Separate and Reassemble: Generative AI through the lens of art and media histories

Separar e remontar: IA generativa através das lentes das histórias da arte e da mídia

LEV MANOVICH[a]

City University of New York. New York, United States of America

ABSTRACT

AI image generation represents a logical evolution from early digital media algorithms, starting with basic paint programs in the 1970s and advancing to sophisticated 3D graphics and media creation software by the 1990s. Early algorithms struggled to simulate materials and effects, but advances in the 1970s and 1980s led to realistic simulations of natural phenomena and artistic techniques. Generative AI continues this trend, using neural networks to combine and interpolate visual patterns from extensive datasets. This method of digital media creation underscores the modular and discrete nature of computer-generated imagery, distinguishing it from traditional optical media.

Keywords: AI image generation, digital media, neural networks, computer graphics, generative AI

RESUMO

A geração de imagens por IA representa uma evolução lógica dos primeiros algoritmos de mídia digital, começando com programas básicos de pintura na década de 1970 e avançando para sofisticados gráficos 3D e softwares de criação de mídia na década de 1990. Os primeiros algoritmos tinham dificuldade para simular materiais e efeitos, mas os avanços nas décadas de 1970 e 1980 levaram a simulações realistas de fenômenos naturais e técnicas artísticas. A IA generativa continua essa tendência, usando redes neurais para combinar e interpolar padrões visuais de conjuntos de dados extensos. Esse método de criação de mídia digital ressalta a natureza modular e distinta das imagens geradas por computador, distinguindo-as da mídia óptica tradicional.

Palavras-chave: Geração de imagens por IA, mídia digital, redes neurais, computação gráfica, IA generativa

[a] Presidential Professor of Computer Science at the City University of New York's Graduate Center and the Director of the Cultural Analytics Lab. He authored and edited 15 books, including Artificial Aesthetics, Cultural Analytics, Instagram and Contemporary Image, Software Takes Command, and The Language of New Media. Orcid: https://orcid.org/0000-0003-0667-7584. E-mail: manovich.lev@gmail.com

DOI: http://dx.doi.org/10.11606/issn.1982-8160.v18i2p7-18

D

Separate and Reassemble

AI IMAGE[1] REPRESENTS a further logical evolution of the process that begins with digital media algorithms in the 1970s and continues in the following decades. The first computer paint programs were created in the 1970s, but could not yet simulate different paint types, brushes, and textured surfaces like canvas (Smith, 2001, 2021). But in the 1990s, software such as Corel Painter (1991–) started to offer these features ("Corel Painter", 2024). Similarly, the first 3D computer graphics algorithms for rendering solid shapes, Gouraud shading (1971), and Phong shading (1973), could not yet simulate the looks of different materials. Later, in the 1970s and 1980s, computer graphics researchers created numerous algorithms to simulate the appearance of various materials and textures, such as cloth, hair, and skin, as well as shadows, transparency, translucency, depth of field, lens flares, motion blur, reflections, water, smoke, fireworks, explosions, and other natural phenomena and cinematography techniques and effects.

Simulating many of these phenomena and techniques requires multiple separate algorithms that were developed over time. Thus, we find distinct sessions devoted to such algorithms with names like Volumes and Materials, Fluid Simulation, or Cloth and Shells in the annual proceedings of Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH), the main conference in computer graphics (CG) field (ACM SIGGRAPH, 2022). As an example, the paper "Predicting Loose-Fitting Garment Deformations Using Bone-Driven Motion Networks" presented in the 2023 conference describes "a learning algorithm that uses bone-driven motion networks to predict the deformation of loose-fitting garment meshes at interactive rates." Another conference paper "Rendering Iridescent Rock Dove Neck Feathers" describes a new approach for modeling and rendering bird feathers; and so on.

In my 1992 article "Assembling Reality: Myths of Computer Graphics" (Manovich, 1992)[2], I have analyzed this fundamental aspect of computer graphics, explaining that "synthetic photorealism is fundamentally different from the realism of the optical media, being partial and uneven, rather than analog":

> Digital recreation of any object involves solving three separate problems: the representation of an object's shape, the effects of light, and the pattern of movement. To have a general solution for each problem requires the exact simulation of underlying physical properties and processes. This is impossible because of the extreme mathematical complexity. . . In practice, computer graphics researchers have resorted to solving particular local cases, developing a number

[1] The terms generative media, synthetic media, or generative AI refer to the process of synthesizing media objects with artificial neural networks. The examples of such objects include text, voice, music, 3D models, datasets, and computer code. The terms generative image, AI Image or visual AI refer to specially synthesized visual objects. These objects can be still images that imitate the appearance and structure of all types of visual media, from photographs to drawings, and also moving images that imitate appearances of animation and video.

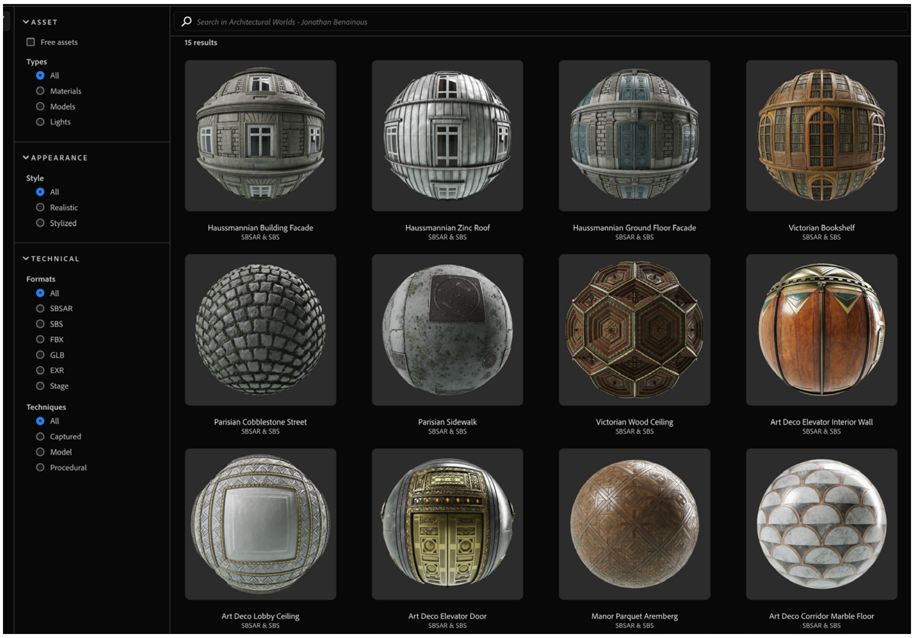
[2] See also "Part 4: The Illusions" (Manovich, 2002).

> of unrelated models for simulation of some kinds of shapes, materials and movements. (Manovich, 1992, pp. 12-14)

In other words, 3D CG disassembles the world, separating objects, shapes, materials, light reflections, textures, movements, and behaviors (Figure 1). During rendering, the effects of multiple algorithms simulating all these aspects are combined together. Thus, *visual representations created using CG are discrete and modular, rather than continuous and "monistic."* This is one of the most important characteristics of CG medium, distinguishing it from lens-based optical image media.

Figure 1

A few from the thousands of materials available in Adobe 3D content creation software



Note. Adapted from author

This logic of separation and recombination also defines next stage of digital media: PC software for media creation and editing. Following its initial release in 1990, Photoshop gradually began to include simulated effects and techniques from various artistic mediums, ranging from darkroom photography to oil painting, within a single program. These effects can be combined in a single digital image. Music software similarly allows users to combine many simulated instruments and multiple effects such as reverb and echo in one composition. Word processing and desktop publishing software separate the physical process of print composition into its basic parts that also can be now recombined—for example, you can take any font and arbitrarily change its size or generate your own font[3].

All of these media software capabilities were first proposed in the 1970s and later realized in the 1980s and 1990s, eventually becoming ubiquitous. AI generative media follows that same logic, although its underlying technical implementation is different. During training, neural networks learn visual patterns characteristic of hundreds of different types of art media, lighting techniques and effects from history of photography and cinematography, and visual signatures of many thousands of historical and contemporary artists, architects, fashion designers, and other creators. A reference website Midlibrary currently lists 367 "artistic techniques" that the Midjourney AI image generator tool can reliably simulate according to the tests conducted by this website team[4]. They range from "albumen print" and "anaglyph" to "wood carving" and "wireframe rendering."

Importantly, a user can include references to multiple techniques and/or multiple creators in a single prompt, potentially generating new types of media effects that did not exist before. Here are examples of such prompts from my own experiments:

- Using multiple artists in one prompt: "18th century very big and detailed panoramic etching showing landscape in the style of *Michael Kaluta, Kawanabe Kyosai, Pieter Bruegel the Elder*, insane detail, cinematic"

- Using multiple artistic media in one prompt: "18th century futuristic infinite museum storage space with art objected on the shelves, snow fall inside the space and fog, wide angle view looking down, 7pm soft evening light, detailed intricate *drawing and etching* with very fine shading, subtle nuanced sombre *color pencils* and fine *pens*"

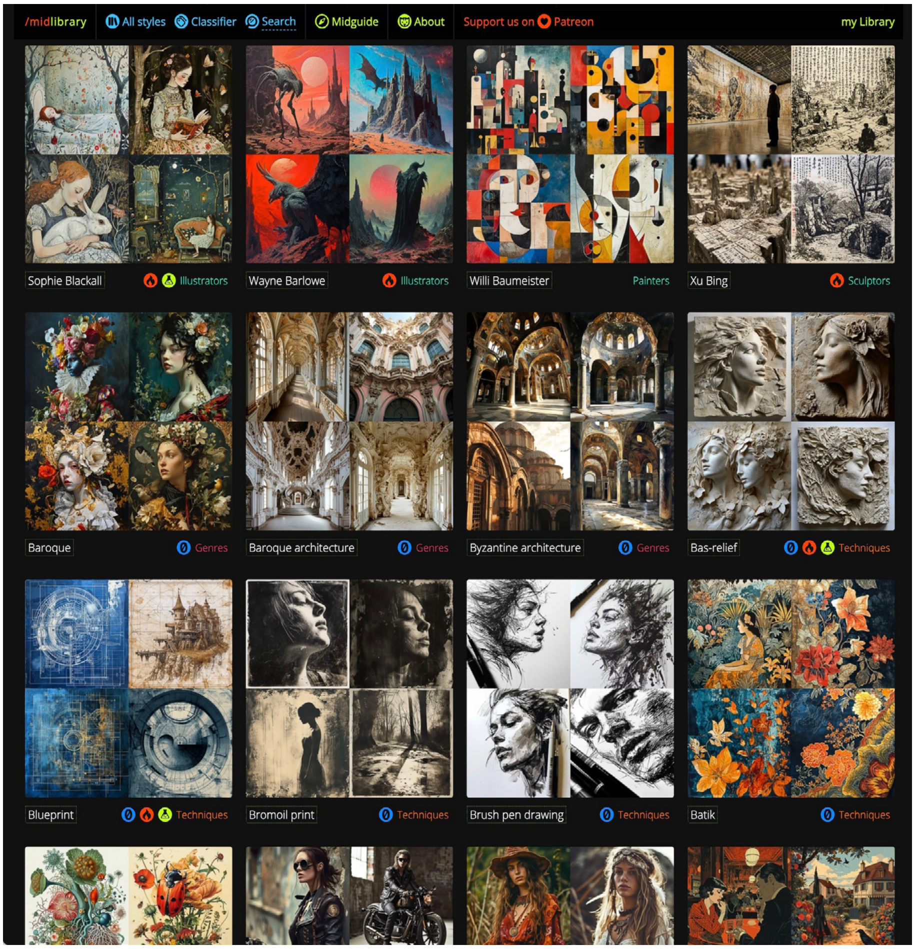
Figure 2 shows screenshot from midlibrary.io showing a few of artistic techniques, art genres, and "styles" of painters, illustrators, architects, photographers, and fashion designers that Midjourney can simulate. At the moment of this writing, the library contains nearly 5000 references. (Captured March 24, 2024).

[3] For the detailed analysis of media software and its conceptual origins, see Manovich (2013).

[4] Retrieved February 25, 2024, from https://midlibrary.io/

Figure 2

Screenshot from midlibrary.io



Note. Adapted from author

The pioneering digital media theorist of 1990s and 2000s William J. Mitchell called this key characteristic of digital media "separate and recombine." In his 1996 book *City of Bits*, he described this process in relation to urban planning:

> Classical architects of the eighteenth and nineteenth centuries handled the task of putting spaces together by creating hierarchies of great and small spaces around axial, symmetrical circulation systems connected to grand, formal entries and public open spaces…functionalist modernists of the twentieth century have often derived their less regular layouts directly from empirically established requirements of adjacency

> and proximity among the necessary spatial elements. But when telecommunication through lickety-split bits on the infobahn supplements or replaces movement of bodies along circulation paths, and when telepresence substitutes for face-to-face contact among the participants in activities, the spatial linkages that we have come to expect are loosened. The constituent elements of hitherto tightly packaged architectural and urban compositions can begin to float free from one another, and they can potentially relocate and recombine according to new logics. (Mitchell, 1996, p. 104)

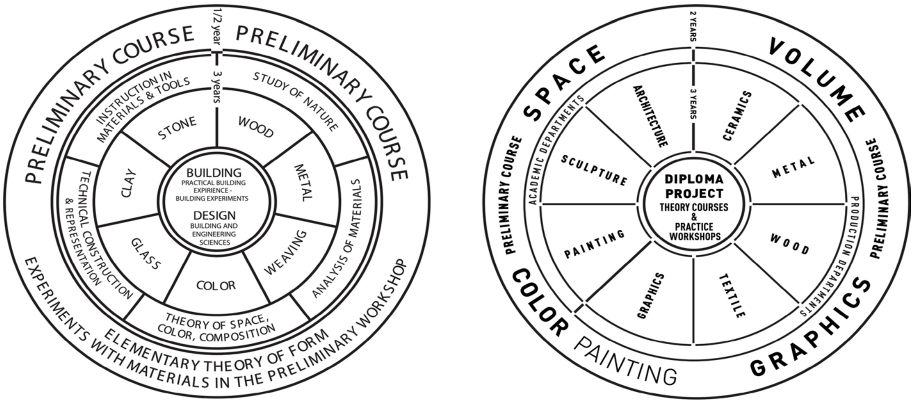
Mitchell's lectures in the 2000s expanded on this formulation, demonstrating how the logic of separation and recombination can be seen in digital media in a variety of ways. Generative AI continues the same logic. A neural network extracts elements and structures from hundreds of millions or billions of images in its training set. They include distinct color palettes, compositions, lighting effects, artifacts of historical photography processes, and so on. When you ask AI image tool to generate new images with specified visual attributes, it does its best to combine (or more precisely, *interpolate* between) appropriate art patterns and effects.

No human historian, theorist, or practitioners of visual art, photography, cinema, or design were ever able to describe all those patterns. In the early 20th century pioneering art historians Aby Warburg and Erwin Panofsky developed the study of iconology. Warburg defines this concept as visual motives that (re)appear in various civilizations and media.

Panofsky used it somewhat differently, referring to symbols and motifs that have existed throughout the history of art. During the same period visual artists and architects disassembled visual arts in a different way, breaking down an image into its basic components and dimensions such as points, lines, planes, two-dimensional forms, color, space, texture, pattern, balance, and equilibrium, among others. While this project of methodical dismantling and creation of new visual languages from these components is central to modernist art and its many-isms, it finds its most methodical development in the curricula of two new schools of art and design. VKHUTEMAS in Moscow (1920–1930) and Bauhaus in Germany (1919–1933) introduced around the same time their Basic Course, in which students were taught how to systematically work with all those elements and dimensions. Instead of drawing from life, painting portraits, or making historical compositions, the students started training by completing exercises with image primitives such as basic shapes, forms, and colors. At VKHUTEMAS, the Basic Course was created in 1920 by Rodchenko, Popova, Ekster, Vesnin, and other faculty members from painting, architecture, and other school areas. In its first iteration, it consisted of a number of workshops such as "Discipline of Synchronized Shapes and Colors", "Plane, Color, and Spatial Design," "Graphic Construction on a Plane Surface," and "Color." It was further transformed during VKHUTEMAS existence. Eventually, four learning sequences were approved for all VKHUTEMAS students: Graphics, Color, Volume, and Space[5] (Figure 3 and 4).

Figure 3

The structures of courses in Bauhaus and VKHUTEMAS. Both curricula begun with the basic course



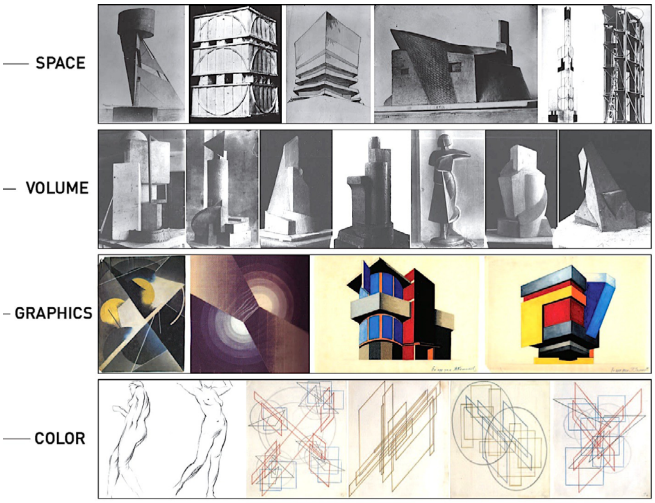
Bauhaus
Teaching Program Diagram, W. Gropius, 1922.

VKhUTEMAS
Teaching Program Diagram, 1923.

Note. Adapted from author

Figure 4

Examples of student exercises at VKHUTEMAS



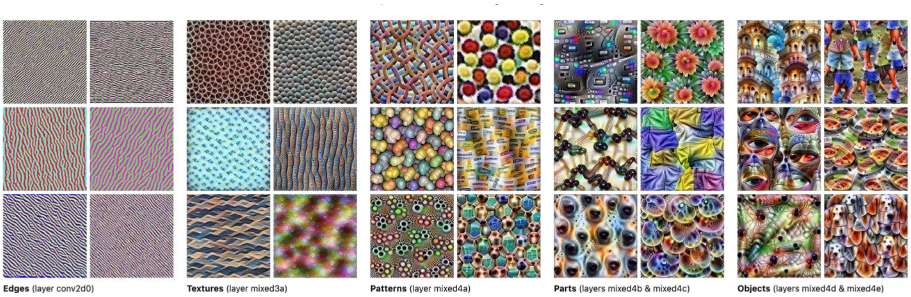
Note. Adapted from Bokov (2014)

[5] See Bokov (2021) and Vkhutemas (2020). The Basic Course at this school was more systematic and comprehensive than a similar course at Bauhaus; it was taught by many different faculty members and lasted two years. Note that VKHUTEMAS was ten times larger than Bauhaus, with 100 faculty and 5000 students during the ten years of its existence versus only 500 students at Bauhaus.

In a certain sense, generative AI models can be said to continue these programs of decomposition and analysis of visual arts that begun in the early twentieth century. Artificial intelligence algorithms extract patterns (or "features") from training data. However, at the moment, we cannot look at billions of parameters in a gigantic generative network and get a neat catalog of all the patterns the network learnt (Podell et al., 2023). In the 2010s when neural networks were simpler and smaller, scientists were able to visualize what their neurons learn. For example, the following visualization shows the features learned by a network trained to recognize objects in photographs (Figure 5): Google Research visualization showing the features learned by progressive layers of a network trained for image recognition. The first layers learn basic features such as edges and textures (see previous page), and subsequent layers learn the appearance of partial and whole objects. A network first learns how to recognize basic features before progressing to object recognition. Unfortunately, the architecture of generative networks that synthesize images prevents us from "looking inside" these networks and visualizing them in the same manner[6].

Figure 5

Google Research visualization by progressive layers of image recognition



Note. Adapted from Olah et al. (2017)

I want to conclude with a relevant quote from my 2018 book *AI Aesthetics* (Manovich, 2018). While at that time deep neural networks were mostly used for media classification and recommendations, with generative AI revolution still four years away, the analysis I developed in the book section called "AI as a Culture Theorist" has become even more relevant today:

> [There is] a crucial difference between an "AI culture theorist" and a human theorist/historian. The latter comes up with explicit principles that describe how a cultural area function…a neural net can be trained to distinguish between works

[6] For the overview of available deep networks visualization methods, see Barla (2024).

> of different artists, fashion designers, or film directors. And it can also generate new objects in the same style. But often we don't know what exactly the computer has learned…Will the expanding use of machine learning to create new cultural objects make explicit the patterns in many existing cultural fields that we may not be aware of? (Manovich, 2018, p. 23)

This theoretical potential to me is one of the most interesting and valuable thing about generative AI—however we will have to wait to see if it may be realized in the future. Visual AI is the fourth significant *data <–> knowledge* effect of the web—a global accumulation of networked hyperlinked cultural content that began to grow quickly after 1993. Although people have been sharing texts and images on the internet since the 1970s, this process has accelerated after 1993, when the first visual browser, Mosaic, was introduced on January 23 of that year.

I have observed several repercussions of the growth of information on the web over the following 30 years. If we wish to situate the development of Visual AI in the early 2020s in this timeline, here are four such effects. Certainly, others can be also named, so this is only one list of techno-cultural development technologies enabled by the web that I am particularly interested in:

1. The first effect is the switch from categorical, hierarchical, and structured organization of information (exemplified by library catalogs and early web directories) to search engines in the late 1990s. There was so much content that organizing it in conventional ways was no longer practical, and search become the new default. Note that *web search is based on a prediction of what will be most relevant to the user* as opposed to giving you a precise and definite answer. Note that generative AI is also predictive—it predicts possible text, images, animation, or music in response to your question or prompt. The regime of absolute certainty, i.e., a truth vs. a lie that is typical for human civilization, is replaced by predictions, as statistics becomes foundation of human sciences in the 20th century, and data science and AI in recent decades.

2. The second major effect is the rise in popularity of data visualization during the 2000s. The field thrives around 2005. As a part of this development, the new field "artistic data visualization" develops in the same decade, along with other new cultural fields: data art and data design. (In our lab we created *Phototrails*, *Selfiecity*, and *On Broadway* in 2012–2014. These were first interactive visualizations of millions

of Instagram images.) If search attempts to find the most relevant items in the giant data universe, visualization tries to show parts of this universe in one image, revealing patterns and connections.

3. The third effect is the emergence of "data science" as the master discipline of the new big data era at the end of the 2000s. While many methods employed in data science have already been available for decades, the rapid increase in unstructured data in the 2000s motivated the development of a separate data science field — the key new profession of the data society. My own version of this stage is "cultural analytics," the idea I came up with in 2005 and worked on for the next 15 years in our lab. Our main method was data visualization, but now applied to large media collections of photos, video, film, manga, magazine covers, Instagram images, etc. I named this method *media visualization*[7]

4. The next, but certainly not the last, effect of the growth of online visual digital content is Generative AI which becomes popular in early 2020s. Dalle-e is released in 2020, MidJourney in 2022, ChatGPT and Photoshop generative fill in 2023, and hundreds of other tools exist today. A bit earlier around 2017, a particular AI method for media generation called GAN became already popular with digital artists.

It is relevant to mention that both Visual AI and Generative AI in general builds on 20 years of work, with the first relevant papers published in 2001. The key idea is to use web content universe as a source of data for machine learning (ML), without labeling it, already appears in the research paper published around that time.

When investigating what kind of pattern is established by these four effects: search is the first method to deal with the new scale of content on the web. Data science focuses on finding patterns, relations, clusters, and outliers in big data, and also predicting future data. Data visualization tries to summarize datasets visually. And now Generative AI explores "big content," yet in another way, by generating new content which combines many patterns from existing media.

To put this differently: Generative AI synthesizes new content that has statistical properties similar to existing content. But it is not a copy of what already exists. AI generates new content (texts, images, animation, 3D models, music, singing, etc.) by interpolating between existing points in the latent space.

[7] See projects at http://lab.culturalanalytics.info.

This space contains numerous patterns and structures extracted by artificial networks from billions of image-text pairs, trillions of text pages, and other large collections of existing human cultural artifacts. AI predicts what could exist between these points in space of patterns. For example, it can predict a "painting" made by artists A, B, C, using techniques D and E, with content F, G, and E, with mood, colors M-N, proportion W, composition K, etc.

Note that the three earlier developments all approach big data by summarizing it. Web search reduces billions of web pages to the top results. Data visualization reduces it to a diagram. Data science reduces it by using summary statistics, cluster analysis, regression, or latent space projection. But Visual AI is doing something new. It also first reduces big data during learning, and then generates new data points.

One way to sum up all this is to say that we moved from probabilistic search to probabilistic media generation: 1999 to 2022. But certainly, generative AI is not the last effect of the existence of web data; others will likely emerge in the future.

REFERENCES

ACM SIGGRAPH. (2022). *SIGGRAPH '22: ACM SIGGRAPH 2022 Conference Proceedings*.

Barla, N. (2024, May 14). *How to visualize deep learning models*. Neptune.ai. https://neptune.ai/blog/deep-learning-visualization

Bokov, A. (2014). *VKhUTEMAS training*. Pavilion of the Russian Federation at the 14th International Architecture Exhibition.

Bokov, A. (2021). *Avant-garde as method: Vkhutemas and the pedagogy of space, 1920-1930*. Park Books.

Corel Painter. (2024, July 5). In *Wikipedia*. https://en.wikipedia.org/wiki/Corel_Painter

Manovich, L. (1992). Assembling reality: Myths of computer graphics. *Afterimage, 20*(2), 12-14.

Manovich, L. (2002). *The language of new media*. MIT press.

Manovich, L. (2013). *Software takes command*. Bloomsbury Academic.

Manovich, L. (2018). *AI aesthetics*. Strelka Press.

Mitchell, W. J. (1996). *City of bits: Space, place, and the Infobahn*. MIT press.

Olah, C., Mordvintsev, A., & Schubert, L. (2017, November 7). Feature visualization: How neural networks build up their understanding of images. *Distill*. https://doi.org/10.23915/distill.00007

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., & Rombach, R. (2023). *SDXL: Improving latent diffusion models for high-resolution image synthesis*. arXiv. https://arxiv.org/abs/2307.01952

Smith, A. R. (2001). Digital paint systems: An anecdotal and historical overview. *IEEE Annals of the History of Computing*, *23*(2), 4-30. https://doi.org/10.1109/85.929908

Smith, A. R. (2021). *A biography of the pixel*. MIT Press.

Vkhutemas. (2020, June 25). *Main course*. https://www.vkhutemas.ru/en/structure-eng/faculties-eng/main-course/

Article received on July 3, 2024 and approved on August 14, 2024