IMAGE
The Interdisciplinary Journal of Image Sciences
37(1), 2023, S. 34-41
ISSN 1614-0885
DOI: 10.1453/1614-0885-1-2023-15448

HERBERT VON HALEM VERLAG

Lev Manovich

AI Image Media through the Lens of Art and Media History

Abstract: I've been using computer tools for art and design since 1984 and have already seen a few major visual media revolutions, including the development of desktop media software and photorealistic 3D computer graphics and animation, the rise of the web after, and later social media sites and advances in computational photography. The new AI 'generative media' revolution appears to be as significant as any of them. Indeed, it is possible that it is as significant as the invention of photography in the nineteenth century or the adoption of linear perspective in western art in the sixteenth. In what follows, I will discuss four aspects of AI image media that I believe are particularly significant or novel. To better understand these aspects, I situate this media within the context of visual media and human visual arts history, ranging from cave paintings to 3D computer graphics.

'AI' as a Cultural Perception

There is not one specific technology or a single research project called 'AI'. However, we can follow how our cultural perception of this concept evolved over time and what it was referring to in each period. In the last fifty years, when an allegedly uniquely human ability or skill is being automated by means of computer technology, we refer to it as 'AI'. Yet, as soon as this automation is seamlessly and fully successful, we tend to stop referring to it as an 'AI case'. In other words, 'AI' refers to technologies and methodologies that automate human cognitive abilities and are starting to function but are not quite there yet. 'AI' was already present in the earliest computer media tools. The first interactive drawing and design system, Ivan Sutherland's Sketchpad (1961-1962), had a feature that would automatically finish any rectangles or circles you started drawing. In other words, it knew what you were trying to make. In the very broad understanding just given, this was undoubtedly 'AI' already.

My first experience with a desktop paint program running on an Apple II was in 1984, and it was truly amazing to move your mouse and see simulated paint brushstrokes appear on the screen. However, today we no longer consider this to be 'AI'. Another example would be the Photoshop function that automatically selects an outline of an object. This function was added many years ago – this, too, is 'AI' in the broad sense, yet nobody would refer to it as such today. The history of digital media systems and tools is full of such 'AI moments' – amazing at first, then taken for granted and forgotten as 'AI' after a while. (In AI history books, this phenomenon is referred to as the 'AI effect'.) At the moment, 'creative AI' refers only to recently developed methods where computers transform some inputs into new media outputs (e.g., text-to-image models) and specific techniques (e.g., certain types of deep neural networks). However, we must remember that these methods are neither the first nor the last in the long history and future of simulating human art abilities or assisting humans in media creation.

From Representation to Prediction

Historically, humans created images of existing or imagined scenes by a number of methods, from manual drawing to 3D CG (see below for explanation of the methods). With AI generative media, a fundamentally new method emerges. Computers use large datasets of existing representations in various media to predict new images (still and animated).

One can certainly propose different historical paths leading to visual generative media today, or divide one historical timeline into different stages – here is one such possible trajectory:

1. Creating representations manually (e.g., drawing with variety of instruments, carving, etc.). More mechanical stages and parts were sometimes carried out by human assistants typically training in their teacher's studio – so there is already some delegation of functions.
2. Creating manually but using assistive devices (e.g., perspective machines, camera lucida). From *hands* to *hands + device*. Now some functions are delegated to mechanical and optical devices.
3. Photography, x-ray, video, volumetric capture, remote sensing, photogrammetry. From *using hands* to *recording information using machines*. From *human assistants* to *machine assistants*.
4. 3D CG. You define a 3d model in a computer and use algorithms that simulate effects of light sources, shadows, fog, transparency, translucency, natural textures, depth of field, motion blur, etc. From r*ecording* to *simulation*.
5. Generative AI. Using media datasets to predict still and moving images. From *simulation* to *prediction*.

"Prediction" is the actual term often used by AI researchers in their publications describing visual generative media methods. So, while this term can be used figuratively and evocatively, this is also what actually happens scientifically when you use image generative tools. When working with a text-to-image AI-model, the neural network attempts to predict the images that correspond best to your text input. I am certainly not suggesting that using all other already accepted terms such as 'generative media' is inappropriate. But if we want to better understand the difference between AI visual media synthesis methods and other representational methods developed in human history, employing the concept of 'prediction' and thus referring to these AI systems as 'predictive media' captures this difference well.

Media Translations

There are several methods for creating 'AI media'. One method transforms human media input while retaining the same media type. Text entered by the user, for example, can be summarized, rewritten, expanded, and so on. The output, like the input, is a text. Alternatively, in the image-to-image generation method, one or more input images are used to generate new images. However, there is another path that is equally intriguing from historical and theoretical perspectives. 'AI media' can be created by automatically 'translating' content between media types. Because this is not a literal one-to-one translation, I put the word 'translation' in quotes. Instead, input from one medium instructs a neural network to predict the appropriate output from another. Such input can also be said to be 'mapped' to some outputs in other media. Text is mapped into new styles of text, images, animation, video, 3D models, and music. The video is converted into 3D models or animation. Images are 'translated' into text, and so on. Text-to-image method translation is currently more advanced than others, but various forms will catch up eventually.

Translation (or mapping) between one media and another is not a new concept. Such translations were done manually throughout human history, often with artistic intent. Novels have been adapted into plays and films, comic books have been adapted into television series, a fictional or non-fictional text was illustrated with images, etc. Each of these translations was a deliberate cultural act requiring professional skills and knowledge of the appropriate media. Some of these translations can now be performed automatically on a massive scale thanks to artificial neural networks, becoming a new means of communication and culture creation. Of course, artistic adaptation of a novel into a film by a human team and automatic generation of visuals from novel text by a net is not the same thing, but for many more simple cases automatic media translation can work

well. What was once a skilled artistic act is now a technological capability available to everyone. We can be sad about everything that might be lost as a result of the automation – and *democratization* – of this critical cultural operation: skills, something one might call 'deep artistic originality' or 'deep creativity', and so on. However, any such loss may be only temporary if the abilities of 'culture AI' are, for example, even further improved to generate more original content and understand context better.

Because the majority of people in our society can read and write in at least one language, text-to-another media methods are currently the most popular. They include text-to-image, text-to-animation, text-to-3D, and text-to-music models. These AI tools can be used by anyone who can write, or by using readily available translation software to create a prompt in a language these tools understand best, such as English. However, other media mappings can be equally interesting for professional creators. Throughout the course of human cultural history, various translations between media types have attracted attention. They include translations between video and music (club culture); long literary narratives turned into movies and television series; any texts illustrated with images in various media such as engravings; numbers turned into images (digital art); texts describing paintings (ekphrasis, which began in Ancient Greece), mappings between sounds and colors (especially popular in modernist art); etc.

The continued development of AI models for mappings between all types of media, without privileging text, has the potential to be extremely fruitful, and I hope that more tools will be able to accomplish this. Such tools would be able be used alone or in conjunction with other tools, and the techniques of using them will be useful both to professional artists and other creators alike. However, being an artist myself, I am not claiming that future 'culture AI' will be able to match, for example, innovative interpretations of Hamlet by avant-garde theatre directors such as Peter Brook or astonishing abstract films by Oscar Fishinger that explored musical and visual correspondences. It is sufficient that new media mapping AI tools stimulate our imagination, provide us with new ideas, and enable us to explore numerous variations of specific designs.

The Popular and the Original

Both the modern human creation process and the predictive AI generative media process seem to function similarly. A neural network is trained using unstructured collections of cultural content, such as billions of images and their descriptions or trillions of web and book pages. The net learns associations between these artifacts' constituent parts (such as which words frequently appear next to one another) as well as their common patterns and structures. The trained net

then uses these structures, patterns, and 'culture atoms' to create new artifacts when we ask it to. Depending on what we ask for, these AI-created artifacts might closely resemble what already exists or they might not.

Similarly, our life is an ongoing process of both supervised and unsupervised cultural training. We take art and art history courses, view websites, videos, magazines, and exhibition catalogs, visit museums, and travel in order to absorb new cultural information. And when we 'prompt' ourselves to make some new cultural artifacts, our own biological neural networks (infinitely more complex than any AI nets to date) generate such artifacts based on what we've learned so far: general patterns we've observed, templates for making particular things (such as drawing a human head with correct proportions, or editing an interview video), and often concrete parts of existing artifacts. In other words, our creations may contain both exact replicas of previously observed artifacts and new things that we represent using templates we have learned, such as color combinations and linear perspective. Additionally, both human and AI models frequently have a default 'house' style (the actual term used by Midjourney developers). If one does not specify a style explicitly, the AI will generate it using this 'default' aesthetic. A description of the medium, the kind of lighting, the colors and shading, and/or a phrase like "in the style of" followed by the name of a well-known artist, illustrator, photographer, fashion designer, or architect are examples of specifications to steer away from this default.

Because it can simulate tens of thousands of already-existing aesthetics and styles and interpolate between them to create new hybrids, AI is more capable than any single human creator in this regard. However, at present, skilled and highly experienced human creators also have a significant advantage. Both humans and artificial intelligence are capable of imagining and representing nonexistent and existing objects and scenes alike. Yet, unlike AI image generators, human-made images can include very particular content, unique miniscule details, and distinctive aesthetics in a way that is currently beyond the capabilities of AI. In other words, today a large group of highly skilled and experienced illustrators, photographers, and designers can represent everything a trained neural net can do (although it will take much longer), but they can also visualize objects and compositions and use aesthetics that the neural net cannot do at this time (or at least has a very hard time to do consistently).

What is the cause of this aesthetic and content gap between human and artificial creators? 'Cultural atoms', structures, and patterns in the training data that occur most frequently are very successfully learned during the process of training an artificial neural network. In the 'mind' of a neural net, they gain more importance. On the other hand, 'atoms' and structures that are rare in the training data or may only appear once are hardly learned or not even parsed at all. They do not enter the artificial culture universe as learned by AI. Consequently,

when we ask AI to synthesize them, it is unable to do so. Due to this, text-to-image AIs such as Midjourney or RunwayML are not currently able to generate drawings *in my style*, expand my drawings by adding newly generated parts, or replace specific portions of my drawings with new content drawn in my style (e.g., perform "outpainting" or "inpainting").[1] Instead, these AI tools generate more generic objects than what I frequently draw or they produce something that is merely ambiguous yet uninteresting. I am certainly not claiming that the style and the world shown in my drawings is completely unique. They are also a result of specific cultural encounters I had, things I observed, and things I noticed. But because they are uncommon (and thus unpredictable), AI finds it difficult to simulate them, at least without additional training using my drawings.

Here we encounter the greatest obstacle we face as creators in using AI generative media. Frequently, AI generates new media artifacts that are more generic and stereotypical than what we intended. This can affect any image dimensions – elements of content, lighting, crosshatching, atmosphere, spatial structure, and details of 3D shapes, among others. Occasionally it is immediately apparent, in which case you can either attempt to correct it or disregard the results. Very often, however, such 'substitutions' are so subtle that we cannot detect them without extensive observation or, in some cases, the use of a computer to quantitatively analyze numerous images. In other words, new AI generative media models, much like the discipline of statistics since its inception in the 18th century and the field of data science since the end of the 2010s, deal well with frequently occurring items and patterns in the data but do not know what to do with the infrequent and uncommon. We can hope that AI researchers will be able to solve this problem in the future, but it seems so fundamental that we should not anticipate a solution immediately.

Subject and Style

In the arts, the relationship between 'content' and 'form' has been extensively discussed and theorized. This brief section does not attempt to engage in all of these debates or to initiate discussions with all relevant theories. Instead, I would like to consider how these concepts play out in AI's 'generative culture'. However, instead of using content and form, I will use a different pair of terms

1 Importantly, other AI models that are open source such as Stable Diffusion make it possible to feed them additional training data supplied by a user. This allows for generation of artistic styles and subjects beyond what the models can do initially. For example, one young Russian artist fine-tuned a Stable Diffusion model on a few dozen images of paintings by Russian conceptual artists such as Ilya Kobakov or Vitaly Komar and Alex Melamid and then generated new images that expand this art tradition.

both of which are more common in AI research publications and online conversations between users: *subject* and *style*.

At first glance, AI media tools appear capable of clearly distinguishing between the subject and style of any given representation. In text-to-image models, for instance, you can generate countless images of the same subject. Adding the names of specific artists, media, materials, and art historical periods is all that is required for the same subject to be represented differently to match these references. Photoshop filters began to differentiate between subject and style as soon as the 1990s, but AI generative media tools are more capable. For instance, if you specify "oil painting" in your prompt, simulated brushstrokes will vary in size and direction across a generated image based on the objects depicted. AI media tools appear to 'understand' the semantics of the representation as opposed to earlier filters that simply applied the same transformation to each image region regardless of its content. For instance, when I used "a painting by Malevich" and "a painting by Bosch" in the same prompt, Midjourney generated an image of space that contained Malevich-like abstract shapes as well as many small human and animal figures like in popular Bosch paintings that were properly scaled for perspective.

AI tools routinely add content to an image that I did not specify in my text prompt in addition to representing what I requested. This frequently occurs when the prompt includes "in the style of" or "by" followed by the name of a renowned visual artist or photographer. In one experiment, I used the same prompt with the Midjourney AI image tool 148 times, each time adding the name of a different photographer. The subject in the prompt remained mostly the same – an empty landscape with some buildings, a road, and electric poles with wires stretching into the horizon. Sometimes adding a photographer's name had no effect on the elements of a generated image that fit our intuitive concept of style, such as contrast, perspective, and atmosphere. But every now and again, Midjourney also modified the image content. For example, when well-known works by a particular photographer feature human figures in specific poses, the tool would occasionally add such figures to my photographs. (Like Malevich and Bosch, they were transformed to fit the spatial composition of the landscape rather than mechanically duplicated.) Midjourney has also sometimes changed the content of my image to correspond to a historical period when a well-known photographer created his most well-known photographs.

According to my observations, when we ask Midjourney or a similar tool to create an image in the style of a specific artist, and the subject we describe in the prompt is related to the artist's typical subjects, the results can be very successful. However, when the subject of our prompt and the imagery of this artist are very different, 'rendering' the subject in this style frequently fails. To summarize, in order to successfully simulate a given visual style using current AI tools,

you may need to change the content you intended to represent. Not every subject can be rendered successfully and satisfyingly in any style. This observation, I believe, complicates the binary opposition between the concepts of 'content' and 'style'. For some artists, AI can extract their style from examples of their work and then apply it to different types of content. But for other artists, it seems, their style and content cannot be separated. For me, these kinds of observations and subsequent thoughts are one of the most important reasons for using new media technologies like AI generative media and learning how they work. Of course, as a media theorist myself, I had been thinking about the relationships between subject and style (or content and form) for a long time, but being able to conduct systematic experiments like the one I described brings new ideas and allows us to look back at cultural history in new ways.