Big Data vs Smart Data on the Generation of Floor Plans with Deep Learning

Ricardo Cesar Rodrigues, Marcelo Kenzo Imagawa, Renan Rubio Koga,
Rovenir Bertola Duarte

State University of Londrina, Brazil
ricardo.rodrigues@uel.br
marcelo.kenzo.imagawa@uel.br
renan.rubiokoga@uel.br
rovenir@uel.br

Abstract. Due to the progressive growth of data dimensionality, addressing how much data and time is required to train deep learning models has become an important research topic. Thus, in this paper, we present a benchmark for generating floor plans with Conditional Generative Adversarial Networks in which we compare 10 trained models on a dataset of 80.000 samples, the models use different data dimensions and hyper-parameters on the training phase, beyond this objective, we also tested the capability of Convolutional Neural Networks (CNN) to reduce the dataset noise. The models' assessment was made on more than 6 million with the Frétche Inception Distance (FID). The results show that such models can rapidly achieve similar or even better FID results if trained with 800 images of 512x512 pixels, in comparison to high dimensional datasets of 256x256 pixels, however, using CNNs to enhance data consistency reproduced optimal results using around 27.000 images.

Keywords: Floor plans, Generative design, Generative adversarial networks, Smart Data, Dataset reduction.

1 Introduction

The creation of floor plans is a key step in architectural design, however, there is not a precise general method or universal tool for the space planning task; even if there are some common features in the early stages, each case (residential buildings, hospitals, schools, etc.) has special requirements (Lobos & Donath, 2010). As a consequence of the design complexity, studies in generative design have focused on the use of Generative Adversarial Networks

(GANs) to automate the process of implementing design rules on generative models, especially in floor plans (Hong et al., 2020).

However even after six years of its creation, the very own inventors of GANs Goodfellow et al., (2020) argue that they are still difficult to train, and suggest that researchers need to design models, costs, or training algorithms to find a quick and consistent optimal convergence, i.e. Nash Equilibrium.

In addition to this problem, GANs are considered a data-hungry technology, because they require hundreds or even thousands of images to train. Moreover, the applications of deep learning techniques in architecture are still in their early beginnings, so naturally, well-balanced training sets are difficult to find, or sometimes they do not even exist (Belém et al., 2019).

However, even if certain types of data are not easily available, when it comes to floor plans we can observe a significant increase in the dimensionality of datasets, such as; CubiCasa5K – 5’000 images (Kalervo et al., 2019); RPLAN – 80’000 images (Wu et al., 2019) and LIFULL HOME’s dataset – 110’000 images (LIFULL, 2016). This makes us – as architects, rethink the way we collect, organize and normalize spatial data in order to ensure a good performance of such techniques.

Considering all these challenges, the present research questions this data-growth, bringing a quantitative comparison between different data dimensions needed to train a GAN model of floor plans. Thus, the aim is to optimize the relationship between the data volume with the quality and time of training on datasets of semantic segmented floor plans.

2 Background

Authors like Mahankali, Johnson and Anderson (2018) explains that by providing homogenous data formats to deep learning models we can store design information as spatial-semantic maps, which can be especially useful in the early design phases of; Housing design (Rahbar et al., 2019), apartments (Chaillou, 2020; Zheng et al., 2020) and hospital layouts (Zhao et al., 2021).

Although several generative methods already exist, such as genetic algorithms, shape grammars etc., in regard to using deep neural networks to generate floor plans, we can also have different generative approaches as well. The differences can be noticed if we take into account not only the method used but also the type of input that the method accepts. These input constraints can be **Graph-based**, which take the form of bubble diagrams as input (Hu et al., 2020; Nauata et al., 2021; Wu et al., 2019), **Language-based**, which takes linguistic descriptions as input to the generative model (Chen et al., 2020; Galanos, 2021), and last but not least **Pixel-based** approaches, which use the pixel color as constrains to the generative model, whereas information like shape, orientation or area could be further determined (Chaillou, 2020; Peters, 2018; Rahbar et al., 2019).

Notwithstanding, there are also approaches which do not use any specific constrain to generate novel floor plans after the training phase is done, like Wasserstein GAN (Newton, 2019), DCGAN (Uzun et al., 2020), and many others. The problem of these approaches rely on the lack of control of the generative process, as Chaillou (2020) explains, when the designer has no control over the generative steps, such models can be seen as black boxes, in this sense, if the approach permits the designer to intervene along the way, this is the ultimate guarantee of the design process quality.

Despite the innumerous pros and cons of each approach, the models trained with the conditional-Generative Adversarial Networks, also known as Pix2pix, have not been taking much advantage of big floor plan datasets. As shown in Table 1, most of them use from 300 to 800 images on training phase.

Table 1. Previous Pix2Pix Models of Semantic Segmented Floor Plans

Author	**Dataset Size**	**Image Resolution**	**Framework**	**Training Time**	**GPU Hardware**
(Zheng, 2018)	800	-	Pix2pix	-	-
(Peters, 2018)	300	256x256	Pix2pix	-	-
(Huang & Zheng, 2018)	100	1024x1024	Pix2pixHD	1,8h	-
(Rahbar et al., 2019)	300	256x256	Pix2pix	-	-
(Zheng et al., 2020)	1279	512x512	Pix2pix	33h	Titan X
(Chaillou, 2020)	800	256x256	Pix2pix	2h	TeslaV100
(Zhao et al., 2021)	100	512x512 & 256x256	Pix2pix	3 ~ 6h	-

Source: Authors

Another obstacle to replicate results is the lack of hyper-parameters as well as the evaluation methods, which differ from one to another. In this sense, we seek to establish a standard evaluation benchmark for this specific generative framework, by dealing with the generation of a huge amount of floor plan images. Once determined good practices on floor plan models with Pix2pix, only then future research may address human-based evaluation criteria like diversity, realism and compatibility (Nauata et al., 2021) or orthogonal design, dimensions of the spaces, proportion of space's area, entrance recognition and logics of space allocation (Rahbar et al., 2019).

3 Materials and Methods

To investigate the relationship between the volume of the dataset with the quality and time of training, an exploratory research was carried out. We divided it into three evaluation phases, each evaluation phase gave subsets of valuable information to change training data and hyper-parameters of the following models, shown in Figure 1.

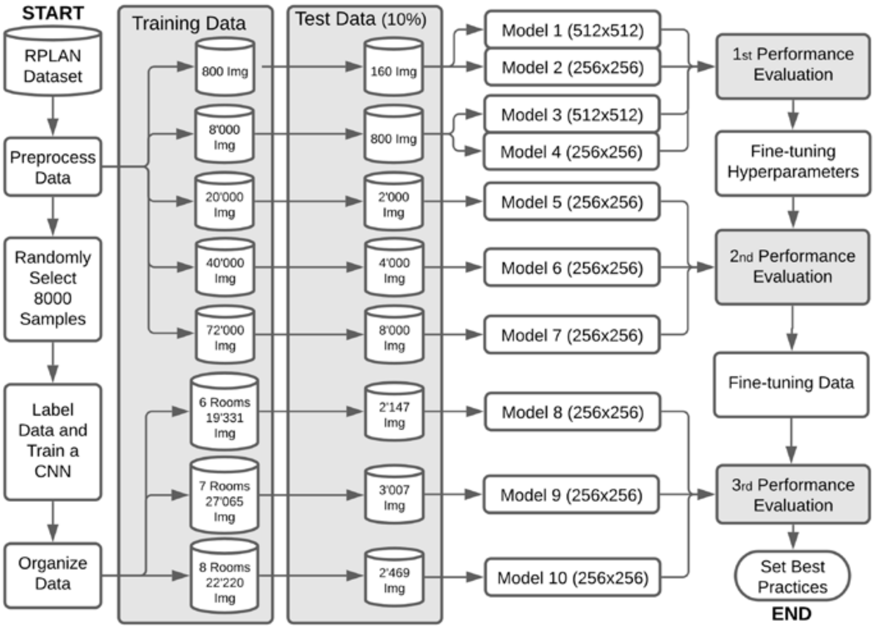

Figure 1. Methodology Diagram. Source: Authors

The models were trained on a single GPU RTX 2060 and used the PyTorch implementation of Pix2Pix (Zhu et al., 2017), with the RPLAN dataset, which consists of approximately 80 thousand multi-channel images of floor plans with rooms labeled by semantic segmentation. We also opted to normalize the room labels with equidistant colors in the RGB spectrum, shown in Figure 2, similar to previous literature experiments. This is because algorithms turn pixel color into information to train the neural network, so keeping mismatched colors on labels helps the human and the machine visualization.

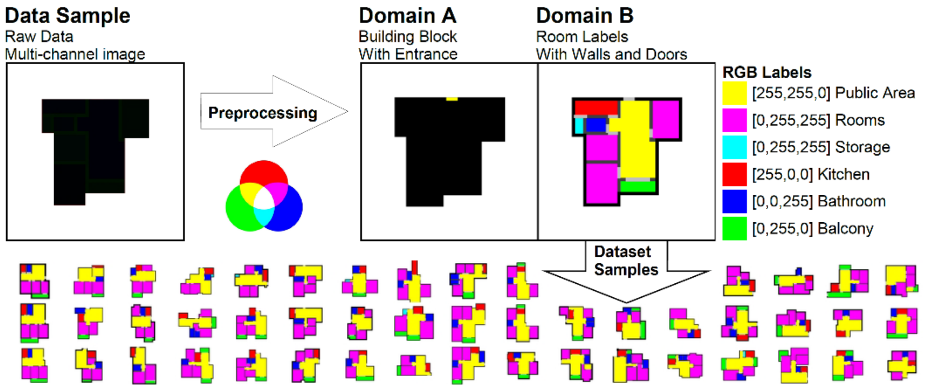

Figure 2. Data Normalization and Labels. Source: Authors

Inspired by Nuha & Afiahayati (2018) approach on trying to reduce dataset sizes, we also divided the RPLAN dataset into similar proportions which were 1%, 10%, 25%, 50%, and 100% of the available data. Similarly, previous experiments of Chaillou (2020), Zheng et al. (2020) and Zheng, (2018) had approximately 800 images, which is 1% of the RPLAN dataset, which makes the comparisons scalable.

According to Nuha & Afiahayati (2018), even if we have high dimensional datasets to train models, their experiments show that by reducing it to the mid-level (around 50,000 images) we can produce competitive, if not better results than the high-level dataset (around 200,000 images). This acknowledgment made us raise two hypotheses:1) Pix2pix models also achieve a point, which more data does not help with the generalization task. 2) If data is structured by the number of rooms before the training phase, results tend to have a faster conversion and return more reliable results.

To test these hypotheses, we ran an algorithm to randomly select 8 thousand images, and manually labeled them into classes of 4, 5, 6, 7 and 8 rooms per floor plan. Then used the ResNet50V Convolutional Neural Network (CNN) to apply the transfer-learning technique, which could further organize the dataset for the last subset of models in the final evaluation phase, despite the fact that the CNN could only achieve 61% of accuracy on test data, the model could split the dataset into very similar amounts to the ones reported in the original paper of the RPLAN dataset made by Wu et al., (2019).

To ensure the replication of results, the models trained until the second performance evaluation phase followed the order on which the dataset is originally organized, however, we highlight that training data and test data are always different data samples for each model.

The models' assessment was done with the PyTorch implementation of the Frétche Inception Distance – FID (Seitzer, 2020), which is one of the main metrics used to analyze the quality of artificially generated images in contrast to the real images, this means that: Low FID scores represent reliable artificially generated images. However, the information that needs to be analyzed in this study is the semantic map, not in the building block, thus, to evaluate the quality of the generated images, the building boundary input is simply ignored.

Other authors like Ibrahim et al. (2021) also used the same method to evaluate Pix2pix results, however, the main difference from their approach to ours is that we calculated FID not only for the model obtained in the last cycle of the training phase, instead of it, we measured FID for every 5 cycles of training, also known as 'epoch'. This permits us to see if the models' performance increase or decrease over time.

Table 2. Models' Hyperparameters

Model	Pixels	Dataset Size	Total Epochs	Lr. Decay	Generator Network	Batch Size	~Total Training Time
M-1	512^2	800	200	100	unet_256	1	7 hours
M-2	256^2	800	400	200	unet_128	1	6 hours
M-3	512^2	8.000	200	100	unet_256	1	61 hours
M-4	256^2	8.000	400	200	unet_128	1	55 hours
M-5	256^2	20.000	100	No Dropout	unet_256	1	42 hours
M-6	256^2	40.000	100	No Dropout	unet_128	4	46 hours
M-7	256^2	72.000	100	No Dropout	unet_128	6	100hours
M-8	256^2	19.031	100	No Dropout	unet_256	1	43 hours
M-9	256^2	27.065	100	No Dropout	unet_256	1	58 hours
M-10	256^2	22.220	100	No Dropout	unet_256	1	45 hours

Source: Authors

Another method used to measure the models' performance during training was the observation of the generator and discriminator loss, which in all cases showed the same behavior, except on the ones with no dropout, which had no significant decrease in the discriminator loss over time, obviously due the no dropout function.

It is important to highlight that Pix2pix was designed for the image-to-image translation task, and it also has two usage functionalities; test and evaluation.

During the test phase, it uses paired images in order to synthesize a new design option given the building boundary and the semantic map. On evaluation, only the building boundary is required, however, even if it uses only one input image, the output will be the same as if it was on paired images. We highlight that using existent data to generate novel data samples expands the idea of solution space. As Rahbar et al. (2019) explain; there is no exact final solution for a design task and there are multiple possible solutions.

4 Results

On the first group of models, it can be observed in Figure 3, that the models trained on 800 samples had better FID scores when trained with images of 512x512 pixels. However, increasing the dataset size up to 8.000 samples did not improve our metrics, as we can see in M3 the best FID remained on the 5th epoch, controversially, M4 showed that smaller resolutions with a quite reasonable dataset size kept stable results during the whole training.

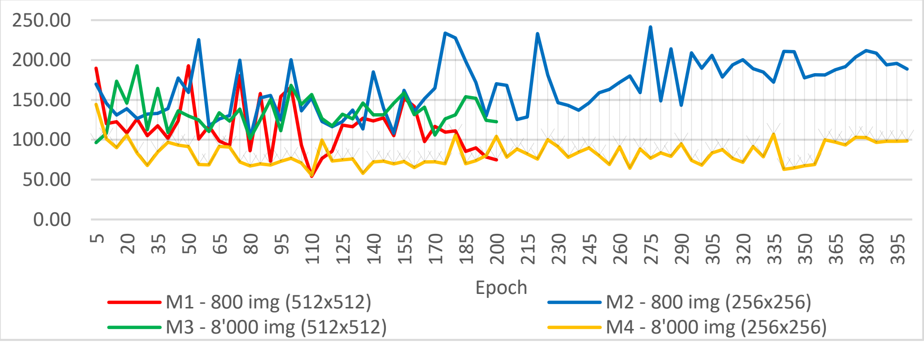

Figure 3. FID Results of the First Evaluation Phase. Source: Authors

As a consequence of the previous results, especially on M3, which had the worst performance, it would not make sense to increase the number of samples in the dataset of 512x512 pixels. Thus, in the next training stage, we kept the 256x256 pixels on the whole dataset, but this time tweaking the generator networks and the dataset dimensionality. The FID results in Figure 4 show that using 20.000 images to train the neural network achieved better results than the models with more data.

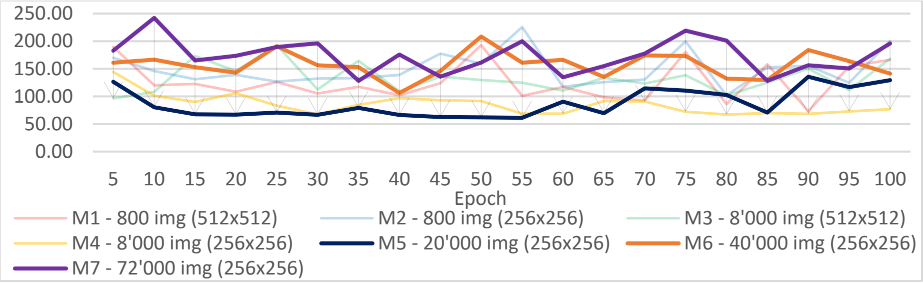

Figure 4. FID Results of the Second Evaluation Phase. Source: Authors

By analyzing the image outputs and the FID over time we assumed that the models trained on high dimensional datasets had so much diversity that the models could not converge to good results, in addition, we also observed on M6 and M7 that increasing batch size and downsampling the input layer on the generator network did not enhance the models' performance, this was done to reduce the training time and dataset noise. At this point of the evaluation phases, the best and most table FID scores were achieved on M5.

On the last evaluation phase, shown in Figure 5, we trained all models with hyperparameters of M5, and also used similar data dimensionalities. Thus, the classes with 4 and 5 rooms were ignored due to the lack of data.

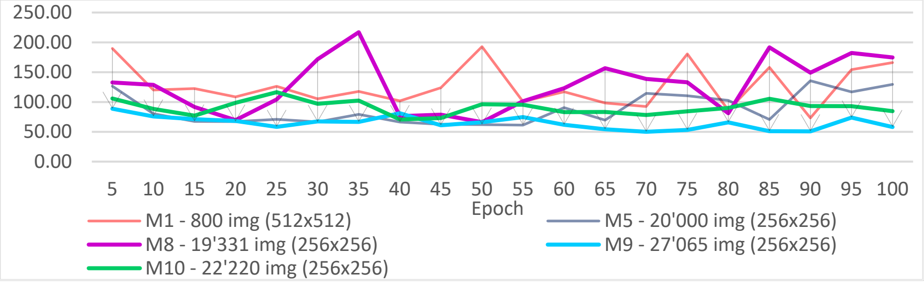

Figure 5. FID Results of the Last Evaluation Phase. Source: Authors

Interestingly, using the CNN as a type of data filter did not make FID get much lower, but if we compare M8 with M10, we can assume that M10 had a higher data complexity due to the number of rooms and even though had good and stable results, this evidences that the generation of floor plans with the Pix2Pix framework will not have their results much affected if the dataset has samples with 1 or 2 more rooms. However, the technique used to reduce the dataset noise showed to be effective on M9 which had optimal results.

In Figure 6 we show random image samples of each model, they demonstrate the machine's ability in using different design strategies to solve the space allocation problem like flipping, rotating and re-arranging the rooms.

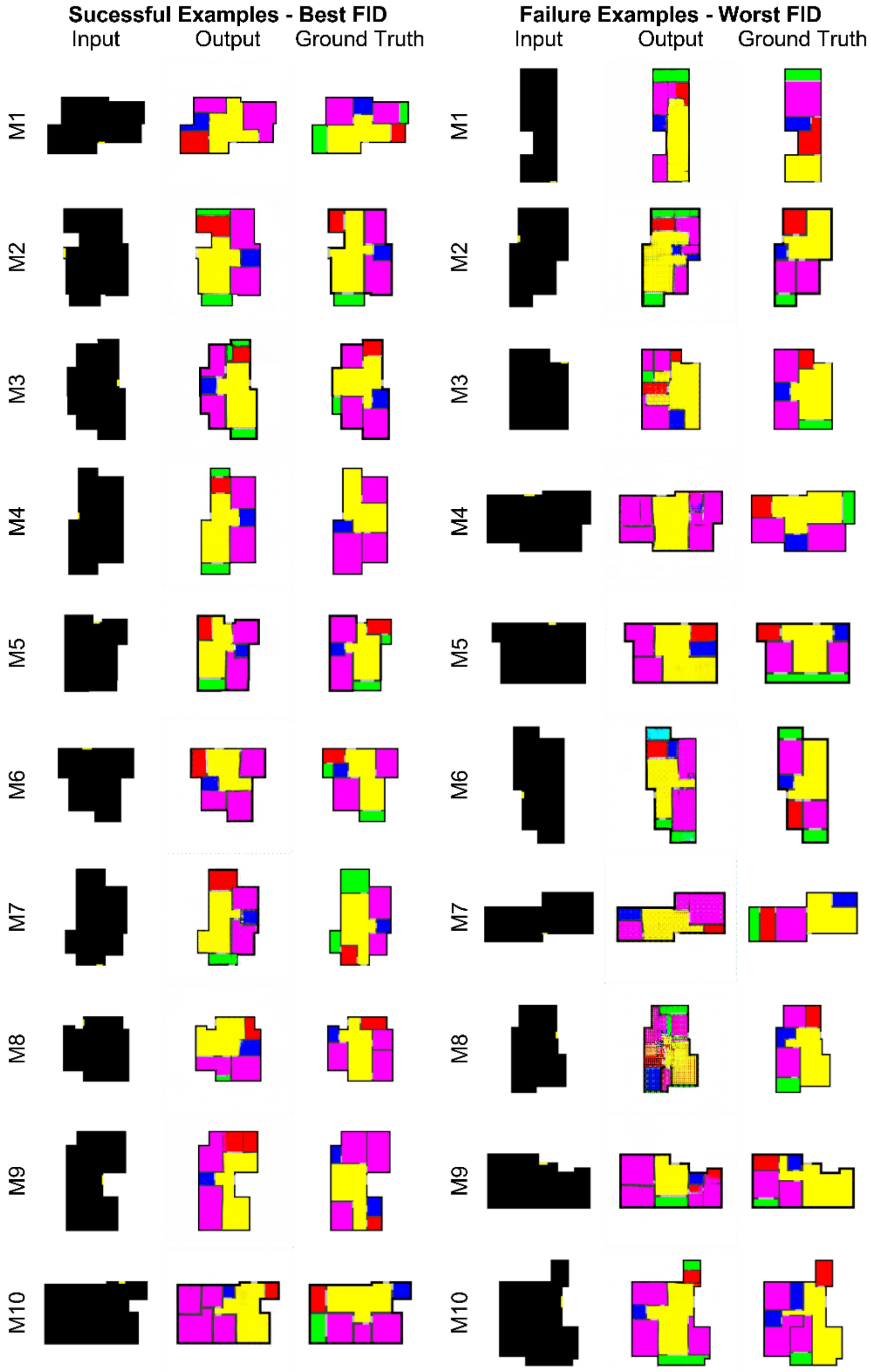

Figure 6. Sample Images of Test Results. Source: Authors

On the worst FID results, images tend to get noisy, or only show schematical diagrams of the space allocation, even though some of them could still re-arrange the space differently from the ground truth. We also observed that the trained neural networks had a difficulty in generating images with few rooms, this is definitely due to the lack of equilibrium on the dataset.

In few cases the output and ground truth were the same, this can be understood as if the model overfitted, or that output and ground truth are both the most likely distribution. Even at the best FID results, at least one room kept the original allocation, especially the public area which happened to repeat most of the times due to the entrance door in the boundary input. Sometimes the models also removed essential rooms like bathroom and kitchen, but also added such rooms in cases which the ground truth did not have them.

5 Conclusion

This research generated and compared almost 6 million images, which computed 463 hours (19 days) just for training models. From the whole process, we can assume that datasets with around 800 images of 512x512 can produce similar or even better results in comparison to high dimensional datasets with lower image resolutions, the advantage of this approach is achieving similar results 5 times faster and with a lot less data. However, this approach does not guarantee stable results during the training phase, thus the best training baseline for Pix2Pix on semantic segmented floor plans is to use around 20.000~30.000 samples to achieve good and stable results, while still maintaining data diversity and a fair assessment due to the number of images.

We also concluded that our first hypothesis is true: For this type of data, scaling dimensionality to around 70.000 images did not make the model achieve better results. Nevertheless, our second hypothesis is false: Structuring the dataset by number of rooms did not have much effect on the models' performance, and sometimes not even divided the floor plan with the right number of rooms, however using a CNN to reduce the dataset noise still had small effects on the results.

The experiments in this study try to enhance the results of the Pix2pix framework on semantic-segmented floor plans. Such images may be useful in the early stages of the design process as they map the spatial distribution through a statistical perspective. However further research is required to investigate the results achieved on different generator networks and combinations of data-augmentation, as well as, novel ways to enhance data-structure and balance the dataset noise before the training phase.

References

Belém, C., Santos, L., & Leitão, A. (2019). On the impact of machine learning. Architecture without architects? *Proceedings of The18th International Conference, CAAD Futures 2019*, *March*.

Chaillou, S. (2020). ArchiGAN: Artificial Intelligence x Architecture. In *Architectural Intelligence* (pp. 117–127). Springer Singapore. https://doi.org/10.1007/978-981-15-6568-7_8

Chen, Q., Wu, Q., Tang, R., Wang, Y., Wang, S., & Tan, M. (2020). Intelligent Home 3D: Automatic 3D-House Design from Linguistic Descriptions Only. *ArXiv*. http://arxiv.org/abs/2003.00397

Galanos, T. (2021). *The Hitchhiker's Guide to Artificial Intelligence: What is AI?* DigitalFUTURES International Doctoral Program of Tongji University. https://www.youtube.com/watch?v=7B9OjyA9uq4&t=3066s

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Hong, T., Wang, Z., Luo, X., & Zhang, W. (2020). State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings*, *212*, 109831. https://doi.org/10.1016/j.enbuild.2020.109831

Hu, R., Huang, Z., Tang, Y., Van Kaick, O., Zhang, H., & Huang, H. (2020). Graph2Plan. *ACM Transactions on Graphics*, *39*(4), 1–14. https://doi.org/10.1145/3386569.3392391

Huang, W., & Zheng, H. (2018). Architectural drawings recognition and generation through machine learning. *Recalibration on Imprecision and Infidelity - Proceedings of the 38th Annual Conference of the Association for Computer Aided Design in Architecture, ACADIA 2018*, *October*, 156–165.

Ibrahim, M. R., Haworth, J., & Christie, N. (2021). *Re-designing cities with conditional adversarial networks*. http://arxiv.org/abs/2104.04013

Kalervo, A., Ylioinas, J., Häikiö, M., Karhu, A., & Kannala, J. (2019). CubiCasa5K: A Dataset and an Improved Multi-task Model for Floorplan Image Analysis. In M. Felsberg, P. Forssén, I. Sintorn, & J. Unger (Eds.), *Image Analysis. SCIA 2019.* (Vol. 11482, pp. 28–40). Springer. https://doi.org/10.1007/978-3-030-20205-7_3

LIFULL, C. L. (2016). LIFULL HOME'S High Resolution Floor Plan Image Data. In *Informatics Research Data Repository*. National Institute of informatics. https://doi.org/10.32130/idr.6.2

Lobos, D., & Donath, D. (2010). The problem of space layout in architecture: A survey and reflections. *Arquitetura Revista*, *6*(2), 136–161. https://doi.org/10.4013/arq.2010.62.05

Mahankali, R., Johnson, B. R., & Anderson, A. T. (2018). Deep learning in design workflows: The elusive design pixel. *International Journal of Architectural Computing*, *16*(4), 328–340. https://doi.org/10.1177/1478077118800888

Nauata, N., Hosseini, S., Chang, K.-H., Chu, H., Cheng, C.-Y., & Furukawa, Y. (2021). *House-GAN++: Generative Adversarial Layout Refinement Networks*. http://arxiv.org/abs/2103.02574

Newton, D. (2019). Deep Generative Learning for the Generation and Analysis of Architectural Plans with Small Datasets. *Blucher Design Proceedings*, *2*, 21–28. https://doi.org/10.5151/proceedings-ecaadesigradi2019_135

Nuha, F. U., & Afiahayati. (2018). Training dataset reduction on generative adversarial network. *Procedia Computer Science*, *144*, 133–139. https://doi.org/10.1016/j.procs.2018.10.513

Peters, N. (2018). *Enabling Alternative Architectures: Collaborative Frameworks for Participatory Design*. Master Thesis, Harvard University Graduate School of Design.

Rahbar, M., Mahdavinejad, M., Bemanian, M., Davaie Markazi, A. H., & Hovestadt, L. (2019). Generating Synthetic Space Allocation Probability Layouts Based on Trained Conditional-GANs. *Applied Artificial Intelligence*, *33*(8), 689–705. https://doi.org/10.1080/08839514.2019.1592919

Seitzer, M. (2020). *pytorch-fid: FID Score for PyTorch* (0.2.0). GitHub. https://github.com/mseitzer/pytorch-fid

Uzun, C., Çolakoğlu, M. B., & İnceoğlu, A. (2020). GAN as a generative architectural plan layout tool : A case study for training DCGAN with Palladian Plans , and evaluation of DCGAN outputs. *ITU A|Z •*, *X*(X). https://doi.org/10.5505/itujfa.2020.54037

Wu, W., Fu, X. M., Tang, R., Wang, Y., Qi, Y. H., & Liu, L. (2019). Data-driven interior plan generation for residential buildings. *ACM Transactions on Graphics*, *38*(6). https://doi.org/10.1145/3355089.3356556

Zhao, C., Yang, J., Xiong, W., & Li, J. (2021). Two Generative Design Methods of Hospital Operating Department Layouts Based on Healthcare Systematic Layout Planning and Generative Adversarial Network. *Journal of Shanghai Jiaotong University (Science)*, *26*(1), 103–115. https://doi.org/10.1007/s12204-021-2265-9

Zheng, H. (2018). Drawing with Bots: Human-computer Collaborative Drawing Experiments. *Learning, Prototyping and Adapting, Short Paper Proceedings of the 23rd International Conference on Computer-Aided Architectural Design Research in Asia (CAADRIA)*, *May*, 127–132.

Zheng, H., Nn, K., Wei, J., & Ren, Y. (2020). Apartment Floor Plans Generation via Generative Adversarial Networks. *25th International Conference Ofthe Association for Computer-Aided Architectural Design Research in Asia (CAADRIA)*, *25*(15), 10.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 2242–2251. https://doi.org/10.1109/ICCV.2017.244