TÉCNICO LISBOA

A dataset for quality assessment of AI generated content

Integrated Project in Telecommunications and Informatics Engineering

Report

Lourenço Completo
Pedro Claro

Telecommunications and Informatics Engineering

Advisor: Prof. João Ascenso

July 2023

Acknowledgments

We would like to dedicate this chapter to express our gratitude to all the individuals and institutions who contributed to the successful completion of this Integrated Project.

First and foremost, we would like to thank our project advisor, Professor João Ascenso. His expertise and support were crucial to the execution of this project. We appreciate his experienced guidance, patience, and encouragement throughout the entire research process. His valuable suggestions and revisions enhanced the quality of this work and contributed to our academic and professional growth.

We would also like to thank the Instituto Superior Técnico for providing us with the opportunity to carry out this Integrated Project. We are grateful for the available resources, which were essential for the development of this research.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Context

Artificial intelligence (AI) image generation technology has witnessed significant advancements over the past years, driven by a rapid growth of deep learning techniques and the increasing availability of vast amount of both computing power and visual data. This improvements are a consequence of the recent development of important AI-image generation techniques such as the use of generative adversarial networks (GANs) and diffusion models (both techniques are briefly described in Section 2.1), empowering computers to learn from massive datasets and generate images that possess remarkable realism. The history of AI generated media can be traced back to the 1950s and 1960s, where computer graphics were used to generate simple patterns and shapes [2]. Following that, different technologies and various approaches (i.e., convolutional neural networks) were continuously developed and applied to this ever-growing field.

Consequently, the present capability for generating realistic images brings into question if current state-of-the-art AI-image generation tools have reached a level of performance where it is difficult for the human perception to distinguish between real images and AI-generated (or synthetic) images (Figure 1.2).

AI image generation can receive as input an image, noise or just a plain text description called *prompt* and, through an advanced machine learning algorithm, creates a synthetic image to illustrate the given input as depicted in Figure 1.1.

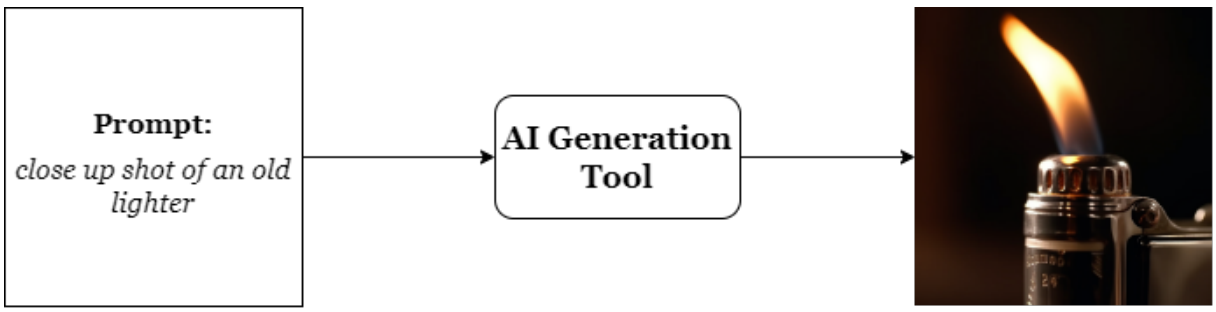


Figure 1.1: AI image generation process.

Given the quality of these synthetic images, and its low generation time comparing to traditional manually crafted photos produced by designers or photographers, this technology could revolutionize several industries, such as e-commerce, marketing, gaming or virtual reality. For example, in the fashion

industry, AI image generation can be used to create clothing designs or style outfits. In the gaming industry, generative AI can be used to develop high-quality characters, backgrounds or environments in a short amount of time which would otherwise take months to build manually. It can also be used in marketing to develop logos, social media posts, designs or any other type of visual project. Meanwhile, artists could also benefit from these tools by aiding them on their creative process and exploring new forms of creative expression.

However, AI-generated images misuse can lead to the proliferation of fake or misleading visual content (Figure 1.3) as this technology can be used as a catalyst to the dissemination of fake news and have severe consequences for the reputation of individuals and their digital identity [3]. This raises future concerns about the authenticity of digital media and how it will be possible to prevent and detect such cases [3, 4].

Figure 1.2: Realistic AI-generated image using Midjourney v5.1. (Credits to `https://twitter.com/nickfloats`)

Figure 1.3: Example of an AI-generated image that potentially spread fake news about Donald Trump, former United States of America.

1.2 Objective and Contributions

The objective of this project is to study how well humans can distinguish between real and synthetic (AI-generated) images and how they perceive them in terms of realism. Additionally, we also studied which categories of synthetic images can be perceived as more realistic. In order to accomplish these goals we divided the work into several steps:

1. Generation of a diverse dataset of synthetic images using state-of-the-art machine learning models.
2. Creation of a standalone application to evaluate the generated visual data inspired in some predefined subjective assessment methodologies.
3. Evaluation of a selected set of visual data by using the developed application and inviting non-experts to evaluate the images, i.e. performing a crowdsourcing-based subjective quality assessment study.

1.3 Structure

The remainder of this document is organized as follows: Chapter 2 presents essential background on AI-based image generation and relevant related work. Chapter 3 presents a detailed description of the process of the whole subjective test and how it was implemented, and Chapter 4 presents the data extracted from the experiment and its thorough analysis. Lastly, Chapter 5 concludes the report with the main observations derived from the results presented in Chapter 4.

Chapter 2

Background and Related Work

This chapter provides background on AI image generation, by presenting an overview of the general concepts behind machine learning techniques in Section 2.1. Next, Section 2.2 lists the tools selected for the used subjective tests and Section 2.3 provides an analysis of the existing methodologies of subjective assessment and lastly, Section 2.4 covers similar existing work.

2.1 General Concepts

Artificial Neural Networks (ANNs) are present in every AI-image generator, being fundamental for this type of technology. These are artificial systems that have the capability to adapt based on input training data. They are inspired by the way the human brain's networks of neurons and are capable of solving problems by detecting specific patterns [5]. The backbone elements of an ANN are its nodes and its connections, which are often organized in layers, as depicted in Figure 2.2. Each node is based on the human neuron and has its own input and output [6]. The input is provided by connections between it and other nodes (dendrites) while its output can be transmitted to other nodes by their respective connections to the initial node (axons). Each connection has its own weight that reflects its significance when calculating the output of the end node, as illustrated in Figure 2.1. Since these weights are adjustable parameters, by modifying them it is possible to train the ANN to solve a given problem, i.e., generate a desired output given some type of complex input. This is achieved by using feedback information obtained from the difference between the generated and expected output, allowing the ANN to adjust itself [5]. The following are the most relevant generative models nowadays:

.

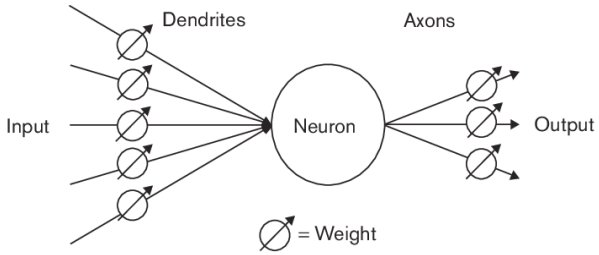


Figure 2.1: Diagram containing a neuron and its input (dendrites) and output (axons) weighted connections.

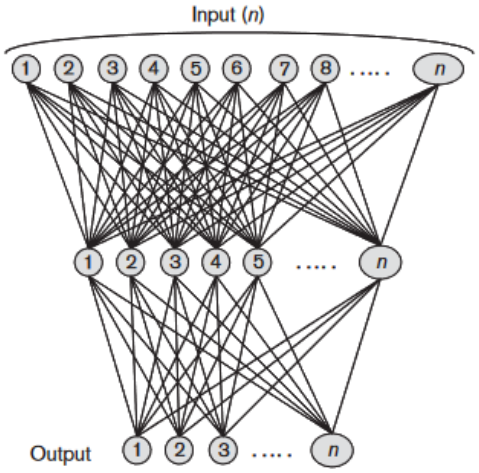


Figure 2.2: ANN general architecture.

- **Convolutional Neural Networks (CNNs)** are a type of ANNs that when used on images can be very useful at detecting and recognizing visual patterns. Its three main types of layers are **Convolutional Layer**, **Pooling Layer** and **Fully-connected layer** [7] (Figure 2.3).

    1. **Convolutional Layer**. A CNN can be composed of multiple convolutional layers where the first one is directly connected to the input. Its primary function is to detect simple features like edges and colors using feature detectors (also known as kernels or filters). These feature detectors are usually a 3x3 matrix of weights that depending on the value of each weight can be used to detect specific image patterns. Assuming that the input is a colored image, it can be represented as a 3D matrix of pixels where each dimension corresponds to each of the RGB channels. A convolutional layer takes this 3D information and iterates through it by performing the dot product between the in use feature detector matrix for a specific image pattern and the target image pixel and its adjacent pixels too. Each one of these iterations performed throughout the image data is a convolution. For each specific used kernel, the convolution final result is a series of dot products that reflect information about the presence of the image pattern that was searched for in the input image. By using multiple convolutional layers besides the one directly connected to the input, a hierarchical analysis can be performed where the first layers are used to detect more lower-level patterns (such as an individual part of a car like a wheel or a headlight, for example) and the last ones for detecting more higher-level patterns (like if the images contains a typical motorcycle, 4-door car or truck with two trailers).
    2. **Pooling Layer**. Used to downsample the resulting data created by the convolutions. Since these operations require a large volume of mathematical operations it is common for a CNN to use few layers of this type with the objective of preserving valuable computing resources.
    3. **Fully-connected layer**. It is responsible for gathering the resulting information about each specific feature taken into account in the previous convolutional layers and for giving a higher-level classification to the input image. Therefore, its task is to produce the output of the CNN in the form of the probability of the input image being part of some category (such as if the image contains a car, a pedestrian or a wall, for example).

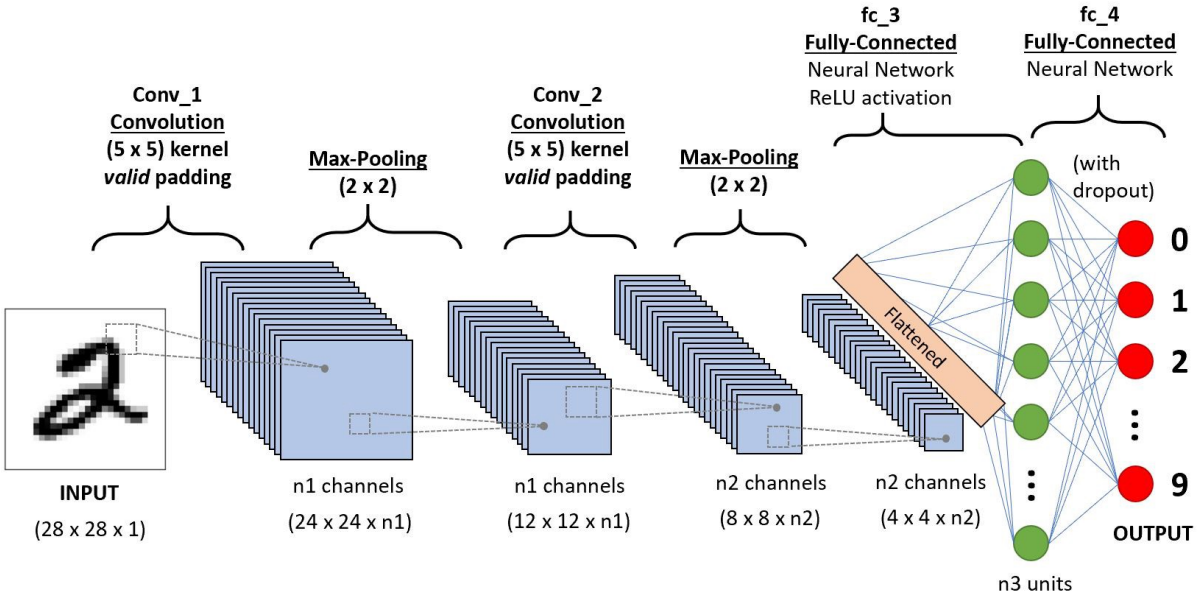


Figure 2.3: Detailed example of a CNN model used for digit recognition with many layers of each type [1].

- **General Adversarial Networks (GANs)** are generative models that consist of two ANNs that compete against each other. The generator network is responsible for the image generation while the discriminator is the network tasked with distinguishing between the real images of the training dataset and the images generated by the generator (Figure 2.4). The general goal of a GAN is to tune the generator so that the discriminator can not correctly distinguish the generated images from the dataset's images. As a consequence, the generator network will learn how to generate realistic images from both a class-conditional, such as the real image's caption present in the training dataset, and noise [8].

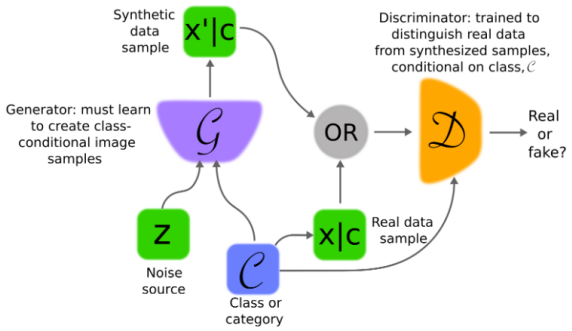


Figure 2.4: Example of a GAN structure.

- **Diffusion Models** are generative models trained to reconstruct images with a certain and controlled amount of white gaussian noise while considering the text input as captions. The text input is tokenized: in other words, each word is numerically codified based on the context of the sentence it belongs to. Therefore, the semantic information of the sentence, i.e. its meaning, is transformed into a numerical format. The end goal is to have a model that can generate an image from pure random gaussian noise and a caption as text input. The following are the two key processes that are responsible for the functioning of this type generative model [9, 10]:

  1. **Forward diffusion process:** for the diffusion model to learn how to reconstruct a noisy

image, first it needs to have various samples with different levels of noise and different levels of image data. These different levels of deconstruction are obtained by starting from a perfect real image from the training dataset and systematically and sequentially adding noise to it (Figure 2.5). This process is performed a total of T number of times and in each iteration the amount of noise added to the image is determined by a schedule. A schedule can operate by adding noise linearly (the simplest mode) or by following some other non-linear function.

2. **Reverse diffusion process:** After having a total of T different versions of the original real image, each with a different amount of noise, the model can be trained based on this data. The goal is to feed the model a t version of the original image (a version of the original image where the forward diffusion process was performed t times and $0 < t < T$), alongside with its corresponding noise ratio and the original image's caption as text conditioning, to predict the $t - 1$ version of the original image (Figure 2.5). Next, we take the prediction from the model, renoise it to a different random noise ratio and feed it back to the model together with the previous text conditioning and the respective previous noise ratio. This iterative process is called the reverse diffusion process and it is how the model learns how to generate images using text as input.

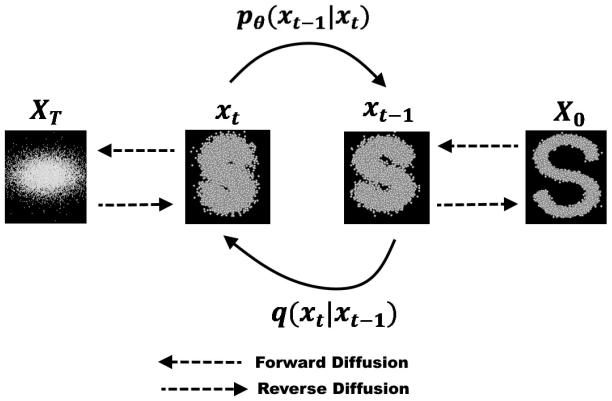


Figure 2.5: Illustration of Forward and Reverse Diffusion processes. The q function represents the Forward Diffusion process and the p_θ function represents the Reverse Diffusion process. X_0 is the original image and X_T is the image after T iterations of noise addition.

2.2 AI Image Generation Tools

The following three tools were used:

Stable Diffusion [1] Released by StabilityAI in 2022, this tool is mainly used to generate detailed images based on text descriptions (*prompts*). It can also be applied to other tasks such as inpainting, outpainting, and image translation. It was developed by researchers from the CompVis Group at Ludwig Maximilian University of Munich and the Runway AI research company. It can run on most consumer hardware with at least 8GB VRAM. The software was trained on images and captions taken from LAION-5B, a publicly available dataset. However, there are limitations to Stable Diffusion, such as the generation of human parts due to the poor data quality of limbs in the LAION database.[2].

[1] `https://stablediffusionweb.com/`
[2] `https://en.wikipedia.org/wiki/Stable_Diffusion`

DALL-E [3] Deep learning model developed by OpenAI in 2021 with the aim of generating digital images from *prompts*. The source code is not available, so large scale image generation is not possible. In this project, DALL-E was accessed through Bing's Image Creator[4]. DALL-E can generate imagery in multiple styles, including photorealistic imagery, paintings, and emojis for a wide variety of arbitrary descriptions. Its limitations are language-based as it sometimes can confuse "A yellow book and a red vase" from "A red book and a yellow vase". The probability of this kind of error rises proportionately to the complexity of the *prompt*[5].

Midjourney [6] Generative artificial intelligence image tool and service created and hosted by Midjourney, Inc. Midjourney can also generate high quality images derived from *prompts*, similar to the other two tools. Midjourney is currently in open beta which started on 2022 and is currently only accessible through a Discord bot, using a *prompt* to which the bot responds with a set of four images. Users may also choose which images they want to upscale[7]. This tool was the backbone of our project as approximately 75% of the images in our dataset were generated by this software. This is because Midjourney is considered to be the best tool for photorealistic images and provides a variety of customization commands and parameters to help users fine-tune the images being generated.

2.3 Subjective Assessment of Images

Image quality assessment is often measured through solid and objective metrics, which are cheap and fast but not always the most reliable. On the other hand, subjective image quality assessment experiments, which are expensive and time-consuming, are much more reliable approach as it is based on the subjective opinion of a large number of subjects. This type of experiments is mainly conducted with high-quality monitors, under controlled environments and lighting. This is called Controlled Environment Subjective Assessment. The main two approaches can be described as:

- **Controlled Environment Subjective Quality Assessment**: The most popular approach to assess image quality is in a controlled environment. This control allows to develop a subjective experiment that are repeatable and unbiased with few variables to disrupt the results. One of the key factors to create this controlled environment is to use top-tier visualization equipment and artificial lighting, which could vary according to the needs of the type of subjective assessment being conducted. Another critical factor is the choice and number of subjects. The Radiocommunication Sector of the International Telecommunication Union (ITU-R) recommends choosing at least 15 participants. The experience of the subjects should also be taken into consideration since there could be an interest in choosing participants that have some kind of knowledge on the matter or no experience at all. Moreover, the subjective test should not take longer than half an hour as that could fatigue the subjects and, consequently, produce unreliable results.

- **Crowdsourcing-based Subjective Quality Assessment**: Subjective image quality assessment can also be performed based on crowdsourcing. With this methodology, the participants conduct the experiment in different types of environments since it is usually deployed on a web server and accessed by the subjects over the Internet. Thus, introduces new conditions (e.g. display) to the assessment but has the advantage that can also be perceived as a more realistic setup.

[3] https://openai.com/research/dall-e
[4] https://www.bing.com/create
[5] https://en.wikipedia.org/wiki/DALL-E
[6] https://www.midjourney.com/home
[7] https://en.wikipedia.org/wiki/Midjourney

Crowdsourcing-based subjective assessment is cheaper and faster and rising in popularity over the past few years, mainly due to the COVID-19 pandemic.

Several subjective assessment methodologies exist and can be separated into two main categories: single stimulus and double stimulus methodologies. What separates them is the amount of stimuli presented to the subjects of the experiment in a given instant. While single stimulus methodologies use a single image, double stimulus use the impairment between two images, shown side by side. These two categories can be described as:

- **Single Stimulus Methodology (SS)** consists on presenting to the subjects a sequence of single images followed by a rating section. The grading scale varies from experiment to experiment but the most used is Absolute Category Rating (ACR) that works as a 1-5 scale where 1 corresponds to ”Bad” and 5 to ”Excellent”. Various grading scales are summarized in Table 2.1. This methodology is a simple and safe choice but leads to results with high variance.
- **Double Stimulus Methodology (DS)** is another type of subjective quality assessment methodology and consists on a side-by-side presentation of a couple of stimuli to rate the difference between both images. This approach takes a longer time in contrast to the single stimulus method despite being generally more accurate. Like in single stimulus, the grading scale can differ as depicted in Table 2.1. Double Stimulus Impairment Scale (DSIS) also known as Degradation Category Rating (DCR) is one of the most used scales and it presents a scale of impairment from 1-5 where 1 corresponds to ”Very annoying” and 5 to ”Imperceptible”.

In this project an hybrid solution was used, using a variant of crowdsourcing in which the external environment was not controlled (e.g. lighting) and was rarely the same. However, the setup was always the same since the test was assigned to the subjects personally. Some recommendations were taken into consideration like the number of participants and the duration of the experiment. The methodology was also a variant of the double stimulus methodology where a sequence of pairs of images - one AI and one real - were shown to the participants but we ask the subject to choose the AI image over the real one. The grading scale was not related to the impairment of the images per se but related to the realism of the chosen image.

2.4 Related Work

In this section, the related work and research are presented. Given that the emphasis of this work is on the analysis of the data extracted from the subjective test, the focus is dedicated to existing studies about subjective tests of synthetic images' quality.

This ever-growing field of AI has attracted researchers from all over the world to thoroughly analyse and understand the technicalities behind AI media generation. Here are some articles that inspired us:

Method	Type	Scale Type	Advantages	Disadvantages
ACR	SS	Discrete	Fast and simple	Scores influenced by the subject's opinion on the content
ACR-HR	SS	Discrete	Allows to remove the variance due to the subjects' personal opinion on the content	Requires a long training procedure to acquaint the subjects with the artifacts
SSCQE	SS	Continuous	Comparable to the continuous grading scale of objective quality metric	Requires a long training procedure to acquaint the subjects with the artifacts
DSIS	DS	Discrete	Not influenced by the subjects' opinion on the content, reliable in evaluating color impairment	Slower than ACR
DSCQS	DS	Continuous	Both the original and the impaired stimuli are graded	Slower than DSIS
DSCS	DS	Discrete	Compares all the different stimuli among themselves	Biggest number of comparisons, bitrate matching is critical

Table 2.1: Summary of the different methods for subjective quality assessment

- **Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2 [11]** describes the history and the evolution of the field of image synthesis and compares three big AI image generation tools - Stable Diffusion, Midjourney and DALL-E 2 - and concludes that, according to the FID metric (performance metric that calculates the distance between the feature vectors of real images and the feature vectors of fake images), Stable Diffusion generates better images than the other two models.
- **Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding [12]** presents a different text-to-image diffusion model - Imagen - and analyses its "unprecedented degree of photorealism and deep level of language understanding". This article also mentions the ethical challenges related to this field, presenting the contrast of progression and augmentation of human creativity to the possible malicious purposes of generative methods (i.e., harrassment, missinformation).
- **On The Detection Of Synthetic Images Generated By Diffusion Models [13]** focuses on the growth of Diffusion Models, how they work and how their generated images can be detected. This article also approaches the possible ethical difficulties regarding AI image generation.

Chapter 3

Subjective Assessment of AI Generated Images

The subjective test's goal was to evaluate the human capability of distinguishing synthetic images from similar real images while also considering the image's category. It was carried out using a local-based PsychoPy Experiment [1], an open-source Python package used to run a wide variety of experiments in the behavioral sciences. The subjective test was comprised of the following phases:

1. **Information**: The participants were provided with information regarding the subjective test's purpose, structure and expected duration.
2. **Participants' Data Collection**: Participants' age and gender data was collected in order to understand the demographic characteristics of the subjective test's participants population.
3. **Main Experiment**: The experiment consisted of 116 sequential test instances, presented in a random order, where each one was composed by a selection and a rating section, presented by this order.
   (a) **Synthetic Image Selection**: At the selection section, two similar images were presented, one being AI-generated (synthetic) and the other being a real image. Both images were presented in random order and had equal size. The participant's task was to select the AI-generated image based on its individual perception. In the meantime a countdown timer with the duration of eight seconds was also displayed on the screen to avoid an excessive decision time. However, it served only as a guideline since when it expired the resulting action was only the displaying of an alert message.
   (b) **Selected Image Evaluation**: At the evaluation section, the image selected in the previous test section appears again. The participant's task was to evaluate its level of realism in a scale of one to five, where one corresponds to "Not realistic" and five corresponds to "Very realistic".

3.1 Generation of Test Data

The dataset for this test consists of 116 AI generated images and 116 real images which were then presented in a random sequence of 116 pairs of images where one is synthetic and the other is a real image that is similar to it.

[1] `https://www.psychopy.org/`

The synthetic images used for our subjective test were produced through the use of the AI image generation tools listed in Section 2.2.

First, the work started with research of which tools would be used, which was done by testing each one while learning the basics of *prompts*. Stable diffusion was our first attempt at generating an AI image. It was quickly realized that generating high quality images that would at least challenge our subjective test's participants was not going to be an easy task. After some images were generated, it was decided to organize our dataset into multiple categories and multiple types of images to broaden the field of assessment, as shown in Table 3.1.

Midjourney was the tool that was often used for image generation since it was the most detailed and versatile while also being very much capable of producing photorealistic images. At first, AI images generated by other users of this software were collected, but after a while, to expand our categories, tailor-made images were obtained with custom *prompts*.

The real images were acquired in a few different ways:

- Some images are the result of a reverse search on Google Images [2] with its synthetic image pair.
- Others were searched for on the Internet through a text description.
- Lastly, some were searched in websites of stock photos(i.e., Pexels[3]) which were then used as source of inspiration to generate the synthetic image from a suitable *prompt*.

Generating a high quality detailed image that represents exactly what the prompt intends to create was not easy for us in the beginning. The learning process was as follows:

1. Learn by watching experienced users generating images.
2. Research the process of creating high quality *prompts* like using keywords(i.e., ultra realistic, professional photography).
3. Trial and error.

We generated numerous synthetic images. Figure 3.1 provides an example of an image for each category.

Figure 3.1: Some of the generated images (one for each category.)

[2] https://images.google.com
[3] https://www.pexels.com

Category	Description
Animals	Living animals in challenging and complex scenarios
Faces	Highly detailed face realistic images
Cars	Various types of cars in various scenarios
Houses (Exterior)	Architecture designs and details
Houses (Interior)	Decoration details and home organization
Food	Images of food with some challenging characteristics (e.g., crispiness, juiciness)
Objects	Close point of views of inanimate objects with condensed details
Landscapes	High depth of field scenarios with spread out details
Sports	Dynamic moving images of sports
Groups of people	Highly challenging images of multiple detailed humans

Table 3.1: Categories.

3.2 Subjective Assessment Platform

The subjective test was developed by programming the graphical user interface. The PsychoPy v2022.2.5 library was used to faciliate the development process. The subjective experiment has three following phases:

1. **Initialization**: At this phase, images' data was loaded while both instructions and a form concerning demographic data (age and gender) were shown to participants. The loading of images' data was done by reading a .xlsx stimulus file that contained information about the path of the images to be shown and which one was the synthetic image, as well as to which image category it belonged to. Once the images' data was loaded, the intra- and inter-test order was randomized, creating the current participant's sequence of test images. Additionally an inter-test variable concerning the number of the participant's correct selections was initialized.

2. **Data Collection**: Here a routine was used to iterate through the randomized tests list created in the previous phase. Each iteration had a selecting step and a rating step. In the selecting step both images (synthetic and real) were presented side-by-side alongside a countdown timer (as depicted in Figure 3.2). Both images could be selected by clicking on them. If the participant had selected the synthetic image, the number of correct answers variable would be incremented and

the program would store in the current experiment's output file that the participant's selection was correct. Additionally, once the countdown timer was finished the text "Make your choice now!" was displayed in the place of the timer (Figure 3.2). Next, in the rating step, the previously selected image was shown alongside with a rating slider and a next button. This prevented the user from misclicking on the rating slider and giving false responses (as presented in Figure 3.4). After the subject rates the image it can advance to the next image by clicking the button next. In this case, the program would store in the current experiment's output file the value of the subject's rating.

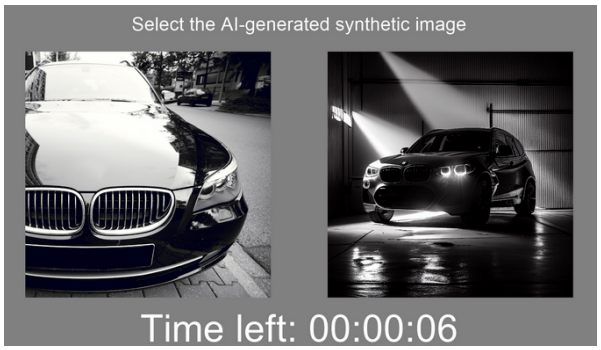


Figure 3.2: Selection phase with timer.

Figure 3.3: Selection phase without timer.

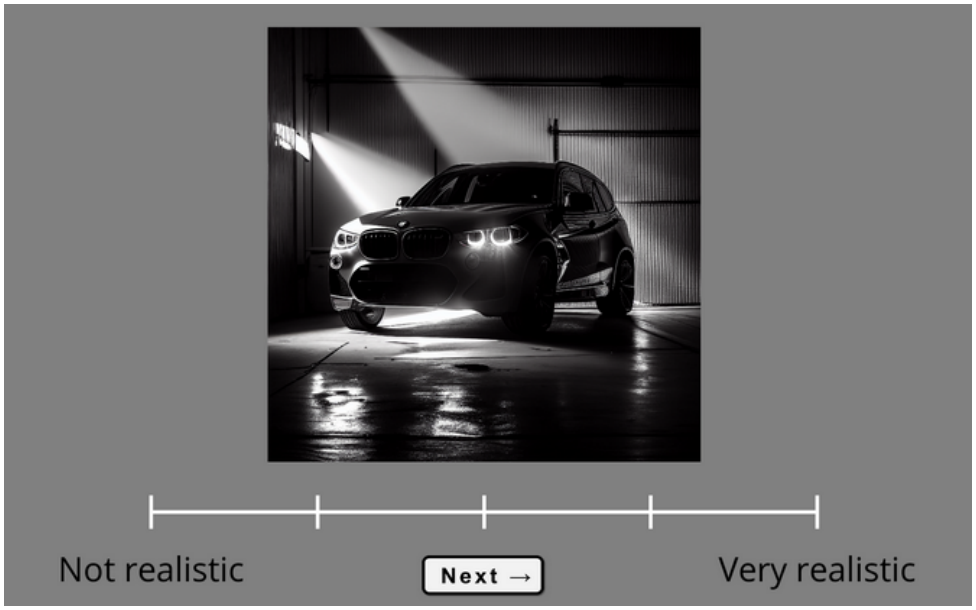


Figure 3.4: Rating phase.

3. **End Results**: In this last phase, the subjective test displayed the number of correct selections versus the number of total selections through the use of the number of the participant's correct answers variable. This was just to inform the subjects of how much success they had in the identification of the synthetic images.

Chapter 4

Subjective Assessment: Experimental Results

The objective of this Chapter is to present the statistical results obtained after the execution of the subjective test mentioned in Chapter 3 as well as how the resulting data was processed. This experiment was performed by a population of 23 participants.

Since the subjective test had both a selecting phase and a evaluation phase, the key statistics used in this Chapter were the following (**Note:** because participants were asked to select the synthetic image from the pair consisting of both a real and a synthetic image we defined that a **correct selection** happens **when the participant selects the synthetic image**):

- **Rate of correct selections for a synthetic image (**RCS_SI**)** [%]:

$$RCS\_SI = \frac{\text{Total number of correct selections of } S_i}{\text{Total number of selection tests performed using } S_i \text{ as the synthetic image}}$$

  where S_i is a synthetic image.

- **Rate of correct selections for a category (**RCS_Cat**)** [%]:

$$RCS\_Cat = \frac{\text{Total number of correct selections of synthetic images belonging to } C_i}{\text{Total number of selection tests performed a synthetic image belonging to } C_i}$$

  where C_i is a category.

- **Mean Opinion Score for a synthetic image (**MOS_SI**)**:

$$MOS\_SI = \frac{\sum_{k=1}^{T_{S_i}} \text{Score given at rating test } k}{\text{Total number of evaluation tests performed using } S_i \text{ as the synthetic image}}$$

  where:

  - S_i is a synthetic image.
  - T_{S_i} equals to the total number of rating tests performed using S_i as the synthetic image with previous selection test correct.

- **Mean Opinion Score for a category (**MOS_Cat**)**:

$$MOS\_Cat = \frac{\sum_{k=1}^{T_{C_i}} \text{Score given at test } k}{\text{Total number of rating tests performed using synthetic images belonging to } C_i}$$

  where:

– where C_i is a category.

– T_{C_i} equals to the total number of rating tests performed using synthetic images belonging to C_i with previous selection test correct.

4.1 Data Processing

We considered that if the synthetic image was as realistic as the real image then the probability of the participant selecting the synthetic image would be equal to 50%. This is the criteria that reflects the best-case scenario, meaning that the synthetic images were indistinguishable from the real images. As a consequence, the worst-case scenario corresponds to the participants always selecting the synthetic image for the same pair.

Therefore, synthetic images with a RCS_SI lower than 45% correspond to non-expected cases and need to be removed. For example, image pairs where the real image had poor quality or was somehow unrealistic. The removed pairs and their respective categories are presented in Figures 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, and 4.7.

(a) Synthetic Image. (b) Real Image.

Figure 4.1: Pair number one.

Figure 4.1 is an example of a pair of synthetic/real images that was removed due to the low RCS_SI. The main reason is the unrealism of the real image, which was heavily processed. It is likely that Figure 4.1b was altered as the eyes appear to have been modified and the face smoothed out with some filters.

(a) Synthetic Image.

(b) Real Image.

Figure 4.2: Pair number two.

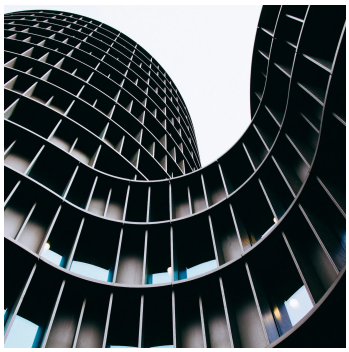
(a) Synthetic Image.

(b) Real Image.

Figure 4.3: Pair number three.

Figure 4.2 and Figure 4.3 are similar cases where it is understandable that Figure 4.2b and Figure 4.3b are perceived as synthetic images due to the smoothness of the lighting and the reflections.

(a) Synthetic Image.



(b) Real Image.

Figure 4.4: Pair number four.

In Figure 4.4 the RCS_SI was lower than anticipated since Figure 4.4b appears to have been subject of post color correction.

(a) Synthetic Image.

(b) Real Image.

Figure 4.5: Pair number five.

As in other pairs of images, Figure 4.5b appears to have been altered due to the unrealistic appearance of the sparks, which could be the justification for the low RCS_SI.

(a) Synthetic Image.

(b) Real Image.

Figure 4.6: Pair number six.

Figure 4.6b presents some unrealistic color saturation which could have led to a RCS_SI.

(a) Synthetic Image.

(b) Real Image.

Figure 4.7: Pair number seven.

Lastly, Figure 4.7b was chosen multiple times due to its poor quality and its unrealism.

For synthetic images with five or less correct selections, i.e. synthetic images that have a good level of realism since participants could not distinguish the real image from the synthetic one, it was necessary to process the scores in a different way. Since these images got very few evaluations it was decided to use the MOS calculated from the subjective ratings given to the real image, i.e. an incorrect selection was previously performed. This action was an attempt to improve the MOS_SI of these synthetic images due to their low representation considering the total amount of data.

4.2 Experimental Results Analysis

The experimental results were obtained after the participation of 23 subjects. The vast majority of participants were aged between 20 and 24 years old (Figure 4.8). Additionally, most participants identified themselves as males (Figure 4.9).

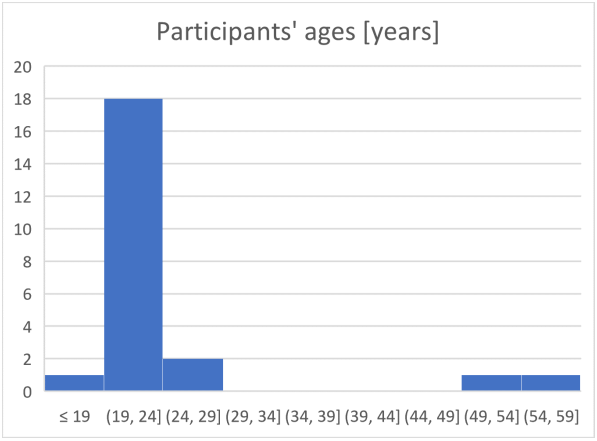


Figure 4.8: Histogram of the participants' ages.

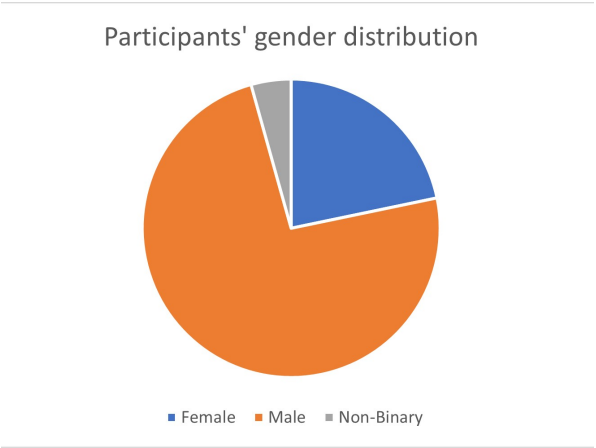


Figure 4.9: Participants' gender distribution.

By analysing the RCS_Cat (Figure 4.10) it can be concluded that this metric varies between 70% and 90% percent, approximately. As a result, using our best-case-scenario criteria defined previously in section 4.1, no category had an outstanding performance. Nevertheless, the categories *Houses(Interior)*, *Houses(Exterior)* and *Objects* were the ones with the best results.

A deeper analysis was also done by calculating the RCS_SI for each individual synthetic image inside each category and thus better understand how does this metric varies inside each category (Figure 4.11). *Animals*, *Faces* and *Sports* are the categories that appear to have a lower variance of results in opposition to categories such as *Food*, *Groups of People*, *Houses(Exterior)*, *Houses(Interior)* and *Objects*. Nevertheless, throughout every category there is at least one image that had RCS_SI above 90%, meaning that participants were able to correctly identify at least one synthetic image in all categories.

The distribution of all calculated MOS_SI (as shown in Figure 4.12) varies between 1.5 and 4.5. However, most values are between 2.5 and 4.25 thus implying that the generated synthetic images were relatively realistic.

By analysing the calculated MOS_Cat (Figure 4.13) it is possible to observe an almost inverse correlation between the MOS_Cat and the RCS_Cat. This further supports our best-case-scenario criteria (defined in section 4.1) that consists of low rates of correct selections and high $MOSs$. Additionally, it is possible to conclude that generated synthetic images seem to perceived as more real when depicting people.

After taking all of these results into consideration, we can conclude that although there is not a high number of synthetic images or categories with outstanding performance there were categories with clearly better results than others. Synthetic images from the category *Objects* had the best performance while the category *Faces* had the lowest performance.

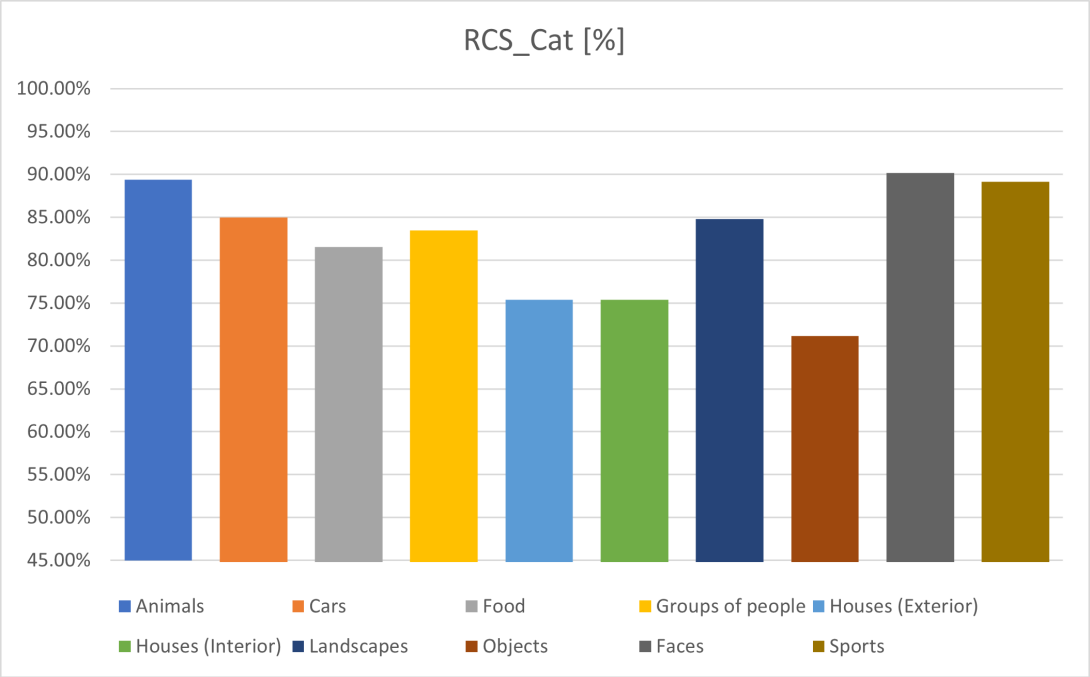


Figure 4.10: RCS_Cat for each category.

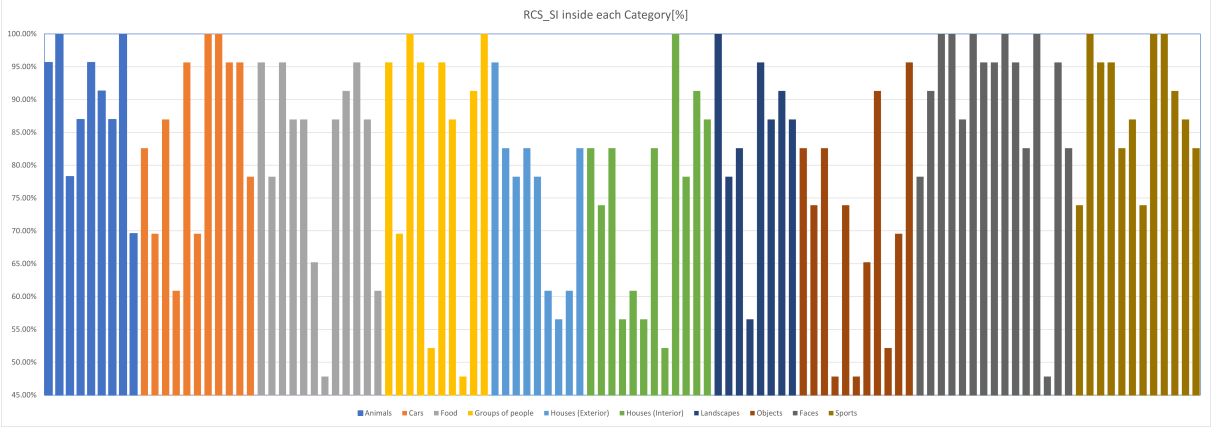


Figure 4.11: RCS_SI inside each category. Each bar corresponds to a single synthetic image.

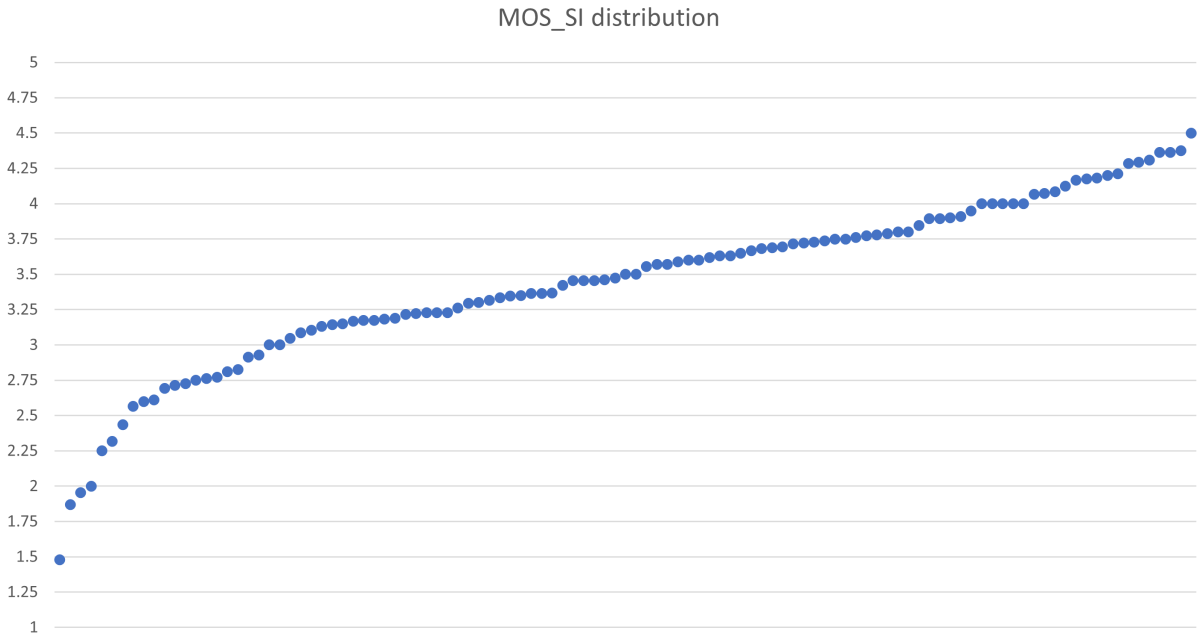


Figure 4.12: Distribution of MOS_SI. Each dot represents a synthetic image.

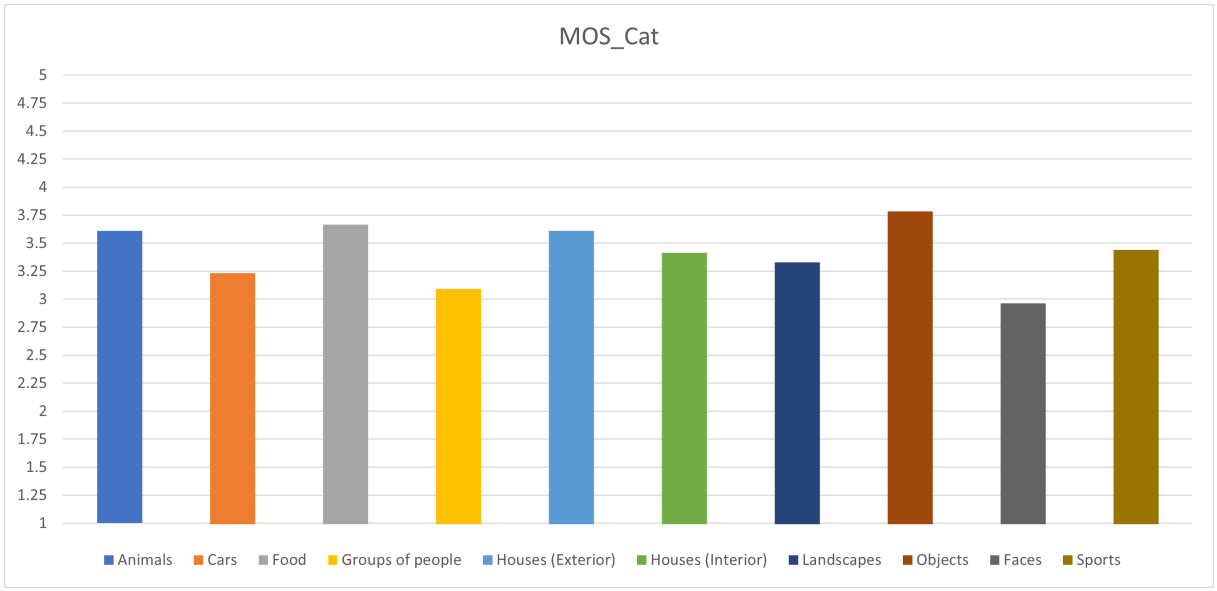


Figure 4.13: MOS_Cat for each category.

Chapter 5

Conclusion

In this project, it was possible to conclude that state-of-the-art AI image generation technologies have reached a point where they can generate images that can be perceived as realistic. However, in terms image realism, the results of this project show that perceived synthetic images realism can vary considerably (Figure 4.12) since only 19.2% of synthetic images had a MOS_SI equal or greater than 4 and most scores were evenly distributed between 2.5 and 4.25, approximately.

Specifically, in what concerns image categories it is possible to conclude that there are some categories that were perceived as more realistic than others. The categories with the highest MOS_Cat were *Objects* and *Food*, while *Animals* and *Houses(Exterior)* had practically equal scores. On the other hand, categories with the lowest MOS_Cat were *Faces* and *Groups of people* (Figure 4.13). Additionally, if the RCS_Cat for each one of the categories used is taken into consideration it can be shown that the ones with the lowest RCS_Cat were *Objects*, *Houses (Exterior)* and *Objects* while the ones the highest were *Faces*, *Sports*, and *Animals* (Figure 4.10). Given these results, it is possible to conclude the following concerning the realism of these synthetic generated images:

- Synthetic images of *Objects* had the best performance in terms of realism since participants could not distinguish the real image from the synthetic image around 28% of the time and this was the category with the highest MOS_Cat.
- When comparing to all studied image categories, AI-image generation tools seem to have difficulty in creating realistic images depicting humans in general.

Bibliography

[1] P. Ratan, 2020. "https://www.analyticsvidhya.com/blog/2020/10/what-is-the-convolutional-neural-network-architecture/" Accessed: 2023-07-10.

[2] R. Kundu, "Ai-generated art: From text to images & beyond [examples]," 2022. https://www.v7labs.com/blog/ai-generated-art Accessed: 2023-07-05.

[3] R. B. H. D. V. G. G. B. Federica Lago, Cecilia Pasquini, "More real than real: A study on human visual perception of synthetic faces," 2021. https://arxiv.org/abs/2106.07226 Accessed: 2023-07-08.

[4] G. Z. G. P. K. N. L. V. Riccardo Corvi, Davide Cozzolino, "On the detection of synthetic images generated by diffusion models," 2022. https://arxiv.org/pdf/2211.00680.pdf Accessed: 2023-07-08.

[5] T. K. Dr. Rama Kishore, "Backpropagation algorithm: An artificial neural network approach for pattern recognition," *International Journal of Scientific & Engineering Research*, vol. 3, no. 6, 2012.

[6] M. B. Enzo Grossi, "Introduction to artificial neural networks," *European Journal of Gastroenterology & Hepatology*, 2008. https://www.researchgate.net/publication/5847739_Introduction_to_artificial_neural_networks Accessed: 2023-06-06.

[7] IBM, "Convolutional neural networks." "https://www.ibm.com/topics/convolutional-neural-networks Accessed: 2023-07-12".

[8] V. D. K. A. B. S. A. A. B. Antonia Creswell, Tom White, "Generative adversarial networks: An overview," 2017. https://arxiv.org/abs/1710.07035 Accessed: 2023-07-07.

[9] C. Luo, "Understanding diffusion models: A unified perspective," 2022. arXiv:2208.11970v1 Accessed: 2023-07-06.

[10] P. A. Jonathan Ho, Ajay Jain, "Denoising diffusion probabilistic models," 2020. https://arxiv.org/pdf/2006.11239.pdf Accessed: 2023-07-07.

[11] A. Borji, "Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2," 2023. "https://arxiv.org/pdf/2210.00586.pdf Accessed on 2023-07-09".

[12] S. S. L. L. J. W. E. D. S. K. S. G. B. K. A. S. S. M. R. G. L. T. S. J. H. D. J. F. M. N. Chitwan Saharia, William Chan, "Photorealistic text-to-image diffusion models with deep language understanding," 2022. "https://arxiv.org/pdf/2205.11487.pdf Accessed on 2023-07-09".

[13] G. Z. G. P. K. N. L. V. Riccardo Corvi, Davide Cozzolino, "On the detection of synthetic images generated by diffusion models," 2022. "https://arxiv.org/pdf/2211.00680.pdf Accessed on 2023-07-09".

Appendix A

Appendix chapter

Número	Categoria	Ferramenta
1_001	Animals	Midjourney v5
1_002	Animals	Midjourney v5
1_003	Animals	Midjourney v5
1_004	Animals	Midjourney v5
1_005	Animals	Midjourney v5
1_006	Animals	Midjourney v5
1_007	Animals	Midjourney v5
1_008	Animals	Midjourney v5
1_009	Animals	Midjourney v5
2_001	Faces	Midjourney v5
2_002	Faces	Midjourney v5
2_003	Faces	Stable Diffusion (local)
2_004	Faces	Stable Diffusion (local)
2_005	Faces	Stable Diffusion (local)
2_006	Faces	Stable Diffusion (local)
2_007	Faces	Stable Diffusion (web)
2_008	Faces	Stable Diffusion (web)
2_009	Faces	Stable Diffusion (local)
2_010	Faces	Midjourney v5
2_011	Faces	Midjourney v5
2_012	Faces	Midjourney v5
2_013	Faces	Midjourney v5
2_014	Faces	Midjourney v5
2_015	Faces	Midjourney v5
2_016	Faces	Midjourney v5
3_001	Cars	Midjourney v5
3_002	Cars	Stable Diffusion (web)
3_003	Cars	Stable Diffusion (web)
3_004	Cars	Stable Diffusion (web)
3_005	Cars	Stable Diffusion (web)
3_006	Cars	Stable Diffusion (web)
3_007	Cars	DALL-E (Bing)
3_008	Cars	DALL-E (Bing)
3_009	Cars	DALL-E (Bing)
3_010	Cars	Midjourney v5
3_011	Cars	Midjourney v5
3_012	Cars	Midjourney v5
4_001	Houses (Exterior)	Midjourney v5
4_002	Houses (Exterior)	Midjourney v5
4_003	Houses (Exterior)	Midjourney v5
4_004	Houses (Exterior)	Stable Diffusion (web)
4_005	Houses (Exterior)	Stable Diffusion (web)
4_006	Houses (Exterior)	DALL-E (Bing)
4_007	Houses (Exterior)	DALL-E (Bing)
4_008	Houses (Exterior)	DALL-E (Bing)
4_009	Houses (Exterior)	Midjourney v5
4_010	Houses (Exterior)	Midjourney v5
4_011	Houses (Exterior)	Midjourney v5
4_012	Houses (Exterior)	Midjourney v5
5_001	Houses (Interior)	Midjourney v5
5_002	Houses (Interior)	Midjourney v5
5_003	Houses (Interior)	Stable Diffusion (web)
5_004	Houses (Interior)	Stable Diffusion (web)
5_005	Houses (Interior)	Stable Diffusion (web)
5_006	Houses (Interior)	Stable Diffusion (web)
5_007	Houses (Interior)	DALL-E (Bing)
5_008	Houses (Interior)	DALL-E (Bing)
5_009	Houses (Interior)	DALL-E (Bing)
5_010	Houses (Interior)	DALL-E (Bing)
5_011	Houses (Interior)	Midjourney v4
5_012	Houses (Interior)	Midjourney v5
6_001	Food	Midjourney v5
6_002	Food	Midjourney v5
6_003	Food	Midjourney v5
6_004	Food	Midjourney v5
6_005	Food	Midjourney v5
6_006	Food	Midjourney v5
6_007	Food	Midjourney v4
6_008	Food	Midjourney v5

6_009	Food	Midjourney v5
6_010	Food	Midjourney v5
6_011	Food	Midjourney v4
6_012	Food	Midjourney v5
7_001	Objects	Midjourney v4
7_002	Objects	Midjourney v5
7_003	Objects	Midjourney v5
7_004	Objects	Midjourney v5
7_005	Objects	Midjourney v4
7_006	Objects	Midjourney v5
7_007	Objects	Midjourney v5
7_008	Objects	Midjourney v5
7_009	Objects	Midjourney v5
7_010	Objects	Midjourney v5
7_011	Objects	Midjourney v5
7_012	Objects	Midjourney v5
8_001	Landscapes	Midjourney v4
8_002	Landscapes	Midjourney v5
8_003	Landscapes	Midjourney v5
8_004	Landscapes	Midjourney v5
8_005	Landscapes	Midjourney v5
8_006	Landscapes	Midjourney v5
8_007	Landscapes	Midjourney v4
8_008	Landscapes	Midjourney v4
8_009	Landscapes	Midjourney v5
8_010	Landscapes	Midjourney v5
9_001	Sports	Midjourney v5
9_002	Sports	Midjourney v5
9_003	Sports	Midjourney v5
9_004	Sports	Midjourney v5
9_005	Sports	Midjourney v5
9_006	Sports	Midjourney v5
9_007	Sports	Midjourney v4
9_008	Sports	Midjourney v4
9_009	Sports	Midjourney v5
9_010	Sports	Midjourney v4
9_011	Sports	Midjourney v5
9_012	Sports	Midjourney v5
10_001	Groups of people	DALL-E (Bing)
10_002	Groups of people	DALL-E (Bing)
10_003	Groups of people	Stable Diffusion (web)
10_004	Groups of people	Midjourney v5
10_005	Groups of people	Midjourney v5
10_006	Groups of people	Midjourney v5
10_007	Groups of people	Midjourney v5
10_008	Groups of people	Midjourney v5
10_009	Groups of people	Midjourney v5
10_010	Groups of people	Midjourney v5
10_011	Groups of people	Midjourney v5
10_012	Groups of people	Midjourney v5