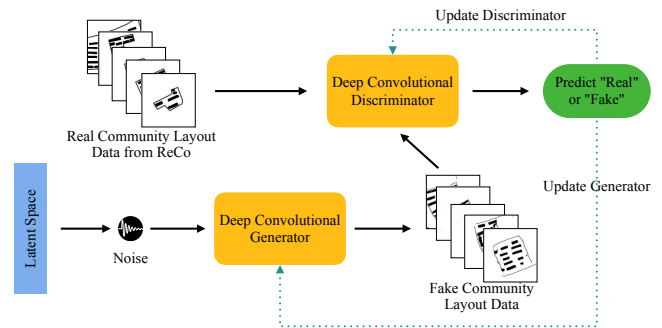


Figure 4: Residential community layout planning generative model architecture.

We trained a DCGAN-based [38] generative model for residential community layout planning by applying ReCo, to demonstrate the applicability of our dataset (our experimental code is redeveloped based on the GitHub repository⁶). The model architecture is illustrated in Fig. 4. ReCo and its three subsets used in the experiments are summarized in Tab. 3. The model was trained for 2K epochs with a batch size of 128 per sub-experiment.

⁶<https://github.com/eriklindernoren/PyTorch-GAN>

Table 2: Datasets comparison. The ReCo Dataset is provided in JSON and GeoJSON formats originally, with support of conversions to commonly used spatial format.

Tasks	Dataset/ Paper	Data type	# of samples	Sampled from	Accessibility
Building floor layout planning	RPLAN [47]	Images	80K	/	Open-source
	LIFULL[27]	Images	5M	/	Open-source
Community layout planning	Dong et al. [13]	Images	1,887	1 city	Private
	Bei et al. [5]	Vectors	1,304+	2 cities	Private
	Yan et al. [48]	Images	2,194	2 cities	Private
	Cheng et al. [9]	Shapefile	1,050	/	Private
	ReCo (Ours)	JSON	37,646	60 cities	Open-source
Urban planning	OSM ¹	*	/	/	Open-source

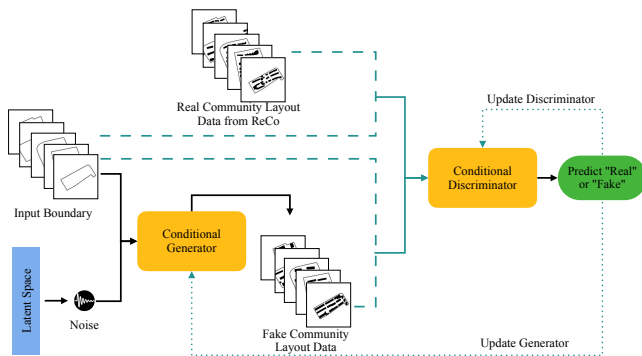
* OSM is considered as raw data and there is no processed public datasets for urban planning tasks.

Table 3: The datasets used in our experiments.

Dataset	Description	Size
city_60	the data from the 60 th city in the ReCo	3,974
h_city_60	randomly sampled from the city_60	2,000
city_40	the data from the 40 th city in the ReCo	2,095
ReCo	the whole ReCo Dataset	37,646

4.2 Boundary constrained residential community layout planning generative model (RQ4,5)

To further meet the design requirements, we took community boundaries as input constraints and trained a cGAN- and pix2pix-based [21, 33] conditional residential community layout planning generative model by applying ReCo (The code is redeveloped based on the GitHub repository⁷). The model architecture is shown in Fig. 5. We used the same datasets in Tab. 3. The model was trained for 2K epochs per sub-experiment.

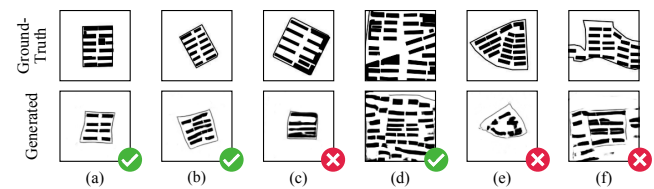
**Figure 5: Boundary constrained residential community layout planning generative model architecture.**

⁷<https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix>

4.3 Results

Model evaluation. To evaluate the performance of the GAN model, it is common practice to use the Fréchet Inception Distance (FID) scores [19, 40], which measures the distance between the distribution of real data and generated data (a smaller FID score means the generated data is closer to the real data). This evaluation method has also been used in HouseGAN++ [35], which is one of the typical models in the field of building floor layout planning.

Generative performance of the model without boundary constraint (RQ1). To demonstrate the model generation performance based on ReCo, we sample the images generated by model trained on the complete ReCo (shown in Fig. 6). As shown in Fig. 6 (a), (b) and (d), we can conclude that part of the generated data has the morphological characteristics of the real data. However, as shown in Fig. 6 (c), there still are communities with uneven building spacing in the generated results. Also, Fig. 6 (e) and (f) show that the existing model performs poorly in the generation of communities with irregular boundaries. We build the boundary constrained model to optimize this situation.

**Figure 6: Examples of ground-truth images and generated images of community layouts. A green mark denotes the preferred design.**

Influence of dataset size (RQ2). As shown in Fig. 7 (a), we observe that the FID score decreases, i.e., the performance of model increases, as the dataset size increases. Besides, it can be found in Fig. 7 (b), that the model trained on the ReCo has better performance than trained only on “city_60”. Similarly, the performance of the model trained on “city_60” is better than trained on “city_40”. These observations indicate that the effect of the dataset size on

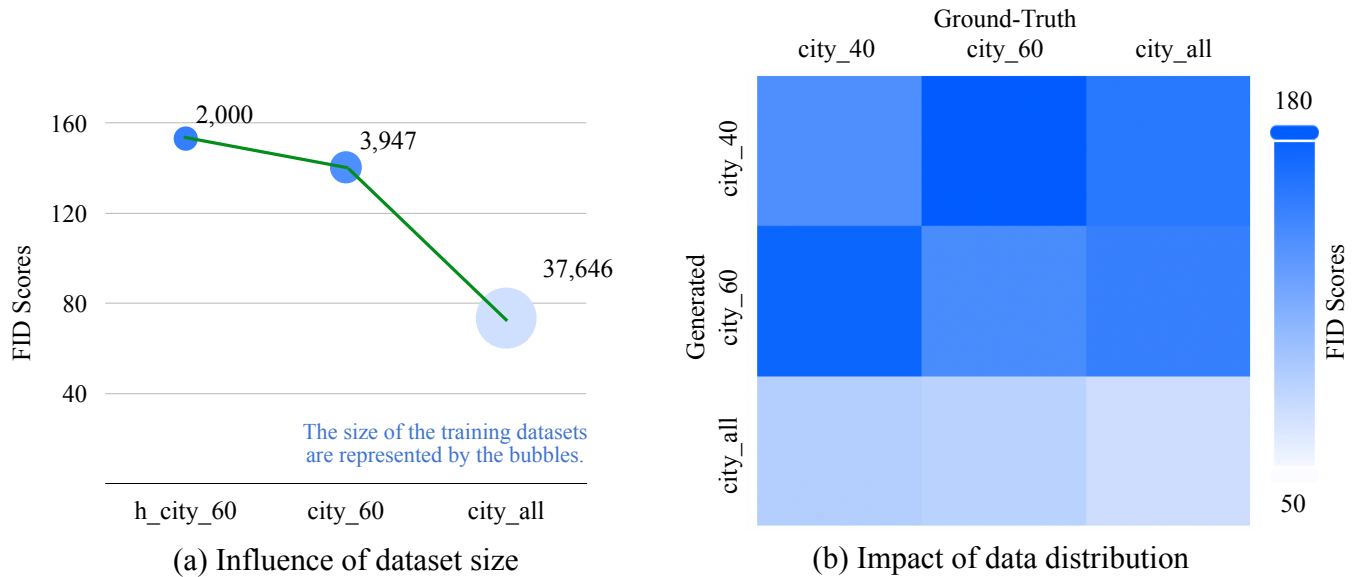


Figure 7: Influence analysis. (FID Scores, the-lower-the-better.)

the experiment, that is, in the case of the same data distribution, the more training data, the better the experimental results.

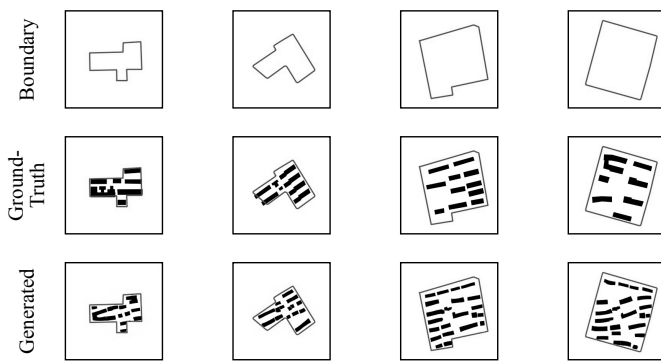
Impact of data distribution (RQ3). As shown in Fig. 7 (b), for the diagonal grid, the score is the lowest in each row. This demonstrates that the FID score between generated data and the corresponding ground-truth data is the lowest. From the perspective of columns, we can see that the lowest scores appear all in the last row, which means the model trained on sufficient data (ReCo) performs better than the model trained on single city data. Moreover, the high scores are seen when comparing data from different cities, e.g., comparing the generated data of model trained by “city_40” to ground-truth data of “city_60”. From these observations, we can conclude that the generated data distributions are closer to the corresponding ground-truth data and there is a certain gap in the data distribution of different cities. This also reflects the diversity of our datasets. However, in the case of sufficient training data, the influence of different training data distribution can be gradually ignored.

Performance of boundary constrained model. (RQ4 and RQ5). We sample images generated by the boundary-constrained model trained on the complete ReCo which is demonstrated in Fig. 8 (a). By comparison with the results in Fig. 2, it can be concluded that the model is able to correctly identify the boundaries of the community and place buildings within the boundary. Furthermore, it performs better in generating communities with irregular boundaries. From the FID scores shown in Fig. 8 (b), we can also conclude that the boundary-constrained model performs better on all four datasets. In practice, community layout planning is also constrained by other indicators, such as FAR, BCR, etc. We speculate that the performance of the model can be further improved in the modeling case with more constraints.

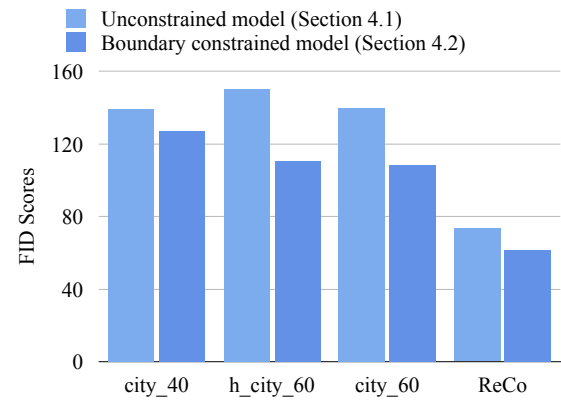
5 DISCUSSION

Limitations. We provide a desensitization dataset ReCo based on coordinate and spatial attribute information, which can be flexibly exported to multiple common spatial file formats. The potential of the ReCo to help researchers build relevant models for the residential community task is also demonstrated in the study. However, there is still some room for improvement in this work. For instance, the Points of Interest (POI) around the community can help researchers obtain contextual information such as functionality, but such information has not been included in the dataset. The height information in our dataset is based on an assumption of average floor height rather than precise building height information. More precise building height information can help researchers to conduct more sophisticated studies. This dataset is somewhat of a mapping of the real-world communities, which means that the dataset should be updated over time. However, most studies are based on historical data, which does not affect research that applies our current dataset. Nonetheless, constantly updating, improving, and maintaining datasets remains a challenge.

Future work. The ReCo Dataset can be extended in the future by collecting and adding more raw data, and can be classified into sub-datasets according to different attributes, such as geographical environment, and community area. Experimental results show that there still are plenty of room to improve the planning effects. This indicates challenges remained in applying ReCo and better models with specific designs are welcome. Our dataset currently only covers residential buildings, we would like to expand the dataset by including other building types, e.g., commercial buildings, and urban complexes, to stimulate more related work, including urban design with different scales and mixed functions. Furthermore, ReCo can also be applied to multiple architectural design tasks, such as obtaining evaluation metrics for designs, and evaluating performance of design schemes or models.



(a) Examples of generated layouts (boundary as input)



(b) Performance comparison

Figure 8: (a) Results of boundary constrained model and (b) model comparison.

6 CONCLUSION

In this paper, we introduce **Residential Community Layout Planning (ReCo)** Dataset, a novel scalable open-source vector dataset related to real-world communities. The current version of the dataset contains 37,646 community layout plans across 60 cities, covering 598,728 residential buildings. The building height information is also included for the extension of 2D information to 3D space. Moreover, we demonstrate the great potential of data-driven models for the automatic generation of community layouts based on our dataset. We expect that our dataset will stimulate extensive research on data-driven approaches for enabling all stages of architecture and urban design.

ACKNOWLEDGMENTS

This work is funded in part by the National Natural Science Foundation of China Projects No. U1936213, No.62176185. This work is also partially supported by the Shanghai Science and Technology Development Fund No. 19DZ1200802, and by the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100) and the Fundamental Research Funds for the Central Universities.

REFERENCES

- [1] Ethem Alpaydin. 2016. *Machine learning: the new AI*. MIT press.
- [2] Gary Anthes. 2013. Deep learning comes of age. *Commun. ACM* 56, 6 (2013), 13–15.
- [3] Imdat As, Siddharth Pal, and Prithwish Basu. 2018. Artificial intelligence in architecture: Generating conceptual design via deep learning. *International Journal of Architectural Computing* 16, 4 (2018), 306–327.
- [4] Fan Bao, Dong-Ming Yan, Niloy J Mitra, and Peter Wonka. 2013. Generating and exploring good building layouts. *ACM Transactions on Graphics (TOG)* 32, 4 (2013), 1–10.
- [5] Weijia Bei, Mingqiang Guo, and Ying Huang. 2019. A spatial adaptive algorithm framework for building pattern recognition using graph convolutional networks. *Sensors* 19, 24 (2019), 5518.
- [6] Yoshua Bengio. 2009. *Learning deep architectures for AI*. Now Publishers Inc.
- [7] Howard Butler, Martin Daly, Allan Doyle, Sean Gillies, Stefan Hagen, Tim Schaub, et al. 2016. The geojson format. *Internet Engineering Task Force (IETF)* (2016).
- [8] Sarvenaz Chaeibakhsh, Roya Sabbagh Novin, Tucker Hermans, Andrew Merryweather, and Alan Kuntz. 2021. Optimizing hospital room layout to reduce the risk of patient falls. *arXiv preprint arXiv:2101.03210* (2021).
- [9] Sun Cheng, Cong Xinyu, and Han Yunsong. 2021. Generative design method of forced layout in residential area based on CGAN. *Journal of Harbin Institute of Technology* 53, 2 (2021), 111–121.
- [10] Artem M Chirkin and Reinhard König. 2016. Concept of interactive machine learning in urban design problems. In *Proceedings of the SEACHI 2016 on Smart Cities for Better Living with HCI and UX*. 10–13.
- [11] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sen Gupta, and Anil A Bharath. 2018. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine* 35, 1 (2018), 53–65.
- [12] Carl Doersch. 2016. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908* (2016).
- [13] Jia Dong, Li Li, and Dongqing Han. 2019. New Quantitative Approach for the Morphological Similarity Analysis of Urban Fabrics Based on a Convolutional Autoencoder. *IEEE Access* 7 (2019), 138162–138174. <https://doi.org/10.1109/ACCESS.2019.2931958>
- [14] Gavrilov Egor, Schneider Sven, Denmark Martin, and Koenig Reinhard. 2020. Computer-aided approach to public buildings floor plan generation. Magnetizing Floor Plan Generator. *Procedia Manufacturing* 44 (2020), 132–139.
- [15] ESRI ESRI. 1998. Shapefile technical description. *An ESRI white paper* 4, 1 (1998).
- [16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).
- [17] S. Hartmann, M. Weinmann, R. Wessel, and R. Klein. 2017. StreetGAN: towards road network synthesis with generative adversarial networks. In *International Conferences in Central Europe on Computer Graphics*.
- [18] Xianjin He, Xinchang Zhang, and Qinchuan Xin. 2018. Recognition of building group patterns in topographic maps based on graph partitioning and random forest. *ISPRS Journal of Photogrammetry and Remote Sensing* 136 (2018), 26–40.
- [19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).
- [20] Ruizhen Hu, Zeyu Huang, Yuhan Tang, Oliver Van Kaick, Hao Zhang, and Hui Huang. 2020. Graph2plan: Learning floorplan generation from layout graphs. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 118–1.
- [21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*.
- [22] Avinash Kumar Jha, Awishkar Ghimire, Surendrabikram Thapa, Aryan Mani Jha, and Ritu Raj. 2021. A Review of AI for Urban Planning: Towards Building Sustainable Smart Cities. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 937–944.
- [23] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [24] Dayi Lai, Chaobin Zhou, Jianxiang Huang, Yi Jiang, Zhengwei Long, and Qingyan Chen. 2014. Outdoor space quality: A field study in an urban residential community in central China. *Energy and Buildings* 68 (2014), 713–720.
- [25] Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, and Tingfa Xu. 2019. Layoutgan: Generating graphic layouts with wireframe discriminators. *arXiv preprint arXiv:1901.06767* (2019).
- [26] Yuxi Li. 2017. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274* (2017).
- [27] Ltd. LIFULL Co. [n. d.]. Lifull home’s dataset. <https://www.nii.ac.jp/dsc/idr/lifull>. Accessed April 8, 2022.
- [28] Lun Liu, Elisabete A Silva, Chunyang Wu, and Hui Wang. 2017. A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. *Computers, environment and urban systems* 65 (2017), 113–125.

- [29] Chongyang Ma, Nicholas Vining, Sylvain Lefebvre, and Alla Sheffer. 2014. Game level layout from design specification. *Computer Graphics Forum* 33, 2 (2014), 95–104.
- [30] Derek Hylton Maling. 2013. *Coordinate systems and map projections*. Elsevier.
- [31] Paul Merrell, Eric Schkufza, and Vladlen Koltun. 2010. Computer-generated residential building layouts. *ACM Transactions on Graphics (TOG)* 29, 6 (2010), 1–12.
- [32] Yufan Miao, Reinhard Koenig, and Katja Knecht. 2020. The development of optimization methods in generative urban design: a review. In *Proceedings of the 11th Annual Symposium on Simulation for Architecture and Urban Design*. 1–8.
- [33] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).
- [34] Nelson Nauata, Kai-Hung Chang, Chin-Yi Cheng, Greg Mori, and Yasutaka Furukawa. 2020. House-gan: Relational generative adversarial networks for graph-constrained house layout generation. In *European Conference on Computer Vision*. Springer, 162–177.
- [35] Nelson Nauata, Sepidehsadat Hosseini, Kai-Hung Chang, Hang Chu, Chin-Yi Cheng, and Yasutaka Furukawa. 2021. House-GAN++: Generative Adversarial Layout Refinement Networks. *arXiv preprint arXiv:2103.02574* (2021).
- [36] Iuliia Osintseva, Reinhard Koenig, Andreas Berst, Martin Bielik, and Sven Schneider. 2020. Automated parametric building volume generation: a case study for urban blocks. In *Proceedings of the 11th Annual Symposium on Simulation for Architecture and Urban Design*. 1–8.
- [37] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. 2021. Atiss: Autoregressive transformers for indoor scene synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 12013–12026.
- [38] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).
- [39] Sven Schneider, Jan-Ruben Fischer, and Reinhard König. 2011. Rethinking automated layout design: developing a creative evolutionary design method for the layout problems in architecture and urban design. In *Design computing and cognition '10*. Springer, 367–386.
- [40] Maximilian Seitzer. 2020. pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid>. Version 0.2.1.
- [41] Jiahui Sun, Wenming Wu, Ligang Liu, Wenjie Min, Gaofeng Zhang, and Liping Zheng. 2022. WallPlan: synthesizing floorplans by learning to generate wall graphs. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–14.
- [42] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.
- [43] Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X Chang, and Daniel Ritchie. 2019. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–15.
- [44] Kai Wang, Manolis Savva, Angel X Chang, and Daniel Ritchie. 2018. Deep convolutional priors for indoor scene synthesis. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.
- [45] Shuo Wang and Xin Yao. 2009. Diversity analysis on imbalanced data sets by using ensemble models. In *2009 IEEE symposium on computational intelligence and data mining*. IEEE, 324–331.
- [46] Shidong Wang, Wei Zeng, Xi Chen, Yu Ye, Yu Qiao, and Chi-Wing Fu. 2021. ActFloor-GAN: Activity-Guided Adversarial Networks for Human-Centric Floorplan Design. *IEEE Transactions on Visualization and Computer Graphics* (2021).
- [47] Wenming Wu, Xiao-Ming Fu, Rui Tang, Yuhang Wang, Yu-Hao Qi, and Ligang Liu. 2019. Data-driven Interior Plan Generation for Residential Buildings. *ACM Transactions on Graphics (SIGGRAPH Asia)* 38, 6 (2019).
- [48] Xiongfeng Yan, Tinghua Ai, Min Yang, and Hongmei Yin. 2019. A graph convolutional neural network for classification of building patterns using spatial vector data. *ISPRS journal of photogrammetry and remote sensing* 150 (2019), 259–273.
- [49] XY Ying, XY Qin, JH Chen, and J Gao. 2021. Generating Residential Layout Based on AI in the View of Wind Environment. *Journal of Physics: Conference Series* 2069, 1 (2021), 012061.
- [50] Lap Fai Yu, Sai Kit Yeung, Chi Keung Tang, Demetri Terzopoulos, Tony F Chan, and Stanley J Osher. 2011. Make it home: automatic optimization of furniture arrangement. *ACM Transactions on Graphics (TOG)-Proceedings of ACM SIGGRAPH 2011*, v. 30,(4), July 2011, article no. 86 30, 4 (2011).
- [51] Xiang Zhang, Tinghua Ai, Jantien Stoter, and Xi Zhao. 2014. Data matching of building polygons at multiple map scales improved by contextual information and relaxation. *ISPRS Journal of Photogrammetry and Remote Sensing* 92 (2014), 147–163.

A DETAILS OF DATA INSTANCE

Instance	Data	Data Format	Example/ Describe
City	Community	Instances	A set of Community instances
Community	_id	String	“61ef8a8b32b5d4672152cf73”
	boundary	Coordinates	A set of 2D coordinates
Building	Building	Instances	A set of Building instances
	City	String	“city_16”
Building	building_id	Int	35710
	floor	Int	3
	coords	Coordinates	A set of 2D coordinates

Table 4: Details of each data instance.