A comparison of the most popular AI text-to-image generation applications

Tom Jacobs and James Boogaard

December 15th 2022
Module: SEAR
Venlo, Limburg, Netherlands

Abstract

Following the increase in popularity of artificial intelligence during the year of 2022, there are now many text-to-image generation applications out there. This is great for the field of artificial intelligence as a whole, however it has made knowing which application is best quite hard for artists who are new to text-to-image generation. This paper looks at four of the most popular generators, with the aim to find out which one is the best for an inexperienced artist.

Contents

List of Figures

1 Introduction

1.1 Context and background

During the year of 2022 there has been a massive increase in the popularity and use of text-to-image generators. Text-to-image generators allow people to provide a text description which the generator then uses to produce images matching the description as closely as possible (Wu 2022). This increase in popularity has led to many different text-to-image applications being made, which on the one hand has given many more options to artists (people who use a text-to-image generator to create and refine an image). However, in our opinion it has also made the landscape more difficult to navigate because of its abundance of choice. This is illustrated by this article which shows that you would have to go through all of the models to find which is the best for each user as this author and most others all base their judgment on their own opinion (Brandon 2022). This paper makes an effort to try and alleviate this problem by picking four of the most used text-to-image generators and comparing them based on their accuracy and process in an objective way. This allows us to help people gain some perspective on which text-to-image generator is best for their needs.
A research paper was published that makes use of surveys to compare models for text-to-image generations (Agnese et al. 2020). This paper aimed to show which text-to-image technique yielded the most realistic results in a wide variety of categories. One issue with this paper was that it was written before the explosion of consumer text-to-image generators happened in 2022, which is why we will instead focus on modern consumer-grade generators which are available today.

1.2 Problem and hypothesis

Our goal during the course of this research paper is to find out what the best text-to-image generator is out of the four we chose, in the case of an artist who is not experienced in the world of text-to-image generation. Our hope is that artists who are new to text-to-image generation will be able to make an educated decision on which generator is best for them on the basis of this paper.
Another question we take a look at is what makes an application good for an inexperienced artist. In answering this question, we will consider: the tools available to an artist, the difficulty in using these tools, the accuracy in generating images, the learning curve of an application and the speed with which the application can generate images. With these points of consideration, it should be possible to score each application on their ease of use, and on their future potential if the artist were to stick with that application.
The text-to-image generators we will be testing during the course of this paper are: Stable diffusion, Midjourney, DallE and Dream by Wombo. We came to these models after looking at several applications in Google Trends (Google 2022). This gave us a good idea of what the landscape looked like, and which applications artists would be likely to choose. You may notice below in Figure 1 that Dream by Wombo is not that popular anymore, however

it was one of the first openly accessible text-to-image applications out there.

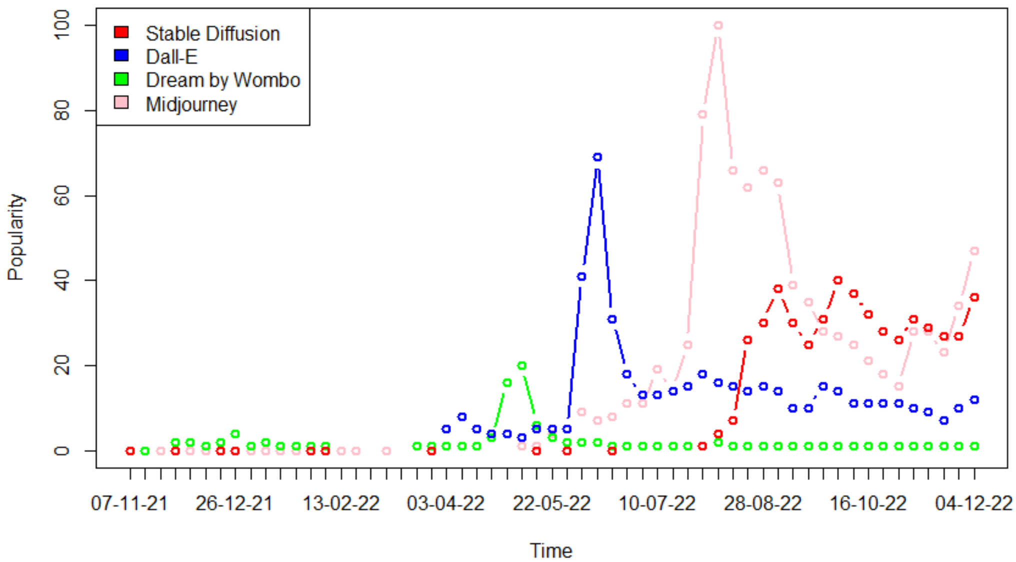

Figure 1: Results from Google Trends showing the popularity of different text-to-image applications

We expect the best text-to-image generator out of the four we chose to be stable diffusion. This is due to the fact that stable diffusion offers its users more tools and techniques like inpainting. Inpainting is an AI technique which allows a user to regenerate a missing part of an image (Jay 2022). This allows the user more freedom to fine-tune their result. Given enough time this will yield a more accurate result.

2 Methods

2.1 Data collection

Our goal during the course of this research paper is to find out what the best text to image generator is, out of the four we chose, for artists who are inexperienced in AI art generation. In order to find out which is the best one, we will first look at the accuracy with which the generator can create images. These images will then be scored using an API developed by DeepAI which can calculate image similarity (DeepAI 2022).

2.1.1 Generating images

We have chosen an image that we find sufficiently covers all the challenging parts of recreating an image using ai art generation such as reflections and complex composition (see Figure 2 below). Subject 1 has experience with Midjourney and Subject 2 has experience with Stable diffusion and neither of them have any experience in the other applications. Both Subject 1 and Subject 2 then use each of the four text-to-image generation applications to try and recreate that image within a given time frame (15 minutes per application). After each iteration we take the generated image and add it to a list of the iterations for that application.

Figure 2: The image the subjects tried to replicate (Pexels 2022)

2.1.2 Scoring generated images

We then run the set of generated images, together with the target image through an image similarity checking API that can give you a numerical value for how close one image is to another (DeepAI 2022). This way we can see how close each iteration is to the image we are trying to recreate. It is able to score the images based on a color histogram and colour location, this is supposed to be resistant to transformations such as scaling and changes in lighting (DeepAI 2018). The number of iterations each application is able to create then gives us insight into how fast the application is and how many iterations you have to do to achieve a certain result. We also gain an idea of how easy to use each application is through this process.

In order to compare the image generation applications effectively we observe a multitude of factors that will indicate the strengths and weaknesses of that application. These factors are: the amount of iterations each subject can generate while trying to create the image within the 15 minute time limit, how easily we were able to access the application and its functionality, how much functionality each application offers and their effects on the final results of our experiment and finally how close each iteration is to the image we are trying to replicate using the ai image comparing API's unit of measurement (the unit is called distance). The closer to zero the distance is, the more similar the images are to each other. A distance of zero means that the two images that are being compared are exactly the same.

2.2 Visualizing data

Once the data is gathered, we can create graphs to compare the applications more easily.

In these graphs the x value is the iteration number for that application and the y value is the result (the distance) for each iteration on the road to trying to recreate an image using these applications. This will allow us to compare the timelines of each subject's attempt at recreating the original image and gives us insight into each applications process of text-to-image generation. The intervening variable in this case would be the amount of experience each subject has with using these applications.

We also want to check each applications average result across both Subject 1 and Subject 2 in order to compare how close each application gets to the original image on average. This will show us if there are any large differences between the applications as a whole, in terms of the accuracy of the images the applications were able to create. If there is, this will play a role in choosing which is the best application.

The last set of important data we want to visually represent is the min and max of each application and how far spread the results of each application are in order to see how reliable each application is.

3 Results

We make use of 4 different web applications in order to get our results that are presented below. Firstly, Stable diffusion which is an open source text-to-image generator and offers a lot of functionality (Stability AI, 2022). Second, Dall-E has a great ease of use, especially for how good the generated images look (OpenAI 2022). Third, Midjourney is the most popular out of our selection and is very similar to Dall-E in our opinion (Midjourney 2022). Finally, Dream by Wombo is the last application and is the oldest out of the applications we are looking at (WOMBO Studios, 2021).

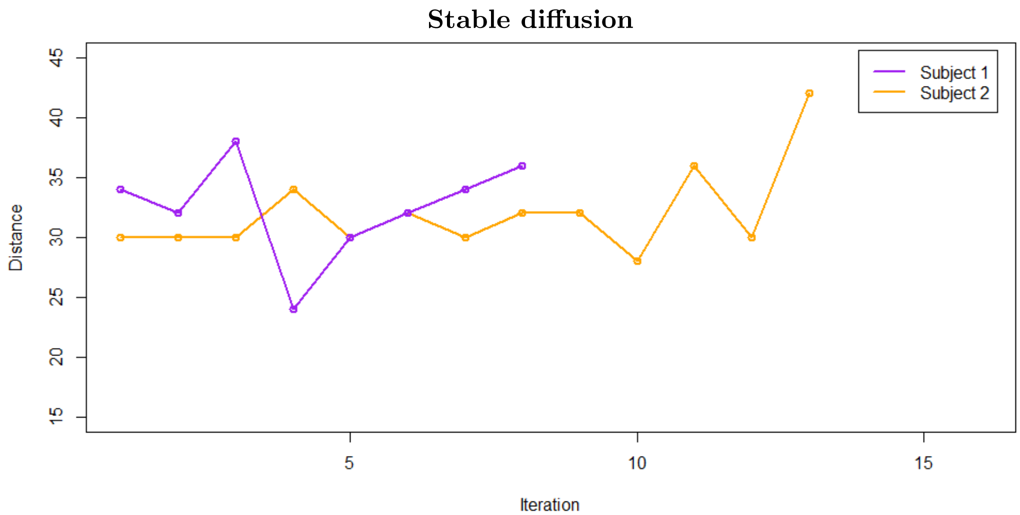

Figure 3: Line graph showing distance of each iteration for both subjects 1 and 2 for Stable diffusion

3.1 Explaining Figure 3 and 4

What is interesting in both Figure 3 and 4 is that the discrepancy between the two subjects in both applications is larger than that of any of the other applications we tested. This is seen with subject 2 having nearly double the number of iterations when compared to subject 1 in Figure 3 and then subject 1 having more iterations than subject 2 in Figure 4. Another point of interest is that the application in Figure 3 has as a small number of iterations when compared to the other applications.

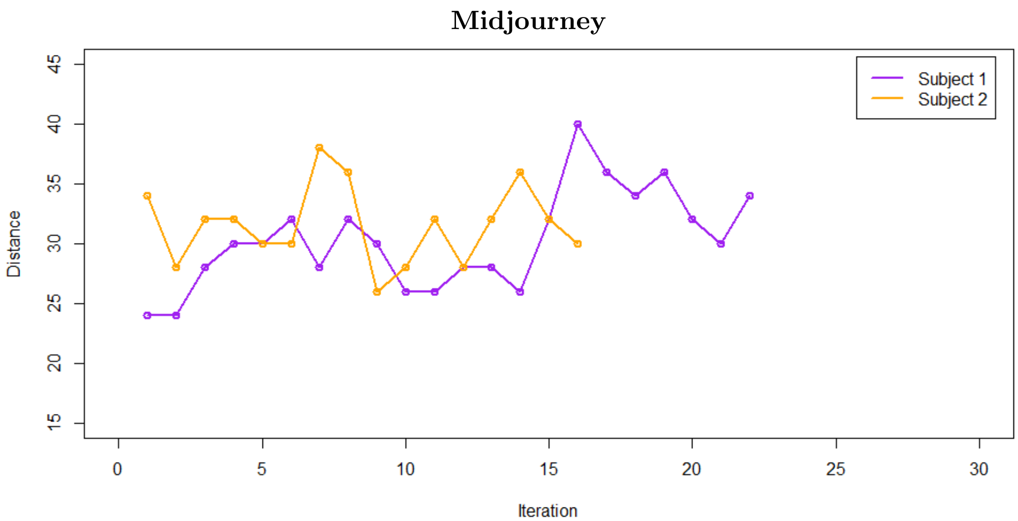

Figure 4: Line graph showing distance of each iteration for both subjects 1 and 2 for Midjourney

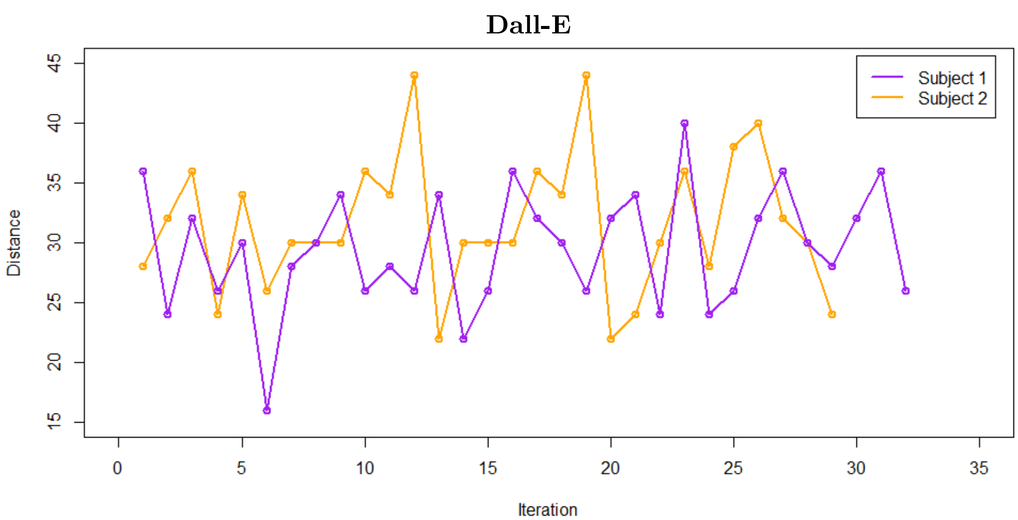

Figure 5: Line graph showing distance of each iteration for both subjects 1 and 2 for Dall-E

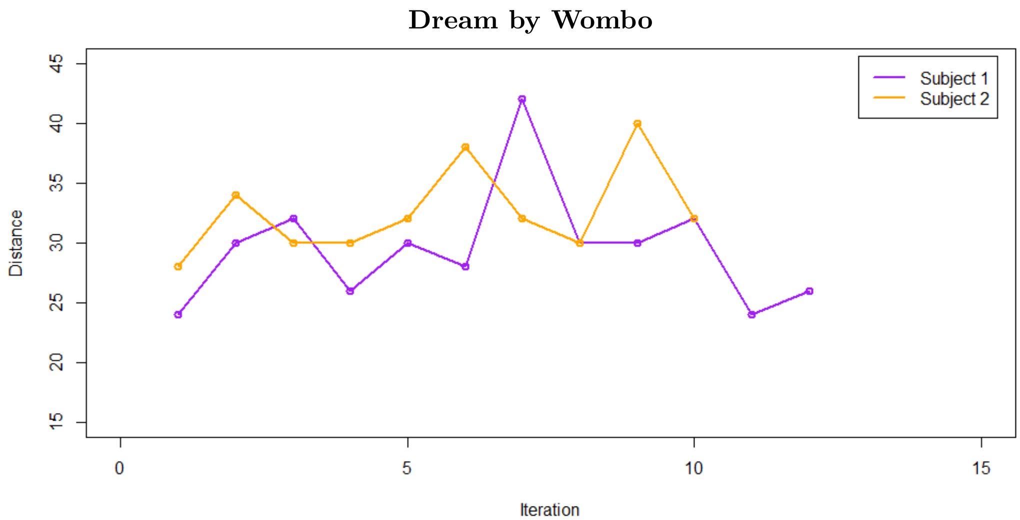

Figure 6: Line graph showing distance of each iteration for both subjects 1 and 2 for Dream by Wombo

3.2 Explaining Figure 5 and 6

In Figure 5 we see that Dall-E has the most iterations out of any of the other applications. When comparing the number of iterations that each subject has in both Figure 3 for Stable diffusion and Figure 4 for Midjourney, the two applications are very far apart. In contrast the number of iterations between each subject is very close together in Figure 5 for Dall-E and Figure 6 for Dream by Wombo.

3.3 Explaining Figure 7

Figure 7 shows that the average distances of all the applications from both subjects 1 and 2 are very close. When comparing the applications Dall-E has the least average distance and stable diffusion has the most average distance but again only by a slight margin.

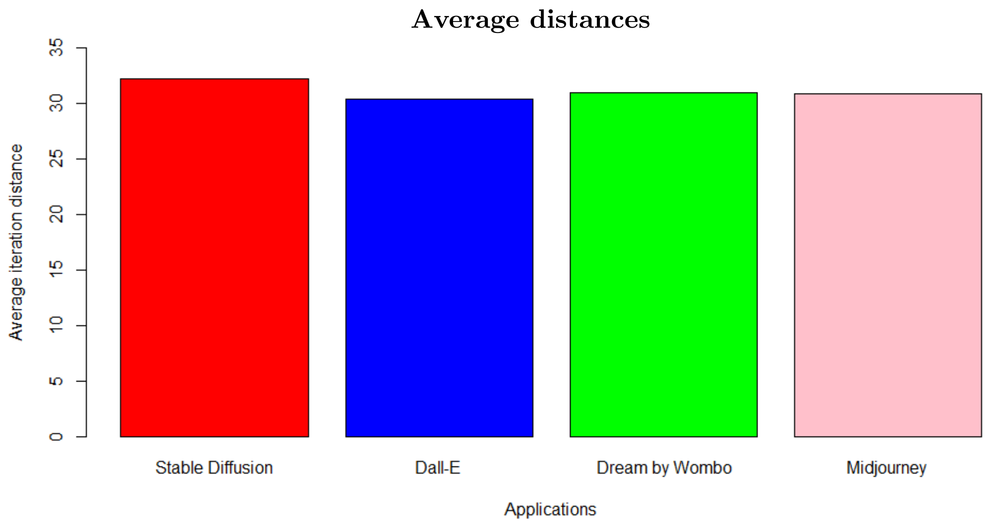

Figure 7: Bargraph showing the average distances of all iterations per application for both subjects 1 and 2

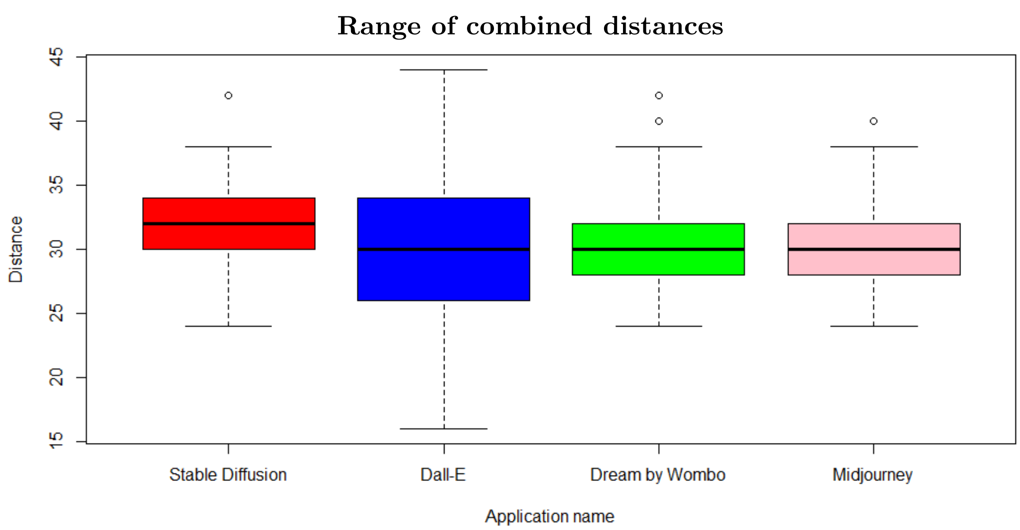

Figure 8: Boxplot showing range of combined distances for both subjects 1 and 2 per application

3.4 Explaining figure 8

In Figure 8 there are two important points of interest. Firstly, is that Dall-E has a massive range between the minimum and maximum distances when compared to the other applications, as well as having the largest interquartile range. This means the spread of the distances is also the largest out of any of the other applications. Second is that Stable Diffusion has the largest median distance.

4 Discussion

4.1 Interpretation

In order to find the best application out of the four we chose we needed to see what actually makes a text-to-image generation application good. This was done by an analysing Figures 3 through 7. The applications shown in Figures 3 and 4 are applications that both subjects have experience in (subject 1 has experience with Midjouney and subject 2 has experience with Stable diffusion). What makes Figure 3 and 4 important is that both the graphs show the difference that having experience in an application makes, and that difference is only in the number of iterations the user is able to generate.

In Figure 5 and 6 we then see two applications that neither subject 1 nor subject 2 have any prior experience in. Using comparative analysis we can then compare both the graphs that the users have experience in (being Figure 3 representing Stable diffusion and 4 representing Midjourney) and the graphs that the subjects don't have experience in (being Figure 5 representing Dall-E and 6 representing Dream by Wombo) and we can see more clearly that when both subjects have the same amount of experience with an applications the amount of iterations the subjects are able to generate is noticeably closer together compared to when there is a discrepancy in experience level.

When we take a look at Figure 7, we can see that even though it is the combinations of the averages of all iterations across both subject 1 and 2 per application, the differences between the bars is very small. This can be attributed to the nature of text-to-image generation being that of a random series of attempts until the user reaches a point where they are happy with the result. This is further proven when you look at any of the patterns for Figures 3 through 6 and how the distance for each application's iterations follows no discernible pattern but rather jumps around sometimes even ending up further away from the original image than some of the earlier iterations. An example of this can be seen in Figure 4 with subject 1.

Finally in Figure 8 we see the spreads of each of the applications. This is important as when generating an image using a text-to-image generation application, the user relies a lot on randomness and luck. Thus, having a larger spread can actually be a good thing because the more spread you have the more varied your results will be which means while you may get worse results sometimes it also means that you may get better results.

4.2 Answering secondary research question

In order to answer the secondary research question, we looked at how different aspects of the applications affected the resulting images that were generated and came to the conclusion that the amount of functionality does not play a role in the quality of generated images with inexperienced users which makes sense as they may struggle to use said functionality in the first place. This can be seen when you take into consideration that stable diffusion has the most functionality but has the worst average iteration quality across both users. We also looked at how easy the application was to gain access to and came to the conclusion that every one of them was very easy to use and gain access to as they were all just on websites, so that also became an irrelevant comparison factor. The accuracy of each of the applications also became less relevant as they were very close to each other, however this doesn't mean that factors relating to accuracy are no longer relevant. This can be seen with the spread of the distances, because in text-to-image generation the user will benefit from more varied results and thus an application that can supply that is usually better for an inexperienced user. In terms of looking at learning curve when judging an application for an inexperienced learner this also became less relevant. We came to this conclusion after looking at Figure 3 to 5 (line graphs) and noticing that having experience doesn't seem to impact image quality but rather just how many images you are able to create.

With all this information we can now say that an application's value to an inexperienced user is not about how good the AI generation technique is nor is it about how much functionality it supplies its user. This is because AI text-to-image generation is very random and relies a lot on luck with our current technology even for users that are experienced with these applications. What we do know is that currently if all the applications rely on luck to some degree it is then better if the application is able to give many iterations because the more attempts you are able to make, the faster you are more likely to get the result you are looking for provided the spread of the results is large enough.

4.3 Answering primary research question

Our primary research question can then be answered now that we know what makes an AI text-to-image application good for an inexperienced user and that is how fast the application is able to generate iterations. With this in mind we can then say that the best application out of the four we chose is Dall-E as it is able to generate the most iterations. On a side note, our experience with it also led to us believing that it offers a very streamline user experience, is easy to access and intuitive to use. Finally, it also provides the most spread out of any of the applications.

4.4 Further research

The ways in which our research can be carried out in an improved manner are to firstly use a bigger sample size in order make the results of the research more convincing. This can be done by using more text-to-image applications and trying to recreate more than just one picture.

Secondly you can use many different ways to compare images rather than just one in order to make the results more accurate. This can be done with surveys where you ask people which image is closer to a human's perspective or by using not just one application for image comparison.

References

Agnese, J. et al. (2020). "A survey and taxonomy of adversarial neural networks for text-to-image synthesis". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.4, e1345. (Visited on 11/23/2022).

Brandon, S. (Dec. 23, 2022). *The Best AI Art Generators (I've Personally Tried Them All!)* URL: `https://samanthabrandon.com/best-ai-art-generators` (visited on 01/23/2022).

DeepAI (Mar. 31, 2018). *A Modified Image Comparison Algorithm Using Histogram Features.* URL: `https://arxiv.org/ftp/arxiv/papers/1804/1804.01142.pdf` (visited on 12/03/2022).

— (2022). *Image Similarity API.* URL: `https://deepai.org/machine-learning-model/image-similarity` (visited on 11/18/2022).

Google (Nov. 10, 2022). *Google Trends.* URL: `https://www.google.com/trends` (visited on 11/10/2022).

Jay (Nov. 6, 2022). *Stable Diffusion Inpainting Tutorial.* URL: `https://hashdork.com/stable-diffusion-inpainting-tutorial/` (visited on 01/26/2023).

Midjourney (2022). *Midjourney.* URL: `https://midjourney.com/home/` (visited on 10/08/2022).

OpenAI (2022). *Dall-E.* URL: `https://openai.com/dall-e-2/` (visited on 11/10/2022).

Pexels (2022). URL: `https://www.pexels.com/` (visited on 11/10/2022).

Stability AI, (2022). *Stable Diffusion.* URL: `https://stablediffusionweb.com/` (visited on 11/14/2022).

WOMBO Studios, (2021). *Dream by Wombo.* URL: `https://dream.ai/` (visited on 11/14/2022).

Wu, Y. (June 22, 2022). *How AI creates photorealistic images from text.* URL: `https://blog.google/technology/research/how-ai-creates-photorealistic-images-from-text/` (visited on 01/23/2022).