A walk through the latent space using computational aesthetics as a compass

Victor Sardenberg[1,2], Igor Guatelli[2], Mirco Becker[1]

[1] Leibniz Universität Hannover, Hannover, Germany
sardenberg@iat.uni-hannover.de; becker@iat.uni-hannover.de
[2] Universidade Presbiteriana Mackenzie, São Paulo, Brazil
igor.guatelli@mackenzie.br

Abstract. Generative Adversarial Network (GAN) models produce a latent space where many new images emerge. These models translate vectors from a latent space of possible designs into actual images, introducing a new degree of variability to the concept of *objectile*. This research proposes applying a computational aesthetics framework to navigate the latent space and present the designer with new images for feeding their imagination. Theories of parts to whole from aesthetics and cognitive psychology are combined with Birkhoff's aesthetic measure and computer vision to predict aesthetic preferences and map the latent space.

Keywords: Artificial Intelligence, Computational aesthetics, Generative design, Computer vision, Latent space.

1 Introduction

Generative Adversarial Network (GAN) models can produce hundreds of thousands of images rapidly and cheaply. These models utilize a low-dimensional vector value to map n-dimensional points to generate images. Varying this vector slightly and continuously allows an observer to visualize how the models can produce interpolations with valid images that transform gradually. Primarily, this technique produces animations where each point/image is sequentially stored as a frame. Such animations are used to comprehend the possible outputs of these models better.

The architectural discourse is developing ways to discuss and use these image-generation models. Because its output is bi-dimensional bitmaps, many criteria of evaluation that architects are familiar with when working on 3D models are impossible to evaluate (e.g., spatial, structural, and environmental). Therefore, this is an opportunity to engage in a meaningful discussion about architectural images in terms of their aesthetics.

This paper explores the latent space walk of generative image models by evaluating each image individually and quantitatively scoring its aesthetics. It utilizes (1) an adapted formula of Aesthetic Measure from G. D. Birkhoff to evaluate how efficiently each image produces an aesthetic feeling and (2) a predicted hedonic response (PHR) model to predict how each image is liked or disliked by an audience. The scoring allows an understanding of what regions of the latent space produce the most visually appealing outputs and presents them to the designer.

As an experiment, a latent space of architectural pavilions is generated using GAN models and navigated by varying their vectors. Since such a space is n-dimensional, the statistic method t-SNE is applied to flatten it into two dimensions and map images that are similar closer and dissimilar farther. Such a lower-dimensional representation is called a latent space map. An artificial neural network trained on the hedonic response of a specific audience is used to predict how each image is preferred in a metric titled predicted hedonic response (PHR). Such a value arranges all images in the latent space map in a third dimension.

The latent space map is presented to the designer so they can navigate it to visualize how each region generates different designs. The designer can interact with it by zooming in and out to understand each region better. They can also filter the images using the PHR to visualize especially the highest-scoring images, allowing it to work as a compass for the latent space map.

This research aims to enable architects to use a computational aesthetics framework to understand the latent space of generative models better and explore design ideas that are counter-intuitive and unexpected with the assistance of image-generating models. This paper approaches the concept of latent space from two directions: as a data compression representation and as a set of possible designs/images.

1.1.1 Latent space as data compression representation

GAN models compress information to learn relevant information about its data points (Tiu, 2020). These models read vectors (collections of numerical values) from the input images and compress them in the latent space. For example, an RGB bitmap of 512 x 512 pixels is represented as a matrix of 512 x 512 x 3 (each channel of RGB), resulting in 786.432 dimensions. Each possible image with this size is represented by a vector with 786.432 dimensions. When it is compressed to the latent space, it is mapped to a lower dimensional vector. Even though the latent space contains fewer dimensions, the number of dimensions in GAN is much larger than three, making it impossible to imagine. There are methods to visualize it, such as t-SNE, a statistical method to map high-dimensional data in two or three dimensions so that similar objects are clustered (Figure 1) (Maaten & Hinton, 2008).

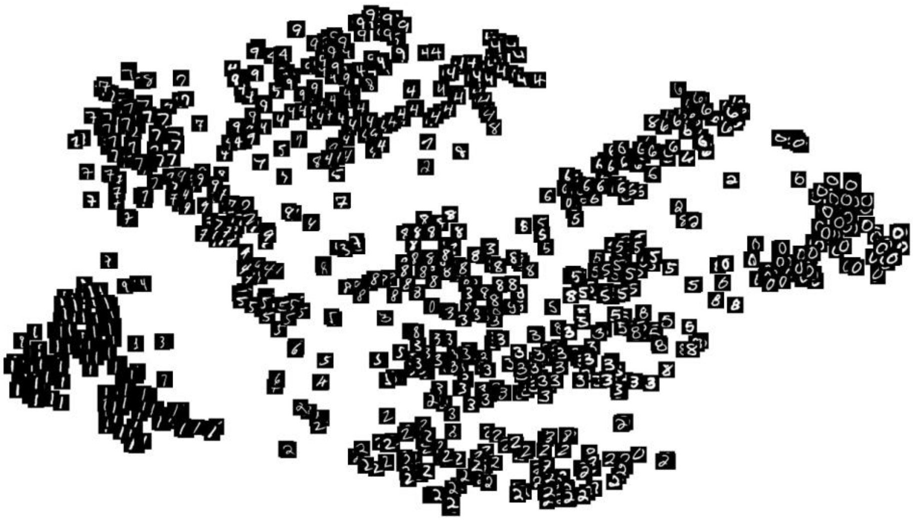

Figure 1. A t-SNE projection of the latent space of a model trained to recognize the numerical value of hand-written numbers. Source: Despois, 2017.

These data compression methods make similar images closer together in the latent space map. When similar images are clustered together, they belong to the same manifold. In data science, a manifold is a group of similar data. Similarity must be understood as objects with certain features expressed in their lower-dimensional vector. When data is reduced to lower dimensions, one can more easily recognize clusters of similarities or manifolds. It is possible to interpolate vectors in the latent space to visualize how they output images that continuously and smoothly transform from one point to another. This interpolation is popularly called latent space walk.

1.2 Latent space as possible designs

The vectors inhabit a GAN model's latent space and are pure virtuality. The latent space encompasses all possible imaginations a model can produce, such as all possible notes in a musical composition. A GAN model starts as pure noise, an undifferentiated field. When presented with a dataset of images, it specializes in creating images that are similar to it. The GAN model actualizes the vector from the latent space and realizes new images, like an instrument translating notes into sound waves. These new images are not fake but virtual possibilities of the latent space. They are extrapolations of all real images that the machine has seen. Finally, they are the machine's hallucinations (del Campo & Leach, 2022).

The latent space is also an *objectile*. The *objectile* is the combination of the words *object* and *projectile* and is a concept demonstrated by Bernard Cache and Gilles Deleuze (Cache, 1995; Deleuze, 1992) of a new status of the object that is not molded or standard. This new status of the object implies continuous variations of form and matter in a temporal modulation. In the 90s, it was associated with the solution space of a parametric model as in all variations of Gramazio Kohler's 2002 mTable (Gramazio Kohler, 2018; van Stralen, 2018). The *objectile* and parametric models deal with the same elements: Parameters, variations, modulations, and associations (Duarte et al., 2017). However, parametric models can only produce extensive variations, lacking the ability to create differences in qualities or intensities. Because the latent space of GANs encompasses images that vary in quality, it is a better example of an object of continuous temporal modulation, continuous folding and unfolding, and caves inside caves. A new degree of variability was introduced from the single standard object to the non-standard parametric solution space. The latent space of GAN models offers a further degree of difference.

According to Deleuze, purely actual objects do not exist: *"Every actual surrounds itself with a cloud of virtual images. This cloud is composed of a series of more or less extensive coexisting circuits, along which the virtual images are distributed, and around which they run"* (Deleuze, 2002). The latent space of GAN continuously renews itself, keeping the output image uncertain and undetermined.

2 Methodology

This chapter describes the (2.1) preparation of the GAN model, (2.2) introduces the background on quantitative aesthetics where this work is situated, (2.3) discusses developments in the computer vision methodology for image analysis, and (2.4) presents how the PHR model was trained.

2.1 GAN model preparation

For producing a latent space, StyleGAN2 was adopted because it is state-of-the-art in unconditional image modeling in both existing distribution quality metrics and perceived image quality (Karras et al., 2020). The model was fed with 3891 images of architectural pavilions that were scraped from Google Images and manually selected to filter only external perspective views. The search terms were: "MoMA PS1 YAP", "Serpentine Gallery Pavilion," "Tallin Architecture Biennale Pavilion," "Architectural Association Pavilion," and "ICD Pavilion." The selection criteria of these search terms were the consistency of building scale and variation in aesthetics. RunwayML cloud computers were used to train the model in 11000 steps, always cropping images at its center. The model scored 60 in FID and is publicly available in RunwayML as 230717_PavilionGAN. The model generated 1000 images (Figure 2) in a sweet spot where they were not too abstract (hard to recognize any architectural concept or element) nor too realistic (presenting finished designs).

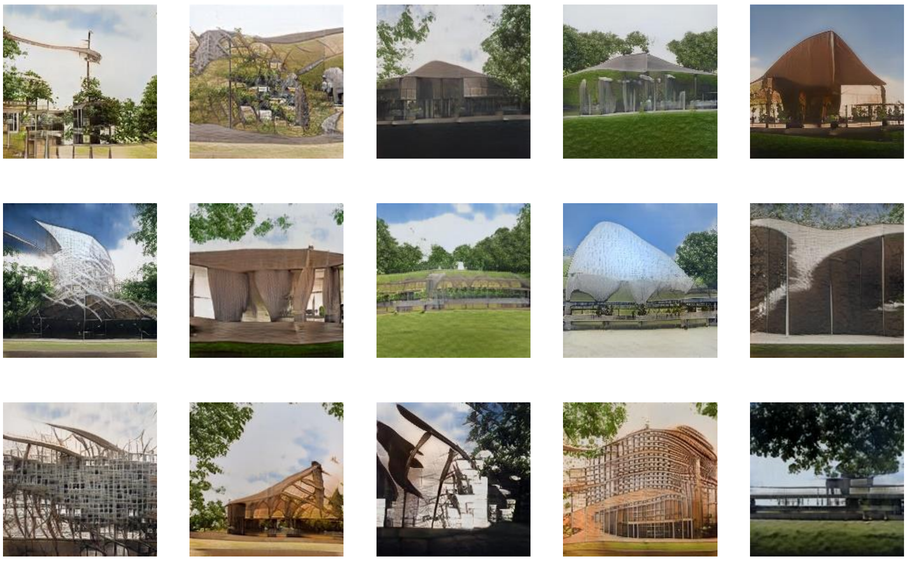

Figure 2. Fifteen images of pavilions that the GAN model generated. Source: Authors, 2023.

2.2 Quantitative and empirical aesthetics

This research is situated in the long tradition of quantitative aesthetics. One can trace its roots in the *kanon*, a lost aesthetics treatise by the Greek sculptor Polykleitos, which is believed to be a *"scientific system of proportion, symmetria, and idealization to create works that he* [Polykleitos] *believed to be the manifestation of beauty and perfection"* (Schuman, 2013). Other precedents of quantitative aesthetics are the painter William Hogarth, who proposed ways of analyzing beauty (Hogarth, 1753), and the philosopher David Hume, who suggested defining standards of taste (Hume, 1757).

In 1860, the field of empirical aesthetics was founded from the work of physician G. T. Fecher, who proposed a formula relating the strength of physical stimuli to its psychological sensation as the following (Fechner, 1965):

$$\gamma = k\,(\log \text{ß}/b) \qquad (1)$$

Where γ is the sensation, k is a constant, and $\text{ß}/b$ is the fundamental value of the stimulus.

In 1933, the mathematician G. D. Birkhoff proposed for the first time a mathematical formula for measuring aesthetic objects in his *"Aesthetic Measure"* book. He aimed *"to bring the basic formal side of art within the purview of the simple mathematical formula defining aesthetic measure"* (Birkhoff, 1933). Birkhoff describes the aesthetic experience as composed of three phases: (1) an effort of attention proportional to the C*omplexity (C)* of the object, (2) followed and rewarded by an aesthetic *feeling (Aesthetic Measure)*, and (3) the realization of particular *Order (O)* in it. Finally, he offered the following formula for the aesthetic measure:

$$\text{Aesthetic Measure} = \text{Order} / \text{Complexity} \qquad (2)$$

The premise is that the aesthetic feeling produced by an object is a rate of elements of order by the effort to perceive this order. It does not necessarily mean that the effect is more substantial, but it produces more associations in relation to the effort to perceive it. For Birkhoff, it is the role of aesthetics to define the elements of order and complexity for each class of aesthetic objects.

2.3 Parts from wholes

The analysis of how parts relate to each other and the whole of a building is adopted to adapt Birkhoff's aesthetic measure formula to architecture. Parts to whole relationship is a keystone to classical architectural aesthetic theories. Some relevant examples are cited ahead with terms associated with parts and whole are in bold:

"[…] Beauty [is] when the appearance of the ***work*** *is pleasing and in good taste, and when its* ***members*** *are in due proportion according to correct principles of symmetry"* (Vitruvius, 1914).

"Beauty is a definite proportional relationship among all ***parts*** *of a* ***thing****, so that nothing can be added, reduced, or changed, without making that thing less deserving of approval"* (Alberti, 2011).

"Beauty will derive from a graceful shape and the relationship of the ***whole*** *to the* ***parts****, and of the* ***parts*** *among themselves and to the* ***whole****, because buildings must appear to be like complete and well-defined bodies, of which one member matches another and all the members are necessary for what is required"* (Palladio, 2002).

Even though this discussion has lost its forefront position in contemporary architecture aesthetics discourse, it has gained attention in cognitive psychology. There are two divergent theories about how humans perceive the world disputing for hegemony: Feature Integration Theory (FIT) and Recognition-by-Components theory (RBC) (Goldstein, 2011). Both approaches defend that we perceive parts (lines, curves, colors for FIT and 3D parts named Geons for RBC) and subsequentially (FIT) or parallelly (RBC) put them together to recognize wholes as objects and environments.

Using algorithms like MSER, computers can also recognize parts in images (Matas et al., 2004). MSER binarizes images across multiple thresholds, and when a region is consistent among them, it is recognized as a part. Sardenberg and Becker used MSER to analyze and quantify architectural images through a diagram of scaled parts - that focuses on the number of individual parts - and a diagram of connectivity graph - that enables machines to grasp how parts relate (Sardenberg & Becker, 2022b).

Machine learning models have been proving more consistent than algorithms for computer vision in recent years. The latest one, with impressive results, is the Segment Anything Model (SAM) (Kirillov et al., 2023). SAM and MSER capture different qualities of architectural images, as seen in Figure 3. SAM consistently segments entire buildings from their context, while MSER captures architectural elements from an overall building. The experience of architecture is a constant reappraisal of parts producing wholes. As a Matryoshka doll, parts and wholes change roles in the perception of architecture. The observer always perceives that a part on a particular scale is a whole on a smaller scale, while what was a whole becomes a part on a larger scale. Therefore, SAM and MSER are complementary methods to understand how parts and wholes interplay in architectural perception.

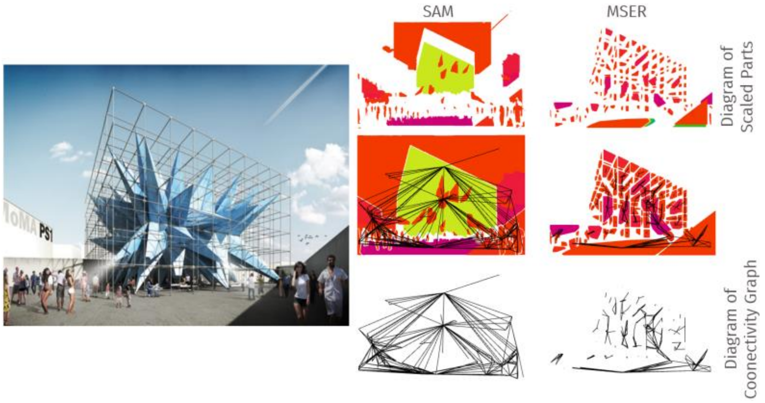

Figure 3. HWKN's Wendy Pavilion and diagrams of scaled parts and of connectivity graph produced using SAM and MSER. Source: HWKN, 2012, and Authors, 2023.

To evaluate the parts recognized by SAM and MSER, we use two diagrams (Figure 3): The diagram of scaled part is used to output the number of parts of an image, and the diagram of connectivity graph is applied to extract the number of connections between all parts that intersect and the length of each vertex that connects their centroids. These values are used in an adapted version of the aesthetic measure formula:

$$\text{Aesthetic Measure} = \frac{\text{Order}}{\text{Complexity}} = \frac{\begin{array}{c}\text{Connection edge length average}\\ \text{x Number of connections}\end{array}}{\begin{array}{c}\text{Number of parts}\\ \text{x} \sqrt{\text{Number of pixels}}\end{array}} \tag{3}$$

Finally, the weight of each term of the formula is calibrated according to a specific audience. The audience's data-gathering details and the formula's calibration are described in Sardenberg and Becker (2022a).

2.4 The predicted hedonic response model

The predicted hedonic response (PHR) model is an artificial neural network model trained to predict the hedonic response of a specific audience. It uses the quantitative outputs of the diagrams of scaled parts and connectivity graph as input neurons. The target output is the average hedonic response of the audience. It was previously tested and proved efficient for navigating a parametric model's solution space and scoring and ranking its designs despite its accuracy not being high (Sardenberg & Becker, 2023).

Table 1. Comparison between the artificial neural networks trained on MSER, SAM, and MSER+SAM.

							Avg.
Average Response (Ground truth)	**3.50**	**4.50**	**7.50**	**6.00**	**7.00**	**3.00**	
ANN MSER	3.01	4.60	5.11	5.98	7.21	7.61	
ANN MSER Accuracy	83%	97%	53%	99%	97%	39%	**78%**
ANN SAM	4.45	7.92	4.77	4.65	5.32	5.75	
ANN SAM Accuracy	78%	56%	42%	70%	68%	52%	**61%**
ANN MSER+ SAM	3.60	4.24	7.95	6.13	7.24	2.88	
ANN MSER+ SAM Accuracy	97%	93%	94%	97%	96%	95%	**95%**

Sardenberg & Becker (2023) argued that the PHR model trained using MSER could increase accuracy by combining it with other models. Considering the better consistency of SAM, it was expected that a new PHR model trained on it would be more accurate. However, compared with the ground truth in Table 1, the PHR model trained on SAM performs worse than the one trained on MSER. However, its accuracy increases substantially when an ANN model is trained using MSER and SAM. After 30,000 steps of training, it reaches an RMSE of 0.32, an R2 score of 0.93, and an accuracy of 95%

3 Results

The experiment described in this paper fosters the application of a computational aesthetics framework to map the latent space of GAN and present it to a designer and enable them to understand its design possibilities better, mainly focusing on counter-intuitive and unexpected concepts and designs. The latent space map allows an intuitive interaction between a designer and the latent space.

3.1 Latent space map

To flatten the 512-dimensional space of the latent space into two-dimensional cartesian space, t-SNE was applied. t-SNE is a statistical method that maps together similar data and maps further dissimilar data. It was used to map each image in a 2D space according to its aesthetic measure, calibrated aesthetic measure, number of parts, number of connections, and connection length average of MSER and SAM. As shown in Figure 4, it successfully clusters similar images.

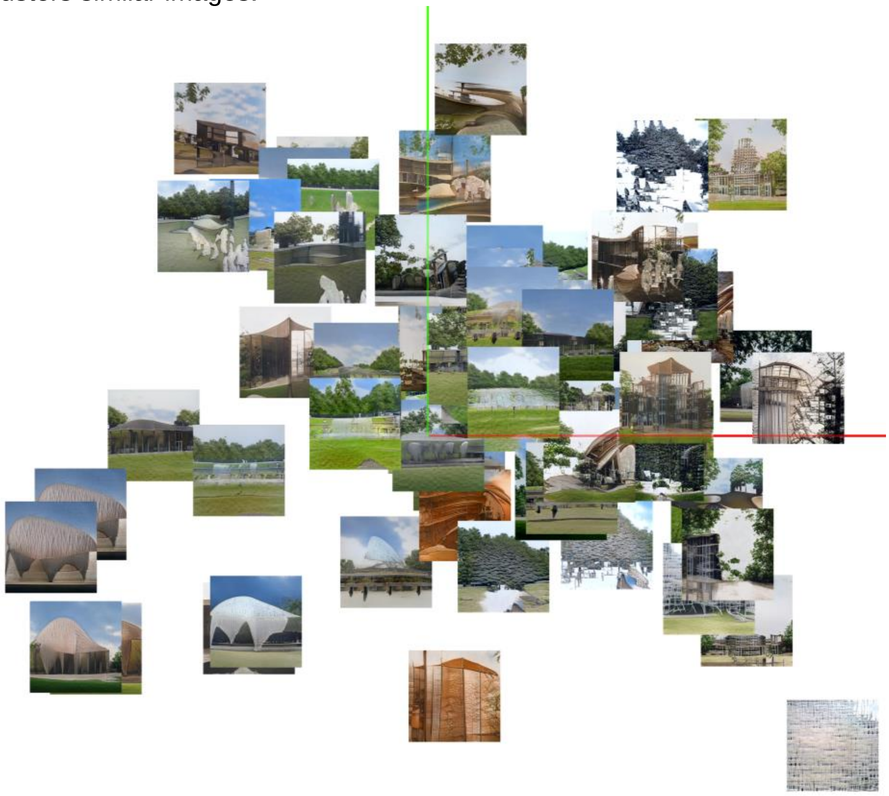

Figure 4. Bidimensional latent space map with 75 designs produced by the GAN model. Source: Authors, 2023.

In three dimensions, the vertical location of the image is defined by the output of the PHR model, which places on top the highest-scoring images (Figure 5). This sorting allows filtering the map to show only the highest-scoring images. This 3D map is interactive and can be zoomed in and out to see different designs in more detail.

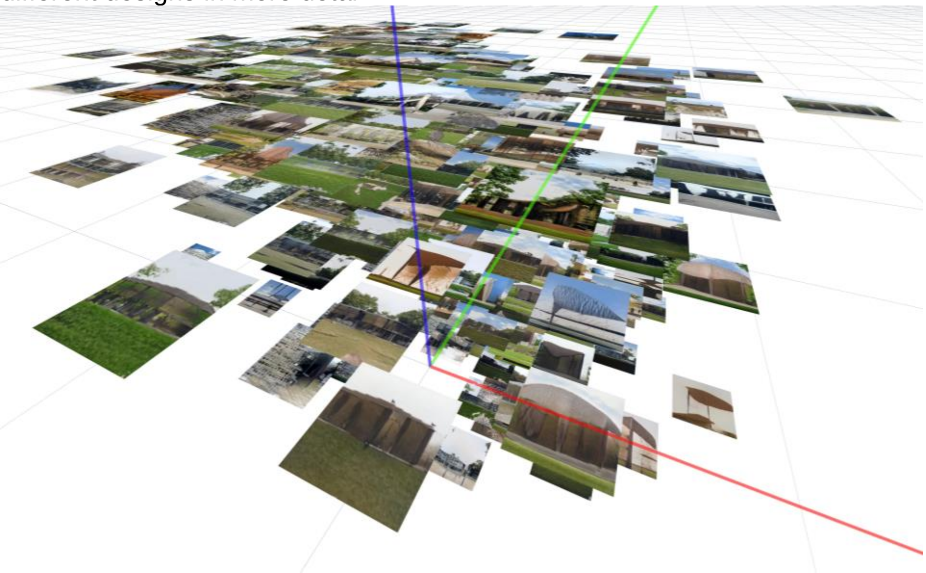

Figure 5. Perspective view of the latent space map of 1000 designs produced by the GAN model. The t-SNE method defines the horizontal green and red axis coordinates, while the PHR model defines the vertical blue coordinates. Source: Authors, 2023

4 Discussion

Generative image models are transforming architectural culture and practice. Text-to-image models like Stable Diffusion have produced architectural images comparable in quality to conventional photo-realistic CGI. However, since these models are trained on existing image datasets, they can only replicate and mimic mainstream architectural ideas. This ability of generative models obliges architects to differentiate themselves to produce new architectural concepts not present in generative models' datasets. Therefore, it is necessary to explore new diagrams of architectural disciplinary problems, like how architectural objects touch the ground, how mass relates to voids, inside to outside, the surface to volume, and parts to the whole. It is through architectural knowledge and imagination that architects can produce works that go beyond the average of a generative model. However, that does not mean that generative models cannot be helpful. Architects can navigate the latent space to explore unfinished, shifting, unfamiliar, counter-intuitive, and strange design concepts. A synonymous of latent interesting for imagination is dormant. The latent space is populated by many dormant architectural concepts and diagrams waiting to be activated by designers exploring it.

The latent space offers the challenge of never freezing the design process in a single, actual, optimized design solution. In traditional paper-based design, the architect explores options to narrow it to a single product. In parametric design, the architect designs an algorithm that produces many variations in extensive quantities, ranked by many quantitative criteria. The latent space of GANs allows the design process to continuously transform its intensive qualities and the latent space map to explore unforeseen design options. The latent space is closer to Cache's original concept of the *objectile*, a status of the object that is a projectile launched into the world with no predefined target.

The latent space demands architects to question again: can architecture perpetually wander its possibilities, or do we always need to choose a moment to freeze it? Can the walk through the latent space delay decisions to allow permanence in the cloud of images? Is the latent space a prolific and unstable reflexive field?

References

Alberti, L. B. (2011). De Re Aedificatoria. In M. F. Gage (Ed.), & B. Mitrovic (Trans.), Aesthetic Theory: Essential Texts for Architects and Designers. W. W. Norton & Company, Inc.

Birkhoff, G. D. (1933). Aesthetic Measure. Harvard University Press.

Cache, B. (1995). Earth Moves: The Furnishing of Territories (M. Speaks, Ed.; Illustrated edition). The MIT Press.

del Campo, M., & Leach, N. (2022). Can Machines Hallucinate Architecture? AI as Design Method. Architectural Design, 92(3), 6–13. https://doi.org/10.1002/ad.2807

Deleuze, G. (1992). The Fold: Leibniz and the Baroque (T. Conley, Trans.; Illustrated edition). University of Minnesota Press.

Deleuze, G. (2002). Dialogues II / Gilles Deleuze and Claire Parnet ; translated by Hugh Tomlinson and Barbara Habberjam. With "The actual and the virtual" / translated by Eliot Ross Albert. (2nd ed.). Columbia University Press.

Despois, J. (2017, February 23). Latent space visualization—Deep Learning bits #2 | HackerNoon. Hackernoon. https://hackernoon.com/latent-space-visualization-deep-learning-bits-2-bd09a46920df

Duarte, R. B., Sanches, M. M., & Lepri, L. S. (2017). Objectile e as "novas pretensões" do projeto paramétrico em arquitetura. Gestão & Tecnologia de Projetos, 12(3), Article 3. https://doi.org/10.11606/gtp.v12i3.134297

Fechner, G. T. (1965). Elements of Psychophysics. In H. S. Langfeld (Trans.), A Source Book in the History of Psychology (pp. 66–75). Harvard University Press.

Goldstein, E. B. (2011). Cognitive psychology (3. ed., international student ed., special ed.). Wadsworth, Cengage Learning.

Gramazio Kohler: MTable [Metaz 01/11]. (2018). https://www.youtube.com/watch?v=CAUL6NosMNc

Hogarth, W. (1753). The analysis of beauty: Written with a view of fixing the fluctuating ideas of taste. By William Hogarth.

Hume, D. (1757). Of the Standard of Taste. In D. Hume (Ed.), Essays Moral, Political, and Literary (pp. 226–249). Libertyclassics (1987).

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and Improving the Image Quality of StyleGAN (arXiv:1912.04958). arXiv. https://doi.org/10.48550/arXiv.1912.04958

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything (arXiv:2304.02643). arXiv. https://doi.org/10.48550/arXiv.2304.02643

Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 9(86), 2579–2605.

Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing, 22(10), 761–767. https://doi.org/10.1016/j.imavis.2004.02.006

Palladio, A. (2002). The Four Books on Architecture (R. Schofield & R. Tavernor, Trans.; Reprint edition). The MIT Press.

Sardenberg, V., & Becker, M. (2023). Aesthetics as a Criterion: Navigating Solution Spaces Utilizing Computer Vision, the Aesthetic Measure, and Artificial Neural Networks. 2023 Annual Modeling and Simulation Conference (ANNSIM), 496–507.

Sardenberg, V., & Becker, M. (2022a). Computational Quantitative Aesthetics Evaluation—Evaluating architectural images using computer vision, machine learning and social media. Pak, B, Wurzer, G and Stouffs, R (Eds.), Proceedings of the 40th eCAADE Conference 2022, Ghent, 13-16 September 2022, Pp. 567–574. http://papers.cumincad.org/cgi-bin/works/paper/ecaade2022_75

Sardenberg, V., & Becker, M. (2022b, October 20). Aesthetic Measure of Architectural Photography utilizing Computer Vision: Parts-from-Wholes. Design Computation Input/Output 2022. Design Computation Input/Output 2022. https://doi.org/10.47330/DCIO.2022.GGNL1577

Schuman, A. (2013). Polykleitos: A Canon of Beauty and Perfection. Student Scholarship. https://digitalcommons.lindenwood.edu/student-research-papers/15

Tiu, E. (2020, February 4). Understanding Latent Space in Machine Learning. Medium. https://towardsdatascience.com/understanding-latent-space-in-machine-learning-de5a7c687d8d

van Stralen, M. (2018). Mass Customization: A critical perspective on parametric design, digital fabrication and design democratization (p. 149). https://doi.org/10.5151/sigradi2018-1770

Vitruvius. (1914). Vitruvius, the ten books on architecture. Cambridge : Harvard University Press. http://archive.org/details/vitruviustenbook00vitr_0