

## Image Similarity Analysis in Generative AI

<sup>1</sup>Choi Haerin, <sup>2</sup>Lee Hyunseok

<sup>1</sup>Master Student, Department of Design, Pusan National Univ., South Korea

<sup>2</sup>Professor, Department of Design, Pusan National Univ., South Korea

E-mail: <sup>1</sup>popo5210@naver.com <sup>2</sup>leehs@pusan.ac.kr

### Abstract

*In Consciousness Explained, Daniel Dennett argued that consciousness is a phenomenon emerging from the complex flow of information in the brain, and to understand it, an objective approach is necessary. While AI is increasingly mimicking human functions, it is difficult to say that AI possesses consciousness similar to humans. However, consciousness is an essential factor for perception, but perception does not necessarily require consciousness. Therefore, this study aims to analyze how similar the way AI, particularly the DALL-E model developed by OpenAI, processes visual information is to the structure of human perception.*

*In the study, new images were generated using the GPT-4 DALL-E model based on five sets of reference images, and the structural similarity between the generated images and the reference images was analyzed using SSIM (Structural Similarity Index Measure). The SSIM scores of the images generated by DALL-E based on the reference images ranged between 0.131 and 0.63. This confirmed that AI learned some degree of the visual patterns from the reference images. However, AI did not generate images that perfectly aligned with human perception, and images that contained complex shapes or fine textures recorded lower SSIM scores. Notably, the AI showed limitations in depicting human portraits, suggesting that AI's perception system is simplified compared to the complexity of human perception structures.*

*This study demonstrated that while the DALL-E model has potential in processing visual information, there remains a clear difference from the complex human perception system. These results suggest that AI still has limitations in mimicking the way humans process visual information, indicating a need for further in-depth research into the independent characteristics of AI perception in the future*

**Keywords:** AI, DALL-E, Visual Information Processing, SSIM, Structural Similarity, Human Perception

## 1. INTRODUCTION

As artificial intelligence (AI) technology continues to develop, new philosophical discourses are emerging. Google CEO Sundar Pichai announced the "AI-First era," and NVIDIA CEO Jensen Huang predicted that there is a high likelihood of forming AI teams in the future.[1] With the advancement of AI technology, new philosophical debates have arisen. In particular, the question of "Can machines have consciousness?" has prompted a fundamental exploration into AI's conscious reasoning, and discussions are still ongoing. David Chalmers, a cognitive scientist at New York University, mentioned that GPT-3 shows signs of consciousness. He said, "When it comes to consciousness, I'm open to the idea that a worm with 302 neurons has

---

Manuscript received: September 7, 2024 / revised: October 25, 2024 / accepted: November 29, 2024

Corresponding Author: [leehs@pusan.ac.kr](mailto:leehs@pusan.ac.kr)

Tel: +82-51-510-2952, Fax: +82-51-512-1741

Master. Student, Department of Design, Pusan National Univ., South Korea

consciousness, so I'm positive about the idea that GPT-3, with 175 billion parameters, might also have consciousness." [2] This emphasizes the complexity and essence of consciousness that we have yet to fully understand and suggests that as AI advances, new discussions about consciousness are necessary. Along with the development of AI, new discourses on consciousness and perception are emerging, and the need to study and evaluate these issues has arisen.

This study aims to analyze how similar AI's method of processing and perceiving visual information is to human perceptual structures and further examine whether AI can form an independent perceptual structure. In order to specifically assess whether AI possesses independent perceptual abilities rather than merely being a mechanical tool for processing data, a clear foundation for such discussions is necessary. In particular, this study investigates whether AI can recognize structural patterns and generate images in a way similar to humans.

For this purpose, this study selected the DALL-E model as a case analysis tool. DALL-E is a representative model that demonstrates multimodal processing capabilities beyond simple text input. [3] Therefore, the structural similarity between the reference images and the images generated by DALL-E is analyzed using SSIM. SSIM is a metric that quantitatively measures brightness, contrast, and structural elements between two images, allowing an evaluation of how similar the AI-generated images are to the visual information.

## 2. THEORETICAL CONSIDERATIONS

### 2.1 Perception, Consciousness and AI

AI is defined as a "program that thinks and makes judgments like a human," but the standard for "judgment like a human" is not clearly defined. The core of the debate is whether AI, through Artificial Neural Networks (ANN), which mimic human neuronal structures, can possess consciousness similar to humans. [4]

Daniel Dennett explains consciousness not as a subjective experience but as an objective information processing procedure. He argues that consciousness is not the observation and control of the world by a singular self, but rather is akin to the "multiple drafts" model, where one of the many parallel streams of information processed in the brain is selected and experienced as a narrative. [5] This emphasizes that consciousness is the result of information processing, not a fixed function of a self or a specific location in the brain.

In his book *Consciousness Explained*, Dennett also describes consciousness as an information-processing process that can be observed objectively from the outside, rather than as a subjective experience. He sees consciousness as a phenomenon arising from the brain's complex flow of information, and argues that understanding it requires objective analysis.

Exploring the relationship between consciousness and perception remains fraught with philosophical and scientific limitations. While we cannot be certain that AI has consciousness, we also cannot definitively claim that it lacks perception. This raises the need to reconsider the definition of consciousness, while still distinguishing between consciousness and perception. Contemporary philosophy maintains a human-centered perspective, but the development of AI increasingly mimics human functions. Dennett's third-person approach to consciousness suggests a method for analyzing consciousness objectively, but since consciousness also involves subjective experience, there are still many limitations. Therefore, we need a more flexible philosophical approach to understanding consciousness and perception.

### 2.2 Generative AI

Artificial Intelligence (AI) is a field of Computer Science that aims to develop technologies enabling computers to reason, learn, and act like humans. [6] Generative AI refers to the creation of new content, such as text, audio, or images, by utilizing existing content. [7] The goal of generative AI is to produce new content that is similar to or indistinguishable from the original. [8] For this purpose, deep learning-based artificial neural network technology called GAN (Generative Adversarial Networks) is used. Machine learning is one method of implementing AI, where algorithms learn from data and independently produce

results. Deep learning, a subset of machine learning, uses deep neural networks to rapidly process vast amounts of data and has driven innovation in various application fields.

The equation below describes the competition between the Generator (G) and the Discriminator (D) in a Generative Adversarial Network (GAN):

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

In Equation (1),  $D(x)$  represents the probability assigned by the Discriminator that  $x$  is real data, while  $G(z)$  refers to the data generated by the Generator using random noise  $z$ . The expectation operator,  $E$ , represents the average probability over the data distribution, with  $p_{data}(x)$  denoting the probability distribution of real data and  $p_z(z)$  describing the distribution of random noise used as input to the Generator. The first term ( $E_{x \sim p_{data}(x)} [\log D(x)]$ ) aims to maximize the probability of the Discriminator correctly classifying real data as real. The second term ( $E_{z \sim p_z(z)} [\log(1 - D(G(z)))]$ ) maximizes the probability of the Discriminator correctly identifying fake data as fake, while the Generator tries to minimize this term to deceive the Discriminator.

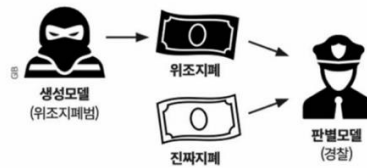


Figure 1. counterfeiter and police(www.gettyimagesbank.com)

To explain GANs, we often use the analogy of a counterfeiter and the police, as shown in Figure 2. The counterfeiter tries to make fake money that looks as close to real money as possible, while the police compare the counterfeit money with real currency to detect the fakes. As this process is repeated, the counterfeiter eventually succeeds in producing fake bills indistinguishable from the real ones.

In the field of image-generating AI, development has progressed beyond GANs to Diffusion models. According to the 2021 paper *Diffusion Models Beat GANs on Image Synthesis*, Diffusion models generate higher-quality images than GANs through a process of adding and removing noise. This process helps avoid the issue of mode collapse, where the generator model in GANs might only learn certain patterns to deceive the discriminator, resulting in reduced diversity of samples. Diffusion models, in contrast, provide better diversity and quality in the generated samples. These diffusion models are applied in popular programs like DALL-E, Stable Diffusion, and Midjourney.

## 2.2 SSIM (Structural Similarity Index Measure)

SSIM is a method used to compare the similarity between images by utilizing three factors: luminance, contrast, and structure. These elements are used for comparison because human visual perception also processes images in a similar manner.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (2)$$

The equation in Figure 3 shows that the numerator represents the luminance and structural correlation between two images, while the denominator reflects the individual luminance and contrast of each image. As a result, the SSIM value ranges between 0 and 1, with values closer to 1 indicating greater structural similarity between the two images.

In this equation(2),  $\mu_x$  and  $\mu_y$  represent the average luminance of the two images, respectively.  $\sigma_x$  and  $\sigma_y$  represent the contrast of the two images, respectively.  $\sigma_{xy}$  indicates the covariance, or the structural similarity between the two images.  $C_1$  and  $C_2$  are constants used to prevent the denominator from becoming zero during the calculation.[10]

### 3. RESEARCH METHOD

In the first stage of this study, reference images to be input into the GPT-4 prompt were selected. These images consist of five sets, each reflecting various themes and visual characteristics. The selected reference images are everyday objects, such as traffic signs, traffic lights, and smartphone icons, with clearly defined visual elements like color, shape, and composition, making them easily recognizable to people.



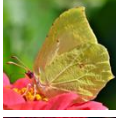




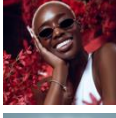
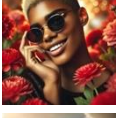
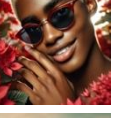
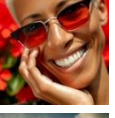
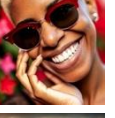
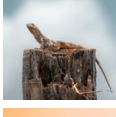

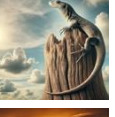







In the second stage, images were generated based on the reference images using GPT-4's DALL-E. DALL-E, developed by OpenAI, is an AI model whose core functions include text-to-image and multimodal capabilities. It creatively generates new images based on the patterns learned from the reference images, functioning in a manner similar to human visual perception processes. The generated images were systematically organized by assigning names such as A, A1, etc., to each set, to enhance the efficiency of coding tasks.

In the third stage, SSIM scores between the reference images and the generated images were calculated and visually analyzed using Visual Studio Code and Python. After loading the images through OpenCV's CV2 library, the color images were converted to grayscale to compare their structural similarity using SSIM. Since SSIM calculations are based on luminance information rather than color, the structural comparison between the images became clearer. During this process, the `calculate_ssim` function was used to evaluate the similarity between the two images.

### 4. DATA ANALYSIS RESULTS

#### 4.1 GPT-4, DALL-E Generated Images

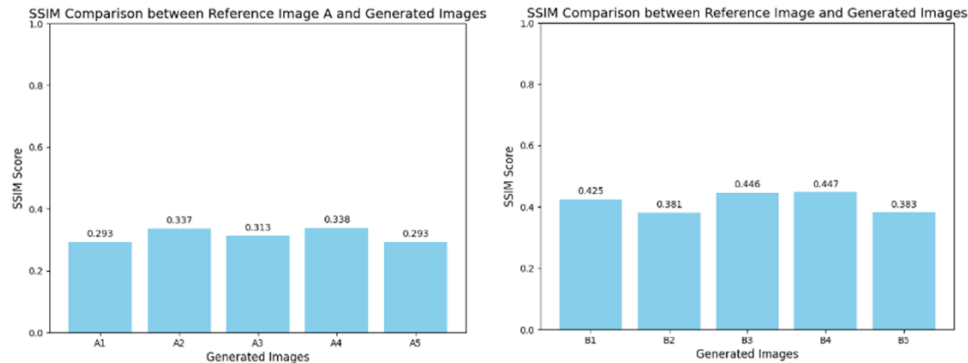
**Table 1. Generated Images**

Set	reference image	1	2	3	4	5
A						
B						
C						
D						
E						

[Table 1] shows the images generated by the DALL-E model after inputting reference images into the GPT-4 prompt. Each set, from A to E, is composed of different themes. The first image in each set is the reference image, followed by the images generated by DALL-E. The generated images are structurally similar to the reference images, but the color and contrast are more vividly expressed. This indicates that

while the DALL-E model reflects visual patterns, it tends to exaggerate color and contrast.

## 4.2 SSIM Evaluation

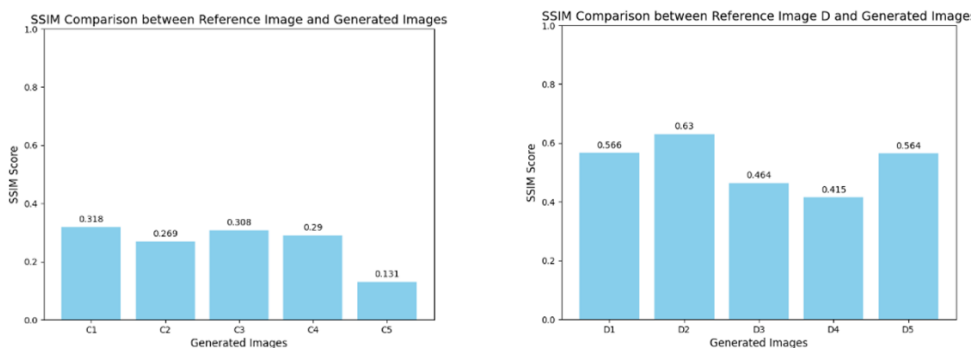


**Figure 4. Set A and Set B**

In the case of Set A, the scores range between 0.293 and 0.338. This analysis suggests that the generated images are partially similar to the reference image. However, since none of the scores exceeded 0.5, it indicates a lack of overall structural similarity. A4 in Set A showed the highest similarity with a score of 0.338, while A1 and A5 scored relatively lower at 0.293. The reference image for Set A is a close-up of the side view of a cat, with distinctive features such as ears, whiskers, and detailed fur. The generated images depicted a similar side view of the cat but with a rounder and softer impression, using brighter lighting.

For Set B, the scores ranged from 0.381 to 0.447. This range suggests that the images share a certain level of structural similarity, though they are not perfectly identical. Similarly, the fact that none of the scores reached 0.5 indicates that there are some structural differences between the generated and reference images. B4 recorded the highest SSIM score of 0.447, while B3 also showed a similar structural similarity with a score of 0.446. B2 and B5 had the lowest SSIM scores, at 0.381 and 0.383, respectively.

The reference image for Set B is a close-up photo of a butterfly sitting on a flower, viewed from the side. The shape of the butterfly's wings resembles the leaves of the plant, and the small details on the wings are visible. The generated image also depicted a close-up of a butterfly on a flower, similar to the reference image, but the butterfly's wings and body had more curved lines compared to the reference image, with details more vividly expressed. Additionally, the type of flower was different, and the direction in which the butterfly was sitting, along with the contrast, was more pronounced compared to the reference image.



**Figure 5. Set C and Set D**

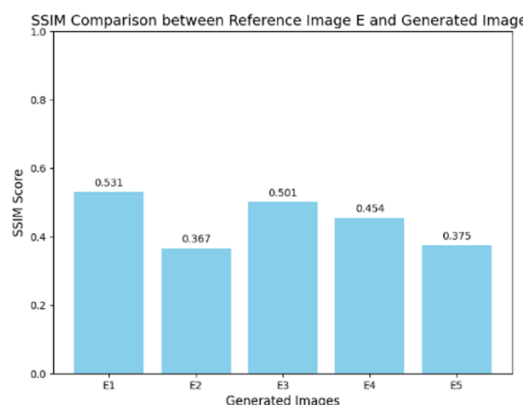
The scores for Set C range from 0.131 to 0.318. Compared to other sets, the SSIM scores are much lower, indicating that the generated images differ significantly from the reference images. C1 and C3 recorded SSIM scores of 0.318 and 0.308, respectively, making them the most similar to the reference images among

the generated images in Set C. C5 scored the lowest, with an SSIM score of 0.131, and it is particularly noticeable that C5 failed to replicate the texture of the original image.

In the reference image of Set C, the person is wearing small black sunglasses, resting their face on their hand while smiling. The background is filled with red plants. The person's face is shown in a close-up, with the composition focusing on the upper body. The generated images are somewhat similar to the reference image, but there are slight differences in the details. Particularly, C5 has a more illustrated quality compared to the other generated images, contributing to its lower similarity score.

For Set D, the scores range from 0.451 to 0.63, making it the set with the highest similarity overall. D2 achieved a score of 0.63, the highest across all sets. D1 and D5 recorded similar SSIM scores of 0.566 and 0.564, respectively. D3 scored 0.464, placing it in the middle of Set D in terms of similarity to the reference image. D4, with a score of 0.415, had the lowest similarity, indicating the least visual resemblance to the reference image.

The reference image in Set D shows a lizard sitting on a tree, with blurred clouds and the sky in the background. The rough texture of the tree's surface and the dry texture of the lizard's skin are visible. The generated images also capture many of these details, though there are differences in composition and the representation of the clouds.



**Figure 6. Set E**

The SSIM scores for E1 range between 0.531 and 0.367. E1 scored the highest similarity in Set E, with a score of 0.531, while E2 scored the lowest at 0.367.

The reference image for Set E shows two people paddleboarding on calm water, with distant mountain ranges in the background. The silhouettes of the figures are relatively small and positioned far away. The lighting gives a diffuse feel, and overall, the contrast between the figures and the background is somewhat low. The generated images also depict two people paddleboarding on water, but in many cases, the figures' genders are more distinctly portrayed, the contrast is more pronounced, and the sky is filled with more clouds, making the background appear more elaborate, resulting in a lower similarity to the reference image.

## 5. RESULT

This study used the DALL-E model to compare and analyze the images generated by AI based on reference images with those that can be recognized by humans using the Structural Similarity Index Measure (SSIM). The results showed that while the images generated by DALL-E demonstrated a certain level of structural similarity in each set, the degree of similarity varied across sets.

A comprehensive analysis of the results for each set revealed that Set A had generally low structural similarity. Set B showed higher similarity than Set A but still recorded low overall structural similarity. Set C had the lowest similarity, while Set D had the highest similarity among the five sets, with the image D2 showing very close resemblance to the reference image. Set E displayed a moderate level of similarity, but it struggled to perfectly replicate complex elements like water reflections and lighting variations.

Images with clear and simple patterns, like those in Sets D and B, recorded relatively high similarity,

suggesting that AI can effectively learn and reproduce simple forms or repetitive patterns. These images exhibited high structural consistency, allowing AI to recognize patterns well. However, for images with intricate details or delicate textures, like in Sets C and A, the AI struggled to recognize and reproduce such complex patterns, indicating that AI does not consistently perform well in processing complex visual information. Set E, where contrast and lighting played a key role, was handled relatively well by the AI, but it still revealed limitations in dealing with complex elements like natural light. This suggests that while AI is proficient at processing simple patterns, it still faces challenges when dealing with complex and dynamic visual elements.

As shown by the results from Set C, AI exhibited limitations in maintaining consistency in certain areas. Thus, it is still too early to conclude whether AI can develop an independent perception system similar to humans. For AI to reach the level of human perceptual ability, further research is needed to go beyond simple structural similarity and enable AI to learn more complex elements.

The limitations revealed in this study suggest that AI's perception system operates in a fundamentally different way from that of humans. This highlights the need for in-depth exploration into how AI can move beyond simple information processing and mimic the complex processes of human perception.

## REFERENCES

- [1] Maeil Business Newspaper, *Now It's AI-First : The Transformation of Google from Mobile-First*, <https://www.mk.co.kr/news/world/7323537>
- [2] Daily Nous, *Philosophers on GPT-3 (GPT-3 and General Intelligence by David Chalmers)*, <https://dailynous.com/2020/07/30/philosophers-gpt-3/>
- [3] P.Hana, *A Case Study On Application Of Text To Image Generator AI DALL·E*, The Treatise on The Plastic Media, Vol.26 No.1, pp.104, (February , 2023)
- [4] Takashi, Imoto, *AI Textbook*, Sungandang, PP. 26, (2023).
- [5] D. Dennett, *Solving the Mystery of Consciousness*, Chapter 4, Beyond the Cartesian Theater to the Multiple Drafts Model, Okdang, 2013
- [6] S.ye Yoon, and others, *A Study of Generative AI Trends and Applications*, The Journal of the Convergence on Culture Technology (JCCT), Vol. 10, No. 4, pp. 607-612, (July, 2024).
- [7] Y.O.Kim and others, *Generative AI Sapiens*, Saengneung Books, pp. 17, (2023)
- [8] Y.O.Kim and others, *Generative AI Sapiens*, Saengneung Books, pp. 23, (2023)
- [9] Samsung SDS, *Consept and Understanding of GAN*, <https://www.samsungsds.com/kr/insights/generative-adversarial-network-ai-2.html>
- [10] Zhou, W., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. *Image Quality Assessment: From Error Visibility to Structural Similarity*. IEEE Transactions on Image Processing. Vol. 13, Issue 4, pp. 600–612. (April 2004)