

CARNEGIE MELLON UNIVERSITY

School of Architecture

College of Fine Arts

Thesis

Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Computational Design

TITLE:

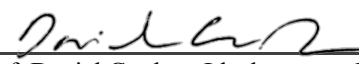
Datasets for Architectural Research:

A review of selected plan-based datasets and their implementations

AUTHOR:

Junjie Xu

ACCEPTED BY ADVISORY COMMITTEE:


Prof. Daniel Cardoso Llach Principal Advisor

August 9, 2023
DATE


Prof. Eddy Man Kim Secondary Advisor

August 9, 2023
DATE


Prof. Jinmo Rhee Third Advisor

August 9, 2023
DATE

Datasets for Architecture Research: A review of selected plan-based datasets and their implementations

Junjie Xu

Abstract

Architectural research increasingly relies on plan-based datasets to support various applications of computational design and machine learning tools, such as structural recognition, plan generation, and 3D scene reconstruction. However, there is a lack of understanding and comparison of existing datasets' characteristics, potentials, and challenges. This thesis aims to address this gap by conducting a selective review of existing plan-based datasets and their implementations for architectural research. The thesis examines the data collection, annotation, analysis, and limitations of selected datasets with their usage and evaluation. The ultimate goal is to suggest best practices and future directions for architectural research using plan-based datasets. This thesis aims to contribute to the current research by providing a better understanding of the potential of these datasets to support architectural research. By providing potential for data format organization and proposing a collaborative data annotation process from the findings of the dataset review, the work can provide insights for researchers and practitioners interested in exploring the possibilities of plan-based datasets for computational design.

Acknowledgment

I would like to express my sincere gratitude to my principal advisor, Prof. Daniel Cardoso Llach, for his guidance, support, and encouragement throughout this thesis. His expertise, criticality, and insights have been invaluable for my research and learning.

I am also grateful to my secondary advisor, Prof. Eddy Man Kim, and my third advisor, Prof. Jinmo Rhee, for their constructive feedback and helpful suggestions in shaping the current narrative of my thesis. I appreciate their time and effort in reviewing my work and participating in my thesis committee.

I also want to thank the School of Architecture and the Computational Design program at Carnegie Mellon University for providing me with the opportunity and resources to pursue this research. I am indebted to my classmates and friends in the Computational Design program for their collaboration, inspiration, and friendship. I have learned so much from them and enjoyed all the conversations with them on various projects. I also want to thank my friends, AUO members, church family, and Beeler gangs. They made important memories in making my CMU years unforgettable, and the list continues. Especially, I want to mention Boyang Lin for the suggestions and support at the lab during the summer and Tingsong Ou for the kind help and advice on the technical implementation of my thesis.

Last but not least, I would like to express my love and gratitude to my beloved family for their unconditional support and encouragement. They have always been there for me in every step of my journey. I dedicate this thesis to them.

Contents

Abstract	2
Acknowledgment	3
Contents	4
1) Introduction	6
a) Deep Learning Models and the Datasets in Architecture	7
b) Thesis Vision	10
c) Thesis Outline	10
2) Background	11
a) Dataset preparation: current practice	12
b) Preliminary Experiments	15
i) Preparation of new datasets	16
ii) Testing on floor plan generation	17
iii) Testing on circulation information	20
c) Dataset Survey in other disciplines	22
3) Hypothesis	26
a) Architecture Plan Datasets Overview	27
i) Availability of Plan Datasets among platforms	27
ii) Data Type Review of Building Plan Datasets	29
iii) Overview of Selected Plan Datasets	32
b) Research Questions	33
c) Research Goal	34
4) Methods: Dataset Survey	35
i) CVC-FP Dataset	36
ii) ROBIN Dataset	42
iii) CubiCasa5k Dataset	46

iv) RPlan	49
v) FloorPlanCAD	54
vi) Jigsaw Plan	58
5) Findings and Discussions	61
a) Potentials for data format organization	62
b) Towards a collaborative data annotation process	65
6) Conclusion	67
a) Contributions	68
b) Limitations	68
c) Future work	68
Bibliography:	70
Appendix	73

1) Introduction

Architectural researchers have grown interested in using plan-based datasets to support generative models and machine learning tools. As researchers have shown, the curation of architectural datasets can profoundly influence the performance and outcome of those tools¹, as well as the richness and diversity of the architectural forms and contexts². However, there is a lack of studies about these architectural datasets despite the considerable potential of the data. Therefore, there needs to be a more comparative understanding of these architectural datasets' characteristics, potentials, and challenges. Moreover, this thesis aims to address this gap by conducting a selective review of existing plan-based datasets and their implementations for architectural research.

a) Deep Learning Models and the Datasets in Architecture

Recently, many computational design studies in architecture have delved into using technology to automate the generation of architectural plans. The recent topic regarding generative models came to the forefront with the release of the pix2pix paper by UC Berkeley (UCB)³. When the tool was invented by a group of UCB researchers earlier as an image-to-image translation tool using a conditional adversarial network, it provided a framework for many pixel-based transformation tasks with testing cases on semantic labels, architectural labels, maps, sketches, etc.⁴ Since then, architectural researchers have also explored the possibility of using data to create architectural floor plan designs. To do this, they often make their floor plan datasets to train their neural networks on. In these datasets, the parameters and labeling are important to the accuracy of the final generated plans. For instance, Hao Zheng and Weixin Huang published a paper examining using GANs for plan generation⁵. In their study, the experimental datasets were separated into two categories: spatial program recognition (created from an input of a rendered interior plan) and interior plan generation (generated from an input of a segmented program layout). The dataset used in this research was

¹ Newton, David William. "Deep Generative Learning for the Generation and Analysis of Architectural Plans with Small Datasets." *Blucher Design Proceedings* (2019).

² Cardoso Llach, Daniel. "Sculpting Probabilistic Spaces: Brief History and Prospects of Machine Learning in Design." In *The Routledge Companion to Artificial Intelligence in Architecture*, edited by Wassim Jabi and Mine Özkar, 1st ed., 3-25. Routledge, 2021.

³ Isola, Phillip, et al. "Image-to-Image Translation with Conditional Adversarial Networks." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017. Crossref, <https://doi.org/10.1109/cvpr.2017.632>.

⁴ As 2. Isola, Phillip, et al. "Image-to-Image Translation with Conditional Adversarial Networks."

⁵ Zheng, H., & Huang, W. (2018). *Architectural drawings recognition and generation through machine learning*. Cambridge: MA, ACADIA

sourced from a house rental platform and included information about the spatial division, specific furniture, and floor furnishings (See Fig. 1).

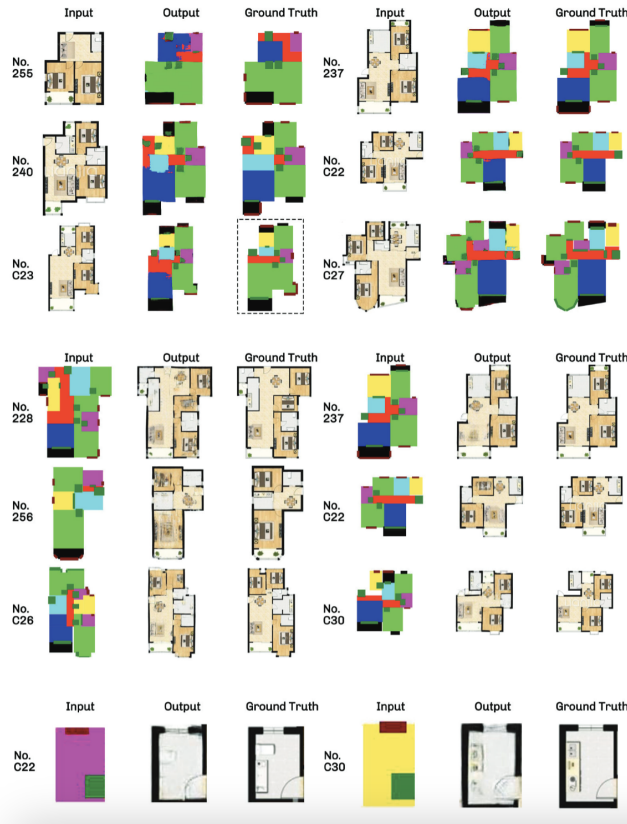


Figure 1. Generative Models by Huang and Zheng in recognizing and generating plans⁶.

Chaillou (2019) demonstrated the floorplan generation using the pix2pix across all different scales of architectural design-furniture, room, and building scale⁷ (See Fig. 2). He labeled the data based on the program of the room, the structural elements, entrances, and window openings. Despite some inaccuracies in architectural elements in the generated “pixelated images,” some plans still show a readable or interpretable set of design outcomes. He also tested datasets with different architectural styles (such as Gothic architectural style) using the same pix2pix platform. His paper pointed out such diverse and imaginary potentials for architects to generate specific design output with a well-curated dataset.

⁶ As 2. Zheng, H., & Huang, W. (2018).

⁷ Chaillou, Stanislas. “ArchiGAN: Artificial Intelligence x Architecture.” (2020).

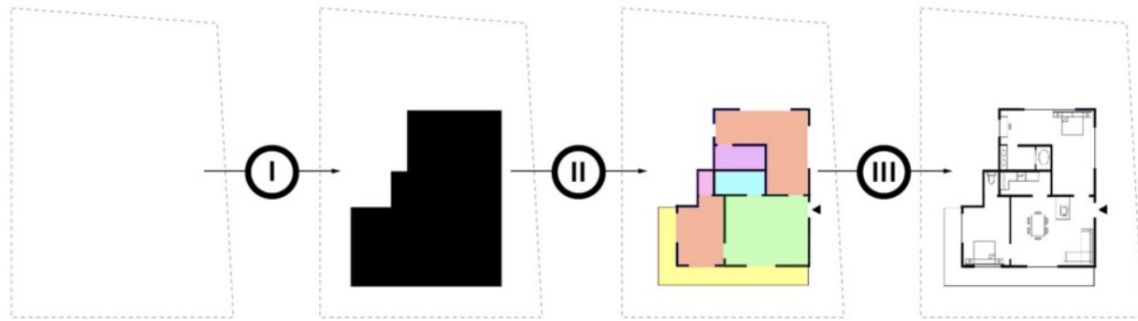


Figure 2. Generative Stack in Chaillou's research ArchiGAN⁸.

In addition to pixel-based methods, other researchers have also developed vector-based approaches. Chen Liu and colleagues introduced Floor Plan Transformation, which converts a rasterized floor plan image into a vector-graphics representation of a 2D plan and an extruded 3D model⁹. This learning-based approach addresses the limitations of pixel-based plans by converting the image into a set of junctions with semantic information, such as door endpoints and wall corners. These junctions are then transformed into simple vectorized elements such as wall lines, door lines, or icon boxes. Jimo Rhee at Carnegie Mellon University also explored the potential of combining building footprint data with Pittsburgh urban contexts as the conditionals of the footprint generation¹⁰. These studies have shown architectural datasets' potential in solving inquiries of architectural and urban questions. Apart from those machine-learning-based models, researchers have recently applied large language models, such as Stable Diffusion models¹¹, to architecture plan generation¹². This research proposes technical-driven solutions based on a task-oriented large dataset, and the existing diverse set of data becomes their base for evaluating the proposed methods.

A paper by Silvio Carta, the head of Art and Design at the University of Hertfordshire, discusses the

⁸ as 4. Chaillou, Stanislas. "ArchiGAN: Artificial Intelligence x Architecture." (2020).

⁹ C. Liu, J. Wu, P. Kohli and Y. Furukawa, "Raster-to-Vector: Revisiting Floorplan Transformation," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2214-2222, doi: 10.1109/ICCV.2017.241.

¹⁰ Rhee, Jinmo & Veloso, Pedro & Krishnamurti, Ramesh. (2020). Integrating Building Footprint Prediction and Building Massing: An Experiment in Pittsburgh. 10.52842/conf.caadria.2020.2.669.

¹¹ Rombach, Robin, et al. "High-Resolution Image Synthesis with Latent Diffusion Models." 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022. Crossref, <https://doi.org/10.1109/cvpr52688.2022.01042>.

¹² Hao Zheng. <https://www.linkedin.com/feed/update/urn:li:activity:7036371613503647745/>. Accessed March 22, 2023

research and inquiry surrounding the field of self-organizing floor plan generation¹³ (Carta et al., 2020). The paper highlights the need for deep learning in design, as classic machine learning approaches to architectural design, such as self-organizing maps (SOM) and genetic algorithms (GA), rely on heuristic algorithms that introduce randomness and can be difficult to control. On the other hand, deep learning can provide an alternative approach that allows for more control over the generated design results through datasets. This thesis presents a new potential for combining generative and rule-based algorithms in architectural practice. As such, the importance of datasets in this field cannot be overstated, and it is crucial to review and understand them to realize their potential fully.

***b)* Thesis Vision**

This thesis aims to fill the gap in the field of architectural research by conducting a selective review of existing plan-based datasets and their implementations. Despite the growing interest in using these datasets to support generative models and machine learning tools, a comparative understanding of their characteristics, potentials, and challenges is needed. Through this examination, this thesis will identify the gaps and limitations of current datasets and propose new potentials for data format organization and collaborative annotation for architectural researchers.

***c)* Thesis Outline**

The thesis is structured as follows: Chapter 2 reviews the background literature on dataset preparation and dataset survey, experimenting with creating a set of floor plan datasets with annotations. Chapter 3 presents the hypothesis and research questions. It proposes that by conducting a selective review of existing plan-based datasets and their implementations, researchers can better understand the potential of these datasets to support architectural floor plan research. Chapter 4 describes the methods and results of the dataset survey, which aims to provide a selective review of existing plan-based datasets and their implementations. Chapter 5 discusses the findings and implications of the survey. It summarizes the main features and gaps of existing plan-based datasets and proposes potentials in data format organization and collaborative data annotation. Finally, Chapter 6 concludes the thesis and suggests future work.

¹³ Carta, Silvio. “Self-Organizing Floor Plans.” *Harvard Data Science Review* (2021): <https://hdsr.mitpress.mit.edu/pub/w1gujxim/release/3> .

2) Background

This section begins with an understanding of the data preparation process of the current practice. Then the section will talk about experiments in creating a set of datasets with labels and using GANs to generate architectural plans. The study indicated that the quality and accuracy of the plans depend on the datasets' size, diversity, and label information. After that, this section looks at the current practice of dataset surveys in other disciplines, which will provide a foundation for the later dataset survey of the thesis.

a) Dataset preparation: current practice

One of the most important questions to ask when discussing architectural datasets is how to prepare them, which is essential to survey the current data preparation practice in this thesis. Thereby, it is crucial to understand the data preparation and construction (data curation) process in the general machine learning field. However, it is important to recognize that researchers use different methods to prepare the specific data of their domain. The approaches discussed in this section may only be one of the many ways of doing that.

The Korea Advanced Institute of Science and Technology (KAIST) research paper surveyed data collection processes for Machine Learning in 2019¹⁴. The article pointed out two reasons for data collection from the data management perspective. One is the need for more labeled training data, especially for recent applications, due to the expense and domain expertise. The other reason is the need for more datasets for deep learning. The current practices divide the data collection into data acquisition, data management, and recent data/model improvement.¹⁵

¹⁴ Y. Roh, G. Heo and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, pp. 1328-1347, 1 April 2021, doi: 10.1109/TKDE.2019.2946162.

¹⁵ As 14. Y. Roh, G. Heo and S. E. Whang,

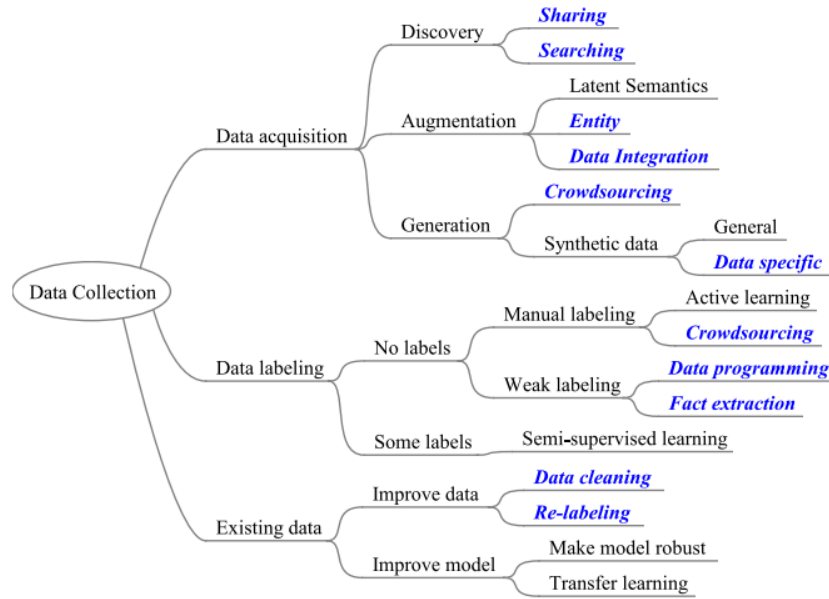


Figure 3. Dataset Collection Diagram for Machine Learning, by Roh, and Heo et al¹⁶.

The diagram (Fig. 3) above shows a detailed diagram of the data collection process for machine learning. The KAIST paper mentioned above-conducted research on the machine learning community (e.g., traditionally ML-heavy fields such as NLP and CV) and the data management community (under data science). The first step is data acquisition, which contains three approaches: data discovery (through sharing collaboratively or searching), data augmentation (to add more external data for current existing datasets), and data generation (to get new datasets from crowdsourced or synthetic datasets when no existing dataset are available for specific application). The table below (table. 2) illustrates lists of data acquisition techniques by their tasks and approaches. And each of the tasks has multiple problems and potentials to look into. For example, the search of datasets currently focuses on the format of tuples, documents, or webpage as the state of the art¹⁷. However, new methods from other fields, such as information retrieval, databases, and semantic web search, can potentially integrate with dataset search solutions (See Table 3).

¹⁶ As 14. Y. Roh, G. Heo and S. E. Whang,

¹⁷ Chapman, Adriane, et al. "Dataset Search: A Survey." The VLDB Journal, vol. 29, no. 1, Aug. 2019, pp. 251–72. Crossref, <https://doi.org/10.1007/s00778-019-00564-x>.

Task	Approach	Techniques
Data discovery	Sharing	Collaborative Analysis [9]–[11]
		Web [12]–[17]
	Searching	Collaborative and Web [18]
		Data Lake [19]–[23]
Data augmentation		Web [24]–[34]
		Deriving Latent Semantics [35]–[37]
		Entity Augmentation [30], [31]
Data generation	Crowdsourcing	Data Integration [38]–[44]
		Gathering [45]–[54]
	Synthetic Data	Processing [49], [50], [55], [56]
		Generative Adversarial Networks [57]–[62]
		Policies [63], [64]
		Image [65]–[71]
		Text [72]–[74]

Table 2. A classification of dataset acquisition techniques by Roh, and Heo et al¹⁸.

	Entity-centric	Information retrieval	Database	Tabular search
Open data publishing platforms (e.g., CKAN, Socrata)		•		
Data marketplaces	•			
Linked data search engines	•	•	•	
Google DSS	•	•		

Table 3. Search Technology used across implementations by Roh, and Heo et al¹⁹.

The data labeling process is tight with data acquisition in many scenarios, on the dataset collection survey paper²⁰. Labeling can be divided into several categories, using existing labels, using crowd-sourcing to label individual examples, or applying active learning, generating less than perfect labels (which the paper called “weak labels”) of larger quantities. The survey also categorized labeling approaches by specific ML task (e.g., classification and regression) and data type (the input and output). A detailed category of labeling techniques can be found in the table below (See Table 4), with techniques indicating more than 50 papers the researcher surveyed.

¹⁸ As 38. Y. Roh, G. Heo and S. E. Whang,

¹⁹ As 41. Chapman, Adriane, et al.

²⁰ As 38. Y. Roh, G. Heo and S. E. Whang.

Category	Approach	Machine learning task	Data types	Techniques
Use Existing Labels	Self-labeled	classification	all	[92]–[96]
		regression	all	[97]–[99]
	Label propagation	classification	graph	[100]–[102]
		classification	all	[103]–[109]
Crowd-based	Active learning	regression	all	[110]
			text	[111], [112]
	Semi-supervised+Active learning	classification	image	[113]
			graph	[114]
			all	[50], [54], [115]–[122]
	Crowdsourcing	regression	all	[123]
Weak supervision	Data programming	classification	all	[3], [124]–[127], [127]–[130]
	Fact extraction	classification	text	[131]–[142]

Table 4. Search Technology used across implementations by Roh, and Heo et al.²¹.

The survey also pointed out some approaches based on re-labeling and cleaning the current data and models²². The approach is applicable especially when new datasets are hard to find for specific applications or adding data could not enhance the model's performance. Although architectural researchers have different approaches and standards for dataset preparation (as mentioned in section 1.1 from the Introduction chapter), the references above of the current practice dataset preparation process will provide a detailed and structured approach for dataset preparation in architectural design.

b) Preliminary Experiments²³

Based on the importance of the preparation process of datasets in the research and design methods using deep learning, this section presents a hands-on experiment for floor plan image generation. The experiment is constructing a set of datasets with detailed labels and generating architectural plans using deep learning. The detailed pipeline and process are documented in the subsections below. The experiments revealed the limitations of the labeled datasets in that the generated plan's quality and accuracy heavily depend on the size and diversity of datasets and the feature or label information on the original data. Therefore, further research on data preparation, annotation, and analysis of existing datasets for architectural plan generation is essential.

²¹ As 38. Y. Roh, G. Heo and S. E. Whang

²² As 38. Y. Roh, G. Heo and S. E. Whang

²³ Part of the experiment continued my collaboration with researcher Jinglun Zhao in the University of Hong Kong. Link from https://issuu.com/hkuarchitecture/docs/prospectus2020_hkudoa_digital_page

i) **Preparation of new datasets**



Figure 4. Overview of prepared datasets in Adobe Illustrator.

In the series of tests based on the pix2pix model, a specific dataset consisting of 80 floor plans was produced with vector-based data to create different sets of planar images (See Fig 4). The dataset is mainly based on CVC-FP dataset²⁴ and re-edited and colored to test the efficacy of zoning colors, circulation, window and door labels, wall divisions, and hatching patterns in the future (See Fig 5).

²⁴ Heras, Lluís-Pere & Terrades, Oriol & Robles, Sergi & Sánchez, Gemma. (2015). CVC-FP and SGT: a new database for structural floor plan analysis and its ground truthing tool. International Journal on Document Analysis and Recognition (IJDAR). 18. 10.1007/s10032-014-0236-5.

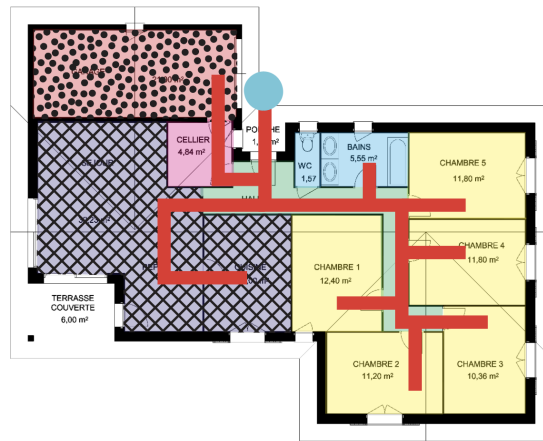


Figure 5. Zoom-in picture of single floor plan data.

ii) Testing on floor plan generation

The process of designing a plan is always the core of architectural training. A well-designed architectural plan can take years of practice for an architect. With the development of Artificial Intelligence (AI) and deep learning, researchers have implemented ways to use GAN to train the computer to generate architectural elements and style. Inspired by Hao Zheng's use of pix2pixHD to generate house plans²⁵. This project tries to examine the effectiveness of GAN in producing architectural plans from three perspectives: 1) generating apartment plans from footprint datasets 2) generating zoning division data from footprint datasets 3) generating apartment plan from zoning division.

Legend of color for the zoning plan representing different function zoom as follows: green for the balcony, yellow for the living room, red for the kitchen, cyan for storage, dark blue for the toilet, and pink for the bedroom.

²⁵ As Zheng, H., & Huang, W. (2018). Architectural drawings recognition and generation through machine learning. Cambridge: MA, ACADIA

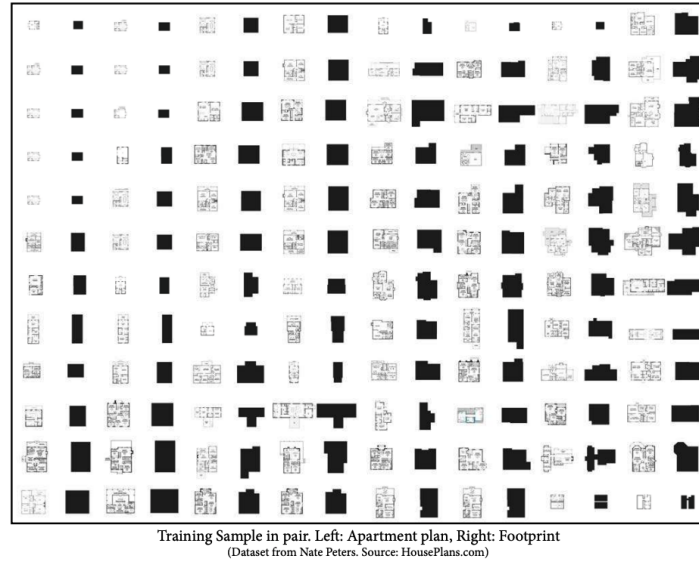


Figure 6. Training Samples in pair with pix2pix.

My experiment aimed to generate pixel images of one plan representation from the other. At the same time, the discriminator is also getting trained to better distinguish fake images (“plan-alike” or “room-zoning alike” images) from real ones. After 150 epochs of training for the whole train set(300 images including data augmentation), I examined the result of the Generator by inputting 100 footprint images (or Room Zoning images) to see the generator’s output (See Fig. 6).

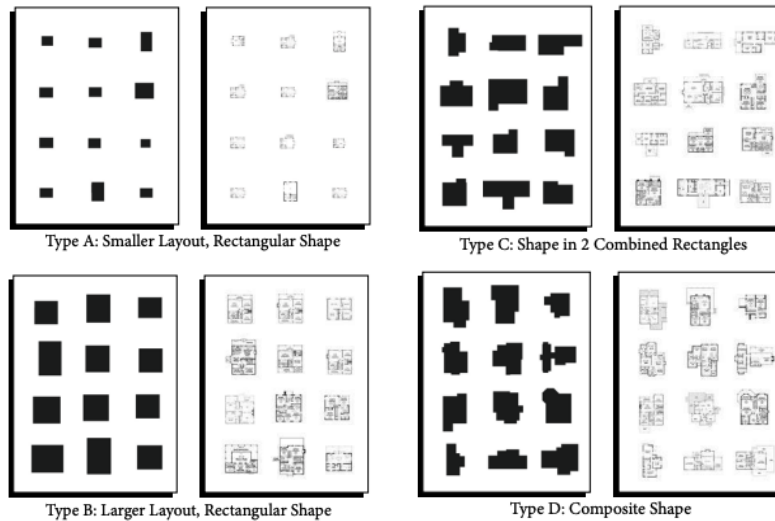


Figure 7. Training Samples in pair with pix2pix.

Considering the inclusive types within the data, I analyzed the generator’s performance by four typologies. The training process from epoch 1 to epoch 150 was documented and shown below. The

generator's result (room zoning or plan) would also be represented together with the input (footprint or zoning) and the ground truth (as the original matching image for input).

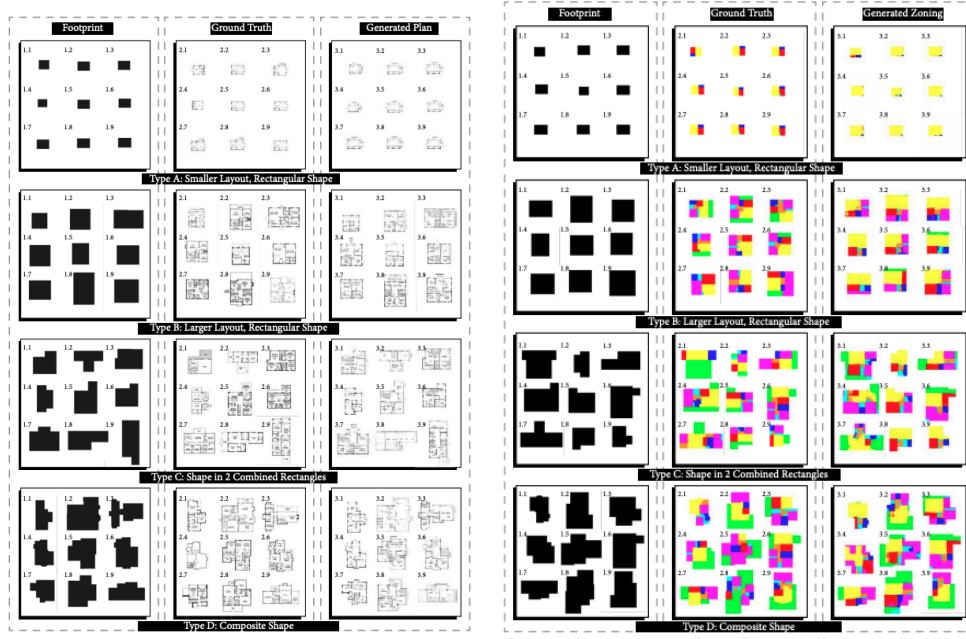


Figure 8. Experiment 1 (left) and 2 (right) for plan generation.

From the performance of the generated plan, Type B, C, and D can get a generally cleaner plan with partitions (See Fig. 7). Type A could be more effective due to the homogeneous result and ambiguous inner layout. But still, the generator sometimes performs poorly on some elongated plans (eg, Type C 3.8, Type D 3.9).

Experiment 2 aims to generate zoning division from the black-and-white footprint. The training process during different epochs is documented every 20 epochs. I examined the result of the generator by inputting 100 Footprint images (or room zoning images) to see the generator's output after 150 epochs of training the GAN (See Fig. 8). The generator of Experiment 2 does not perform so well in general. But like Experiment 1, Type B, C, and D perform better than Type A according to the result on the right. The result can be cleaner, but for some cases (e.g. Type C 3.7, Type D 3.5), the separation of different room zoning is unclear.

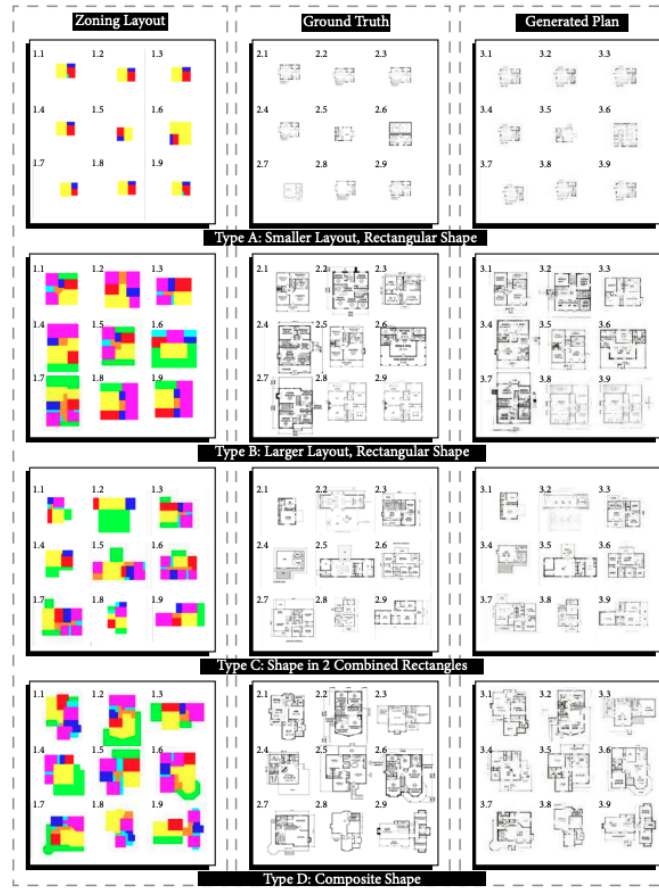


Figure 9. Experiment 3 for plan generation.

Experiment 3 aims to enhance the generator to produce “plan-like” images from room zoning (or “generated room zoning” from Experiment 2) (See Fig. 9). I examined the result of the generator by inputting 100 Footprint images (or room zoning images) to see the generator’s output after 150 epochs of training the GAN. The Generator’s result of Experiment 3 is not generally effective in deepening the room plan from the zoning layout. But still, some of the elements’ shapes, such as the table in the living room or the bed in the bedroom, could be found in some cases. The notations of the original data set may influence the result. These preliminary experiments could create further potential for various datasets, such as the transformation process from sketches to planar design outcomes.

iii) Testing on circulation information

The primary objective of Experiment 4 is to generate "Room-Zoning alike" images from Footprint

with circulation (See Fig. 10). This enhancement aims to enable architects to interact with the neural network using more detailed information, such as circulation, thereby allowing the generated plans to be more flexible and personalized to meet specific requirements. The training process during different epochs is thoroughly documented and evaluated every 20 epochs. The effectiveness of the generator is evaluated by inputting 100 Footprint images (or Room Zoning images) and examining the generator's output after 150 epochs of training the GAN.

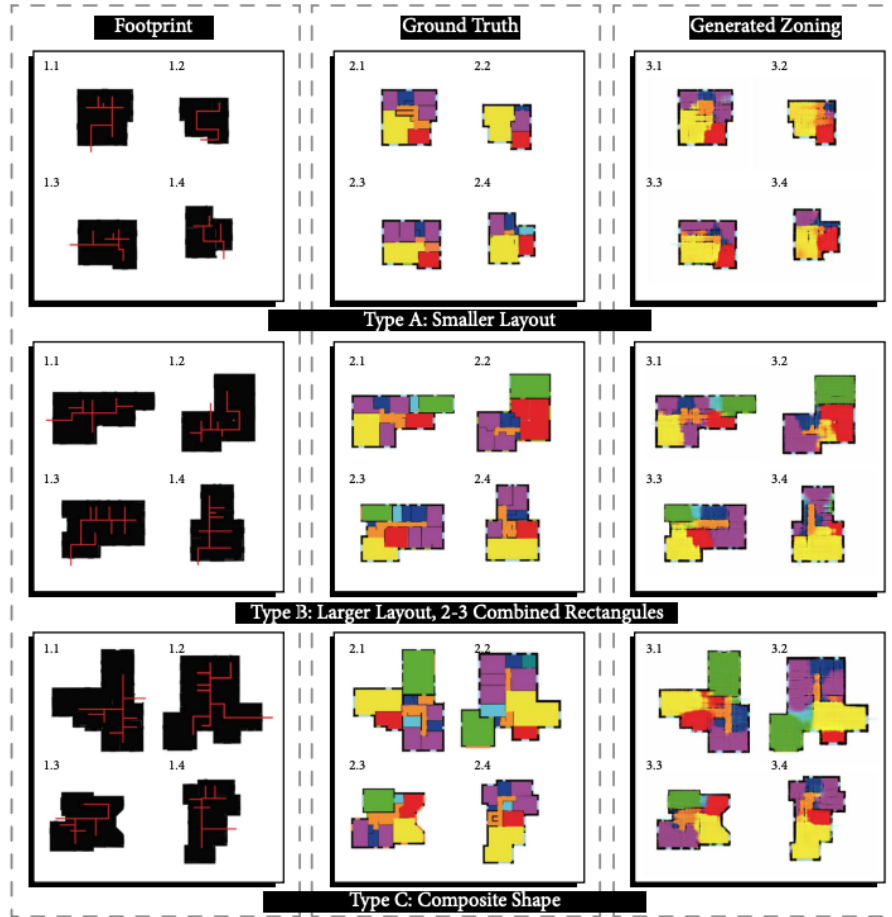


Figure 10. The training dataset for circulation labels.

To improve the clarity of different zoning areas, wall partitions are added (in black) to the ground truth, leading to a more precise separation of zoning areas than in experiment 2. However, the generated zoning outcome still needs more clarity, and the interior partitions must be made apparent. Further experimentation can be done to explore new ways of marking the circulation, such as marking circulations as doors to each room, to improve the quality of the generated zoning output. These

findings provide valuable insights and pave the way for future research to improve the accuracy and precision of neural network-generated room zoning plans using circulation information.

The preliminary experiments tested the effectiveness of GANs for generating architectural plans from different inputs. It revealed the limitations of the quality and diversity of the generated plans, the influence of the data set annotations, and detailed and specific information from the datasets. The experiments suggested that further research is needed due to the limitations of data preparation, annotation, and analysis for architectural datasets. This experiment demonstrated the need for more comprehensive and diverse datasets, consistent and meaningful annotation, and evaluation measures for surveying architectural datasets.

c) Dataset Survey in other disciplines

In many machine learning or deep learning-related architectural research, curating datasets is becoming a common problem for architectural researchers. The experiments above indicate the necessity for surveying floor plan datasets in architecture, which aims to identify existing datasets and their characteristics and propose new methods for data collection, annotation, and analysis. While curating the data for preliminary research on plan-based datasets, the limitations of dataset preparation came into reality. For my experiment, the baseline dataset (CVC-FP²⁶ floor plan) might need to be updated, and the size might need to be revised for conducting a new task. The similar difficulties this experimentation faced in curating datasets are also commonly observed in other works. Such challenges may result from the wide range of specific data types and tasks required by architecture and computational design research.

Previous works in computational design have explored and curated architectural datasets, including the Computational Design Laboratory (CoDeLab) under the Master of Science in Computational Design (MSCD) program at Carnegie Mellon University (CMU). For example, In the cases of Hao Zheng's work in commercial datasets, they discussed in detail the labeling principles for making those datasets into pix2pix models²⁷. In one of the MSCD cohort's research about 3D form analysis of wood

²⁶ As Heras, Lluís-Pere & Terrades, Oriol & Robles, Sergi & Sánchez, Gemma. (2015).

²⁷ As Zheng, H., & Huang, W. (2018).

churches by Michael Hasey in 2022²⁸, the author spent a significant portion of the content talking about constructing the dataset using a 3D construction method named NeRS²⁹ (Neural Reflectance Surfaces). Cecilia Ferrando's work³⁰ collected a curated dataset of floor plans labeled according to their spatial properties and privacy levels to train a neural network model to detect spatial patterns. However, the pipeline requires significant prior work in image processing and parameter settings. In Jinmo Rhee and Pedro Veloso's work on type investigation for high-rise building forms,³¹ They discussed their dataset collection and processing methods to encode building forms into tensors. Preparing new datasets may be hard to avoid for new task-specific research in architectural research. Still, efforts and experiences can be accumulated into knowledge with a selective survey of architecture datasets.

Hence, I looked at how previous works in other fields conducted their dataset review. It should be recognized that different datasets in each area may vary in data size and scope. However, there should be some relevant methodology that can inspire the dataset review in architecture design. In a paper on dataset survey regarding time series datasets for deep learning forecasting³² researchers conducted paper screening from Web of Science and Papers with Code. They extracted 53 public datasets from above (seen in Fig 11). The selection process was based on the number of references (more than ten citations) and keyword search (such as "deep learning" as the constraint). To balance out the newly published papers and datasets with lower chances of acquiring 10 citations, the paper reduced the restrictions to 5 citations for papers from last year before its publication. The paper also conducted a thorough review of related surveys in the field of time series forecasting and listed and examined several metrics in the characteristics of datasets, including references, easy access, comparability, cross-domain, datasets statistics, and analysis. It also studies the domain distribution for the public datasets in the field. The paper also used statistical measures to calculate the Euclidean distance between each pair of datasets and identified clusters based on the statistical values.

²⁸ Hasey, Michael (2022). The 3D Form Analysis of Regional Architecture Using Deep Learning: A Case Study of Wood Churches from the Carpathian Mountains. Carnegie Mellon University. Thesis. <https://doi.org/10.1184/R1/21291672.v1>

²⁹ Zhang, J.Y., Yang, G., Tulsiani, S., & Ramanan, D. (2021). Neural Reflectance Surfaces for Sparse-view 3D Reconstruction in the Wild. In Conference on Neural Information Processing Systems. arXiv: 2110.07604 [cs.CV]

³⁰ Ferrando, C. (2018). Towards a Machine Learning Framework in Spatial Analysis. Master's thesis, Carnegie Mellon University, Pittsburgh, PA.

³¹ Rhee, J. and Veloso, P. Type Investigation in the Form of High-Rise Building Using Deep Neural Networks. Computer-Aided Architectural Design. Design Imperatives: The Future is Now 2022, Volume 1465 ISBN : 978-981-19-1279-5

³² Hahn, Yannik & Funken, Tristan & Meyes, Richard & Meisen, Tobias. (2023). Time Series Dataset Survey for Forecasting with Deep Learning. Forecasting. 5. 315-335. 10.3390/forecast5010017.

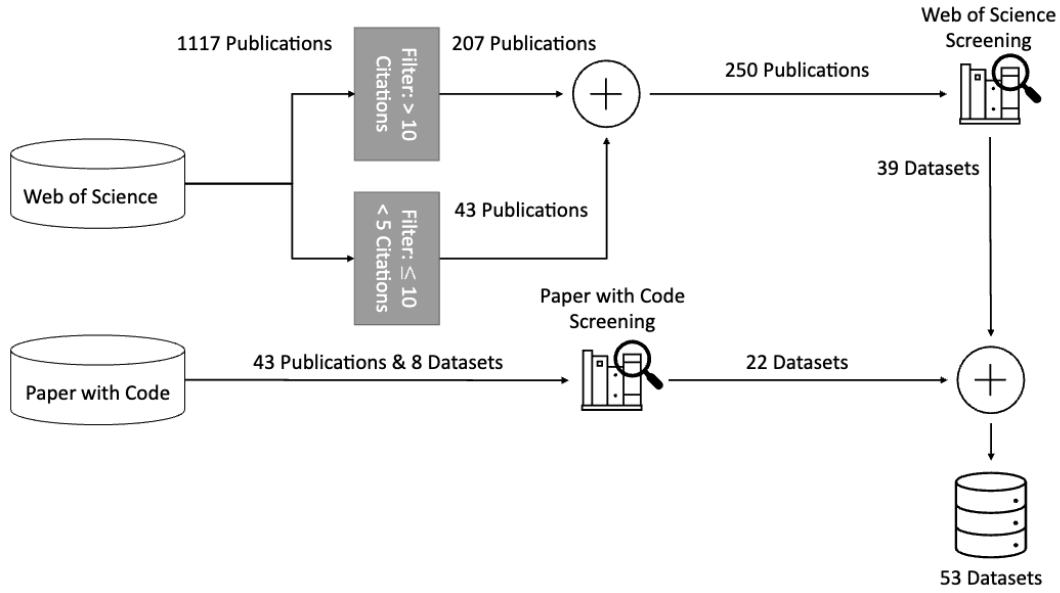


Figure 11. Search process of public datasets for dataset review by Hahn et al.³³

In areas closer to architectural datasets, Steve et al.'s paper³⁴ has compared the existing datasets regarding 3D indoor datasets based on the annotation type and attributes of the datasets. (See Table 1). According to their analysis, the datasets were divided into real plans (from real-world practice) and synthetic plans (artificially generated data to mimic real data's characteristics). Data type, layout formats and applications were surveyed across 17 indoor datasets. It also divides the applications into object detection (as O), scene understanding (as U), structured 3D modeling (as S), layout estimation (as L), depth information (as D), and floor plan information (as F). The research provides a potential direction for comparing the datasets that this paper analyzed and offers new directions for creating new datasets for floor plans in the future.

³³ IAs 27. Hahn et al.

³⁴ Cruz, Steve, et al. "Zillow Indoor Dataset: Annotated Floor Plans With 360o Panoramic and 3D Room Layouts." arXiv.org, Cornell University, 30 Oct. 2020, <https://arxiv.org/abs/2010.14977>.

	Name	Data type	# of images	Layouts	Applications
Real	Gibson [47]	RGB-D pano	572 scenes	–	O U D
	Replica [41]	CAD	18 scenes	C, M, N-M	O U D F
	SceneNN [17]	RGB-D	100	–	O U D
	CRS4/ViC ⁴	RGB pano	191	C, M, N-M	O U L F
	SUN3D [48]	RGB-D	415	–	O U D
	PanoContext [52]	RGB pano	514	C	O U L
	2D-3D-S [2]	RGB-D pano	552	–	O U D
	LayoutNet [55]	RGB-D pano	1,071	C	O U D L
	NYU-Depth V2 [29]	RGB-D	1,449	–	O U D
	ScanNet [11]	RGB-D video	1513	–	O U D
	MatterportLayout [56]	RGB-D pano	2295	C, M	O U D L
	Matterport3D [7]	RGB-D pano	10,800	–	O U D
	ZInD (Ours)	RGB pano	71,474	C, M, N-M	S L F
Synthetic	SunCG [40]	CAD	45, 622	C, M, N-M	O U D F
	Structured3D [54]	CAD	196k	C, M, N-M	O U D S F
	SceneNet [22]	CAD	20M	C, M, N-M	O U D F
	InteriorNet [22]	CAD	22M	C, M, N-M	O U D F

Table 1. Overview of 3D indoor datasets³⁵.

Surveying datasets in other disciplines has shown the established methodologies for conducting dataset reviews. The methods can inspire architecture researchers to conduct an in-depth survey of plan-based datasets. This survey will help researchers better understand the existing datasets and propose new potentials for future research on plan-based datasets.

³⁵ As Cruz, Steve, et al.

3) Hypothesis

The Previous experiments documented above highlighted the potential benefits of surveying architectural planar datasets. The main research question of this thesis is: How can existing plan-based datasets and their implementations support architectural research? To answer this question, the thesis proposes the following hypothesis: By conducting a selective review of existing plan-based datasets and their implementations, we can better understand the characteristics, potentials, and challenges of these datasets for architectural research. The study will help us identify the gaps and limitations of current datasets, suggest best practices and standards for data collection, annotation, and analysis, and propose new directions and opportunities for future plan-based datasets.

a) Architecture Plan Datasets Overview

It is essential to understand the currently available datasets for architecture plan information. However, when researchers are obtaining architectural data, they mostly try to compare the performance of a particular model over a specific dataset (such as HouseGAN++³⁶ by Nauta et al. using RPlan³⁷), or they prepare a specific dataset to testify the application of certain machine learning tools (such as Nathan Peter's studies³⁸ using housing plans from Runway Data³⁹). There are not enough reviews of available datasets in architecture or, even more specifically, architecture plan datasets. An overall view of architecture datasets is becoming essential to review existing datasets in architecture datasets comprehensively.

i) Availability of Plan Datasets among platforms

As an essential part of the search process, this research surveyed several popular platforms⁴⁰ for the

³⁶ Nauata, Nelson et al. "House-GAN++: Generative Adversarial Layout Refinement Network towards Intelligent Computational Agent for Professional Architects." *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021): 13627-13636.

³⁷ Wenming Wu, Xiao-Ming Fu, Rui Tang, Yuhan Wang, YuHao Qi, and Ligang Liu. Data-driven interior plan generation for residential buildings. *ACM Transactions on Graphics (TOG)*, 38(6):1–12, 2019.

³⁸ Peters, Nathan. (2017). Master thesis: "Enabling alternative architectures: Collaborative frameworks for participatory design." Cambridge, MA: Harvard Graduate School of Design

³⁹ <https://runwayml.com/>

⁴⁰ <https://towardsdatascience.com/the-top-10-best-places-to-find-datasets-8d3b4e31c442>

dataset to find plan datasets, including PapersWithCode⁴¹ (a popular web-based platform used by hosting code, datasets, and evaluation metrics for various machine learning tasks), GoogleDatasetSearch⁴² (a search engine developed by Google to help researchers and data analysts find publicly available datasets), DataGov⁴³ (platform by the US government provides access to a wide variety of government datasets), DataHub⁴⁴ (a centralized system for data storage, definition, and delivery) and Kaggle⁴⁵ (an online platform that hosts data science competitions). Some platforms share open-source datasets in strategic efforts (e.g., Datahub). Some are more comprehensive platforms to review research outcomes (e.g., PapersWithCode) and hold challenges and collaborations for data science communities (e.g., Kaggle).

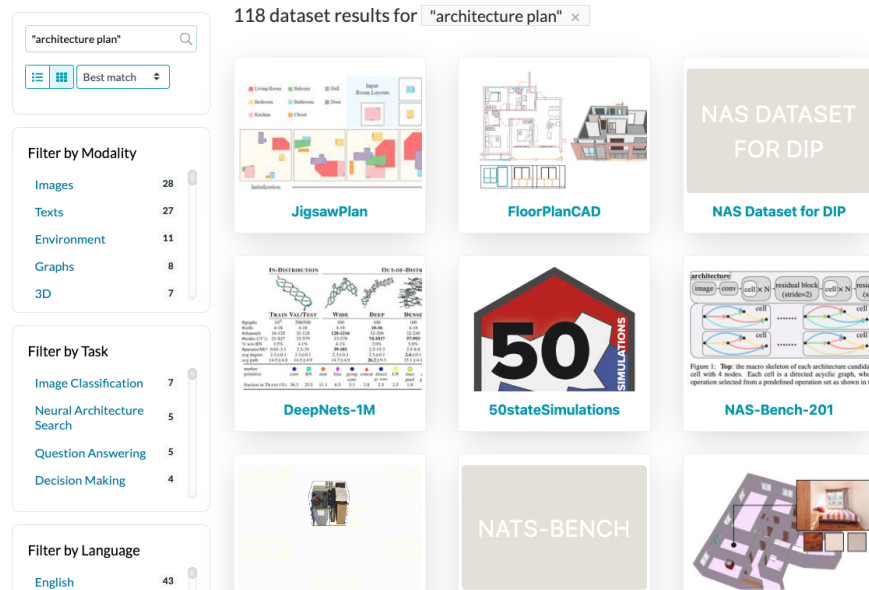


Figure 12. Search result for “Architecture plan” in the PaperWithCode platform.

Among those public dataset platforms, however, not every platform contains data regarding architectural plan data. When searching for the keywords “Building Plan” or “Architecture Plan”, comprehensive platforms such as PaperWithCode may turn out to have more results that are closer to

⁴¹ <https://paperswithcode.com/datasets>

⁴² <https://datasetsearch.research.google.com/>

⁴³ <https://data.gov/>

⁴⁴ <https://datahub.io/docs/about>

⁴⁵ <https://www.kaggle.com/datasets>

architectural plans that contain image information (such as JigSawPlan⁴⁶, See Fig. 12). But for some other platforms such as Kaggle and DataHub, there is much less relevant data available. That shows that the search of architectural datasets can be vital for curating a review of plan datasets, especially regarding the specificity of plan datasets for architectural research.

ii) Data Type Review of Building Plan Datasets

To train machine learning models, researchers may require specific architecture datasets, including image-based, CAD-based, and graph-based models. However, an overview analysis of available datasets is necessary. The most commonly known architectural datasets are images based (e.g., JigsawPlan⁴⁷), CAD-based (such as FloorPlanCAD), and graph-based models. However, a more thorough analysis of the collective datasets needs to be surveyed.

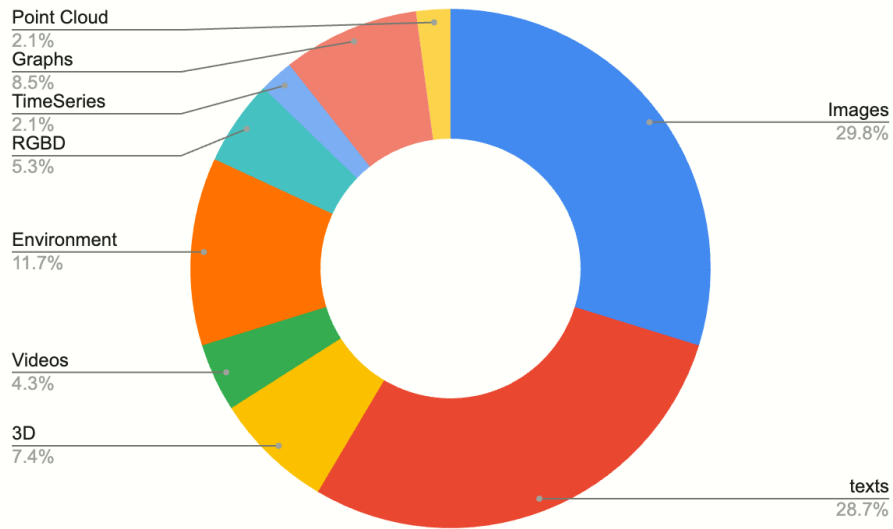


Figure 13. Result of searching “Archi Plan” on PaperWithCode Website Filter by Modality.

⁴⁶ Hosseini, Sepidehsadat et al. “JigsawPlan: Room Layout Jigsaw Puzzle Extreme Structure from Motion using Diffusion Models.” *ArXiv* abs/2211.13785 (2022): n. pag.

⁴⁷ As 38.

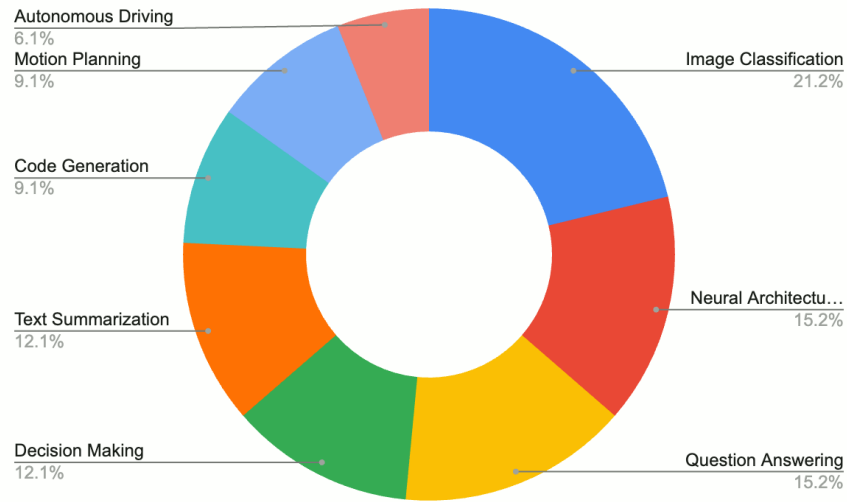


Figure 14. Result of searching “Archi Plan” on PaperWithCode Website Filter by Task.

When searching for “architecture plan,” there are 119 dataset results from the PaperWithCode platform. Among the search results, most of the data are image or text-based data, occupying nearly 60% of the datasets (See Fig. 13). Among the image-based plan datasets, JigsawPlan⁴⁸ has the most significant number of room layouts and floor plans for more than 98,000 houses. The dataset was a result of using diffusion models for the conditional generation process, taking room layouts and room types as input and predicting the location and orientation of each room. The dataset contains rooms with 16 different room types in different colors, with notations of doors, as compared with another dataset RPlan (See Fig. 14). In the classification, datasets such as FloorPlanCAD⁴⁹, which is a real-world CAD-based data of more than 10,000-floor plans, was also considered as an image-type dataset. Some image-based datasets are also connected to 3D scene reconstruction, such as Rent3DD++, which converts floor plans to 3D scenes⁵⁰. Similar contributions are also addressed in the Plan2Scene model, turning the floor plan and associated photos of the residence into 3D mesh

⁴⁸ as 15. Hosseini, Sepidehsadat et al.

⁴⁹ Fan, Zhiwen, et al. “FloorPlanCAD: A Large-Scale CAD Drawing Dataset for Panoptic Symbol Spotting.” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2021. Crossref, <https://doi.org/10.1109/iccv48922.2021.00997>.

⁵⁰ Vidanapathirana, Madhawa, et al. “Plan2Scene: Converting Floorplans to 3D Scenes.” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021. Crossref, <https://doi.org/10.1109/cvpr46437.2021.01059>.

models with textures⁵¹ (See Fig. 15)

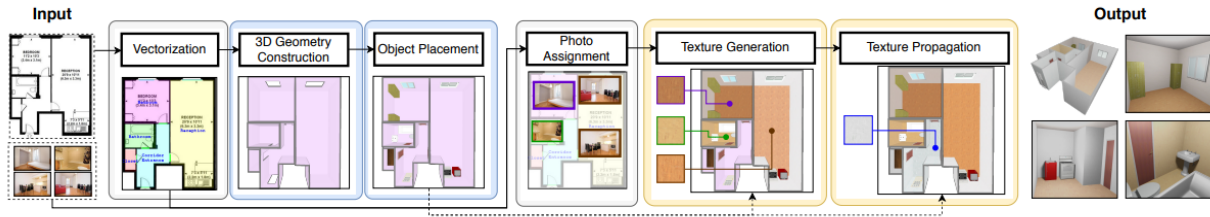


Figure 15. Pipeline for Plan2Scene model.⁵²

Most of the text-based datasets on these dataset platforms are not very close to architectural discipline, and some are totally irrelevant. But there are some exceptions. Heri-Graphs proposed workflows to construct multi-modal datasets with graph-based ML tasks regarding heritage information among the datasets more relevant to building plans.⁵³ Among other text-based results from the search, the Coverage Path Planning (CPP) datasets containing text-based data are introduced for remote sensing and planning⁵⁴. In research by Prof. Cardoso-Llach about using data visualization for contemporary design practices⁵⁵, the text-based dataset of architectural coordination data was used to trace the design ecologies of large architectural projects and to reveal hierarchy, growth, clusters, spatiality, and relations of the design process.

From the different disciplines of the search results (See Fig. 16), most of the tasks are from image classification, neural architecture search, and question-answering, which are not closely related to architectural plan datasets discussed in the scope of this study. The ones that are closer to the

⁵¹ Vidanapathirana, Madhawa, et al. “Plan2Scene: Converting Floor Plans to 3D Scenes.” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021. Crossref, <https://doi.org/10.1109/cvpr46437.2021.01059>.

⁵² As 22

⁵³ Bai, Nan, et al. “Heri-Graphs: A Dataset Creation Framework for Multi-Modal Machine Learning on Graphs of Heritage Values and Attributes with Social Media.” ISPRS International Journal of Geo-Information, vol. 11, no. 9, Aug. 2022, p. 469. Crossref, <https://doi.org/10.3390/ijgi11090469>.

⁵⁴ Apostolidis, Savvas D., et al. “Cooperative Multi-UAV Coverage Mission Planning Platform for Remote Sensing Applications.” Autonomous Robots, vol. 46, no. 2, Jan. 2022, pp. 373–400. Crossref, <https://doi.org/10.1007/s10514-021-10028-3>.

⁵⁵ Cardoso Llach, Daniel. “Tracing Design Ecologies: Collecting and Visualizing Ephemeral Data as a Method in Design and Technology Studies.” DigitalSTS: A Field Guide for Science & Technology Studies, edited by Janet Vertesi and David Ribes, Princeton University Press, 2019, pp. 451–468. <https://press.princeton.edu/books/hardcover/9780691187075/digitalsts>.

“building plan” category belong to Motion Planning (as HouseExpo⁵⁶), while other more commonly used ones remain unlisted on the “task” list.

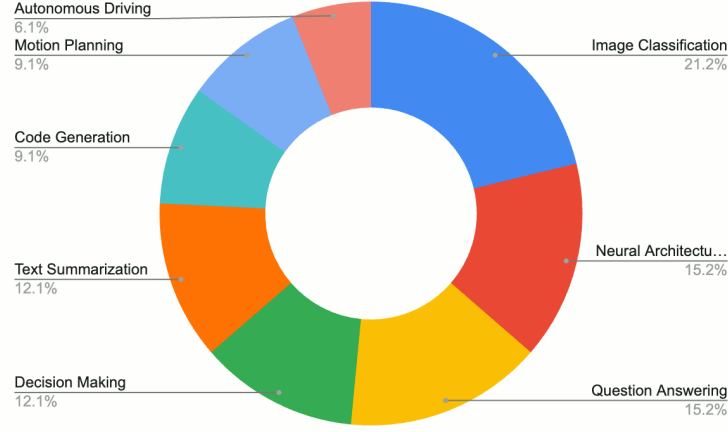


Figure 16. Result of searching “Archi Plan” on PaperWithCode Website Filter by Task.

iii) Overview of Selected Plan Datasets

As the search result cannot gather the exact result for research among building plans, preparing a list of existing datasets is becoming essential for a detailed study of the dataset collection. From the provided results on Paper with Code and Web of Science, we focused on the research published after 2010 and looked at papers with more than ten references. Similar to the methods for the time series dataset survey⁵⁷ mentioned in section 2.4, the survey relaxes the constraint for datasets published within three years to 5 references. Based on the selected results on the two platforms, the dataset is then manually determined by whether they are close to the meaning of “floor plan” in reality (since some appeared results may not be related to the floor plan at all, such as Dex-Net⁵⁸ from Paper with Code). And the selection process focuses on the datasets documented in academic publications. Therefore, this thesis examines the selection of 12 datasets closely related to architectural and engineering research taking into consideration the year, the author, and the size of the datasets (See

⁵⁶ Li, Tingguang et al. “HouseExpo: A Large-scale 2D Indoor Layout Dataset for Learning-based Algorithms on Mobile Robots.” *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2019): 5839-5846.

⁵⁷ I'm 27. Hahn et al.

⁵⁸ <https://paperswithcode.com/dataset/dex-net-2-0>

Table 2).

NAME	YEAR	Link	Author	Conf	Amount
FloorPlans Symbols	2011	paper	CS Researcher	GREC	1000
CVC-FP	2013	Paper	CS Researcher	IJDA	245
ROBIN	2017	paper	CS Researcher	IAPR	510
Zheng and Huang	2018	website	Real Estate	ACADIA	100
CubiCasa5k	2019	Paper	CS Researcher	CVPR	500
RPlan	2019	paper , code	Archi Researcher	SIGGRAPH	60,000
Pix2pix-floorplans	2020		Archi Researcher		103
LIFULL HOME	2020	website	Real Estate		5,310,000
HouseExpo	2020	paper , code	CS Researcher	IROS	35,126
FloorPlanCAD	2021	paper , code , dataset	Archi Researcher	CVPR	15,000
Zillow Plan	2021	paper , code , dataset	Real Estate	CVPR	2737
JigsawPlan	2022	paper , code , dataset	CS Researcher	CVPR	98,780

Table 2. Organized chart for selected plan datasets.

***b)* Research Questions**

The research hopes to answer the questions: How can existing plan-based datasets and their implementations support architectural research? It can be further elaborated into three sub-questions:

- How do existing plan-based datasets vary in their data sources, types, structures, annotations, applications, and evaluations, and what are the implications of these variations for architectural research?
- How can architectural researchers use existing plan-based datasets to generate, analyze, evaluate, and visualize architectural plans with computational design and machine learning tools?
- What are the best practices and standards for collecting and annotating plan-based data, and what are the new directions and opportunities for creating more diverse plan-based datasets for architectural research collaboratively?

***c)* Research Goal**

This review will contribute to the current study of architectural research by providing a better understanding of the potential of these datasets to support research in the field. Additionally, this work will provide insights into how these datasets can be leveraged to improve architectural research and inform the development of new directions for data collection, annotation, and analysis.

4) Methods: Dataset Survey

This section will provide a detailed review of the selected datasets. The six datasets are chosen among the 12 in Table 1 based on the availability of open source data and the detailed description that talks about the whole curation process of the data. Datasets used unofficially from commercial websites (such as the Lianjia dataset from Zheng and Huang's paper⁵⁹) were not surveyed due to copyright reasons. Some other datasets come from other untraceable sites (such as the use of the pix2pix plan from the Runway platform by Nathan Peters⁶⁰). The section discussed answers several criteria for the datasets:

What is the purpose of these datasets?

What is the source of the datasets?

What is the preparation process for these datasets?

What are the data type and data structure of the datasets?

What are the significance and limitations of the datasets?

For plans that contain image data, how can we see the quality and topological distributions of the datasets using statistical tools such as T-SNE⁶¹ (a data visualization tool to visualize and evaluate high-dimensional data)

i) CVC-FP Dataset

The first surveyed dataset is CVC-FP dataset⁶². The dataset collects 1200 actual architectural plans from various sources and styles. It contains actual floor plans with structural ground truth (as the actual structure information) for automatic floor plan analysis. Universitat Autònoma de Barcelona (UAB) researchers in Spain curated the floor plan dataset. The dataset was applied and referenced in many

⁵⁹ As Zheng, Hao., & Huang, Weixin. (2018). Architectural drawings recognition and generation through machine learning. Cambridge: MA, ACADIA

⁶⁰ As Peters, Nathan. (2017). Master thesis: "Enabling alternative architectures: Collaborative Frameworks for participatory design."

⁶¹ <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>

⁶² De las Heras, Lluís-Pere, et al. "Statistical Segmentation and Structural Recognition for Floor Plan Interpretation." International Journal on Document Analysis and Recognition (IJDAR), vol. 18, no. 3, 2015, pp. 221-237.

later papers, including Sharma's deep architecture for analyzing and retrieving building floor plan⁶³ and Liu's work transforming floor plans into vector-graphic representations from raster⁶⁴. The paper provided a dataset of real floor plans (CVC-FP⁶⁵) with annotations for architectural objects (such as wall, window, room, door) and structural relationships⁶⁶. The plan was generated using a ground-truthing method SGT, a web application for users to define their object classes and properties and label those objects and relations using bounding boxes or polygons. The web app also provides users with version control and user control for a more accessible definition and generation of ground truth, allowing collaborators to work with the database definition and ground truth construction (See Fig 17 and 18).

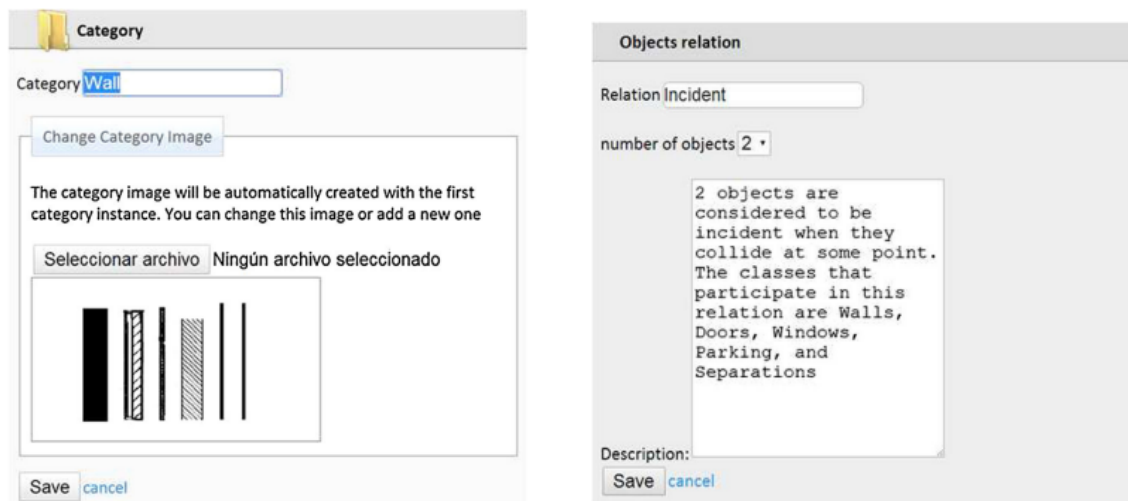


Figure 17. Windows for creating new categories (left) and new relations (right) for SGT tool⁶⁷.

⁶³ D. Sharma, N. Gupta, C. Chattopadhyay and S. Mehta, "DANIEL: A Deep Architecture for Automatic Analysis and Retrieval of Building Floor Plans," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 420-425, doi: 10.1109/ICDAR.2017.76.

⁶⁴ C. Liu, J. Wu, P. Kohli and Y. Furukawa, "Raster-to-Vector: Revisiting Floorplan Transformation," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2214-2222, doi: 10.1109/ICCV.2017.241.

⁶⁵ As 49

⁶⁶ As 37. De las Heras, Lluís-Pere, et al. "Statistical Segmentation and Structural Recognition for Floor Plan Interpretation."

⁶⁷ As 37.

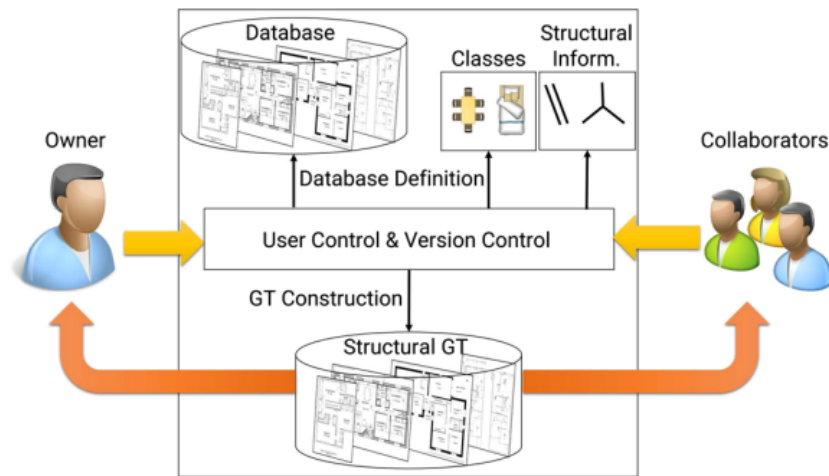


Figure 18. Diagram of SGT Tool Architecture with version and user control⁶⁸.

To provide the ground truth data for the floor plan, the researchers invited a group of architects to express their needs for auto-interpretation applications. The ground truth information includes the location of planar elements and structural relations for floor plan analysis. The research provides a detailed review of the object and structural labeling conventions. Building elements (walls, doors, windows, parking doors, and separations) were marked by polygons of the areas (See Fig 19). The ground truth analysis also studies the functions of those architectural elements and how these factors would affect the labeling outcome. The SGT tools used attributed graphs to describe dependencies between those labeled elements for structural relationships. Nodes in the graph represent annotated elements, and attributed edges represent the contextual relations between objects. The paper considers different types of relations, including incident, surround, neighborhood, access, and surrounding perimeter. (See Fig 20)

⁶⁸ As 37.



Figure 19. Labeling of rooms (in blue) with wall, door, and window (in yellow and green)⁶⁹

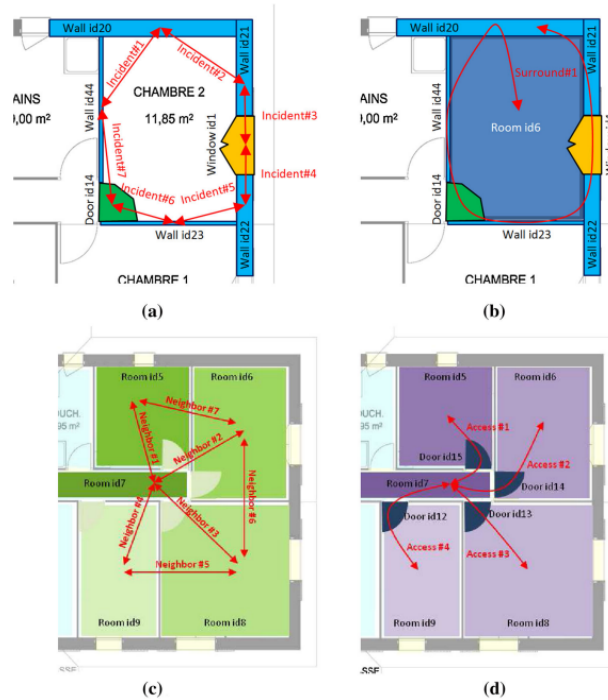


Figure 20. Examples of different structural relationships among annotated objects⁷⁰

The preparation process of the dataset started from collecting documents of real-world floor plans. Beginning with the SCANPLAN⁷¹ project, many datasets are curated for floor plan analysis research. The CVC-FP images come from four different datasets: Black, Textured, Texture2, and Parallel dataset

⁶⁹ As 37.

⁷⁰ As 37.

⁷¹ SCANPLAN PROJECT: <http://www.eurekanetwork.org/project/-/id/4462>. As cited by 37

(See Fig 21). A comparison regarding the structural information, the symbolic meanings, and textural information is listed in the table below (See Table 3). The diversity in the different attributes of the database will enhance the robustness of the technique in applying it to real-world scenarios. The total 122 images from the above four datasets then pass through the SGT tool to identify structural information and relations using the web application with nine people working on graphics recognition. And the revised dataset is then checked by one person for correctness and consistency.

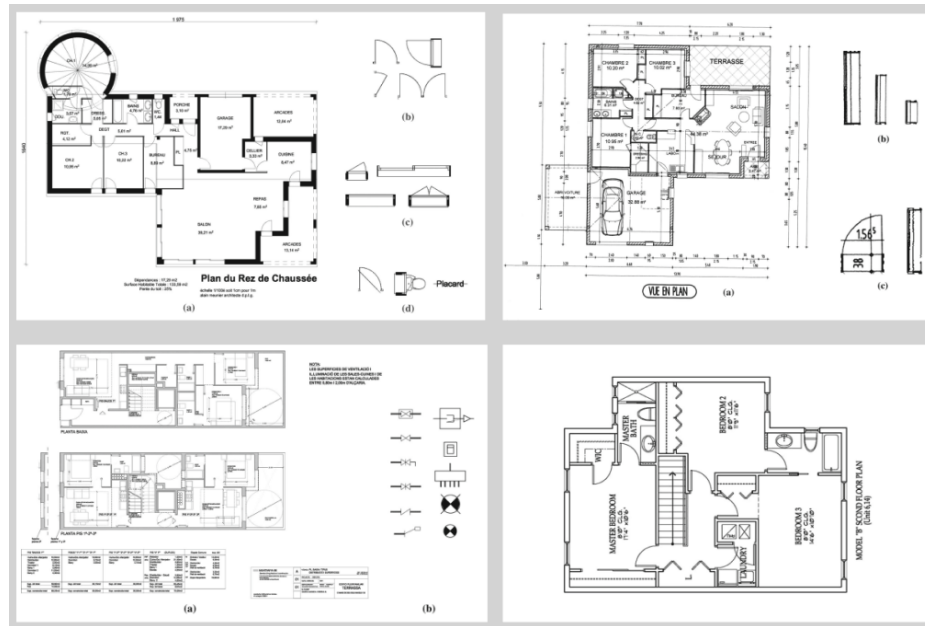


Figure 21. Examples of Black (Left upper), Textured (right upper), Textured2 (left lower), and Parallel (right lower) datasets.⁷²

⁷² As 37.

Dataset	Structural Information	Symbolism	Textual Information
Black Dataset	90 binary images of good quality and high resolution ($2,480 \times 3,508$ or $3,508 \times 2,480$ pixels)	Walls are modeled by black lines of three different thicknesses, doors by quarter-circle arcs, windows by various shapes, and rooms by polygons	Contains room names, areas, dimensions, and titles
Textured Dataset	10 grayscale images of poor quality and varying resolutions (from $1,098 \times 905$ to $2,218 \times 2,227$ pixels)	Walls are modeled by two parallel lines with a diagonal or dotted pattern in between, doors by quarter-circle arcs, windows by rectangles, and rooms by polygons	Contains room names, areas, dimensions, and legends
Textured2 Dataset	18 grayscale images of high resolution ($7,383 \times 5,671$ pixels) that belong to a single building of six floors	Walls are modeled by two parallel lines with a diagonal pattern in between, doors by quarter-circle arcs, windows by rectangles, and rooms by polygons	Contains room names, areas, dimensions, legends, and specific architectural information for each floor
Parallel Dataset	4 binary images of good quality and high resolution ($2,550 \times 3,300$ pixels) that belong to two detached houses for sale	Walls are modeled by simple parallel lines, doors by quarter-circle arcs, windows by rectangles, and rooms by polygons	Contains room names and areas

Table 3. Comparison table between 4 datasets from CVC-FP datasets⁷³.

The CVC-FP dataset stores the ground truth in the SVG (Scalable Vector Graphics) file in metadata information. The format includes general information about image dimensions, labeled instances, and class names within the dataset. It also stores the list of elements from the floor plan image, with object

⁷³ As 37.

labels and its descriptive text-line, the polygon of extremity points, and relations between objects.

The dataset has significance in providing a benchmark to compare methods for floor plan analysis, such as wall segmentation, room detection, symbol recognition, and structural and semantic understanding. Furthermore, it sets an important example for evaluating and comparing different plan analysis methods in this research field. The ground-truthing tool (SGT) also allows to define and label object classes and properties and to generate SVG files for research. However, there are still limitations. The number of images from four sources may not cover the vast diversity of floor plans. The manual annotation process can be subjective and can include errors and inconsistencies.

Meanwhile, due to the limited number of datasets, it does not show any distinct features of the normal distribution when the whole dataset was trained on a VAE⁷⁴ (variational auto-encoder) network (See Fig. 22). And it is relatively hard to see clear patterns from plotted T-SNE results, as the images do not contain apparent feature distributions (See Appendix 1 and 2).

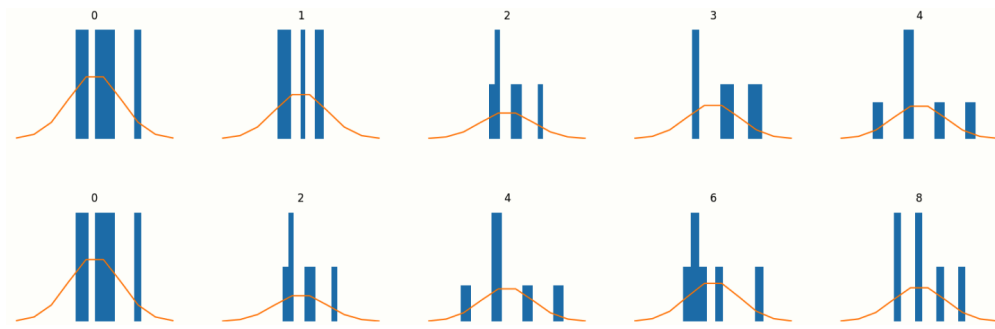


Figure 22. Latent Space Distribution of CVC-FP dataset

ii) **ROBIN Dataset**

The coming dataset is the ROBIN dataset published in 2017⁷⁵. The dataset contains 510 real-world architectural plans for plan retrieval tasks (to find similar floor plans from a large database from a query floor plan image, See Fig 23). The dataset is categorized into three broad layouts that vary from the number and types of rooms within the floor plans. Moreover, each class among the three is categorized

⁷⁴ Kingma, Diederik P., and Max Welling. "Auto-Encoding Variational Bayes." Machine Learning Group, Universiteit van Amsterdam, 2013. <https://arxiv.org/abs/1312.6114>

⁷⁵ Sharma, Divya, et al. "DANIEL: A Deep Architecture for Automatic Analysis and Retrieval of Building Floor Plans." 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Vol. 1. IEEE, 2017.

into ten sub-categories based on the shape of the planar layout. Furthermore, the annotation methods of the ROBIN dataset follow the standards for symbol recognition and spotting systems indicated by previous research by Valveny et al.⁷⁶ The standards for architectural floor plans were based on object type identification and relations. The objects include walls, doors, windows, rooms, and furniture with polygons to enclose the bounding boxes (See Fig 24). The relations have incident, surround, neighbor, and access. The design of the ROBIN plan aims to suit the preferences and needs of property buyers to search for large repositories of similar plans. The data was labeled to adapt to different plan analysis research, including symbol spotting, room segmentation, and layout matching.

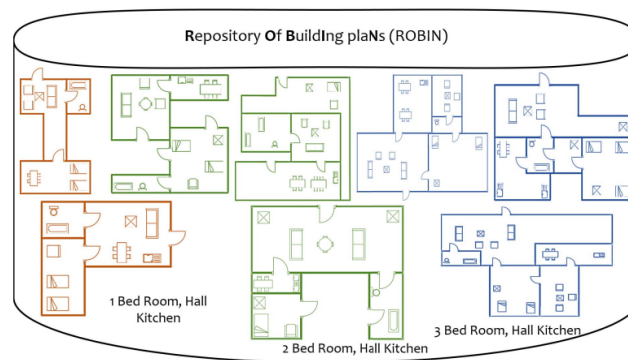


Figure 23. Visual representation of ROBIN Dataset, with different colors to represent different room categories.⁷⁷

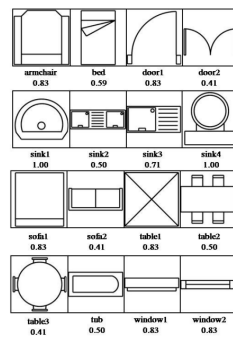


Figure 24. Thumbviews of architectural elements labels with scales.⁷⁸

⁷⁶ Delalandre, Mathieu, et al. "CVC-FP and SGT: a new database for structural floor plan analysis and its ground truthing tool." *International Journal on Document Analysis and Recognition (IJDAR)* 17.1 (2014): 19-36.

⁷⁷ As 46, Sharma, Divya, et al. "DANIEL: A Deep Architecture for Automatic Analysis and Retrieval of Building Floor Plans".

⁷⁸ As 48, Delalandre, Mathieu, et al. "CVC-FP and SGT: a new database for structural floor plan analysis and its ground truthing tool".

The dataset was created due to a need for a public benchmark database for floor plan analysis and retrieval research. The previous datasets, such as SESYD⁷⁹ and CVC-FP⁸⁰, are relatively small, less similarity, less inter-class similarity (overlapping classes), or too specific for particular tasks of plan analysis. Therefore this CVC-FP paper presents the large and diverse dataset ROBIN and a deep architecture for finding similar layouts (DANIEL). DANIEL is a deep learning method for automatic analysis and plan retrieval. It consists of a feature extraction module and a retrieval module. (See fig 25) Higher features from plan images, including global layout and local details) are extracted by a convolutional neural network (CNN) from the feature extraction module. CNN learns a hierarchical representation of the visual content for plans. Then the retrieval module compared the extracted features of a query floor plan with those of the floor plans within the database. The module returns a ranked list of similar floor plans based on similarity scores. (See Fig 26) The modules learn the similarity features by calculating the triplet loss function, encouraging similar plans to have similar feature representations.

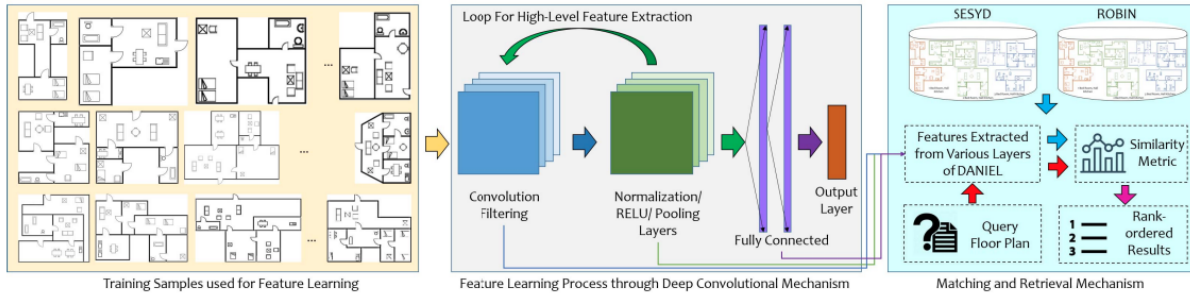


Figure 25. Deep Architecture of Daniel with feature extraction, feature learning and matching and retrieval mechanism.⁸¹

⁷⁹ M. Delalandre, E. Valven, T. Pridmore, and D. Karatzas, “Generation of synthetic documents for performance evaluation of symbol recognition & spotting systems,” IJDAR, vol. 13, no. 3, pp. 187–207, 2010.

⁸⁰ As 37.

⁸¹ As 46, Sharma, Divya, et al. “DANIEL: A Deep Architecture for Automatic Analysis and Retrieval of Building Floor Plans” .



Figure 26. Ranked order retrieval result of DANIEL for three query floor plans from ROBIN dataset (Top 5 rank ordered plans for each query).⁸²

The ROBIN plan dataset provides a benchmark for future research and allows researchers to compare their methods with existing approaches and develop new ways to improve those. The dataset also enhances the diversity and complexity of real-world floor plans. It also demonstrates the effectiveness of the DANIEL framework. Meanwhile, limitations of ROBIN datasets still exist. There may need to be more than a limited number of plans to represent the diversity of real-world plans. The three broad classes and ten sub-classes based on the room number, type, and layout shape may not capture all relevant factors reflected in property buyers’ decisions.

The latent space distribution of the ROBIN plans (See Fig. 27) does not follow the normal distribution line very closely, but it is much closer than the distribution of CVC-FP plans. It shows that the ROBIN data is more representative than the CVC-FP plan. And it may be due to the more significant number of datasets and broad categories of layouts that the plans have covered. However, it seems complicated to find regularities from the TSNE distribution (See Appendix 3) and plotted visualization (See Appendix 4), possibly due to the rich categories and subcategories of the dataset layout shapes.

⁸² As 46, Sharma, Divya, et al. “DANIEL: A Deep Architecture for Automatic Analysis and Retrieval of Building Floor Plans” .

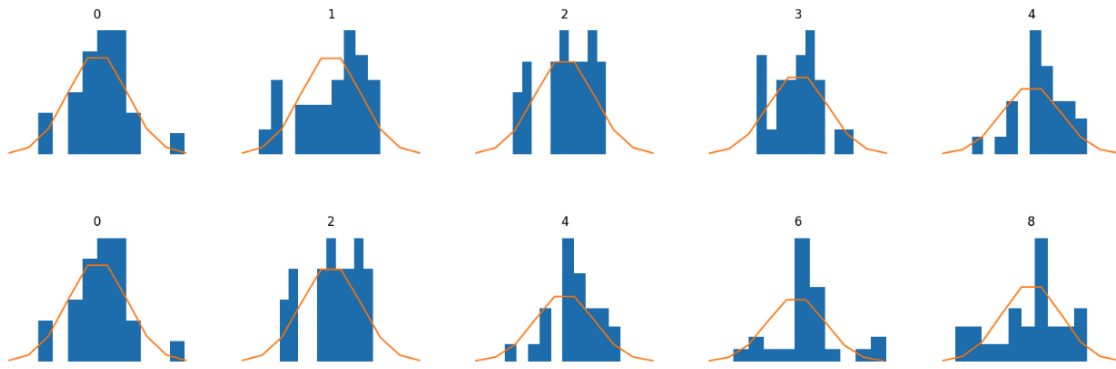


Figure 27. Latent Space Distribution of ROBIN dataset.

iii) CubiCasa5k Dataset

The CubiCasa5k dataset⁸³ contains 5000 samples annotated into over 80 plan object classifications, including rooms, doors, walls, windows, and fixed furniture. It aims to create a rich toolset for automatic floor plan image analysis. It comes from a versatile labeling process using polygons for object separation. The dataset came from an online floor plan vectorization pipeline⁸⁴ that operates on Finland's real estate marketing material conversations. The pipeline was developed by CubiCase⁸⁵, a firm that provides 2D and 3D floor plan digitization for real estate professionals. The pipeline aims to encourage the development of accurate and automated models for real estate from the research community. An extensive set of 15,000 floor plans were collected and reviewed to select 5000 with manual annotations. The 5000 plans are divided into high-quality architectural (3732 plans, with high precision and detail), high-quality (992 plans, less detailed but still good quality), and colorful (276 plans, with colors but fewer details than the former ones). The semantic and geometric annotations in SVG format were done by human annotators trained for the task. Moreover, the time for single image annotations varies from 5 to 120 minutes which varies by the complexity of the source and floor levels⁸⁶. An annotation protocol was applied to format the annotation order and utilize all information.

⁸³ Kalervo, Ahti, et al. "CubiCasa5K: A Dataset and an Improved Multi-Task Model for Floorplan Image Analysis." *Lecture Notes in Computer Science*, 2019, pp. 28–40. Crossref, https://doi.org/10.1007/978-3-030-20205-7_3.

⁸⁴ <http://cubitool.cubi.casa.s3-website-us-west-2.amazonaws.com/?config=customize&rl=2&loc=na&id=8000&color=000000>, from <https://www.cubi.casa/>

⁸⁵ <https://www.cubi.casa/> Accessed Jul 12, 2023.

⁸⁶ As Kalervo, Ahti, et al. "CubiCasa5K: A Dataset and an Improved Multi-Task Model for Floorplan Image Analysis."

A Two-stage Quality assurance process was applied for better annotation accuracy and consistency.

Together with the dataset, the paper presents a new method based on an improved multi-task convolutional neural network. (See Fig 28) To restore lost information from rasterized floor plan images, CNN can perform a web search, segmentation, localization, and post-processing tasks. The network architecture follows a previous work⁸⁷ based on ResNet-152⁸⁸ (pre-trained with ImageNet⁸⁹). It uses an encoder-decoder architecture combining residual blocks and skip connections. (See Fig 29) The segmentation tasks (including rooms, icons, and openings) were combined with heatmap regression (including wall junctions, icon corners, and opening endpoints). Post-processing steps include inference of wall, room, icon polygons and doors/ windows to get the finalized SV representation.



Figure 28. Example result using CubiCasa5k dataset and pipeline provided. (left to right, original floor plan, SVG label, automatic prediction)⁹⁰

⁸⁷ Liu, C., Wu, J., Kohli, P., Furukawa, Y.: Raster-to-vector: Revisiting floorplan transformation. In: The IEEE International Conference on Computer Vision (ICCV) (2017)

⁸⁸ He, K., Zhang, X., Ren, S., Sun, J.: Deep Learning for Image Recognition. Proc. CVPR pp. 770–778 (2016)

⁸⁹ Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. CVPR. pp. 248–255 (2009)

⁹⁰ As 55. Kalervo, Ahti, et al.

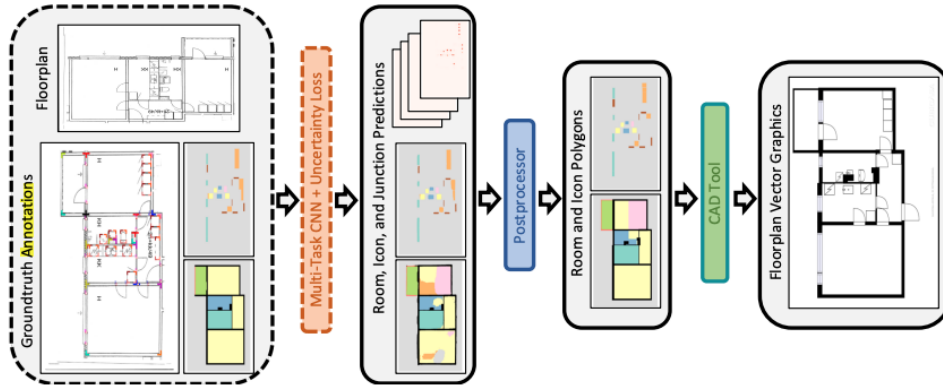


Figure 29. proposed pipeline for automatic floorplan parsing⁹¹

The CubiCase5k dataset is a novel benchmarking dataset for floor plan analysis. The annotations of the dataset are more accurate than previous datasets due to objects' precise shape and direction and the annotation protocol and quality assurance process. It also contains improved variability for apartment types and drawing styles. But more diversity and representation of this dataset is needed considering most of the floor plans are from real estate markets of Finland region.

The latent space distribution of the colorful set from CubiCasa5k does not represent a normal distribution (See Fig 30). It might be due to the limited image data in this specific “Colorful” dataset (around 350 plans). The TSNE dots distribution showed no apparent clustering but only a linear pattern when the perplexity value is close to 10 (See Appendix 5). The plotted scattered image shows that most datasets share the square footprint, while those with lower x-value and higher y-value (on the left side of the line) are more toward vertical orientation. (See Appendix 6).

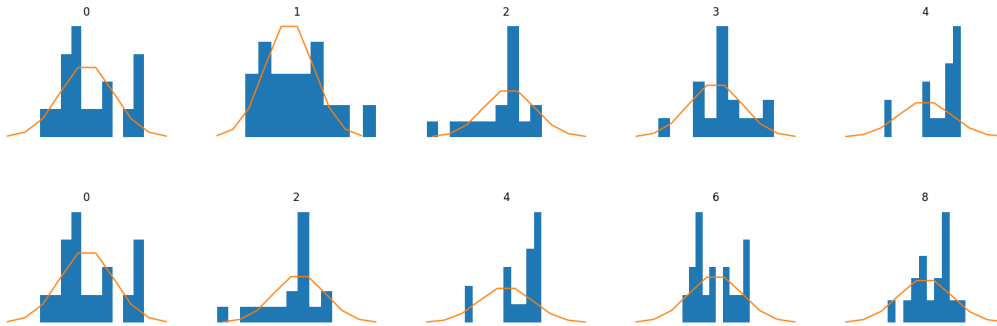


Figure 30. Latent Space Distribution of CubiCasa5k “Colorful” dataset.

⁹¹ As 55. Kalervo, Ahti, et al.

iv) RPlan Dataset

The RPlan⁹² provided real-world residential plans in the Asian estate market. The large-scale dataset contains more than 80K floor plans with labeled rooms and walls. (See Fig. 31) Within the dimension of 18 meters by 18 meters, each plan contains a vector-graphic representation with semantic and geometric annotations at the pixel level.

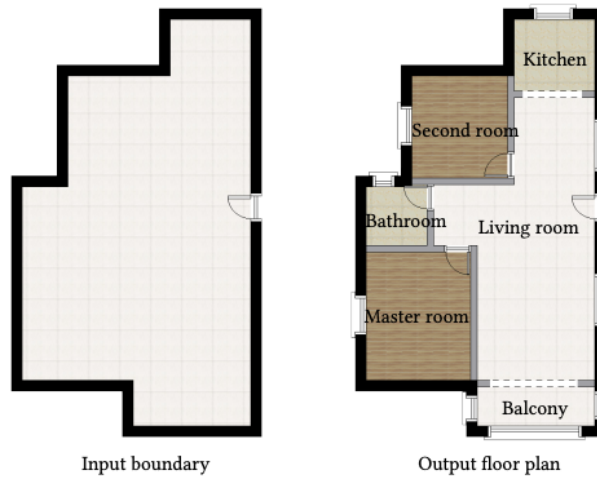


Figure 31. Floor Plan of a residential building generated by the paper's approach⁹³.

The data preparation process involves data collection, filtering, annotation, and representation as the four essential steps. (See Table 4) The researchers used a web crawler to get more than 120K floor plans from the estate market in SVG format. Then they filter the data by applying specific empirical rules and observations (including total area, total number and proportion of rooms, presence of living room, and others) to remove non-standard data. This data filtering process helps maintain the dataset's quality and reliability. The annotation process was executed by labeling each element manually among the 13 room types and three auxiliary labels, assisted by a semi-automatic tool that extracts the semantic information from SVG format. (See Fig. 32) Then, the data was represented as four-channel images that included the inside mask, boundary mask, the semantics of rooms and walls, and information to separate rooms with the same labels. (See Fig 33 and Table 5) That also standardizes the data format and data type of the final RPLAN dataset.

⁹² Wu, Wenming, et al. "Data-driven Interior Plan Generation for Residential Buildings." ACM Transactions on Graphics, vol. 38, no. 6, Nov. 2019, <https://dl.acm.org/doi/10.1145/3355089.3356556>.

⁹³ As 78. Wu, Wenming, et al.

Filtering Process	Input	Output	Manual labor	Selection process or criteria
Data collection	Real-world residential buildings in Asia	More than 120K floor plans with vector-graphics representation	Manual collection of floor plans from online sources	No specific criteria, from Real Estate websites
Filtering	More than 120K floor plans with vector-graphics representation	More than 80K floor plans that satisfy five requirements (total area, presence and proportion of living room, etc.)	No manual labor involved	Five requirements based on the observation of real floor plans
Annotation	More than 80K floor plans with vector-graphics representation	More than 80K floor plans with semantic annotations at the pixel level for rooms and walls	Manual labeling of room types and wall positions for each floor plan	No specific criteria
Representation	More than 80K floor plans with semantic annotations at the pixel level	More than 80K floor plans with simplified representation for rooms and walls using small squares and masks	No manual labor involved	Simplification based on the convenience of training and sampling

Table 4. Comparison table between four data filtering processes for RPLAN⁹⁴

⁹⁴ As 78. Wu, Wenming, et al.

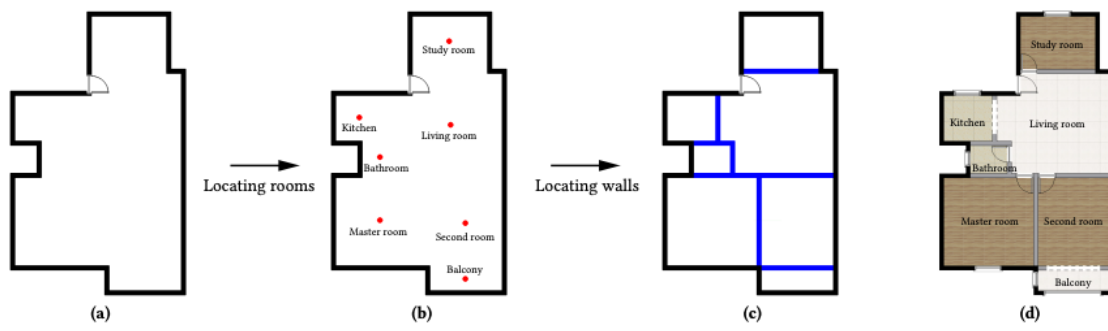


Figure 32. Overview of the RPLAN preparation workflow.⁹⁵

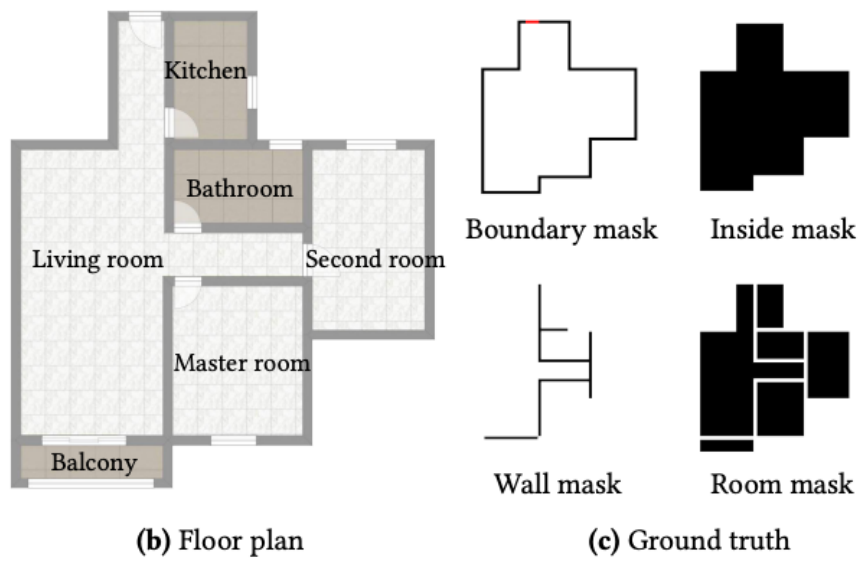


Figure 33. Floor Plan and Ground truth of RPLAN dataset⁹⁶.

⁹⁵ As 78. Wu, Wenming, et al.

⁹⁶ As 78. Wu, Wenming, et al.

Channel	Input	Output	Information Contained	Data Format
Inside Mask	Building boundary	Binary image	Indicates whether a pixel is inside or outside the building boundary	2D array of 0s and 1s
Boundary Mask	Building boundary	Binary image	Indicates whether a pixel is on the boundary of the building	2D array of 0s and 1s
Semantics of Rooms and Walls	Floor plan with labeled rooms and walls	Integer image	Each pixel is assigned an integer value indicating its semantic label (e.g., living room, bedroom, kitchen)	2D array of integers
Extra Information	Floor plan with labeled rooms and walls	Integer image	Used to distinguish between different rooms with the same labels	2D array of integers

Table 5. Comparison table between four data filtering processes ⁹⁷

The paper proposed a two-stage method to imitate the human design process to learn from the RPLAN dataset with the annotations. The first stage predicts the living room's position; other rooms can be added by an encoder-decoder network and a continuing network. (See Fig. 34) The second stage uses another encoder-decoder network to predict wall locations on the previous room allocations. Post-processing methods are used later to vectorize the architectural elements, including doors, windows, and walls, which define the geometry, connection, and function of different rooms in the plan.

⁹⁷ As 78. Wu, Wenming, et al..

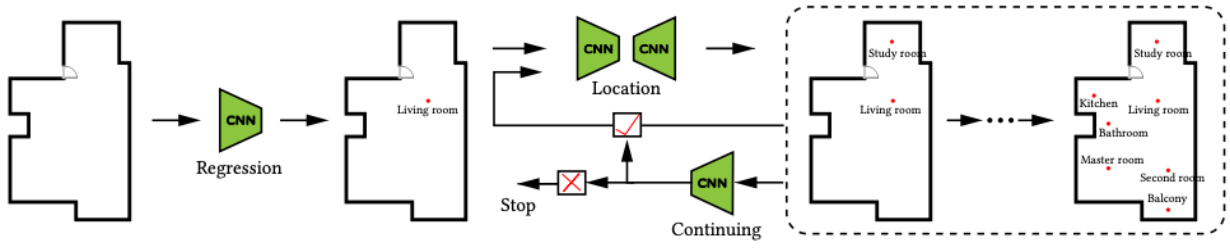


Figure 34. Floor Plan and Ground truth of RPLAN dataset⁹⁸.

The dataset provided extensive benchmark data for evaluating machine learning models for floor plan analysis. As a dataset retrieved from the real estate market, removing user and floor plan privacy information is essential for RPLAN. The dataset and the method are evaluated through extensive experiments and user studies, demonstrating the proposed approach's feasibility, efficacy, and plausibility. The results show that the technique can generate realistic and diverse floor plans comparable or superior to existing methods or human-created ones within four seconds per plan on average.

However, as the dataset contains only Asian floor plans, the dataset may not represent the diversity of other complex buildings and floor plans in other regions. The latent space distribution histogram results from the image datasets (See Fig. 35) are closer to a normal distribution, as indicated by the orange line, compared to previous datasets (such as CubiCasa5k). Since the dataset has more than 60k plans and evolves manual work with human knowledge and prior experience, it shows that the dataset could be more representative of the plans judging by its latent space distribution. The T-SNE plotted dot-distribution results showed that most of the data is distributed in the middle horizontal line (See Appendix 7). The plotted sample result (See Appendix 8) indicates that the datasets with higher x-value are towards a horizontal axis, and those with lower x-value are more vertically oriented. The datasets at the top and bottom are more towards an L shape, with the datasets with lower y-value leaning towards the orientation in the “7” shape.

⁹⁸ As

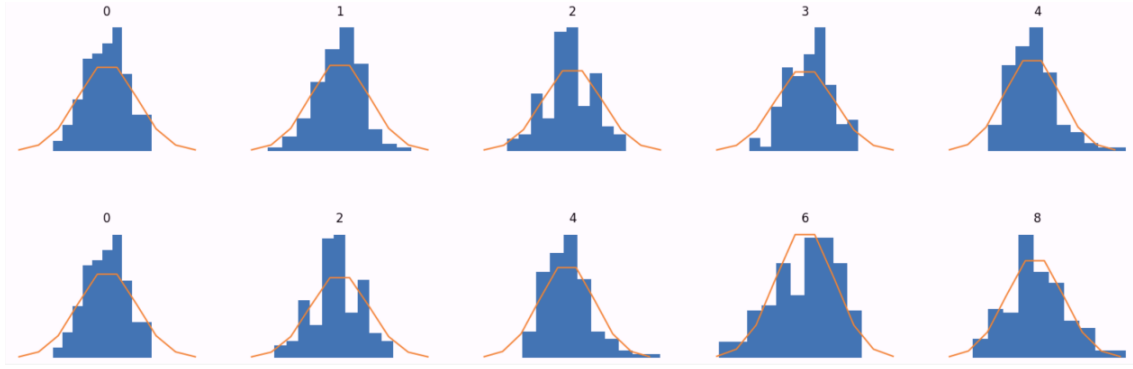


Figure 35. Latent Space Distribution of RPLAN dataset

v) **FloorPlanCAD Dataset**

The FloorPlanCAD dataset is a vector graphics-based CAD drawing containing more than 15,000 real-world floor plans.⁹⁹ The source covers a wide range of architectural types with more than 100 projects of different structural complexity levels, including schools, hospitals, residential, and shopping malls. It also provides a detailed representation and annotation style for CAD drawings. The paper discussed several challenges of the existing datasets and the current architecture practice. Existing datasets have smaller data samples and contain limited CAD styles and appearances. They also do not provide detailed line-grained or panoptic annotations. They do not contain instance and semantic detections of symbols as recognition problems. There are limitations to current architectural practice as well. The file format that current architectural practice uses needs to be standardized. As a result, different sources of elements may be represented by different styles and appearances of symbols.

The dataset was selected from CAD drawings from public repositories of architectural practice. And the initial dwg format was parsed into standard svg format (see Table 6), making the data more accessible and easy to manipulate with a broader range of software. After this stage, the data has geometric information and structural information only. Then the floor plan was labeled using line-grained annotations with 35 detailed object categories (see Fig 36). The annotation includes both thing instances (such as doors, windows, furniture, etc. See Fig 37) and stuff semantic (such as wall

⁹⁹ Fan, Zhiwen et al. "FloorPlanCAD: A Large-Scale CAD Drawing Dataset for Panoptic Symbol Spotting." *arXiv*, 2021, <https://arxiv.org/abs/2109.13353>. Accessed 7 July. 2023.

and parking), and each category is divided into subcategories. Each graphical entity is labeled with a semantic class with an instance index to distinguish one from another. Bigger plans are also cropped into smaller ones for a more straightforward process. Apart from the format conversation, data obfuscation is added to remove personal or commercial information from the data. To further protect privacy, a 10m x 10 m block was used to select 30% of the blocks into the dataset.

Data Type	dwg	svg
Data Structure	Binary	XML-based
Function	Stores 2D and 3D design data and metadata for CAD software packages	Stores 2D vector graphics with support for interactivity and animation
Information Stored	Design data, including geometric primitives, text, dimensions, hatches, etc.	Vector graphics data, including geometric primitives, text, gradients, etc.
Advantages	Native file format for several CAD software packages; widely used in the industry; supports complex design data	Open standard; widely supported by web browsers and other software tools; human-readable and easy to manipulate

Table 6. Comparison table between dwg and svg charts

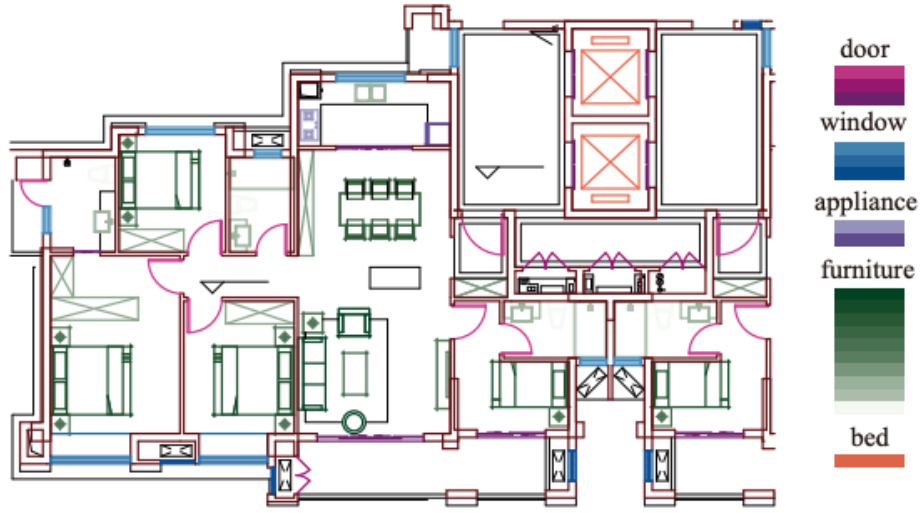


Figure 36. FloorPlanCAD with instance and semantic annotations¹⁰⁰.

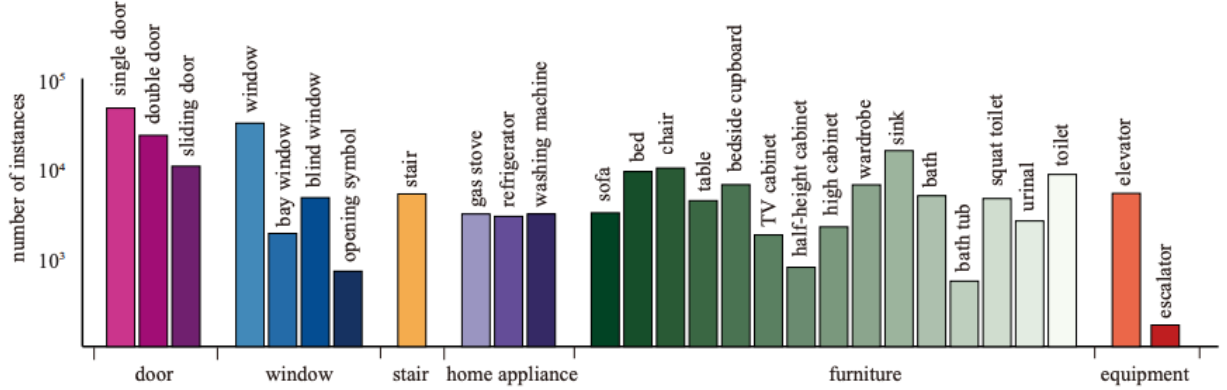


Figure 37. Number of annotated instances for 28 thing classes and categories¹⁰¹.

The paper also proposed a deep learning model PanCADNet regarding panoptic spotting problems trained on the FloorPlanCAD dataset (See Fig 38). The problem requires semantic information within graphical and textural symbols from CAD. The task is challenging due to the diversity and density of symbol types and the interactions between symbols and other drawing elements¹⁰². The model inputs a floor plan and outputs a panoptic segmentation map, assigning a semantic instance index to each pixel¹⁰³. The network consists of a CNN backbone, a graph convolution head, and a detection head.

¹⁰⁰ As 85 Fan, Zhiwen et al.

¹⁰¹ As 85 Fan, Zhiwen et al.

¹⁰² Chen, Zhihao, et al. “CADTransformer: Panoptic Symbol Spotting Transformer for CAD Drawings.” arXiv preprint arXiv:2107.12299 (2021).

¹⁰³ As 85 Fan.

The CNN captures Euclidean features from CAD input, and the graph convolutional head calculator non-euclidean features through topology connections. Symbol spotting was achieved by the detection head, which is based on Faster R-CNN¹⁰⁴. The training process first constructed a graph from CAD drawings using nodes as graphical elements and edges as adjacency-based or parallelism-based connections. Then the CNN backbone calculator geometric features (length, normalized center, and textural features) feed into the graph convolution head for GCN features of the node¹⁰⁵. The detection head takes features from the backbone as inputs and generates bounding boxes with index and probability scores for each instance. And the loss was calculated from the GCN head and detection head to train the whole PanCADNet.

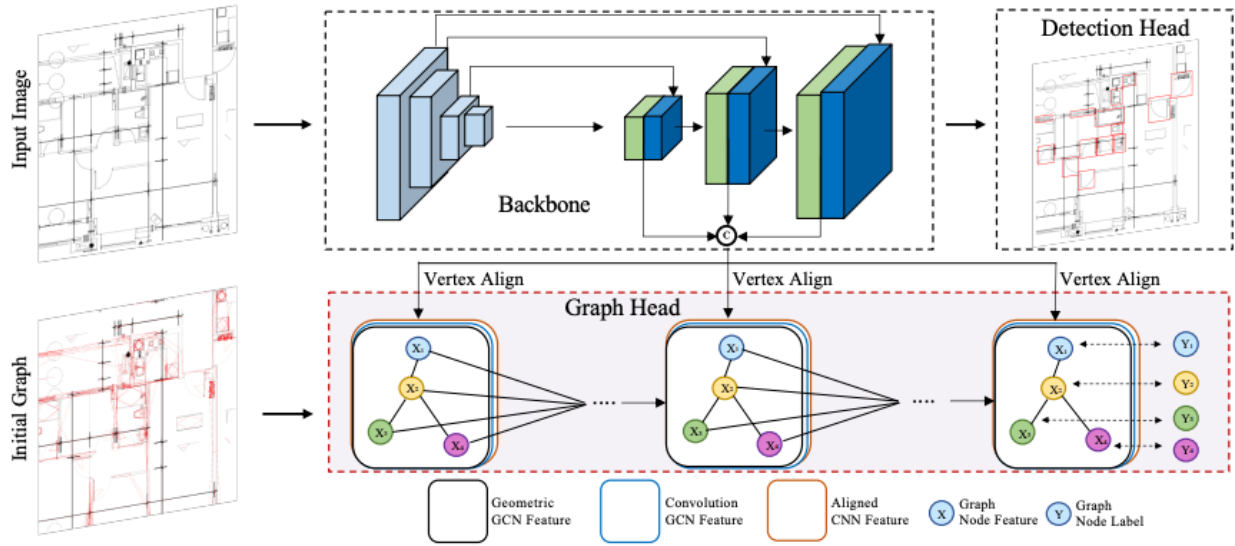


Figure 38. The network architecture of proposed PanCADNet¹⁰⁶.

The FloorPlanCAD dataset marks its significance as the first large real-world CAD dataset in line-grained annotated format. Labeling thing instances and stuff semantics facilitates the evaluation of future methods for panoptic symbol spotting. Its contribution to the field of automatic perception of

¹⁰⁴ Alireza Rezvanifar, Melissa Cote, and Alexandra Branzan Albu. Symbol spotting on digital architectural floor plans using a deep learning-based framework. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 568–569, 2020.

¹⁰⁵ Lumin Yang, Jiajie Zhuang, Hongbo Fu, Kun Zhou, and Youyi Zheng. Sketch Gcn: Semantic sketch segmentation with graph convolutional networks. arXiv preprint arXiv:2003.00678, 2020

¹⁰⁶ As 85 Fan, Zhiwen et al.

CAD drawings can further lead to efficient 3D modeling approaches in the AEC industry¹⁰⁷, which can save labor-intensive tasks by a great leap.

vi) Jigsaw Plan Dataset

The Jigsaw Plan is a dataset with more than 98,000 single-story houses and apartments.¹⁰⁸ The new dataset focuses on the Extreme Structure from Motion (E-SfM) problem research. The E-sfM task takes the room layout and the room types as input and predicts room position and orientation¹⁰⁹ (as Fig 39). The dataset was presented with a new Diffusion Model-based framework that solves the problem using generative progress.

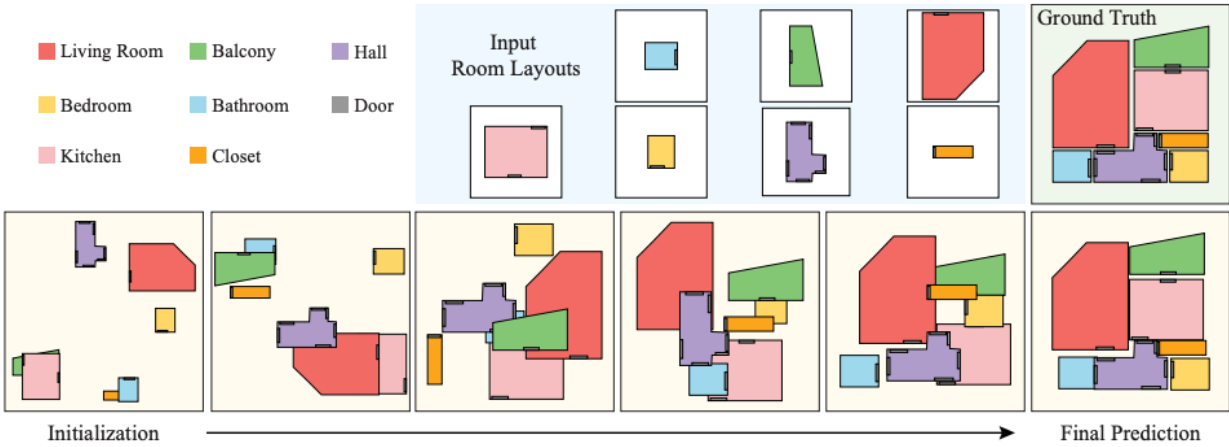


Figure 39. Example of E-sfM task, from the initial combination of layouts to prediction.¹¹⁰

The dataset was represented in vector graphics, using polygon curves and line segments to define shapes and attributes of planar elements (rooms and doors). Room types were associated with each room represented by a one-hot vector to the Diffusion Model. Sixteen room types were labeled in the JigsawPlan (See Fig 40). The pipeline was prepared from an Augmented Reality production pipeline that allows users to define the room corners for room layouts. The entire plans are normalized into squares of 256 x 256 pixels and scaled by a factor between 0.8 and 1 to avoid the different scale factors

¹⁰⁷ As 85. Fan, Zhiwen et al.

¹⁰⁸ Hosseini, Sepidehsadat, et al. "JigsawPlan: Room Layout Jigsaw Puzzle Extreme Structure from Motion using Diffusion Models." arXiv preprint arXiv:2211.13785v1, 2021. <https://arxiv.org/abs/2211.13785v1>.

¹⁰⁹ As 86. Hosseini, Sepidehsadat, et al.

¹¹⁰ As 86. Hosseini, Sepidehsadat, et al.

for the longer and shorter extent of the plan.

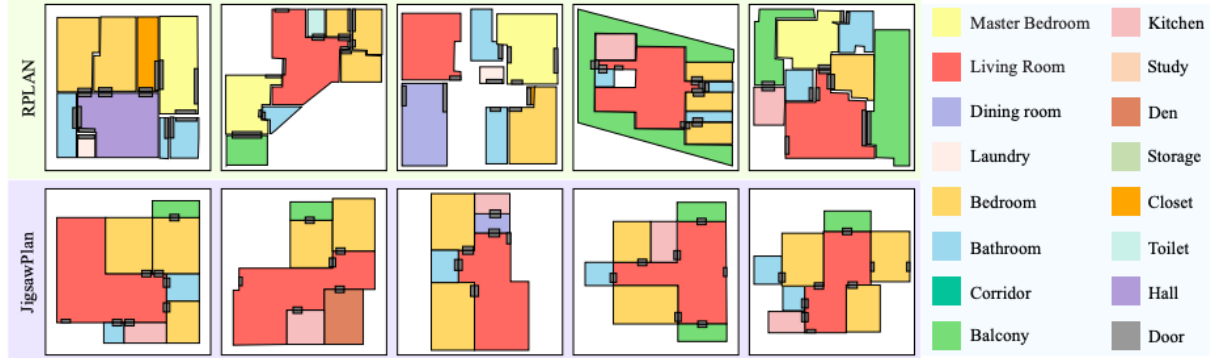


Figure 40. Samples from JigsawPlan and RPLAN by room types.¹¹¹

The JigsawPlan dataset sets a realistic benchmark for the proposed diffusion model (DM) to learn and evaluate the E-sfM problem's performance. The approach uses DM to solve the E-SfM problem as a conditional generative process, using the input room types and shapes as conditions. (See Fig. 41) To generate layout arrangements similar to the ground truth, the DM model uses a forward process to add Gaussian noise to the layout arrangement and a reverse process that infers the noise and recovers the arrangement with the original room condition. The performance of DM is evaluated by Mean Positional Error (MPE) in pixels¹¹² of the room and Graph Editing Distance (GED)¹¹³ in the edits of fixing the graph.

¹¹¹ As 86. Hosseini, Sepidehsadat, et al.

¹¹² Mohammad Amin Shabani, Weilian Song, Makoto Odamaki, Hirochika Fujiki, and Yasutaka Furukawa. Extreme structure from motion for indoor panoramas without visual overlaps. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2021

¹¹³ Nelson Nauata, Sepidehsadat Hosseini, Kai-Hung Chang, Hang Chu, Chin-Yi Cheng, and Yasutaka Furukawa. Housegan++: Generative adversarial layout refinement network towards intelligent computational agent for professional architects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13632–13641, 2021.

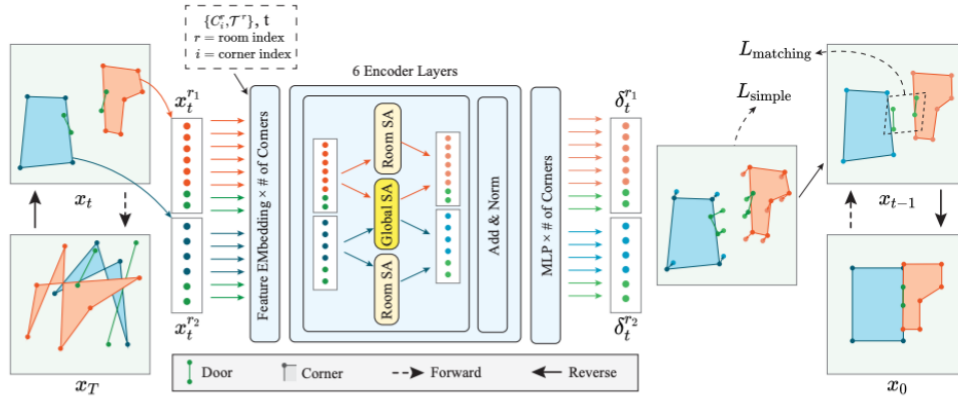


Figure 41. Diffusion model structure proposed for E-SfM.¹¹⁴

The dataset sets a new standard and evaluation metric for E-SfM problems with clear and detailed annotations. It can test the robustness and scalability of methods within this specific research field, which can be applied to indoor mapping and floor plan reconstruction with mobile devices. The rich information about shapes and attributes of rooms and doors also sets up a new paradigm for datasets to train generative models for synthesizing realistic indoor scenes. However, some things could be improved regarding the diversity of data sources. The floors only contain those from single-story houses and apartments, which does not reflect the complexity of the real-world dataset. The dataset may not count for possible errors based on the Manhattan assumption¹¹⁵, which simplifies the estimating of 3D structure and rotation of scenes from scans into flat surfaces aligned with three main axial directions.

¹¹⁴ As 86. Hosseini, Sepidehsadat, et al.

¹¹⁵ Guo, Haoyu, et al. "Neural 3D Scene Reconstruction with the Manhattan-world Assumption." arXiv, vol. 2205, no. 02836v2, 2022, pp. 1-10. <https://arxiv.org/abs/2205.02836>.

5) Findings and Discussions

a) Potentials for data format organization

This part explores the future potential of curating datasets by looking into each data type.

Comparative analysis of datasets will give us a more in-depth understanding of the existing data information and annotation for floor plan datasets. Inspired by the previous of indoor datasets by Cruz et al.¹¹⁶, the table below (Table 7) shows the data type and annotation information across the surveyed datasets for this thesis, with the potential for future architecture researchers.

Dataset	Image Data	Vector Data	CAD Info	Architectural Elements	Text Files	Potential use for Architectural Research
CVC-FP	Yes (PNG, JPG, TIFF)	Yes (SVG)	No	Yes (Walls, doors, windows, etc.)	No	To synthesize new floor plans that meet specific requirements or preferences, using generative models that learn from the existing examples in the dataset.
ROBIN PLAN	Yes (Binary)	No	No	Yes (Room labels, furniture positions)	No	To analyze the semantic and structural features of floor plans and evaluate the performance of symbol recognition, room detection, and layout retrieval systems
CubiCasa5k dataset	Yes (PNG, JPG)	Yes (SVG)	No	Yes (Rooms, walls, doors, windows, furniture)	Yes (CSV, JSON)	To develop and evaluate methods for automatic floorplan analysis and reconstruction
RPLAN	Yes (Image)	Yes (SVG)	No	Yes (Rooms, walls, doors, windows)	Yes (JSON, TXT)	To train and test novel algorithms for synthesizing realistic and functional floor plans, or to compare different design criteria and metrics for assessing the quality of floor plans.
FloorPlan CAD	Yes (PNG)	Yes (SVG)	Yes	Yes (Doors, windows, furniture, appliances, walls)	Yes (JSON)	To apply generative design and machine learning methods to create novel and optimized floor plans based on the dataset
JigsawPlan	Yes (Raster images for semantic segmentation and occupancy maps)	Yes	No	No	No	To enable novel applications such as automated floorplan design, layout optimization, virtual reality, and interior decoration
Future Potentials	combinational ways of annotations, such as labeling masks	combining image data and vector data closely	standardize the format and structure for CAD files			NA

Table 7. Comparison chart with the additional surveyed datasets

¹¹⁶ As Cruz, Steve, et al. Zillow Indoor Dataset: Annotated Floor Plans With 360o Panoramic and 3D Room Layouts. arXiv.org, Cornell University, 30 Oct. 2020, <https://arxiv.org/abs/2010.14977>.

For image-based data, combinational annotations, such as labeling masks, can be further explored for future research. Combining the underlying mask layers and pixel layers is a possibility. From Table 26, most of the datasets this thesis surveyed contain image data. Although their formats may vary between PNG, JPG, or TIFF, they all have rich information that is easily understood. Other formats may not easily capture the image's features (such as hatches and color codings). As mostly accessible architecture plans online are image-based information (See Fig 42), and images are more accessible for the general public to comprehend, making it used widely in real-estate websites. This type of real-world image plan (as in ZIND, LIFULL, CVC-FP, and Cubi-Casa datasets) is also commonly used for tasks such as floor plan analysis and object recognition in architecture research. However, other formats of image-based datasets also represent only certain planar information, such as specific mask information for boundaries, rooms, and walls (as RPLAN). Architectural researchers often neglect these layers of information. Still, they can provide valuable information for image-based research, such as labeling the functional zonings of different architectural elements (e.g., rooms and walls) for image generation GANs (such as the preliminary experiment part of my thesis background part using the pix2pix model). It can also contribute to the new exploration of Stable Diffusion models that provides different examples of labeling in image format (e.g., in Zheng's post¹¹⁷). The extraction and reuse of image-based floor plans as mask information is a potential step for future exploration.

Figure 42. Result for “Architecture plan” in Google Search.

There is also the potential to combine image data and vector data closely. Image data requires relatively more storage and processing power. However, vector-graphic information can represent floor plans and architectural elements in a compact and easily-analyzable format. It can provide precise measurements together with scalable representations. Processing with vector-based data reduces the computational resources required for tasks such as object recognition and semantic segmentation with better accuracy in the results. However, most datasets analyzed in this paper store their annotations in vector-graphics format. Some vector-graphic annotations are combined with rasterized plan images to provide rich datasets for feature extraction-related problems such as automatic plan image analysis (e.g., CasaCubi5k dataset). Combining image- and vector-based data will encourage future research for symbol spotting, architecture element segmentation (such as room segmentation), and 3D reconstruction, with a more standardized annotation format for plans. It will save researchers and architects time working across the two types of plan data and help researchers analyze the spatial and functional relationships among annotated labels.

CAD datasets also have the potential to standardize the format and structure for future research. Among all the other vector-graphic information with annotations such as bounding boxes and labels, CAD (Computer-Aided Design) file is a special vector graphic data commonly used in the AEC industry. The standardization of industrial firms such as Autodesk allows the prevalence of CAD files to capture 2D and 3D features of physical space. However, from the surveyed datasets, it was not thoroughly studied and used by most researchers. Part of the reason may be that most of the datasets prepared are from a computer vision or robotics background. Those datasets may have limited applicability for architectural researchers where the CAD file is a more precise interpretation and significant format of the design output. And the proposal of FloorCADPlan has indeed boosted the innovative architectural research focusing on areas such as automated floor plan generation (e.g., HouseGAN++¹¹⁸). The availability of open-source CAD-based plans has attracted a new discussion about computational agents that bridges the gap between professional architects and researchers from technical backgrounds. However, the CAD files in the AEC industry share different standards by firms or even by teams within the same firm. Therefore, standardizing the format and structure of CAD data is important for researchers to incorporate new CAD files into the dataset. It can be expected that more accurate CAD-based datasets will benefit future architecture researchers in collaborating with architects and the AEC industry.

¹¹⁸ As 34. Nauata, Nelson et al. “House-GAN++: Generative Adversarial Layout Refinement Network towards Intelligent Computational Agent for Professional Architects.”

b) Towards a collaborative data annotation process

The surveyed datasets discussed the data curation process in detail. For example, the RPLAN discusses the overall data collection, filtering, annotation, and representation process. However, the research requires a lot of time for data annotation. And some datasets can be applied for various tasks (e.g., CubiCasa5k and RPLAN can be used for floor plan image analysis and retrieval). Therefore, it is important to propose a framework for data annotation to provide flexibility for multiple tasks and applications for plan-based datasets.

To foster a more collaborative annotation process, the first step is to ensure a consistent and hierarchical annotation style that uses a common taxonomy among project researchers in specific disciplines. For example, the FloorPlanCAD researchers propose to choose 35 object classes when curating the dataset, and the listed categories and subcategories should be taken into consideration when future researchers are thinking about creating another CAD dataset for similar purposes or purposes that may have some overlap.

Secondly, data types, research tasks, and the domain should be flexible and adaptable (as indicated in 5.1). It can provide more customization and flexibility for a collaborative annotation process. For instance, if my research hopes to study the effectiveness of image-based generative models for different labeling styles, having the image data and image-based annotation from RPLAN and CVC-FP is essential. And with more available tools that can convert vector-graphic information (as in ROBIN and CubiCasa5k dataset) into more informative image-based ones, the same domain task can engage more diverse data to test the performance result.

Thirdly, adapting current state-of-the-art tools in collaborative labeling is also important. In the surveyed datasets, researchers have provided examples of engaging multiple users and experts (RPLAN) using tools with version control (CVC-FP) and quality assessment (CubiCasa5k). However, other tools for collaborative editing of raster and vector images simultaneously include CERVI¹¹⁹ (as Fig 43). Some tools support text annotations (such as TeamTat¹²⁰), allowing researchers to track task completion. Adapting new tools to the need for floor plan dataset annotation would be very beneficial

¹¹⁹ Bath, Ulrike et al. "CERVI: collaborative editing of raster and vector images." *Vis. Comput.* 38 (2022): 4057-4070.

¹²⁰ Rezarta Islamaj and others, TeamTat: a collaborative text annotation tool, *Nucleic Acids Research*, Volume 48, Issue W1, 02 July 2020, Pages W5–W11, <https://doi.org/10.1093/nar/gkaa333>

to collect, curate and share data among research communities collectively.

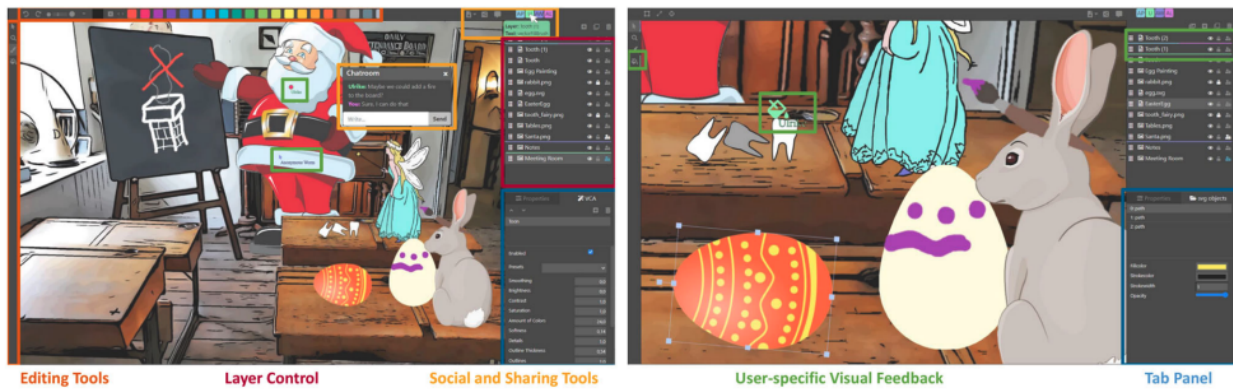


Figure 43. Example of CERVI GUI with editing tools and visual feedback from users¹²¹.

¹²¹ As Bath, Ulrike et al. “CERVI: collaborative editing of raster and vector images.”

6) Conclusion

***a)* Contributions**

The thesis provides a selective and systematic review of 6 existing plan-based datasets and their implementations for architectural research. The survey identifies the characteristics, potentials, and challenges of existing plan-based datasets, such as data sources, data types, data structures, data annotations, data analysis, and data applications. The work identifies the gaps and limitations of current plan-based datasets in architecture research and discusses the challenges and difficulties of obtaining, preparing, and annotating plan-based data. Therefore, the thesis suggests best practices and recommendations for architectural researchers regarding new potentials in data format organization and collaborative data annotation. The work can help researchers better understand the plan-based datasets and their implementations and inspire new research questions and methods for architectural research.

***b)* Limitations**

This thesis conducted a selective review of existing plan-based datasets and their implementations but only covered a fraction of the available datasets in architecture design and research. In the meantime, the survey focused on the characteristics, annotations, potentials, and challenges of plan-based datasets. Still, it needed to evaluate the performance or quality of the models or methods that used these datasets. Some qualitative analyses using dimension reduction tools (see Appendix 1 - 8) may not reflect the characteristics of the datasets provided or may only suggest the typology of image data of the complex dataset. Meanwhile, the survey is limited to the datasets containing floor plan information and its implementations. It may not address other datasets or methods relevant to architecture research (such as 3D data, panoramic data, etc).

***c)* Future work**

In the future, there is potential to expand the data format of the datasets regarding the architecture field. The survey can extend the dataset to include more plan-based datasets from different sources, domains, and formats, such as CAD, BIM, or GIS data, to provide a more comprehensive and diverse overview of the available data for architectural research. There is still the possibility of developing new pipelines to incorporate different datasets for data collection, annotation, and analysis. Although this thesis tried to categorize and compare the data formats, preparation process, and pipelines of the data, evaluating

the design datasets is still a problem. So it would be essential to set reasonable metrics and benchmarks combined with user studies to assess and compare the quality of different architectural datasets.

Bibliography:

Alireza Rezvanifar et al. Symbol spotting on digital architectural floor plans using a deep learning-based framework. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 568–569, 2020.

Apostolidis, Savvas D., et al. “Cooperative Multi-UAV Coverage Mission Planning Platform for Remote Sensing Applications.” *Autonomous Robots*, vol. 46, no. 2, Jan. 2022, pp. 373–400. Crossref, <https://doi.org/10.1007/s10514-021-10028-3>.

Bai, Nan, et al. “Heri-Graphs: A Dataset Creation Framework for Multi-Modal Machine Learning on Graphs of Heritage Values and Attributes with Social Media.” *ISPRS International Journal of Geo-Information*, vol. 11, no. 9, Aug. 2022, p. 469. Crossref, <https://doi.org/10.3390/ijgi11090469>.

C. Liu, J. Wu, P. Kohli and Y. Furukawa, "Raster-to-Vector: Revisiting Floorplan Transformation," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2214-2222, doi: 10.1109/ICCV.2017.241.

Cardoso Llach, Daniel. “Tracing Design Ecologies: Collecting and Visualizing Ephemeral Data as a Method in Design and Technology Studies.” *DigitalSTS: A Field Guide for Science & Technology Studies*, edited by Janet Vertesi and David Ribes, Princeton University Press, 2019, pp. 451-468. <https://press.princeton.edu/books/hardcover/9780691187075/digitalsts>.

Cardoso Llach, Daniel. “Sculpting Probabilistic Spaces: Brief History and Prospects of Machine Learning in Design.” In *The Routledge Companion to Artificial Intelligence in Architecture*, edited by Wassim Jabi and Mine Özkar, 1st ed., 3-25. Routledge, 2021.

Carta, Silvio. “Self-Organizing Floor Plans.” *Harvard Data Science Review* (2021). <https://hdsr.mitpress.mit.edu/pub/w1gujxim/release/3>.

Chaillou, Stanislas. “ArchiGAN: Artificial Intelligence x Architecture.” (2020). <https://www.semanticscholar.org/paper/ArchiGAN%3A-Artificial-Intelligence-x-Architecture-Chaillou/53c1aed0d6c043f40e5353f40d744bbe0bb454f6>

Chapman, Adriane, et al. “Dataset Search: A Survey.” *The VLDB Journal*, vol. 29, no. 1, Aug. 2019, pp. 251–72. Crossref, <https://doi.org/10.1007/s00778-019-00564-x>.

Chen, Zhihao, et al. “CADTransformer: Panoptic Symbol Spotting Transformer for CAD Drawings.” arXiv preprint arXiv:2107.12299 (2021).

Cruz, Steve, et al. “Zillow Indoor Dataset: Annotated Floor Plans With 360o Panoramic and 3D Room Layouts.” arXiv.org, Cornell University, 30 Oct. 2020, <https://arxiv.org/abs/2010.14977>.

D. Sharma, et al, "DANIEL: A Deep Architecture for Automatic Analysis and Retrieval of Building Floor Plans," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 420-425, doi: 10.1109/ICDAR.2017.76.

Delalandre, Mathieu, et al. "CVC-FP and SGT: a new database for structural floor plan analysis and its ground truthing tool." *International Journal on Document Analysis and Recognition (IJDAR)* 17.1 (2014): 19-36.

Delalandre, Mathieu, et al. "Generation of synthetic documents for performance evaluation of symbol recognition & spotting systems," *IJDAR*, vol. 13, no. 3, pp. 187–207, 2010.

De las Heras, Lluís-Pere, et al. "Statistical Segmentation and Structural Recognition for Floor Plan Interpretation." *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 18, no. 3, 2015, pp. 221-237.

Deng, Jia et al. Imagenet: A large-scale hierarchical image database. *CVPR*. pp. 248–255 (2009)

Fan, Zhiwen et al. "FloorPlanCAD: A Large-Scale CAD Drawing Dataset for Panoptic Symbol Spotting." *arXiv*, 2021, <https://arxiv.org/abs/2109.13353>. Accessed 7 July. 2023.

Ferrando, Cecelia. (2018). Towards a Machine Learning Framework in Spatial Analysis. Master's thesis, Carnegie Mellon University, Pittsburgh, PA.

Guo, Haoyu, et al. "Neural 3D Scene Reconstruction with the Manhattan-world Assumption." *arXiv*, vol. 2205, no. 02836v2, 2022, pp. 1-10. <https://arxiv.org/abs/2205.02836>.

Hahn, Yannik et al. (2023). Time Series Dataset Survey for Forecasting with Deep Learning. *Forecasting*. 5. 315-335. 10.3390/forecast5010017.

Hao Zheng. <https://www.linkedin.com/feed/update/urn:li:activity:7036371613503647745/>. Accessed March 22, 2023.

Hasey, Michael (2022). The 3D Form Analysis of Regional Architecture Using Deep Learning: A Case Study of Wood Churches from the Carpathian Mountains. Carnegie Mellon University. Thesis. <https://doi.org/10.1184/R1/21291672.v1>

He, Kaiming et al. Deep Learning for Image Recognition. *Proc. CVPR* pp. 770–778 (2016)

Heras, Lluís-Pere & Terrades, Oriol & Robles, Sergi & Sánchez, Gemma. (2015). CVC-FP and SGT: a new database for structural floor plan analysis and its ground truthing tool. *International Journal on Document Analysis and Recognition (IJDAR)*. 18. 10.1007/s10032-014-0236-5.

Hosseini, Sepidehsadat et al. "JigsawPlan: Room Layout Jigsaw Puzzle Extreme Structure from Motion using Diffusion Models." *ArXiv* abs/2211.13785 (2022).

Isola, Phillip, et al. "Image-to-Image Translation with Conditional Adversarial Networks." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017. Crossref, <https://doi.org/10.1109/cvpr.2017.632>.

Kalervo, Ahti, et al. "CubiCasa5K: A Dataset and an Improved Multi-Task Model for Floorplan Image Analysis." *Lecture Notes in Computer Science*, 2019, pp. 28–40. Crossref, https://doi.org/10.1007/978-3-030-20205-7_3.

Kingma, Diederik P., and Max Welling. "Auto-Encoding Variational Bayes." Machine Learning Group, Universiteit van Amsterdam, 2013. <https://arxiv.org/abs/1312.6114>

Li, Tingguang, et al. "HouseExpo: A Large-scale 2D Indoor Layout Dataset for Learning-based Algorithms on Mobile Robots." *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2019): 5839-5846.

Lumin Yang et al. Sketch Gcn: Semantic sketch segmentation with graph convolutional networks. arXiv preprint arXiv:2003.00678, 2020

Nauata, Nelson, et al. "House-GAN++: Generative Adversarial Layout Refinement Network towards Intelligent Computational Agent for Professional Architects." *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021): 13627-13636.

Newton, David William. "Deep Generative Learning for the Generation and Analysis of Architectural Plans with Small Datasets." *Blucher Design Proceedings* (2019).

Peters, Nathan. (2017). Master thesis: "Enabling alternative architectures: Collaborative Frameworks for participatory design." Cambridge, MA: Harvard Graduate School of Design

Rhee, Jinmo & Veloso, Pedro & Krishnamurti, Ramesh. (2020). Integrating Building Footprint Prediction and Building Massing: An Experiment in Pittsburgh. 10.52842/conf.caadria.2020.2.669.

Rhee, Jinmo and Veloso, Peloso. Type Investigation in the Form of High-Rise Building Using Deep Neural Networks. *Computer-Aided Architectural Design. Design Imperatives: The Future is Now 2022*, Volume 1465 ISBN: 978-981-19-1279-5.

Roh, Yuji. et al. "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1328-1347, 1 April 2021, doi: 10.1109/TKDE.2019.2946162.

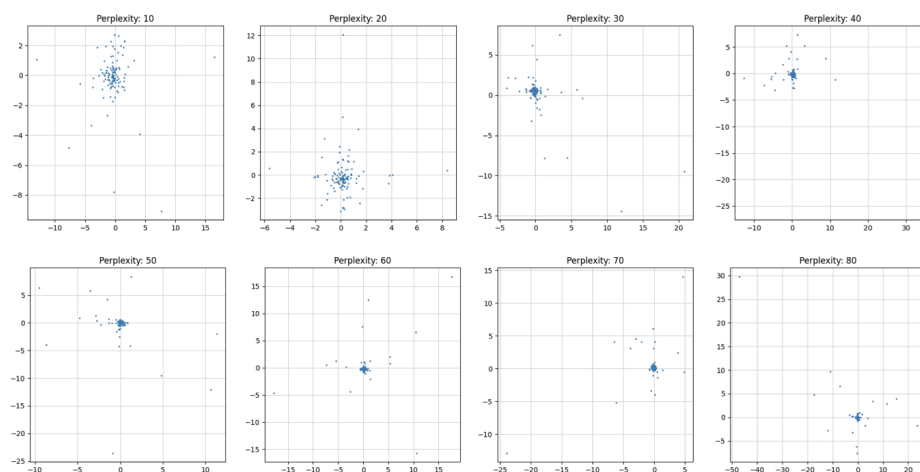
Wenming Wu, Xiao-Ming Fu, Rui Tang, Yuhao Wang, YuHao Qi, and Ligang Liu. Data-driven interior plan generation for residential buildings. *ACM Transactions on Graphics (TOG)*, 38(6):1–12, 2019.

Wu, Wenming, et al. "Data-driven Interior Plan Generation for Residential Buildings." *ACM Transactions on Graphics*, vol. 38, no. 6, Nov. 2019, <https://dl.acm.org/doi/10.1145/3355089.3356556>.

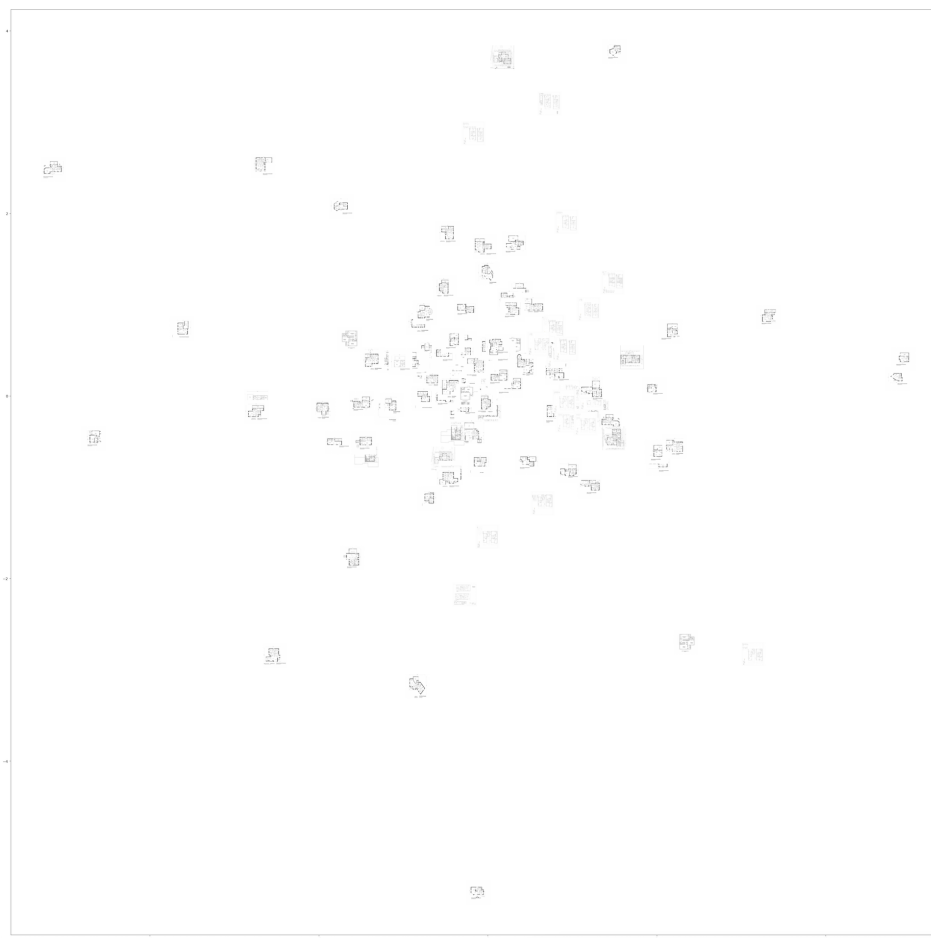
Zhang, Jason. Et al. (2021). Neural Reflectance Surfaces for Sparse-view 3D Reconstruction in the Wild. In *Conference on Neural Information Processing Systems*. arXiv: 2110.07604 [cs.CV]

Zheng, Hao., & Huang, Weixin. (2018). *Architectural drawings recognition and generation through machine learning*. Cambridge: MA, ACADIA

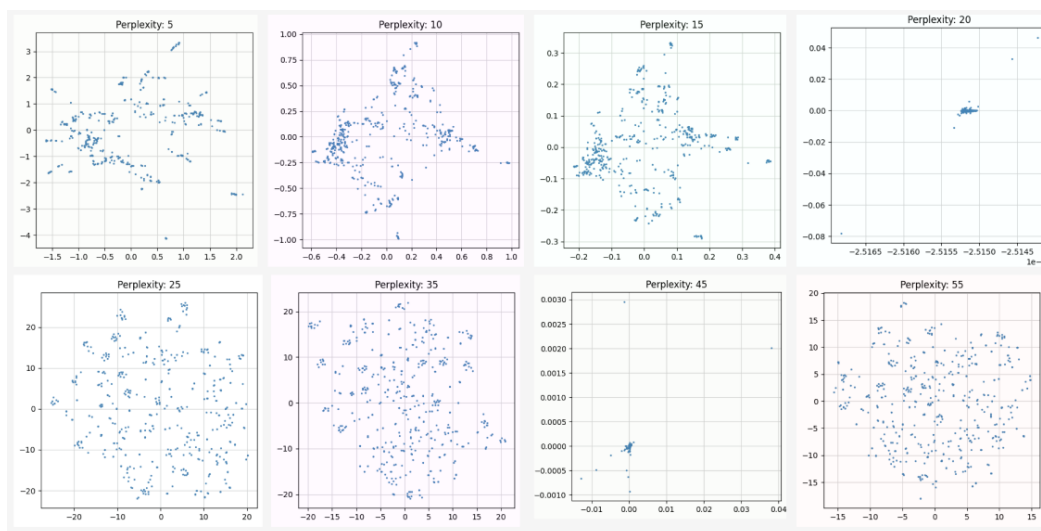
Appendix



Appendix 1. T-SNE Distribution of CVC-FP dataset



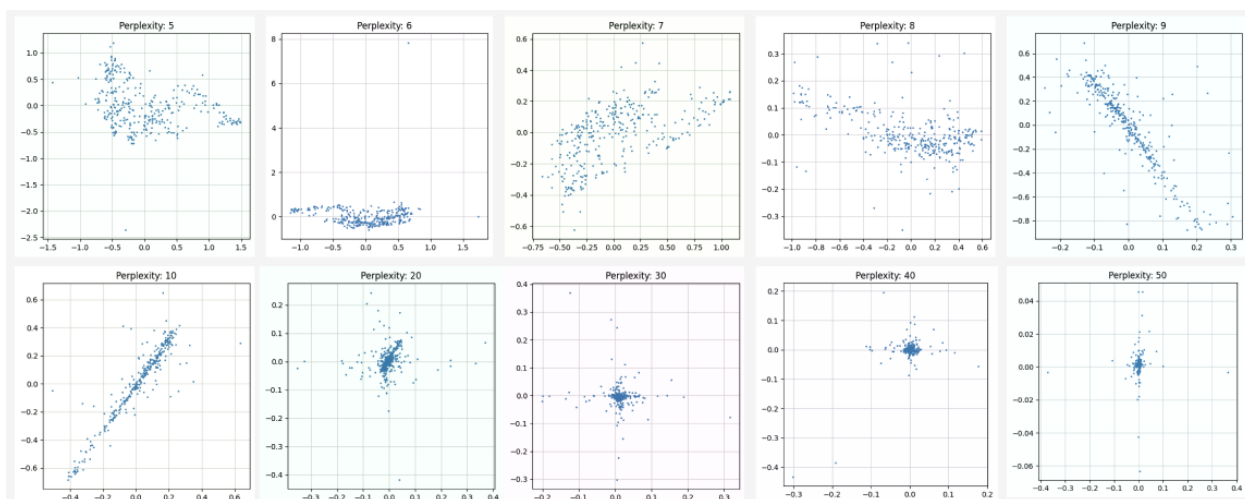
Appendix 2. Visualization of T-SNE results for CVC-FP datasets.



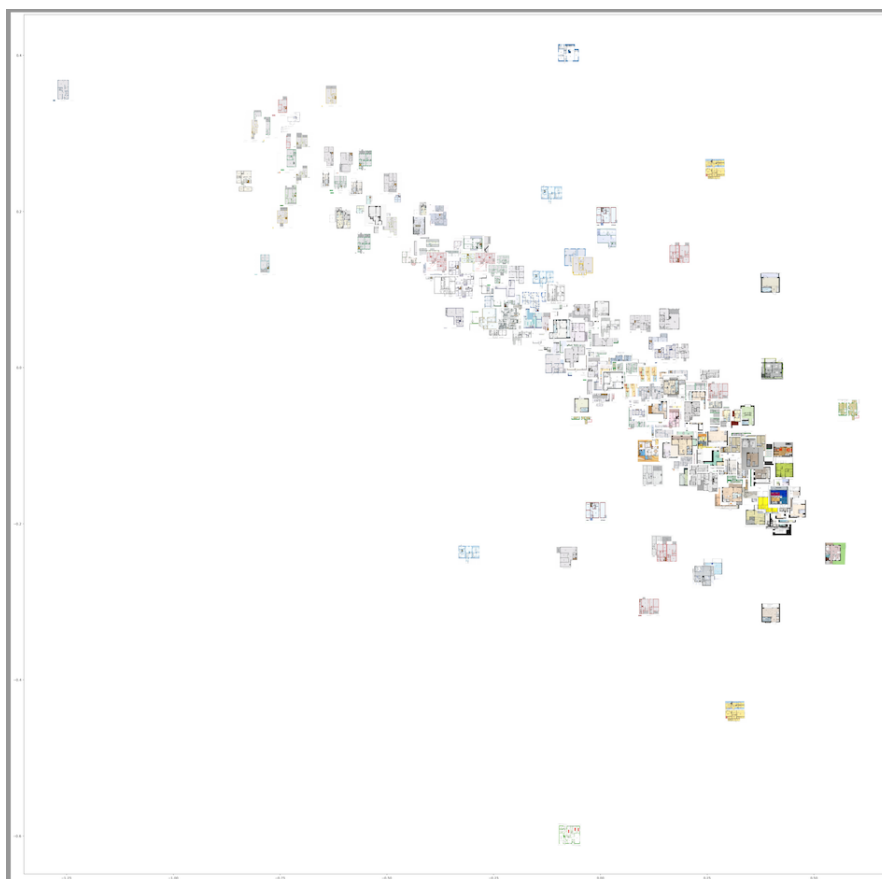
Appendix 3. T-SNE distribution of ROBIN datasets.



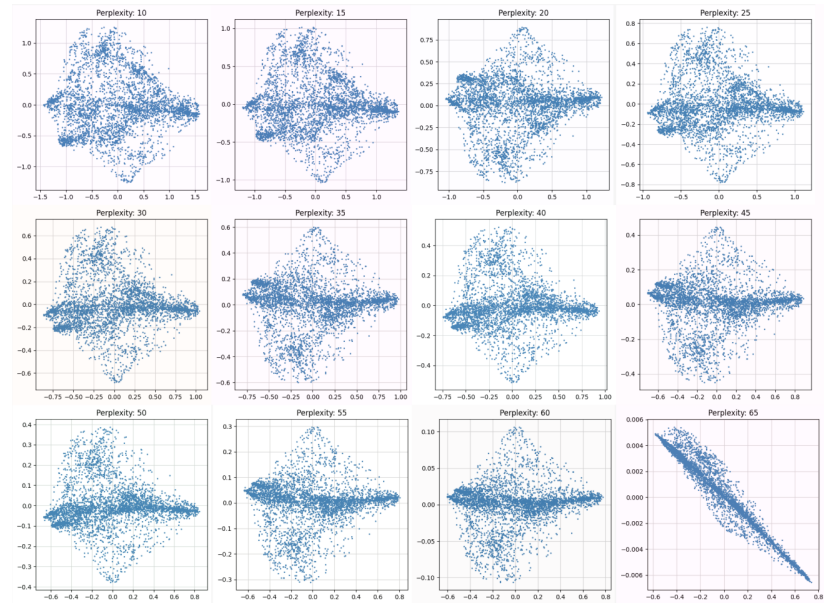
Appendix 4. Visualization of T-SNE results for ROBIN datasets.



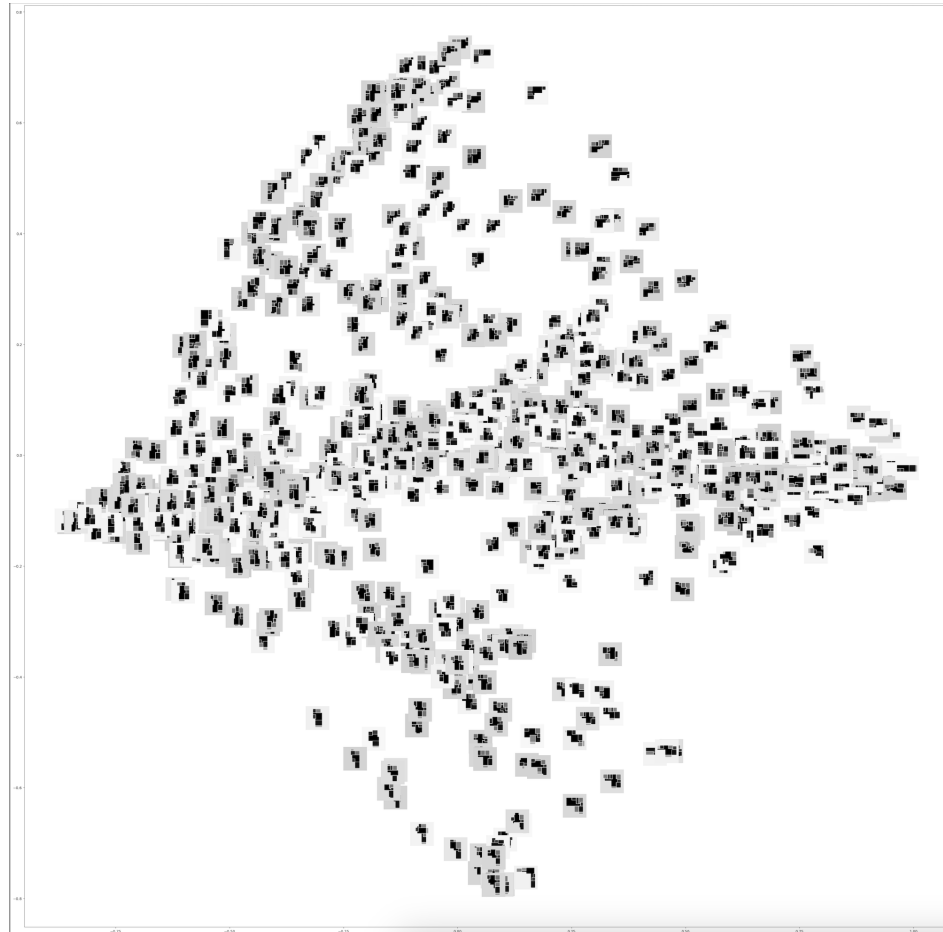
Appendix 5. T-SNE Distribution of CubiCasa5k “Colorful” dataset



Appendix 6. Visualization of T-SNE result for selected CubiCase5k “Colorful” datasets.



Appendix 7. T-SNE Distribution of RPLAN dataset



Appendix 8. Visualization of T-SNE result for selected RPLAN datasets.