ACTA·GRAPHICA
ISSN: 0353-4707
e-ISSN: 1848-3828
ZNANSTVENI ČASOPIS ZA TISKARSTVO I GRAFIČKE KOMUNIKACIJE
JOURNAL FOR PRINTING SCIENCE AND GRAPHIC COMMUNICATIONS

Scientific article

Exploring the photorealistic capabilities of AI image generators: Stable Diffusion, DALL-E mini and dream by WOMBO

Aleksandra Bernašek Petrinec[1], Ivan Papiga[2], Katja Milković[3], Silvio Plehati[1]

[1]Tehničko veleučilište u Zagrebu, 10 000 Zagreb, Croatia

[2]Tehničko veleučilište u Zagrebu, 10 000 Zagreb, Croatia - student

[3]Europapier Adria d.o.o., 10 000 Zagreb, Croatia

* Correspondence: abernasek@tvz.hr

Abstract: *In today's digital age, artificial intelligence plays a key role in transforming the way we create and perceive art and photography. Now it is possible to generate imagined visual content with input of just a few words of text. Rapid technological advancement has led to significant progress in the field of artificial intelligence, which is increasingly being used in various aspects of human life. For recent years, AI tools have become progressively popular in academic circles, particularly for testing various algorithms in the creation of scientific and professional experiments. This paper explores the artificial intelligence generators Stable Diffusion, DALL-E mini, and Dream by WOMBO to study the hypothesis whether high-quality photorealistic images can be created with simple input, without additional iterations.*

Keywords: *artificial intelligence, photography, Stable Diffusion, DALL-E mini, Dream by WOMBO*

1. Introduction

Artificial intelligence represents the ability of a device to mimic human actions. It is a branch of computer science that develops the ability of computers to use a form of intelligence to perform specific tasks. Artificial intelligence is based on neural networks, which enable the comprehension of vast amounts of information. The function of an intelligent system involves interaction with the human world or other systems, refining, gathering knowledge based on experience, critical thinking, task-solving, contemplation, planning, and similar actions [1, 2, 3]. It is divided into strong and weak artificial intelligence, where weak artificial intelligence includes all generators devised today, while strong artificial intelligence implies intelligence equal to or stronger than human intelligence, artificial superintelligence with the potential to develop consciousness [4, 5].

To enhance this branch of science, scientists are developing new tools such as machines, devices, and applications that function and gather knowledge like humans. The intention is for the machine to develop its intelligence and educate, inform and evolve on its own. With every day technological advancements, significant improvements have been observed in the development of artificial intelligence. Scientists and companies strive to construct an informal computer architecture consisting of chips that mimic the human brain in terms of analytical thinking and intuitiveness. Such a system would ensure the formation of a real context of a large database, and such computers are currently developed to resemble the brain to obtain information for a better understanding of the human brain [2, 6, 7].

Artificial intelligence has revolutionized the way we perceive and practice the creation of art and photography. Traditionally, the processes of creating photographs required the use of a camera, while artworks were the result of physical work with artistic tools and canvases. The arrival of AI has enabled the creation of photorealistic images and artworks without the need for traditional methods. Generating images is becoming easier, where just entering a few words of text is enough to create a wanted image.

The testing conducted in this paper depends on many parameters, and the results are usually unpredictable due to the imperfection and underdevelopment of the entire process. There are many artificial intelligence generative programs, all still in the developmental stages, with this branch of science always being upgraded with new knowledge and information. There are free versions, which are attractive due to their availability, while the paid versions have the advantage of output quality, more precision, created databases that are controlled, maintained, and contain more data for creating samples, and thus create more accurate results [8, 9].

Due to its popularity and intrigue, this technology attracts many scientists from all branches, as well as people who like to experiment and develop their own ideas. The system is evolving at an exponential pace and with artificial intelligence's great capacity to learn, analyze and generate vast amounts of content. This has resulted in significant advancements in all areas, including photography generation, which is the subject of study in this paper [10, 11].

2. Experimental part

The rapid technological advancement in recent years has led to significant development in the field of artificial intelligence, which is increasingly integrated into different aspects of human life. Academic circles are particularly interested in using artificial intelligence tools to test various algorithms in scientific and professional experiments. As the number of artificial intelligence users grows, so does the number of platforms that enable artificial intelligence generation [12, 13, 14]. For the purposes of this study, Stable Diffusion, DALL-E mini and Dream by WOMBO programs were used, with Stable Diffusion and DALL-E mini being free, while Dream by WOMBO has a free trial period for testing.

Stable Diffusion is a so-called "deep learning" and "text-to-image" artificial intelligence model based on the diffusion model, or a latent variable generative model. It has high-quality, realistic images and tries to maintain features like edges and textures [12]. It is primarily used for generating images containing many details conditioned solely on textual descriptions. This platform generally produces good results and includes tools for further refinement of results. DALL-E mini is an open-source artificial intelligence developed by programmer Boris Dayma. This model generate detailed and realistic images. Since the model is trained on unfiltered internet data, there is a possibility of obtaining results containing various stereotypical representations against minority groups [12]. However, considering the limited resources, the platform still yields solid results. Dream by WOMBO is an artificial intelligence generator of artistic works. It can turn any input into an artistic image or photograph. It creates unique and abstract artwork, but the detail clarity is not main focus [12]. Unlike Stable Diffusion and DALL-E mini platforms, Dream by WOMBO is an application designed for creating artificial intelligence on mobile devices. It is very user-friendly and has many image editing features as well as various styles in the form of templates [15, 16, 17].

This research provides a deeper insight into the capabilities of artificial intelligence in generating photorealistic images and explores the application of three different generative models: Stable Diffusion, DALL-E mini, and Dream by WOMBO [12, 13]. Through the experimental part of

the study, we analyzed the performance of each model in creating photographic images based on input data. The aim of the study was directed towards generating photorealistic images of requested segments through simple inputs. Three different inputs were designed for testing, which were used as input data for three different platforms. There were no iterations and results that were shown are made from just one input. Each description consisted of three simple inputs in English, as follows:

First Test: "Girl in white standing in nature."
Second Test: "Photo of a guy in the woods with the axe."
Third Test: "Girl, red hair, blue eyes, red lips."

2.1. First Test

Example 1:

Program: Stable Diffusion.
Input: "Girl in white standing in nature."
Style: photorealistic

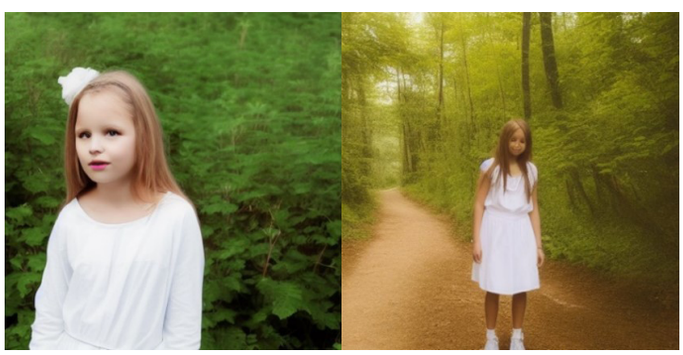

Figure 1a and 1b First Test: Example 1, Stable Diffusion program, untrained model, input: "Girl in white standing in nature."

We notice poor performance of the faces in both images of Example 1. In Figure 1a, we observe oddly positioned eyes on the face as well as lips that appear plastic. Meanwhile, in Figure 1b, the face is completely distorted. Due to the poorly executed segments, it is easy to discern that these depictions are the result of artificial intelligence. Although the nature, clothing, and hair are decently executed, the overall impression of the images are poor.

Example 2:

Program: DALL-E mini
Input: "Girl in white standing in nature."
Style: Photorealistic

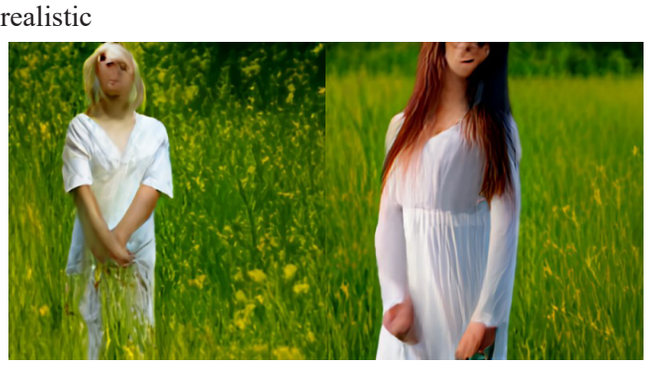

Figure 2a and 2b First Test Example 2. Program DALL-E mini, untrained model, input: "Girl in white standing in nature."

In example number 2, both representations in Figure 2a and 2b are blurred and hardly recognizable, indicating deficiencies in the image generation process. The rendering of the human body is highly unnatural, as is the face, which is almost entirely unrecognizable due to distortion and unclear contours. Additionally, we notice that the nature in this case is poorly generated, manifested by a lack of detail and ambiguity in shaping environmental elements. Instead of a realistic depiction, nature resembles a vague sketch, which is an unusual occurrence considering that natural scenes are typically depicted with satisfactory amount of detail in other instances.

Example 3:

Program: Dream by WOMBO
Input: "Girl in white standing in nature."
Style: photorealistic

Figure 3a and 3b First test Example 3. Program Dream by WOMBO, untrained model, input: "Girl in white standing in nature."

In the third example, we observe a moderate performance in rendering the requested input. Nature is generated very well, with details and textures that provide an authentic impression. However, the face exhibits classic faults characteristic of algorithmic models, such as oddly positioned eyes or oversized and irregularly spaced teeth and other facial features. Despite these shortcomings, the extremities and torso are relatively well executed, contributing to the overall impression of the image.

When comparing all three examples from the first test, we notice that only example 2 exhibits unrealistically rendered nature, while natural elements are relatively well depicted in the other examples. In all cases, deficiencies in facial rendering are evident, indicating attempts by artificial intelligence rather than a photorealistic image. The general problems across all image generation programs lie in reproducing human faces, resulting in incomplete or inaccurate portrayal of facial features. These shortcomings often reveal that the images are created through algorithmic processes rather than being presented as authentic photographic works.

2.2. Second Testing

Example 1:

Program: Stable Diffusion
Input: "Photo of a guy in the woods with the axe."
Style: Photorealistic

Figure 4a and 4b Second Testing: Example 1. Stable Diffusion Program, untrained model, input: "Photo of a guy in the woods with the axe."

In the second testing, for example 1, we can observe an average execution of the natural environment. While the human body has all the necessary segments, we notice certain issues with the rendering of the arms, which are deformed and irregularly shaped in both cases. Additionally, the segment held by the subject is poorly executed and difficult to recognize due to problems with the arms.

Example 2:

Program: DALL-E mini
Input: "Photo of a guy in the woods with the axe."
Style: Photorealistic

Figure 5a and 5b Second Testing: Example 2. DALL-E mini Program, untrained model, input: "Photo of a guy in the woods with the axe."

Example 2 in the second testing showcases two extremely poorly executed renditions of the requested input. In both images, the subject's entire body is distorted, significantly compromising recognizability and aesthetics. In image 5a, the deformation is so pronounced that the subject is hardly recognizable, further complicating the interpretation of the scene. The elements held by the subjects in images 5a and 5b do not resemble the requested object, creating additional confusion and hindering the assumption of what the depiction represents.

Example 3:

Program: Dream by WOMBO
Input: "Photo of a guy in the woods with the axe."
Style: Photorealistic

Figure 6a and 6b Second Testing: Example 3. Dream by WOMBO Program, untrained model, input: "Photo of a guy in the woods with the axe."

In the second testing of example 3, we notice an improved rendition of the body, which is created almost flawlessly, representing a significant advancement compared to previous examples. However, problems arise with facial rendering, reducing the overall quality in images 6a and 6b. Both images in this example have compelling issues in the eye area, where irregularities and deficiencies in detail are evident.

All three examples from the second testing encounter difficulties in rendering the additional object that the subject holds, with none of them clearly depicting an axe. This lack clearly indicates the algorithms' challenges in interpreting specific description requirements and their ability to effectively reproduce objects according to those descriptions. Additionally, a significant problem is the generation of how the subject holds the object. In all examples, there is deformation or inaccurate depiction of the arms, further complicating identification and understanding of the scene. Often, hands are missing or depicted with additional or lack of fingers, resulting in unrealistic or even abstract representations of the subject. The combination of these issues makes it difficult to achieve an authentic depiction and recognize the requested content.

It is particularly significant to note that none of the tested generators successfully executed the task according to the input "Photo of a guy in the woods with the axe." This is a crucial indicator that current algorithms are unable to adequately interpret complex description requirements and successfully reproduce scenes involving specific objects or situations within just one iteration.

2.3. Third Testing

Example 1:

Program: Stable Diffusion
Input: "Girl, red hair, blue eyes, red lips"
Style: Photorealistic

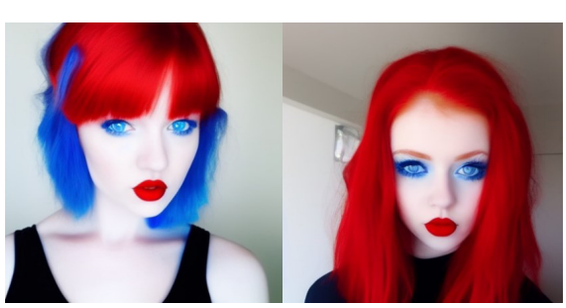

Figure 7a and 7b Third Testing: Example 1. Stable Diffusion Program, trained model, input: "Girl, red hair, blue eyes, red lips"

In the first example of the third testing, the focus is on the human face, where we notice an improved execution of this otherwise complex segment. The subject's face is depicted realistically but with almost flawless porcelain-like skin reminiscent of a doll. Although such an effect may result from image retouching in some of the known graphic programs. The lips and nose are quite well-rendered, with nicely shaped details in both images. However, the true challenge lies in depicting the eyes, which represent a sort of focal point of the image, as well as in rendering the subject's hair. The eyes are shown with unrealistic colors and slight blurriness, revealing the artistic intervention of artificial intelligence and indicating that the image is generated rather than authentic. The subject's hair also exhibits some signs of unnaturalness in color and texture. These shortcomings, although less obvious, still indicate the challenges artificial intelligence faces in reproducing details and authenticity in generated images.

Example 2:

Program: DALL-E mini
Input: "Girl, red hair, blue eyes, red lips"
Style: Photorealistic

Figure 8a and 8b Third Testing: Example 2. DALL-E mini Program, trained model, input: "Girl, red hair, blue eyes, red lips"

In the second example of the third testing, the obtained results are unsatisfactory in almost every segment. The eyes, lips, nose, and even the face exhibit deformations and deficiencies in detail. The subject's face in image 8b appears distorted and hardly recognizable, significantly undermining the overall impression of the image. The hair appears artificial and inauthentic, almost like a drawing. The only segment that can be considered correctly rendered is the facial skin, considering example 1 of the third testing, where the face had a porcelain smoothness. However, even that segment is not entirely perfect, lacking naturalness in skin texture and tones.

Example 3:

Program: Dream by WOMBO
Input: "Girl, red hair, blue eyes, red lips"
Style: Photorealistic

Figure 9a and 9b Third Testing: Example 3. Dream by WOMBO Program, trained model, input: "Girl, red hair, blue eyes, red lips"

Of all the generated examples in the third testing, only example 3 matches the quality of a real depiction considering the overall composition of the image. However, when the segments are viewed separately, elements that are poorly executed are revealed, diminishing the overall impression of authenticity. The eyes are blurry and do not have the natural color of human blue eyes. The facial skin in image 9b appears pale and anemic, and the lips appear drawn, lacking detail and texture.

All three-testing resulted in outcomes that had unrealistic segments. They displayed the most difficulties in facial rendering, indicating challenges that image-generating algorithms encounter, especially when generating facial features.

The best results are evident in the first testing, image 3a of example 3, where the nature and the girl are generated quite realistically. Similarly, in the third testing, example 3, image 9a, with the girl with red hair, looks surprisingly real. This demonstrates that, although there are shortcomings and challenges in reproducing details, algorithms can achieve a surprisingly high level of authenticity in certain situations without iterations.

The shortcomings in individual segments indicate the need for further improvement in image-generating algorithms to achieve greater authenticity and precision in reproducing details. The complexity of the scene and the specific requirements of the description may have resulted in incomplete or inadequate interpretation by the algorithm. The lack of details and the vagueness in depiction of nature may be due to insufficient training data or a lack of diversity in the dataset on which the model was trained.

3. Conclusion

In all examples, there are issues with generating certain image segments. The nature and environment are mostly beautifully and realistically rendered, although in rare cases, the landscape appears unreal, almost like a drawing. The human body is a segment that is partially well-executed. The shape of the torso, head, legs, and arms is usually solid, but problems arise when generating facial features and hands and fingers. Irregularities in facial rendering are most evident in the eye and mouth area. The eyes remain blurred and "plastic," while the largest problem lies in the mouth and depiction of teeth, which can be oversized or oddly positioned in the oral cavity. In most cases, the hair is nicely and realistically rendered, showing shifts and efforts made by the generators to make the images and all segments on them as realistic as possible. When studying the examples, it's important to consider that all tested software solutions are still in development, with much

smaller sample bases for pattern creation compared to paid ones. Additionally, the examples were not created on trained models, which, of course, increases the scope for generating lower-quality images and unwanted content. Since each example was obtained from only one iteration, and no interventions were made on any of the samples, the results are not so poor. It is assumed that through a certain number of iterations, some examples would achieve satisfactory image quality with realistic representations.

In the context of research in the field of artificial intelligence, such shortcomings are important as they emphasize the need for further development of algorithms and techniques to improve the quality of generated images. Analyzing these shortcomings provides valuable insights into the challenges faced by artificial intelligence models in interpreting and creating visual content, which can serve as a foundation for future research and improvement of image generation technology.

There are differences in the tested artificial intelligence generators, including the information they use as a basis to create and reproduce results. However, much investment is being made in this branch of science, and predictions suggest that it will soon be possible to generate images without errors and with very realistic representations. Currently, all three generators still have considerable room for improvement. However, their application will have a wide range of applications in fields such as art, design, digital production, and education.

This research provides a foundation for further testing in the field of artificial intelligence and its application in creating photorealistic images. The continued development of generative models could lead to new innovations in art and technology, opening a new possibility for creative expression and technical advancement.

4. References

[1] Prister, V. (2019) Umjetna inteligencija Media, Culture and Public Relations, Vol. 10, No. 1, 67-72, ISSN 1333-6371, https://doi.org/10.32914/mcpr.10.1.7

[2] Turning, A. M. (1950) I.—computing machinery and intelligence Mind, Volume LIX, Issue 236, October 1950, Pages 433–460, https://doi.org/10.1093/mind/LIX.236.433

[3] Caruana R. Multitask learning (1997) Machine learning. Volume 28(1):41–75

[4] Umjetna inteligencija, Hrvatska enciklopedija, dostupno na: https://www.enciklopedija.hr/clanak/umjetna-inteligencija, (2024-01-29)

[5] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[6] Xu, Y.; Liu, X.; Cao, x.; Huang, C.; Liu, E.; Qian, S.; Liu, X.; Wu, Y.; Dong, F.; Qiu, C.; Qiu, J.; Hua, K.; Su, W.; Wu, J.; Xu, H.; Han, Y.; Fu C.; Yin, M.; Liu, M.; Roepman, R.; Dietmann, S.; Virta, M.; Kengara, F.; Zhang, Z.; Zhang, l.; Zhao, T.; Dai, J.; Yang, J.; Lan, L.; Luo, M.; Liu, Z.; An, T.; Zhang, B.; He, X.; Cong, S.; Liu, X.; Zhang, W.; P. Lewis, J.; M. Tiedje, J.; Wang, Q.; An, Z.; Wang, F.; Zhang, L.; Huang, T.; Lu, C.; Cai, C.; Wang, F. J.; Zhang, (2021) Artificial intelligence: A powerful paradigm for scientific research, The Inovation, Volume 2, Issue 4, https://doi.org/10.1016/j.xinn.2021.100179

[7] Anić, N.; Anić, P. Umjetna inteligencija kao segment strategije (2020) National security and the future Vol. 21, No. 3, 117-138, https://doi.org/10.37458/nstf.21.3.4

[8] Europski pristup umjetnoj inteligenciji, dostupno na: https://digital-strategy.ec.europa.eu/hr/policies/european-approach-artificial-intelligence, (2024-01-30)

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017) Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

[10] Wu, C., Visual chatgpt: talking, drawing and editing with visual foundation models. (2023.) arXiv preprint arXiv:2303.04671

[11] Ljubić Klemše, N. (2023) Umjetna inteligencija - razvoj i primjena, priručnik za početno učenje Bjelovarski učitelj : časopis za odgoj i obrazovanje, Vol. 28 No. 1-3
[12] Shetty, M.; Sheikh, H.; Sharma, P.; Shrivastava, K.; Gonsalves J. (2024) A Comparative Analysis of AI Image Generation Models: Stable Diffusion, Dall-e and Dream by WOMBO, International journal of creative research thought, Vol. 12, Issue 2, ISSN: 2320 – 2882
[13] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2021) DALL-E: Creating Images from Text. https://doi.org/10.48550/arXiv.2102.12092
[14] Wang, T., Zhang, R., Zhu, J. Y., Zhang, X., Wang, C., & Qi, H. (2021) High-Fidelity GAN Inversion: Learning to Generate Images from Unseen Views. https://doi.org/10.48550/arXiv.2109.06590
[15] Joynt, V.; Cooper, J.; Bhargava, N.; Vu, K.; Kwon, O H.; R. Allen, T.; Verma, A.; I. Radaideh, M.; A Comparative Analysis of Text-to-Image Generative AI Models in Scientific Contexts: A Case Study on Nuclear Power, dostupno na: https://arxiv.org/html/2312.01180v1, (2024-02-14)
[16] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022) High-Resolution Image Synthesis with Latent Diffusion Models. https://doi.org/10.48550/arXiv.2112.10752
[17] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. https://doi.org/10.48550/arXiv.1611.07004