PSEUDO-CROSS-MODAL TRANSLATION

Bridging Architectural Plan and Perspective through a pix2pix Network

YUJUN MAO[1], WENZHE PENG[2] and TAKEHIKO NAGAKURA[3]
[1]*Researcher Computational Design.*
[2,3]*MIT Department of Architecture.*
[1]*arch.yjmao@gmail.com, 0009-0008-8762-3388*
[2]*pwz@alum.mit.edu*
[3]*takehiko@mit.edu*

Abstract. Architectural pedagogy often segments designs into diverse representation forms like plans and renderings. With AI's growing influence on early design through GANs, Midjourney, and Stable Diffusion, there remains a gap in translating between diverse architectural representations, a phenomenon we term 'Pseudo-cross-modal Translation', indicating the indirect transformation between non-analogous architectural representations. Addressing this, our research hypotheses a practical need and actionable possibility to translate architectural plans into perspective renderings via neural networks, exploiting the information differences between them. We navigate this intricate translation utilising a pix2pix network of which the dataset encompasses plans with designated view cones and corresponding rendered perspectives. The training data are sampled from the model of Mies van der Rohe's Barcelona Pavilion and its variations. Evaluations through perceptual surveys, which incorporate modifications in information complexity of plans, illuminate the neural networks' nuanced capability to bridge plans and perspectives under various conditions. Our results not only validate this translation but also spotlight the computational statistics' latent potential in deciphering unseen spatial features from the variance between plans and perspectives. This work unveils a novel method for generating architectural imagery, promoting a holistic spatial understanding.

Keywords. Plan, Perspective, pix2pix, Architectural Representation.

1. Introduction

The comprehension and development of spatial designs through conventional architectural pedagogy, which employs various representational forms like plans and perspectives, can be challenging especially for novices. Research, including Nagakura's experiments (Nagakura and Sung, 2014), indicates the intricacy of constructing a mental spatial image from isolated representations and the necessity of

ACCELERATED DESIGN, Proceedings of the 29th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2024, Volume 1, 179-188.

establishing a back-forth process of understanding between various representations conveying information with distinct characteristics.

Prior to the modern computer-aided design tools, translation between various architectural representations throughout history is articulated either through laying out them in attempts to establish simple spatial associations or by the techniques of perspective to represent three-dimensional objects in two dimensions. Computer graphics technology has enabled instant perspective projection from a three-dimensional geometric model. Modern computer-aided design tools, such as Building Information Modelling (BIM), further provide convenience for the aforementioned issue by integrating different media into a unified symbolic data structure (Coons, 1963). However, the construction of such integration on the one hand is time consuming requiring professional skills and on the other hand, it can potentially compromise design creativity due to its deterministic nature - In actual design thinking, a single design representation can be linked to various possibilities in another representation. There remains a gap in translating between different architectural representations in a non-symbolic, analogous manner which we termed “Pseudo-cross-modal Translation', such that facilitating design creativity. And with the rising impact of AI on early architectural design stage, particularly through Generative Adversarial Networks (GANs) to the more recent Midjourney and Stable Diffusion recognized for their intensive iterations of prompt to image translation, it is possible to integrate various architectural representations into a computational statistical structure.

Figure 1. Examples of Pseudo-cross-modal Translation

In response, our research postulates both a pragmatic necessity and a feasible potential for translating architectural representations through neural networks, exploiting the disparities of information between them. Specifically, this research project focused on transforming architectural plans into perspective renderings using a modified pix2pix neural network, addressing the challenge of 'Pseudo-cross-modal Translation'. (Figure 1). Through training on pairs of notational plans with corresponding perspective renderings, the network not only converts one representation to another, and further provides the possibility to go beyond the rigid

correspondence of symbolic data structure in the current computer-aided design tools allowing flexibility and ambiguity to occur. This research aims to foster creativity catering to a more nuanced spatial understanding through moving from a symbolic to an analogous mode of representation translation and to benefit architects by democratising the design evolution process to various stakeholders.

2. Related Works

Translation between various representations to better depict a three-dimensional world in two-dimensional space has a long history. From pre-renaissance cartography art that associates views from two or three different angles on the same flat surface to the rediscovery of perspective technique, by Filippo Brunelleschi at the beginning of the 15th century, such desire of depiction throughout history is motivated by socio-technical development at different stages. On the one hand, the technique of lacking a unified viewpoint is inherited in laying out different representations to establish simple spatial associations, for instance Palladio's aligned architectural drawings to document Temple of Mars the Avenger (Palladio, 1570) and 'Analytique' drawings in Beaux arts tradition (d'Espouy, 1905). On the other hand, the perspective method that directly reflects human visual experience is later developed into more advanced descriptive geometry, which employed by Ferdinando Galli-Bibiena to depict a stage design for a scena par angolo translating a floor plan to an interior perspective drawing (Vesely, 2004).

After entering the era of computer graphics and with the development of rendering algorithms, it became an easy task to produce photorealistic images using three-dimensional models. However, such translation between representations is time-consuming and the requirement for drafting techniques have changed to the skills of modelling technique and the understanding of rendering software. To respond to this, Takehiko Nagakura's research projects such as 'Space Barcoder' (Nagakura, 1998), 'Digitarama' (Nagakura, 1997), 'Deskrama' (Nagakura and Oishi, 2006) and the more recent 'Ramalytique' (Nagakura and Sung, 2014) utilising contemporary technology explore the translation between various representations through the spatial arrangement approach, which is efficient in synchronisation sense and makes it user-friendly even for novices.

With the advance of artificial intelligence, Generative Adversarial Network (GAN) (Goodfellow et al., 2014) has gained significant attention in the arts and design for its remarkable ability to learn and generate creative products. Specifically, pix2pix as a conditional GAN (cGAN), provides general-purpose solutions to image-to-image translation problems (Isola et al., 2018). Previous works relevant to the research purpose of translation between representations includes floor plan generation (Huang and Zheng, 2018) of which the focus was on the conversion from a land plot to a furnished plan, structural analysis during the iterative design phase (Rossi and Nicholas, 2018) by substituting analytical finite element analysis (FEA) modelling with cGAN predictions, and exploration of cGAN's capacity in generalising different perspective of the urban street views through translation from depth map (Steinfield, 2019).

3. Methodology

In this study, we focus on a significant and impactful issue of Pseudo-cross-modal Translation: transforming plans into perspective renderings. The Pseudo-cross-modal Translation from architectural plans into corresponding perspective renderings is formulated as a task which fine-tunes and tests a customised version of pix2pix neural network using plan-rendering image pairs (Figure 2). Here, we describe the stages of data preparation and neural network training.

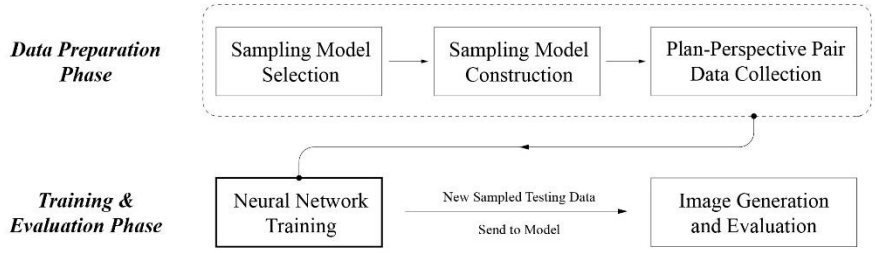

Figure 2. The workflow diagram of Pseudo-cross-modal Translation

In the data preparation stage, we first select Mies van der Rohe's Barcelona Pavilion as the base of data sampling models and create its six variations through a discretized and randomly recombined approach. Then, 1,500 pairs of plan-perspective are sampled from each variation based on a labelling principle. Finally, 10,500 pairs of samples are processed to serve as the training data for a pix2pix neural network.

In the neural network training stage, we employ the collected data pairs to train a pix2pix neural network capable of generating perspective rendering images. The particular implementations of the pix2pix model are discussed in detail below.

3.1. DATA PREPARATION

The majority of the data preparation work required by this project consists of four steps: sampling model selection, sampling model construction, label colour coding, and plan-perspective pair data collection, elaborated below.

3.1.1. Sampling Model Selection

The Barcelona Pavilion designed by Ludwig Mies van der Rohe in 1929 is an important building in the history of modern architecture, known for its continuous space and spectacular use of extravagant materials. It is referenced by Immanuel Koh to illustrate Kengo Kuma's concept about 'particlised architecture' (Kuma, 2012). And according to Koh, the particlised interpretation of Barcelona Pavilion is reinforced by its use of pilotis and plinth, as a formal means to isolate their architecture from the ground, especially when viewed as images (Koh, 2023). Based on this interpretation, in this research, we decide to select Barcelona Pavilion as the basis for the construction of sampling models.

3.1.2. Sampling Model Construction

In order to achieve a good neural network performance, training data with adequate visuospatial quality is the key, thus it is necessary to acquire different variations of Barcelona Pavilion to be sampled for training. The Barcelona Pavilion model is firstly broken down into 18 discrete elements with a 3 by 6 subdivision, which is an alternative reading resonate with the concept of 'particlised architecture' allude to a mechanic form of statistical seeing. The distinction of this manipulation is that figure-and-ground and part-to-whole are completely abandoned. Each decomposition unit is interchangeable without any predefined rules of assemble, where parts could be randomly recombined within the original grid to form a possible whole (Koh, 2023). In this sense, Barcelona Pavilion and its variations conceptually can be regarded as a 'training set' that embodies a hidden range of likely shapes within the same architectural vernacular. We ultimately end up with 6 more variation of the Barcelona Pavilion models. The 7 models, including the Barcelona Pavilion itself, are then used for the plan-perspective pair data collection in later stages (Figure 3).

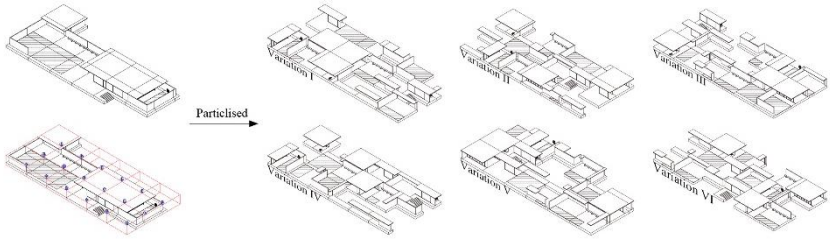

Figure 3. The diagram of sampling model construction

3.1.3. Label Colour Coding

The training of a pix2pix requires pairs of input and output images to learn the translation patterns. And from the perspective of a user of a trained pix2pix network, we offer an input image that conforms to some mapping convention and receive in return a synthetic image that results from the transformation of this input. In such a translation, the key is the correspondence between the input and output elements. Hence, the design of label colour coding to assign various unique RGB colours to each kind of architectural element on the plan becomes particularly important. Figure 4 shows ten colours assigned to represent green alpine marble wall, roman travertine marble wall and floor, transparent glass, grey frosted glass, area covered with carpet, pool with green alpine marble floor, pool covered with cobblestone, roof covered area and steel framework respectively leaving the background with pure white. The selection of colour was based on the principle to differentiate the vertical architectural elements with horizontal ground and canopy as far as possible. Each sampled plan was marked with an isosceles triangle representing the view cone of the camera in yellow of which the middle line represents the viewing direction, and the side lines represent the field of view of the camera in the corresponding perspective view. The view cone is positioned at three-quarters along the central axis of the sampled plan, enabling the

assessment of environmental factors behind the camera, like reflections and shadows.

Figure 4. The diagram of label colour coding

3.1.4. Plan-Perspective Pair Data Collection

Since the objective is to translate architectural plans into their corresponding perspectives, a collection of plans-perspectives pairs is utilised to construct a dataset for training a pix2pix neural network. Our dataset is composed of 10,500 plan-perspective pairs. And the dataset pairs are sampled from Barcelona Pavilion and its particlised variation based on the labelling principle illustrated in previous section.

Our study utilises a Grasshopper program to automatically sample architectural plans. It captures plan elements in a specified location with appropriate range, rotation, and colour-coded labels for elements. Figure 5 illustrates the crucial steps of this process: rotation and cropping. This approach deconstructs the building architecturally, aligning it with the camera's viewpoint for uniform information capture. We ensure the sampled plan rotates to keep the camera facing upward during sampling. Cropping is used to enhance the neural network's efficiency in translation and its capacity for generalising and inferring missing elements, a concept we delve into in the evaluation section. We set the cropping size to 10x10 metres, balancing coverage with generalisation capabilities. For perspective renderings, they are generated in V-Ray, using the camera positioned and oriented as per the plan sampling.

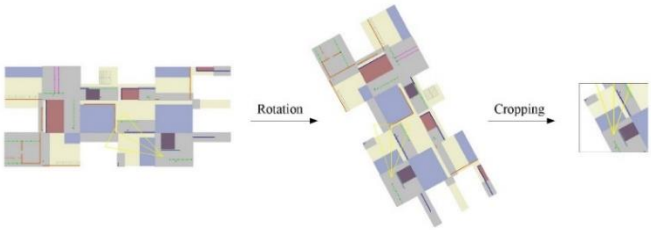

Figure 5. The diagram of rotation and cropping manipulations

3.2. NEURAL NETWORK TRAINING

This project employs the PyTorch version of pix2pix (256px x 256px) architecture. And to accommodate our purpose of experiment, some tunings are applied. We trained the network for 200 epochs with a start learning rate of 0.002 and a batch size of 1. The training, using a T4 GPU on Google Colab with Python 3, takes about 33 hours for a dataset of 10,500 images. Once trained, the neural network, following the pix2pix architecture, generates synthetic RGB images from input cropped-plans. The results and performance evaluation of this network are detailed in the subsequent section.

4. Evaluation of Trained Neural Networks

Three types of studies of the trained neural network are conducted to respectively evaluate its efficacy in translation, its ability to generalise, and its potential for creative outputs. Since the aim of this study is to explore the potential feasibility of leveraging neural networks to achieve an analogous mode of representation translation moving from a symbolic one, a perceptual survey rather than a quantitative method is adopted for evaluation.

A perceptual survey enables the comprehension of human opinions, facilitating the assessment of a crucial aspect of the proposed system: producing results that align closely with human perceptions of space. In this research, six volunteers are invited and divided into two groups - architectural professionals and novices. Each participant takes three questionnaires corresponding to three methods of the trained neural network exploration. Every questionnaire consists of twenty single-choice survey questions; therefore, sixty questions are taken in total. All participants will receive an overview of the research motivation, methodology, and the high-level objectives of each perceptual survey, along with an introduction to the labelling principles before participating in the survey. The following survey question is posed to the participants: "Which of the four perspectives in the right column filling into the blank area in the left best reflects its relationship with the plan above?" Only one of the four images in the right column is generated by the plan as input feeding to the trained neural network while the other three are generated by random input. The labelling principal is placed on the left side of each page for reference.

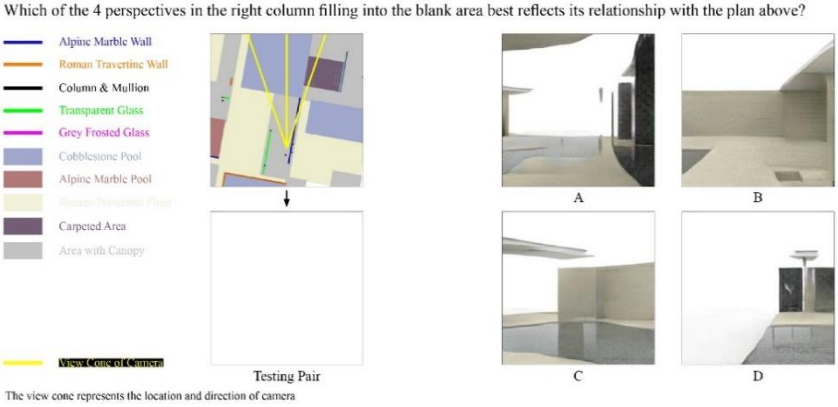

Figure 6. The format of perceptual survey questions

4.1. STUDY 1: TRANSLATION EFFICIENCY OF NETWORKS

The first study focuses on assessing the translation efficiency of a system that converts architectural plans into renderings without direct training on these inputs. Initial results, as shown in Table 1, indicate effective transformation, confirmed by a perceptual survey where both professional and novice groups performed similarly, with most participants only answering 2-3 questions incorrectly out of 20. An example survey (Figure 8) analyses the system's ability to differentiate between interior and exterior perspectives, identifying interior spaces through specific spatial elements like a roof or

view cone placement. Despite some discrepancies, such as the absence of transparent glass or inaccurate carpet depiction, the system effectively translates plans into spatially coherent perspectives, highlighting its proficiency in interpreting architectural information without complete 3D designs.

Survey I	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	Correct	Wrong
Professional.1	√	×	√	√	√	×	√	√	√	√	√	√	√	√	×	√	√	√	√	√	**17**	3
Professional.2	√	×	√	√	√	×	√	√	√	√	√	√	√	√	√	√	√	√	√	√	**18**	2
Professional.3	×	×	×	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	**17**	3
Novice.1	√	√	×	√	×	√	√	×	√	√	√	√	√	√	√	√	×	√	√	√	**16**	4
Novice.2	√	√	√	×	√	×	√	√	√	×	√	√	√	√	√	√	√	√	√	√	**17**	3
Novice.3	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	×	√	√	√	**19**	1

Table 1. The results of the perceptual survey I

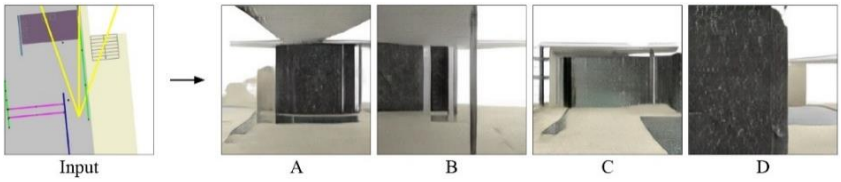

Figure 7. An example of the perceptual survey I

4.2. STUDY 2: GENERALISATION ADAPTABILITY OF NETWORKS

In the second study, plans based on the Barcelona Pavilion were used to assess the network's ability to adapt learned styles to new spatial configurations. The results, detailed in Table 2, revealed that professionals generally outperformed novices, with the latter group making more errors on average. An example survey (Figure 9) involved analysing a plan, where a view cone under a canopy pointed towards a small pool, helping to eliminate certain options. Despite some inconsistencies, such as the absence of grey frosted glass and the relative position of a green wall and pool, the chosen option best represented the spatial configuration, illustrating the neural network's effectiveness in generalising styles to new design contexts.

Survey II	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	Correct	Wrong
Professional.1	√	√	√	×	√	√	√	√	×	√	√	√	√	√	√	√	√	√	√	√	**18**	2
Professional.2	√	√	√	√	√	√	√	√	×	×	√	√	√	√	√	×	√	√	√	√	**17**	3
Professional.3	√	√	√	√	√	√	√	√	×	×	√	√	×	√	√	√	√	√	√	√	**17**	3
Novice.1	√	×	×	√	√	×	√	√	√	√	√	√	×	√	√	√	√	√	×	√	**15**	5
Novice.2	√	√	√	×	√	√	√	×	×	×	√	√	√	√	√	√	√	√	√	√	**16**	4
Novice.3	×	√	√	√	√	√	√	√	×	×	√	√	√	√	√	×	×	√	√	√	**15**	5

Table 2. The results of the perceptual survey II

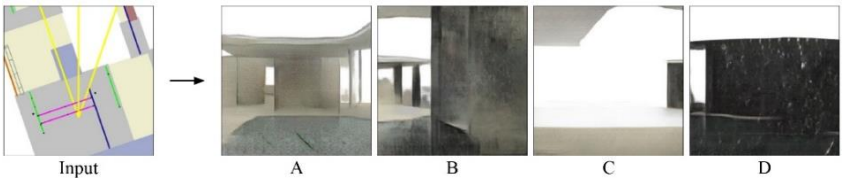

Figure 8. An example of the perceptual survey II

4.3. STUDY 3: CREATIVE FLEXIBILITY OF NETWORKS

The third study examines how trained neural networks handle creative tasks like cropping, colour alterations, and element removal, assessing their ability to infer missing details from the training dataset. Results in Table 3 show that professionals outperform novices in recognizing the network's inference capabilities. A detailed analysis of one survey (Figure 10) reveals swapped labelling of pool colours and marble floor, with a view of water and roof. Despite some confusion, option 'C' is identified as most accurate, showcasing the network's adeptness at managing missing information under complex input variations.

Survey III	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	Correct	Wrong
Professional.1	√	√	√	×	√	×	√	√	×	√	√	√	√	√	×	√	√	√	√	√	**16**	4
Professional.2	√	√	√	√	√	√	√	√	√	√	√	√	√	√	×	√	√	√	√	√	**19**	1
Professional.3	√	√	√	√	√	√	√	√	√	√	√	√	√	×	×	√	√	√	√	√	**18**	2
Novice.1	√	×	√	√	√	×	×	√	√	×	√	√	√	×	√	√	√	√	√	×	**14**	6
Novice.2	√	×	√	×	√	√	√	√	×	√	√	√	×	√	×	×	√	√	√	√	**16**	4
Novice.3	√	√	√	×	√	√	√	√	×	√	√	√	×	√	×	√	√	√	√	√	**16**	4

Table 3. The results of the perceptual survey III

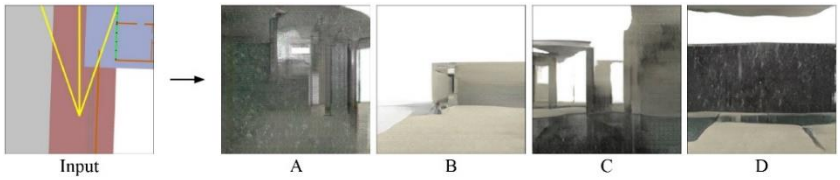

Figure 9. An example of the perceptual survey III

5. Conclusion and Reflection

This project explores the potential of pix2pix neural networks, a form of conditional Generative Adversarial Networks (cGANs), for transforming architectural plans into perspective renderings. The Pseudo-cross-modal Translation system developed in this study simplifies the design process by converting plans into visual perspectives without complex 3D modelling, offering a more accessible and creative approach to architectural design, particularly for novices. Our method stands out from conventional Computer-Aided Design techniques by its ease of use, adaptability to various design layouts, and its potential to enhance creativity, making it an invaluable brainstorming tool that democratises design capabilities and fosters innovative thinking in early-stage design.

The current state of research in generative design, particularly in translating various architectural representations, still faces limitations in terms of efficiency, accuracy, and diversity. The existing pix2pix model struggles with fine details in translation, such as columns and glass, suggesting a need for more robust or custom neural networks, possibly diffusion-based, for enhanced accuracy. Furthermore, the network's limited capability in generalising learned styles highlights the potential of multi-modal neural networks to improve efficiency and diversify outputs. Additionally, expanding the translation capabilities beyond just architectural plans to perspectives, including

reverse translation and other spatial representations, would further streamline and inclusivize the design process.

The progress in AI and its understanding of spatial design significantly impacts not only architecture but also fields like virtual reality, game development, filmmaking, and robotics. This democratises design, allowing greater involvement from amateurs and reducing misunderstandings in design communication. Emerging technologies could shift architects' roles from routine tasks to developing design frameworks.

References

Coons, S. A. (1963, May). An outline of the requirements for a computer-aided design system. In *Proceedings of the May 21-23, 1963, spring joint computer conference* (pp. 299-304).

d'Espouy, H. (1905) *Greek and Roman Architecture in Classic Drawings*. Translated by Henry Hope Reed., 1999. New York: Dover Publications.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. "Generative Adversarial Nets." In *Advances in Neural Information Processing Systems*, 2672–80.

Huang, W., & Zheng, H. (2018). Architectural drawings recognition and generation through machine learning. In P. Anzalone, M. Del Signore, & A. J. Wit (Eds.), *Recalibration: on imprecision and infidelity: Proceedings of the 38th Annual Conference of the Association for Computer Aided Design in Architecture*, ACADIA 2018 (pp. 156-165).

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017) Image-to-image Translation with Conditional Adversarial Networks, 217 *IEEE Conference on Computer Vision and Pattern Recognition* (pp.1125-1134). Available at: https://doi.org/1.119/CVPR.217.632

K. Kuma (2012), *Anti-object: the dissolution and disintegration of architectures*, Reprint., vol. 2. London: Architectural Association Publications.

Koh, I. (2023). Architectural sampling: three possible preconditions for machine learning architectural forms. *Architecture Intelligence*, 2(1), 7.

Nagakura, T. (1997) DIGITARAM. In: Ken Sakamura and Hiroyuki Suzuki, ed. 1997. *The Virtual Architecture*. Tokyo: Tokyo University Digital Museum.

Nagakura, T. (1998) Gushikawa Orchid Center. 17th Exhibition of Winning Architectural Models and Drawing. *SD Review*, December 1998, pp.36-38.

Nagakura, T. and Oishi, J. (2006) Deskrama. In: *Proceedings of ACM SIGGRAPH 1998, Emerging technologies*. Article No. 6. New York: ACM New York.

Nagakura, T., & Sung, W. (2014, November). Ramalytique: Augmented reality in architectural exhibitions. In *Conference on Cultural Heritage and New Technologies 19th Proceedings* (pp. 3-5).

Palladio, A. (1570) *The Four Books on Architecture*. Translated by Robert Tavernor and Richard Schofield., 1997. Cambridge, Massachusetts: MIT Press.

Rossi, G., & Nicholas, P. (2021). Encoded Images: Representational protocols for integrating cGANs in iterative computational design processes. In *Acadia 2020 Distributed Proximities: Proceedings of the 40th Annual Conference of the Association for Computer Aided Design in Architecture* (Vol. 1, pp. 218-227).

Steinfeld, K. (2019, October). Gan Loci. In *Proceedings of 39th Conference of the Association for Computer Aided Design in Architecture: Ubiquity and Autonomy* (pp. 392-403).

Vesely, D. (2004). *Architecture in the age of divided representation: the question of creativity in the shadow of production*. MIT press.