SKETCH WITH ARTIFICIAL INTELLIGENCE (AI)

A Multimodal AI Approach for Conceptual Design

YIFAN ZHOU[1] and HYOUNG-JUNE PARK[2]
[1,2]*School of Architecture, University of Hawaii at Manoa*
[1,2]*{yifanz|hjpark}@hawaii.edu*

Abstract. The goal of the research is to investigate an AI approach to assist architects with multimodal inputs (sketches and textual information) for conceptual design. With different textual inputs, the AI approach generates the architectural stylistic variations of a user's initial sketch input as a design inspiration. A novel machine learning approach for the multimodal input system is introduced and compared to other approaches. The machine learning approach is performed through procedural training with the content curation of training data in order to control the fidelity of generated designs from the input and to manage their diversity. In this paper, the framework of the proposed AI approach is explained. Furthermore, the implementation of its prototype is demonstrated with various examples.

Keywords. Artificial Intelligence; Stylistic Variations; Multimodal Input; Content Curation; Procedural Training.

1. Introduction

Various media have been employed in the process of making conceptual design developments: sketches, drawings, physical or digital models, textual descriptions, and so on. The importance of sketches to delineate a design concept or its description has been highly appreciated by masters in architecture (Graves, 1977, Schildt, 1989, Moore, 2000). Furthermore, diversity of the media involved in the conceptual design augments the creativity and productivity of a designer. Sketches and textual descriptions have been instrumental for applying the implicit (tacit) and explicit knowledge in the conceptual design.

Based upon the premise that Artificial Intelligence (AI) is able to recognize qualitative patterns in data, various AI applications in design exploration have been developed for generating the design variations and suggesting the alternatives to meet given design goals with enhancing the adaptability, productivity, and quality of given input. The common processes of utilizing the AI applications are 1) data collection, 2) establishing artificial neural network (ANN) for machine learning, and 3) performing the training processes through the networks. As an ANN with multiple layers of generation and discrimination in deep machine learning process, Generative Adversarial Network (GAN) is developed for discerning

PROJECTIONS, Proceedings of the 26th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2021, Volume 1, 201-210.

subtle patterns in the given dataset. This pattern recognition extends the application of AI to the generation of 2-D and 3-D design alternatives (Russell and Norvig, 2002, Goodfellow et al., 2016).

Current AI-based design applications show their potential to learn from the dataset and extrapolate its learning into the creation of new alternatives in many design-related disciplines. The applications in architecture have been developed for the generation of 2-D images in plan and facade, and for their stylistic variations. With Pix2pixHD, a modified version of GAN, the neural network is trained to recognize the architecture plan drawings and generate the architecture plan based on the input color pattern diagram (Huang and Zheng, 2018). Chaillou further applies Pix2pix with supervised learning of selected architectural precedents to generate the variations of the floor plans of the precedents according to their program changes (Chaillou, 2019). Also, four stylistic variations (Baroque, Manhattan Unit, Suburban Victorian, Row-House) of the plans have been generated. The stylistic generation of a specific architect's residential plan and facade is performed using GAN (Newton, 2019). With a deep convolutional neural network (DCNN) model, the classification of various design projects is performed according to different architects (Yoshimura et al., 2019). Furthermore, the application of AI in design has been employed for augmenting a user's imagination by generating semi-abstract 'visual prompts' as emergent shapes for a human designer to use them as a starting point of design development (Schmitt and Weiß, 2018). Also, Google quickdraw research showcases the ability of artificial intelligence to automatically recognize various sketch inputs by the users and generate coherent sketch drawings (Ha and Eck, 2017) and could offer drawing refinement suggestions through the auto draw platform (Motzenbecker and Phillips, 2017).

In this paper, sketches and textual descriptions in conceptual design process are employed as multimodal inputs for generating AI's reproduced visions through a proposed procedural training as the visual prompts for a user's design inspiration. 70,000 architectural images are collected from the Internet, and 10,000 images from the collection are labeled according to architect, building category, and so on. Pix2Pix with a U-net structure (Isola et al., 2017), a conditional generative adversarial network (cGAN), is combined with procedural training for generating the variations of its predicted image. The stylistic variations have experimented with 1) sketch & single textual attribute, and 2) sketch & multiple textual attributes. The outcomes of the proposed training model are also compared to two existing GAN models in terms of the image quality and the versatility in the stylistic variations (Isola et al., 2017, Zhao et al., 2019) within the given dataset.

2. Multimodal Interactive AI Application

The proposed multimodal AI approach is composed of three parts: 1) Dataset, 2) Content Curation + Procedural Training, and 3) Sketch & Textual Input + Predicted Image Output.

2.1. DATASET

Different from the traditional Von Neumann computer system, machine learning is a science of training and learning models based on data (Samuel, 1959). The quantity and quality of the initial dataset are critical for the reliability of the outcomes from machine learning. A small dataset may lead to overfitting problems and a poor-quality dataset leads to poor outcomes.

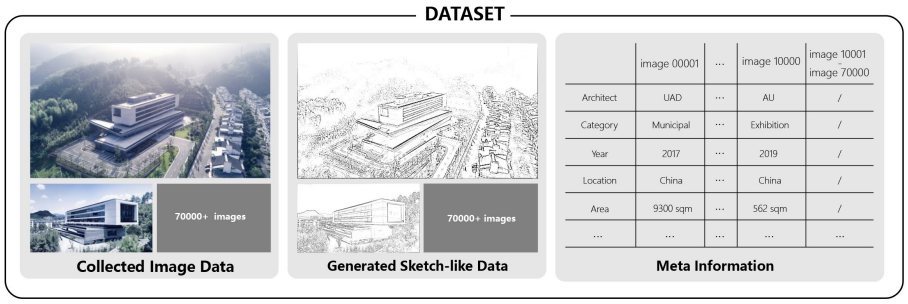

Figure 1. Dataset diagram.

In this paper, the pair of image datasets are prepared. One is the dataset of the images collected from the websites of various architect offices (Foster + Partners, GMP, Perkins and Will, Zaha Hadid Architects) by a python scripted web crawling system. The total number of images reaches 70,000. The other is the sketch-like line drawings paired with the collected images. In order to mimic the sketch input of an architect, the paired sketch-like line drawings are generated using XDog: Extended Difference-of-Gaussians (Sangkloy et al., 2017, Winnemöller et al., 2012). Though generated images using XDog are somehow different from what human sketches are, they still represent similarity to human sketches which often focus on the silhouette and outline of the objects. On the premise that the XDog images could somehow represent human sketches, at this moment, this research uses XDog images to represent human sketches. Human sketch uniqueness like distortion, deformation, and personal style of the sketch will be studied based upon the extended collection of human sketches in the future.

70,000 pairs of images are prepared for updating the parameters of the generator and discriminator at phase one training procedure of the proposed approach. 10,000 out of 70,000 initial pairs are labeled with textual attributes such as architect, building type, project year, project size, project location, and filtered through phase two training procedure. The dataset is randomly split into a 99% portion for training and a 1% portion for validation.

2.2. CONTENT CURATION + PROCEDURAL TRAINING

Our proposed procedural training is performed in a unique dataset pipeline with content curation. The proposed training is divided into two procedures. Both training procedures are performed using Pix2Pix with U-net structure. Phase one training procedure is to make sure the generator and discriminator get a global

understanding of the translation from sketch-like data to colored image. The phase one training takes all the images into the Generator and Discriminator and does the backpropagate(Rumelhart, Hinton et al. 1986). The parameters of the generator and discriminator like weights and biases are updated according to the loss calculated on the training dataset. After 20 epochs of training, parameters are restored. Phase two training procedure is to empower the generators with stylistic variations presented in Section 3. During the phase two training, a dataset filtering algorithm in the dataset pipeline shown in Figure 2 is introduced to curate the data input of the neural network. The dataset filtering algorithm divides the dataset into several categories according to textual attributes. Replicas of the restored parameters of the phase one training are prepared for each curated category. Parameters in each replica are only updated based on the corresponding curated dataset. After 100 epochs of training, parameters of each replica are restored. Currently, the curation of the dataset takes five different attributes (architect, building category, project size, project year and project location) into consideration. The combination of the attributes is scalable such as (architect), (architect + building category) and so on. Therefore, with the extension of the dataset and attributes, more complex textual training is possible.

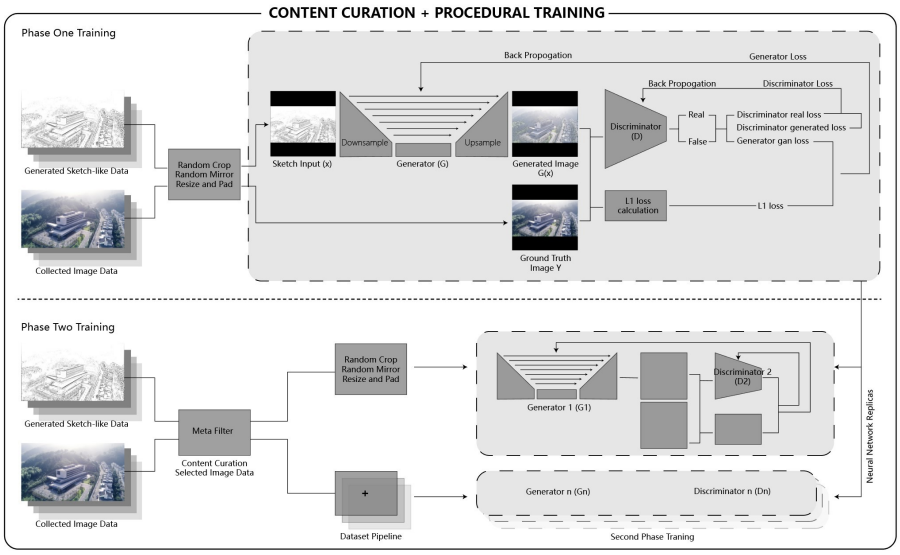

Figure 2. Content curation and procedural training.

Both collected image data and its sketch-like line drawing data are standardized as a [256,256,3] tensor for the input of the generator network. Through the down sample process and up sample process with skip connections, the generator G could output a [256,256,3] tensor as a predicted colored image G(s). The discriminator D takes the generated image G(s) or the ground truth image as the input [256,256,3] tensor and makes true or false judgment through a series of down sample processes. Through the adversarial training process, the generator tries to

fool the discriminator and the discriminator tries to tell the predicted image from the ground truth image. The phase one training procedure uses the recommended learning rate 2*10-4 and the phase two training procedure does a lower learning rate to ensure the stability of the general GAN training process. The updated parameters (weights and biases) through the backward propagation at the phase one training procedure will be recorded as the starting parameter for the phase two training procedures. And the parameters of the proposed approach are restored separately according to the curated data in the phase two training procedure. Loss metrics and inference results in both training dataset and validation dataset will be monitored during the training process to make sure if the training is functional.

2.3. SKETCH & TEXTUAL INPUT + PREDICTED IMAGE OUTPUT

The multimodal inputs (sketch and text) in the conceptual design process are available for generating stylistic variations of a predicted input image with Pix2Pix (Isola et al., 2017) with U-net structure through the content curation and two-step procedural training. Currently, contour-like line drawings or freehand sketches are used as a sketch input. Textual inputs are defined by the content curation of the 10,000 images with labeling. The supported attributes of the curation include architect, building category, and the combination of the architect and building category. Different textual attributes will lead the sketch translation into various results. The results are shown in section 3 Stylistic Variations.

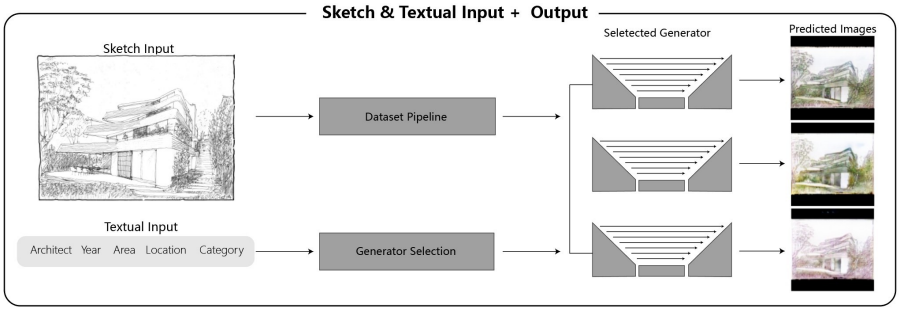

Figure 3. Sketch and textural input and output.

3. Stylistic Variations

The stylistic variations of the input sketch are generated with 1) single attribute of the textual input and 2) multiple attributes. The first experimentation of the single attribute of the textual was performed with the “Architect” attribute. The second was with “Building Type.” “Architect” and “Building Type” attributes are combined for the experimentation of the multiple attributes. Also, the different detail levels of the same input sketch are tested for the stylistic variations. The outcomes of the proposed training model within the given dataset are compared to two existing GAN models as a benchmark.

3.1. SINGLE ATTRIBUTE: ARCHITECT

In this single attribute stylistic variation experimentation, the authors take four representative architects (architecture offices) for instance, which are Foster + Partners, GMP, Perkins and Will, and Zaha Hadid Architects. The generator of the 4 attributes is trained on the 70000 images during the first phase of the procedural training process. For the second phase training, each of them is trained with selected data with the corresponding architect attribute information. As shown in Figure 4, different architect attributes lead to the various predicted styles accordingly. Visual prompts from different color inclinations, material combination preferences, and lighting effects could be observed in the following results. Multiple predicted results provide the user with quick conceptual design direction references.

Figure 4. Single attribute: architect.

3.2. SINGLE ATTRIBUTE: BUILDING TYPE

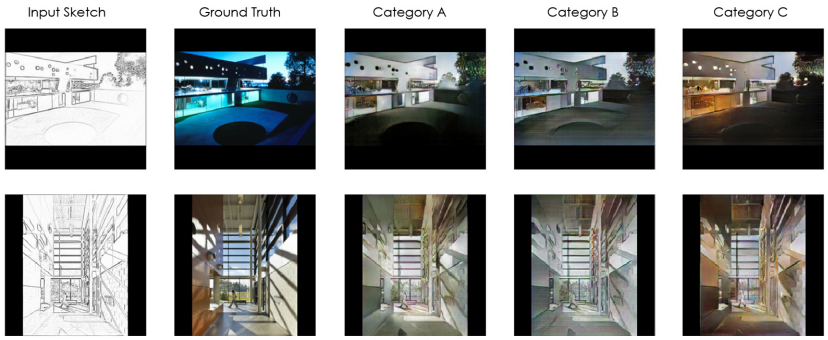

Figure 5. Single attribute: building type.

At the early stage of architectural design development, the comparative approaches of building types also provide an innovative starting point for testing different

architectural nuance. Manipulating the building type information could lead the design into different directions. Three typical categories are selected for highlighting the proposed approach's potential in predicting the colored image with the textual input. In figure 5, category A represents the house interiors and apartment interiors. Category B represents renovation projects. Category C represents hotels and restaurants. The emergent visual prompts from the changes could be also observed from the selected results.

3.3. MULTIPLE ATTRIBUTES: ARCHITECT + BUILDING TYPE

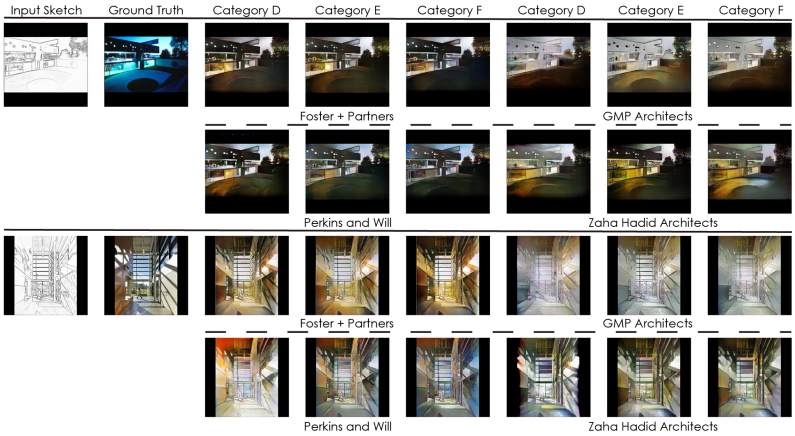

Figure 6. Multiple attributes: architect + building type.

The multiple attributes (Architect & Building Type) are performed with four architect choices and three building types. The outcomes show the scalability of multiple attributes that allows a user to build more complex textual input combinations with extended sets of attributes in the future. The versatility of the stylistic variations is demonstrated with category D (housings & hotels), category E (cultural building), and category F (pavilions and others) in Figure 6.

3.4. BENCHMARK

The existing machine learning models such as Pix2pix with U-net structure (Isola et al., 2017) and GAN techniques (Zhao et al., 2019) perform well on the CelebA dataset (Liu et al., 2015), including the images of prototypical figures such as human faces. This research reproduces the network and training strategy employed in the previous researches (Isola et al., 2017) (Zhao et al., 2019) and compares their results with the one from our method. The pix2pix network (Isola et al., 2017) is conditioned on the input image only. Thus, the network itself does not take the textual information as an input. Its predicted image's variation is led by dropout function in the U-net structure. As shown in Figure 7, its image quality is sharp while the variation led by its dropout function does not cover enough range of

design references. The GAN network structure (Zhao et al., 2019) takes the textual input in addition to a conditional image input. Figure 7 shows the versatility of the GAN network in making variations. However, the sharpness quality of its outcomes does not reach the one from the result of the pix2pix network. The proposed approach takes multimodal inputs such as image and textual information. The textual information is used to select the parameters of generator G to empower its stylistic generations. The comparison made in Figure 7 shows the advantage of our method in terms of the quality and versatility in the stylistic variations within a given dataset.

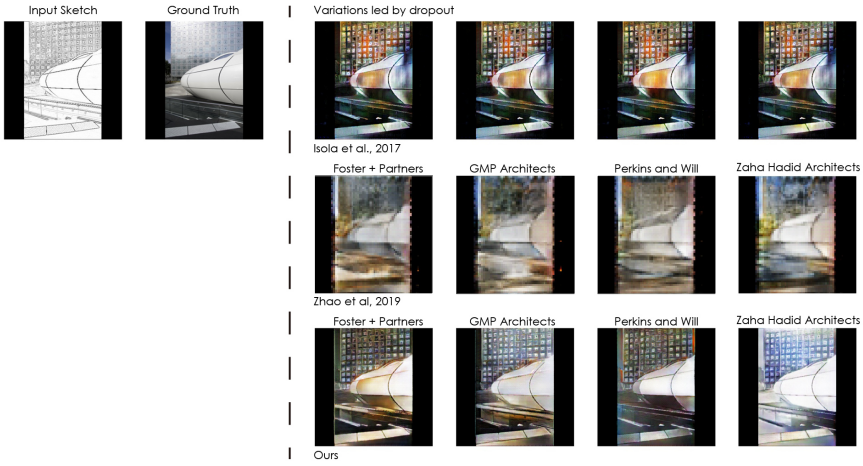

Figure 7. Benchmark with Pix2Pix (Isola et al., 2017) and GAN (Zhao et al., 2019).

3.5. DETAILS OF SKETCH INPUT

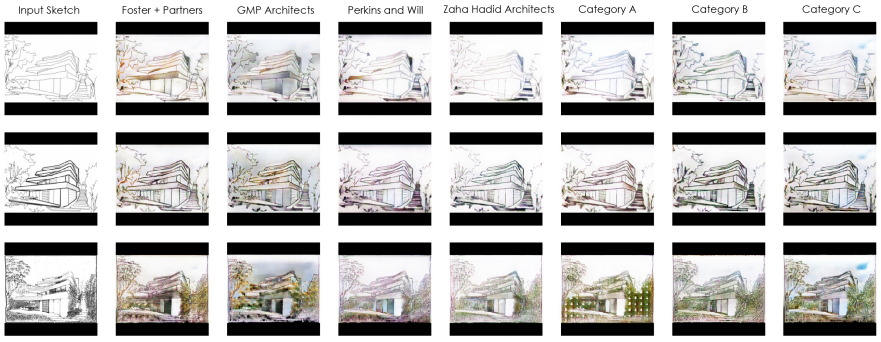

Figure 8. Comparison of sketch detail's influence on the predicted outcomes.

In real practice, architects might not come up with detailed sketches during the conceptual design phase. The different detail levels of sketch inputs are tested for finding the influence of the input details on the predicted images. As shown in Figure 8, in order to fully utilize the generator of the current approach, more detailed sketch input is encouraged for generating convincing images or more human sketches are needed to be collected in the future.

4. Discussion

In the process of making conceptual design developments, design inspiration comes from unexpected design changes as an emergent shape or ambiguous outcomes as a visual prompt (Mothersill and Bove, 2019). This proposed multimodal approach allows a user to engage in the design process with making stylistic variations as its feedback. In terms of conceptual design development, AI augmented visual prompts could act as a quick validation of the conceptual sketch, which could be used in the design discussion within the team or with the client. The versatility in the stylistic variations of the proposed approach could help architects with conceptual inspiration with multiple possible choices suggested by AI when they draw sketches at the early stage of design.

The two phases procedural training models with the content curation employed in this multimodal approach provides a way to manage the quality and diversity of the generated variations within the given dataset. The benchmark results show that the improvements from two existing GAN models in terms of quality and versatility in the stylistic variations have been realized in the proposed approach.

Figure 9. Predicted image based on sketch.

The results on the rough hand drawings in Figure 9 will be further improved by increasing the size of the dataset for the pairs of hand-drawing sketches and colored images in the future. Building up a platform for collecting the hand drawing sketches and colored images pair will be developed further.

References

Chaillou, S.: 2019, *AI + Architecture | Towards a New Approach*, Master's Thesis, Havard Graduate School of Design.

Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y.: 2016, *Deep learning*, MIT press Cambridge.

Ha, D. and Eck, D.: 2017, A Neural Representation of Sketch Drawings, *arXiv preprint*, arXiv:1704.03477.

Huang, W. and Zheng, H.: 2018, Architectural Drawings Recognition and Generation through Machine Learning, *Proceedings of the 38th Annual Conference of the Association for Computer Aided Design in Architecture (ACADIA)*.

Isola, P., Zhu, J.Y., Zhou, T. and Efros, A.A.: 2017, Image-to-image translation with conditional adversarial networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Liu, Z., Luo, P., Wang, X. and Tang, X.: 2015, Deep learning face attributes in the wild, *Proceedings of the IEEE international conference on computer vision*.

Michael, G.: 1977, Necessity for drawing-tangible speculation, *Architectural Design*, **47**, 384-394.

Moore, K.: 2000, Between the Lines: drawing, creativity and design, *Environments by Design*, **3**, 35-58.

Mothersill, P. and Bove, V.M.: 2019, Beyond Average Tools. On the use of 'dumb'computation and purposeful ambiguity to enhance the creative process, *The Design Journal*, **22**, 1147-1161.

Newton, D.: 2019, Generative Deep Learning in Architectural Design, *Technology| Architecture+ Design*, **3**, 2.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J.: 1986, Learning representations by back-propagating errors, *Nature*, **323**, 533-536.

Russell, S. and Norvig, P.: 2013, *Artificial intelligence: a modern approach*, Pearson Education Limited.

Samuel, A.L.: 1959, Some studies in machine learning using the game of checkers, *IBM Journal of research development*, **3**, 210-229.

Sangkloy, P., Lu, J., Fang, C., Yu, F. and Hays, J.: 2017, Scribbler: Controlling deep image synthesis with sketch and color, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Schildt, G.: 1989, *Alvar Aalto, the Mature Year*, Rizzoli, New York.

Schmitt, P. and Weiß, S.: 2018, The Chair Project: A Case-Study for using Generative Machine Learning as Automatism, *32nd Conference on Neural Information Processing Systems (NIPS 2018)*.

Winnemöller, H., Kyprianidis, J.E. and Olsen, S.C.: 2012, XDoG: an extended difference-of-Gaussians compendium including advanced image stylization, *Computers & Graphics*, **36**, 740-753.

Yoshimura, Y., Cai, B., Wang, Z. and Ratti, C.: 2019, Deep learning architect: classification for architectural design through the eye of artificial intelligence, *International Conference on Computers in Urban Planning and Urban Management*.

Zhao, J., Xie, X., Wang, L., Cao, M. and Zhang, M.: 2019, Generating photographic faces from the sketch guided by attribute using GAN, *IEEE Access*, **7**, 23844-23851.