Abstract

Text-to-image generators powered by deep learning have gained immense popularity in recent years. These generators possess the capability to transform written descriptions into visual representations. Theology is a discipline that involves the study of the divine, human existence, and reality, drawing upon sacred texts and philosophical thought. In this thesis, AI is utilized to produce images based on Bible passages, which are then compared to human-made art from the same passages. Machine learning, specifically CNN, is employed to assess the differences between AI-generated art and human-made art, while also analyzing the religious and aesthetic aspects of the pieces.

Contents

Chapter 1

Introduction

In the era of Artific Compariial intelligence (AI) and machine learning, generative models have been developed to generate realistic images from textual descriptions. These models are called text-to-image generators, including DALL·E 2, Midjourney, and Stable Diffusion. More discussion arises as they become more popular for generating visually stunning images from vividly described prompts.

AI-generated art has been argued to lack human attributes such as creativity, originality, subjectivity, emotional depth, context, cultural significance, intention and conceptualisation. Human art often showcases originality, personal expression, and emotional depth, drawing upon individual experiences and cultural contexts. It demonstrates technical proficiency, subjective interpretation, and conceptual intent; these attributes are not the strongest in AI-generated art. However, AI can explore new aesthetic possibilities, challenge traditional norms, and push the boundaries of creativity. While AI-generated art has its merits, it currently falls short of capturing the holistic depth and intentionality that human-created art embodies.

The prompt given to the AI is the building block of the image generated. The prompt leads to choices in artistic style and objects chosen to depict in the artwork. Analysing and exploring the images generated by the prompt can raise the discussion on how humans would interpret the prompt compared to the AI. Looking at biblical prompts as input and comparing human artwork to AI-generated artwork is a discussion the thesis will build around.

The Bible is a highly interpretive text with diverse interpretations and layers of meaning influenced by theology, cultural contexts, and individual perspectives. It invites reflection, study, and discourse. The Bible often serves as believers' spiritual and emotional anchor, providing comfort, inspiration, and a sense of community. It evokes deep emotions and fosters personal connections with the divine. This invites artists to showcase the traits prominent in human artwork when portraying a given text from the Bible in an artwork. However, while the biblical text encompasses a rich descriptive level in many areas, it is essential to acknowledge that there are instances where it lacks explicit or detailed descriptions, for example, physical description, spatial and temporal

details, and visual depictions. These details are usually incorporated in the artwork but are left to the artist's interpretation and knowledge to encompass.

Comparing the human Bible reference art pieces to AI-generated art pieces can create different art interpretations from the text, as seen in Fig 1.1. This shows the story of Babel from the bible created by Pieter Breugel on the left, and images 1 to 4 are produced using text-to-image generators on the same story.

Figure 1.1: Tower of Babel (Genesis 11:1-9)

For this thesis, we considered biblical text as the focal point of the input. The biblical text allows us first to have a source of text that many artists have artistically portrayed. Since the Bible is a religious text, the interpretation and basis of the art produced incorporates all the human art characteristics described before.

This presents the opportunity to formulate semantically relevant image analysis on the data to learn about the characteristics and traits of the art produced. This thesis aims to understand images by comparing human art to AI by analyzing and producing an evaluation. This feat is done by creating a pipeline that analyses images from 3 different text-to-image generators. The pipeline contains a model for scoring Object Detection and a model that classifies sentiment pixels.

Exploring the differences in art can help discover new patterns and meaningful relations among the models and human art based on the same prompt. This presents the opportunity to formulate research questions to evaluate the success of the task at hand.

The research questions and their sub-research questions are as follows:

RQ1: How can we generate biblical images from text-to-image generators?

RQ2: How can we evaluate the images generated by text-to-image generators?

We perform the evaluation by asking the following three sub-research questions:

SRQ2A: How can we use object detection to evaluate the accuracy of generated images?

SRQ2B: Which aspect(s) of sentimental values can be evaluated automatically for the generated images?

SRQ2C: What features can we observe for the generated images regarding religion and aesthetics?

To answer RQ1, we create a database with biblical images created by passing biblical prompts through AI. This is followed by creating a fully automated pipeline that compares the Human detection (Object detection) and Pixel sentiment score (Sentimental classification) to that of Renaissance paintings chosen for this Thesis. This comparison will create a score; the higher the value, the more difference between human art and AI art. The score is done by comparing each human-made image against the AI-generated images produced by the generators. The score will be a combination of human detection and sentimental scores. The Thesis then visually analyzes the images' aesthetic and religious accuracy and discusses the attributes' implications.

This thesis makes the following scientific contribution:

A large dataset of 7,116 images from 9 texts-to-image generators (DALL-E 2, Midjourney, and Stable Diffusion 7 different variants). A survey of painting analysis and labelling with the data extracted as a gold standard for comparison. The Thesis consists of four aspects: accuracy evaluation, sentimental analysis, religious analysis, and aesthetic analysis. A reflection of the performance of the generators gives us insight into human evaluation compared to the pipeline.

This thesis is organised as follows: Chapter 1 Introduction, chapter 2 Related Work where we discuss previous works on aspects of the thesis. Chapter 3 Data the data collection and selection are discussed. Chapter 4 Methodology, the introduction of the pipeline and means to score the results. Chapter 5 Automated Evaluation, where the pipeline results are evaluated. Chapter 6 Religious and Aesthetic Analysis, the discussion of the religious and aesthetic attributes from the images generated. Chapter 7 Discussion, we discuss the findings and limitations of the thesis. Chapter 8 Conclusion and Future Work, finalizing the thesis by concluding the research questions and future directions of this study.

Chapter 2

Related Work

Automated artwork analysers have mainly been done by focusing on classifying aspects such as artist classification [17][5], genre classification [2] as well as style [28]. The main product of the analysis has given rise to algorithms applying CNNs for these tasks, which give successful results in the classification of artworks. Studies such as Karayev et al. [16] demonstrated the effectiveness of CNN models used to evaluate the artwork achieving better classification than human participants on the art style classification. The art analysis has also proven successful in other classifications using CNNs [13][9][4]. These classifications show the increase in an analysis done on artworks. Pattern and object detection in artwork has also been done with work from E. J. Crowley and A. Zisserman. where they created a successful artwork object detector that classified objects found within paintings [8]. Object detection has also been used for finding patterns within artworks [27].

These researches have allowed the ability to evaluate images in pipeline form. Work such as from Gjorgji Strezoski and Marcel Worring, who created a pipeline focusing on object detection, genre detection, school classification, creation period estimation, style classification and colour analysis. Lead to an understanding of how computer science can be applied to art and art history, giving insight into the large-scale comparison of datasets containing artwork [29].

The thesis tries to incorporate the same methodological approach of using a pipeline of machine learning models (CNNs) to evaluate the images generated by AI against human artwork. Focusing on an aspect such as the age and gender of characters within the artwork.

Age and gender have had studies focus on photographic images, not on artistic humans. Studies such as from Gil Levi and Tal Hassner focus on the prediction of the image on humans using CNN models. The result shows promise in the applicability of this method on images [20]. Other studies on age and gender classification yield the same success [14] [1] [25]

These studies give insight into the artworks but do not touch on specific attributes this thesis looks at.

To compare differences, research has been done to evaluate human artworks

against AI artworks[12] in survey form. Our approach is to aid the human evaluation factor. The pipeline will provide a method to compare a large set of images against a singular image. The models used in the pipeline are trained CNNs, making the thesis similar to studies that use machine learning to study and evaluate artwork [6] or images [1]. This carries benefits such as moving away from human subjectivity. Image analysis models create a method of understanding and interpreting the artwork without human bias. Deep learning models such as object recognition and object classification can assist in analyzing some components of the AI-generated art in retrospect to the human. This creates the possibility of analyzing images using attributes not enticed by aspects such as emotion and viewpoints of artwork but looking at objects present and pixel brightness as a comparison leading to a non-bais approach to image analysis.

Most biblical studies using machine learning focus mainly on the text[24] or only on biblical art[23]. Using machine learning or AI generative studies for biblical studies has few implementations. Thus making this thesis the first to inc operate automated evaluating of AI-generated images against human-made artworks from a biblical prompt.

Chapter 3

Data: Prompts, Images, and the Survey

The data collection has two parts. The first is the choice of prompt for the generators, a selection of biblical passages. The second part is the image generator for the dataset. The dataset section answers RQ1 and also is needed for the pipeline, which will use 2 CNN models(Model 1 and Model 3) to evaluate the AI images against the human-made images. The models will be discussed further in Chapter 4.

3.0.1 Prompt selection

This section prompt can be viewed in AppendixA. Prompts were selected to answer the RQ2 and the sub-questions.

Prompt 0 Expulsion from Paradise

Prompt 0 (Genesis 4:23-24) focuses on two human characters, Adam and Eve. It also mentions cherubim, which might be classified as a person in the detector. This prompt was selected since it identifies at least two human characters(one male and one female). The prompt is small compared to the others. Therefore we can evaluate how AI generators can evaluate short biblical passages regarding a male and a female character. In addition, the sentiment from the prompt is negative since it talks about the expulsion of Adam and Eve.

The prompt will allow us to answer RQ2 by accounting for the accuracy of incorporating human artwork compared to AI-generated painting.

SRQ2A and SRQ2B are answered by using model 1 and model 3, which can answer both sentimental and object detection to answer the SRQs—focusing on human detection, age and gender prediction and sentimental prediction.

Prompt 1 Tower of Babel

Prompt 1 (Genesis 11:1-9) focuses on the tower of Babel. The prompt mainly focuses on a negative sentiment; therefore, the result will focus on the sentimental value of the images generated.

The prompt will focus on answering a comparison of the sentimental value of human artwork to AI-generated artwork, which focuses on SRQ2B.

Prompt 2 Binding of Isaac

Prompt 2 (Genesis 22:9-14) focuses on Abram and his son. The prompt has two sentimental values, negative at the start, followed by positive towards the end of the prompt. Therefore, we do not consider the text for in-depth sentimental analysis. The text does bring the opportunity to focus on age and gender detection. Since the father and son characters lead to the assumption of two characters being of different ages, the prompt will focus on object detection by finding at least two characters of different ages.

Prompt 3 The Last Supper

Prompt 3 (Mark 14:12-25) is the most popular prompt chosen, "The last supper". With this prompt, we can analyse the number of people detected, age and gender detection and sentimental value. Since there is an exact number of people, model 1 can evaluate the number of people detected and also the age and gender of them. The characters mentioned in the model are male characters. This will evaluate the male detection of model 1. This can answer RQ2, looking at SRQ2A and SRQ2B.

Prompt 4 Moses Found

Prompt 4 (Exodus 2:5-9), is a prompt that targets female characters and helps evaluate the generator's ability to detect mainly female characters from model 1. There is also a mention of a minimum of 4 characters present in the prompt. Mainly focusing on answering SRQ2A and SRQ2B.

3.1 Image generation

For this thesis, the data collection is a contribution to the field. The thesis produced 7,116 images from 3 texts-to-image generators; DALL-E 2, Midjourney and Stable Diffusion. Stable Diffusion has seven variants leading to 9 image generators producing these images. The data was obtained by creating automation for all three text-to-image generators. Images can be located on ImageKit[1] and Google drive [2]

[1]Images produced `https://imagekit.io/dashboard/media-library/L2RhdGE?view=LIST`.
[2]Images produced `https://drive.google.com/drive/folders/0AHlL9nE6GybgUk9PVA`.

Table 3.1: Table of AI generator used with detail.

Generator	Version	Link	#Images
DALL-E 2	V1 beta	openai.com/v1/images/generations	500
Midjourney	–v 5.1	discord.gg/midjourney	616
Stable Diffusion	V1.4	SG161222/Realistic_Vision_V1.4	1000
Stable Diffusion	V1.5	runwayml/stable-diffusion-v1-5	1000
Stable Diffusion	V1.4	CompVis/stable-diffusion-v1-4	1000
Stable Diffusion	V2.1	stabilityai/stable-diffusion-2-1	1000
Stable Diffusion	V1.1	prompthero/openjourney	1000
Stable Diffusion	V1.1	nitrosocke/Ghibli-Diffusion	500
Stable Diffusion	V2.0	dreamlike-art/dreamlike-photoreal-2.0	500

The images were produced through an automated process where the prompt was fed repeatedly into the generators. For DALL-E 2, the size of the prompt exceeded the character limit. Thus, prompts 1 and 3 were reduced by using NLTK Library.[[10]]. The automated process reduced human error and allowed images to be reproduced using the same automation process. Creating this automated process led to creation of the database to answer RQ1.

3.1.1 Artwork selection

The artwork chosen can be viewed in Appendix B. The biblical artwork chosen to compare to the AI-generated images are paintings from the European and North America Renaissance period. These paintings have similarities in the context of the biblical and in art style. Since the AI generator produces images from the same network, we try incorporating the same behaviour with the human artwork. Choosing a time period narrows down the sample group of biblical art to more similar like-minded artists. In addition, choosing early Western civilisation also brings up similar cultures and societal norms that the artist experience. Even though this type of thinking is flawed and not the most robust academic approach, it is an easy solution to have art that is linked to a degree following a similar concept. The picking of the paintings of Renaissance art was done by considering the visibility of characters and accessibility to give the art a fair possibility for the machine learning models to evaluate it.

3.2 Survey

The survey followed the gold standard protocol except for random sampling since individuals who understand the Bible were chosen. This choice was taken so aspects such as implied objects could be asked, and the individuals would have a broader story scope, thus giving a more in-depth analysis of implied objects. In addition, there is a section where we ask the participant to give inputs to labels we could have missed concerning the text. Participants' understanding of the text will give more religious accurate labelling information.

Each piece of art was surveyed by at least 5 annotators. The survey includes the following: 1) text analysis 2) human detection (e.g. the number of people) 3) weather classification: list of weather present 4) animals classification: list of animal families to choose from 5) environment classification: list of environment features that might present 6) Labels we might not have considered. The survey had multiple purposes. The first was to measure the consensus of people when asked to analyze the picture. The second was to check the difference between model prediction and people prediction. The survey's findings for the human detection section will be discussed further in Chapter 7. [3]

[3]Survey results can be found here `https://drive.google.com/drive/folders/15RMSd6QnSt6VU6byMDmaO73iPbbwH4Fz?usp=drive_link`

Chapter 4

Methodology

The full scope of the project can be seen in Figure 4.1. The yellow label shows the aspect implemented in this thesis. The pipeline to answer RQ2 is in the dotted automated image analysis model. The thesis explores a pipeline of evaluating AI-generated images from biblical prompts by considering the AI generator as a black box. Considering all the images from the generators as output, we evaluated the characteristics of human objects and the sentimental value of the images.

The thesis does not dive into the AI image generator, not looking at the algorithm or architecture of how the image is produced. The main focus is on evaluating the pipeline and the scores it produces. We consider evaluating the art in 2 sections; human object detection and sentimental value of the image (either positive or negative sentiment). From there, we create a score per image by comparing it to the Renaissance paintings (base paintings) per human detection and sentimental value criteria.

Object recognition and classification models have become popular in analyzing images and producing results that can describe attributes within the image. CNN is one of the most common algorithms used to evaluate images. A dataset such as Imagenets[26], which is a large hand-labelled dataset that has supported the rise of the effectiveness of the CNN models. Its contribution has allowed models such as AlexNet [18], GoogLeNet [3], ResNet50 [11] and others to create models prominent in object recognition and classification with images. These models are valuable in obtaining information across different applications.

Using CNN models will help create an automated model which will be able to evaluate large image datasets. In our case, we compare images produced by AI to the base paintings. Evaluating specific attributes classified by these models makes comparison feasible and provides a new approach to evaluating biblical art.

These models produce evaluation reports that span over different criteria. The pipeline expands on previous works on the models used in the evaluation. The pipeline will try incorporating different attributes, such as object recognition for human detection and pixel analysis for sentimental prediction. In

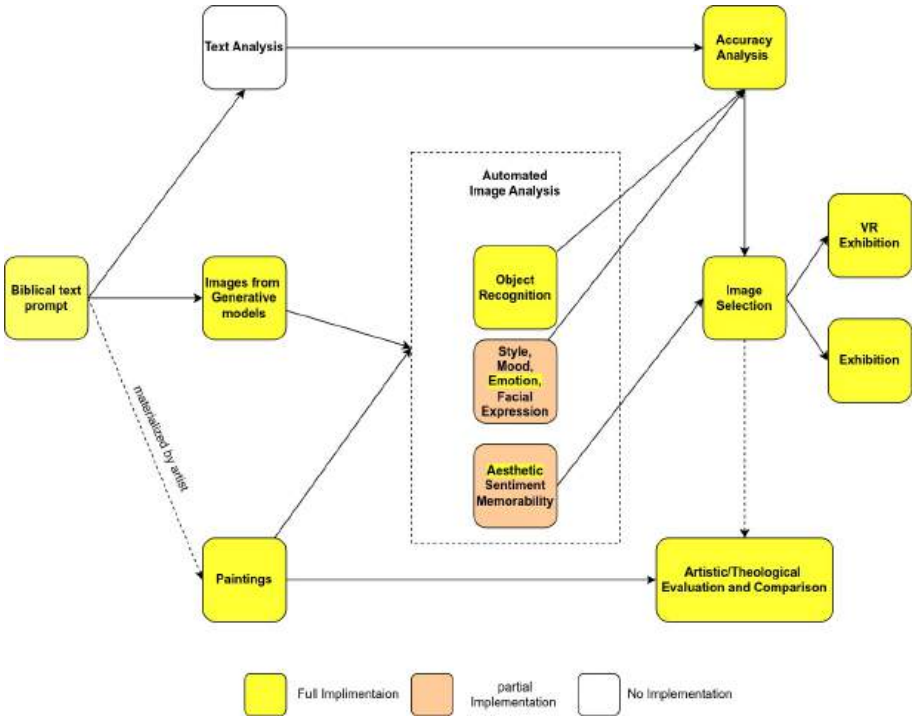


Figure 4.1: Pipeline of process

addition, considering AI-generated images as artwork will allow for aesthetic evaluation of the images produced.

Using a pipeline of CNN models will give an insight into RQ2 and its sub-questions. In addition, it will allow us to evaluate the sentimental and object detection models.

4.1 Research Design

We will conduct research to assess widely-used text-to-image generators by utilizing object recognition and detection tools. Table 4.1 displays the machine learning tools that we will utilize. Models 1 and 2 are capable of identifying and analyzing the characters' age, gender, and presence in the images and artworks. This will address the object detection and recognition segment of the pipeline and will answer SRQ2A. To answer SRQ2B, we will consider the sentimental value of the images by utilizing model 3, which examines the brightness of the pixels (whether they are dark or light in colour) and predicts the sentiment.

The evaluation of model output will be evaluated by comparing the results against the base paintings.

The comparison will be by producing a scoring formula that brings perspec-

Table 4.1: Table Showing the Models Used and Respective Information

Generator	Target	Model	Database
Human Recognition (Model 1)[30][20]	Age, Gender, Number	Mask R-CNN, ResNet-50 [11] + FPN[21] LeNet-5[19]	COCO[22], Cityscape[7], ImageNet [26]
Human Recognition (Model 2)	Age, Gender, Number	Cameralyze	Cameralyze
Sentimental Classification (Model 3)	Positive or Negative Value	AlexNet[18]	Twitter Dataset[31]

tive on the differences as well as an evaluation of the CNN models.

One way to understand the thesis is to see it as an experiment where the prompt is the independent variable and the human and AI-generated images are the dependent variables. The thesis will analyze the outcomes of the generated images using a pipeline. The main focus of the thesis will be the pipeline's results and the discussion around them. The generator specifications and pipeline CNN models' weights and architecture are constant variables. The pipeline's result will be a score that compares the ability of human-made art to AI-generated art. Additionally, the pipeline will provide insight into the performance and limitations of CNN models for the experiment.

The methodology process is split into CNN models for the pipeline and data analysis.

4.2 Deep Learning Models for Pipeline

The aim of the pipeline is to provide answers to SRQ2a and SRQ2b by utilizing Models 1, 2, and 3. The thesis has automated the pipeline process to create a comparison using a simple method. In the next section, we will examine the architecture of the models and the aspects they assess.

4.2.1 People Detector Model 1

Model-1 combines 2 existing pre-trained CNN models, Detectron2 [30] and the Age Gender [20] Model.

Detectron2

Detectron2 is a Mask R-CNN with a ResNet50 and FPN (Feature Pyramid Network) as the backbone. It uses Mask R-CNN and extends the Faster R-CNN model by masking to achieve pixel-wise segmentation. The Mask R-CNN is four layers of 3x3 convultion applied to a 14 x 14 input feature map, then passes through a deconvelution layer which transforms it using a 2x2 kernel and ends with a 1 x1 convultion network that predicts the mask logits. This model is used for mapping the segmentation and is trained on the COCO dataset with

8 categories and the Cityscape dataset and predicted using the backbone model. We only identify the human label to be found from a target image.

The confidence score of 80 per cent was used to predict the humans and returned the array of bounding boxes for each person detected. This model returns the number of people detected in the image.

Age and Gender Detector

For age and gender detection, a custom CNN was developed by Gil Levi and Tal Hassner [20] based on LeNet-5, which consists of 3 convolutional layers and 2 fully connected layers with a few neurons. Each layer of the CNN is followed by ReLu and normalization before being passed on to the next. At the same time, the fully connected layers contain both 512 neurons, which, after passing both layers, are mapped to the final phase classes of age or gender. The Imagenet dataset contains 1,281,167 training images, 50,000 validation and 100,000 test images.

The network produced an age prediction in the form of a range of minimum and maximum age predictions. The average was taken from this and stored as the predicted age.

4.2.2 Cameralyze Model 2

Cameralyze is a commercial platform that uses pre-built models for Artificial Intelligent tasks. Age and Gender detectors were used from the platform as a verification process to compare against model 1. Camaralyze provides an API which was used to collect data on the dataset. This model is mainly to evaluate Model 1.

4.2.3 Model 3 Sentimental Model

The Sentimental Recognition uses a CaffeNet CNN architecture [15], an AlexNet-styled network, composed of five convolutional layers and three fully-connected layers where the base model is trained on a Twitter dataset. It compared the brightness of the pixel in the network with the model, tending to map brighter pixels to more positive sentiment. The model uses Caffe Classifier to classify images with a value between [0, 1]. The value represents negative and positive, respectively. The mapping of the sentimental value is done using Caffe Network. The network uses 2 kernels that produce a 8 x 8 prediction map that fits over the selected image to represent the fully convolutional network.

To obtain different sentimental scores, the CaffeNet architecture was turned into a fully convolutional network. Since no additional training was needed other than the rearrangement of weights. The result led to a 8 x 8 prediction map containing 64 patches of the images, with each patch having a sentimental score. Since the average of the prediction of the 8x8 map is done per patch, the result is not equivalent to the convolutional network, which takes the whole image for the prediction, not patch-by-patch evaluation, therefore, carrying a different

result. The reason to incorporate the two different versions of the sentimental analysis is first to be able to compare the whole image in comparison to the human images. The second architecture allows us to compare 64 patches in AI-generated images to patches generated in the same location in human-made art. This gives us the result of comparing sections of the image's sentimental value. The sentimental model helps us answer RQ2, in particular, SRQ2B.

4.3 Data Analysis

The models in each section of the pipeline produce a prediction result of its respective target. To answer RQ2, the results are compared against the artistic human art piece of the same prompt. The comparison uses mathematical formulas to set the prediction number in an interval [0, 1]. For each model 1 and 3, the scores are added and divided by two to keep it in the [0, 1] format. The formulas can be seen in appendix D.

4.3.1 Data Analysis: Number, Age and Gender

For Model 1, each value is turned into a score. The score differs from the human artistic paintings referred to as base paintings. Firstly model 1 separates the score into 4 sub-scores: Number of people detected, Number of Males detected, Number of Females detected and Number of ages detected. Each sub-score is transformed into an interval, then added up and divided by 4 (number of sub-scores).

Number of people detected

The Number of people detected per image is turned into an interval by taking the maximum detected number of people and turning it into the range of the number seen in appendix D.1. Therefore each prediction will be in the range of the interval. For example, if we take 12 humans detected in the image as the maximum, another given image with four humans detected will have a score of $4/12 = 0.33$. The same is done for the base, and the difference in score is calculated.

Number of Males and Number of Females

The same approach is for gender, where the maximum detected males and females detected represented their interval range. The number of detected Males and females from the base then subtracts from the number of detected. This creates two scores of values between 0 to 1 for male-recognised and female-recognised, as seen in appendix D.2.

Age score

The age detection was split into age categories in intervals of 10 from 0 to 100 since the model had a maximum age of 100 and a minimum of 0. Each interval accounts for the number of people detected in the age group. The difference from the base is subtracted from the number of people detected per age interval. For example, if one person is detected from 0-10 for the base and 3 for the AI-generated image for that interval, 3-1 = 2 is the difference. All intervals' differences are tallied and divided by the maximum number of people detected. This creates an interval score between 0 and 1. The score is then combined and divided by 4 to create the score for model 1. This formula can be seen in appendix D.3.

4.3.2 Sentimental Data Analysis

Full Image

The classifier gives an interval between 0 and 1. The score represents the prediction of the full images passed through the CaffeNet architecture. The score is then subtracted from the score of the base image to create a value representing the difference in the images as seen in appendix D.4.

8x8 Map

The classifier gives an interval between 0 and 1 for each section of the image. An image contains 64 patches. For each patch, the sentimental score is taken. Each patch is compared to the human artwork, and a difference is scored. Each individual patch is tallied up with the others and then divided by 64 to create an interval score of comparison in the [0,1] range. This formula can be seen in appendix D.5.

Chapter 5

Automated Evaluation

The pipeline evaluation [1] will look at the result produced by the pipeline per prompt and then evaluate each model in the pipeline on the prompt. The score produced will be from model 1 and model 3. The score will then be evaluated using graphs, and the focus point of each prompt will be mentioned concerning the results found.

5.1 Prompt Evaluation

5.1.1 Prompt 0 Expulsion from Paradise

The focus of prompt 0 was looking at how Model 1 incorporates the characters of Adam and Eve, focusing on the number of people detected by the prompt and their respective gender. The prompt sentimental score given by model 3 is also considered since focusing on the sentiment is also essential in the given prompt.

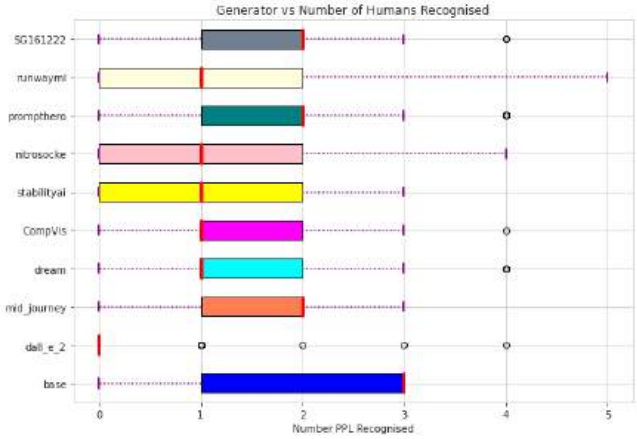
Table 5.1: Prompt 0 score

Generator	Model 1 Score	Model 3 Score	Pipeline Final Score
Midjourney	0.125519	**0.104197**	**0.114858**
SD[CompVis]	0.125975	0.115730	0.120853
SD[prompthero]	**0.117925**	0.129476	0.123700
SD[stabilityai]	0.132212	0.120008	0.126110
SD[dream]	0.125775	0.129355	0.127565
SD[nitrosocke]	0.127562	0.129496	0.128529
SD[SG161222]	0.126137	0.157378	0.141758
SD[runwayml]	0.141788	0.159858	0.150823
DALL-E	0.204475	0.166011	0.185243

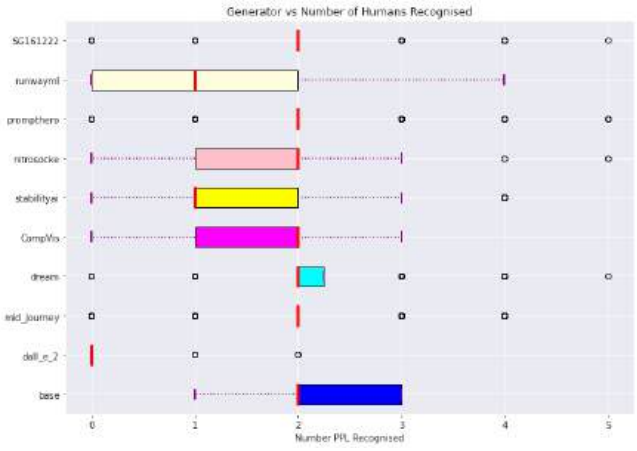
[1] pipeline code `https://github.com/ChiefGitau/bacholar_project`.

For prompt 0, the base paintings have the closes representation to Midjourney, which scored the lowest in the sentimental value difference. Stable Diffusion Prompthero scored the best for human object detection in model 1.

Human detection and classification from model 1 score have small differences, with the highest score being 0.2044475 and the lowest 0.117925. This indicates that most of the generators have similar object numbers and sentimental scores to the base paintings. We can see in fig 5.10b the human detection distribution and the models. It shows the similarity in the mean scores for most of the AI-generated images, with all the AI other than DALL E having ranges that coincide with the base paintings. Model 2 replicates this behaviour seen in 5.10b, showing that the detection is consistent in both models.



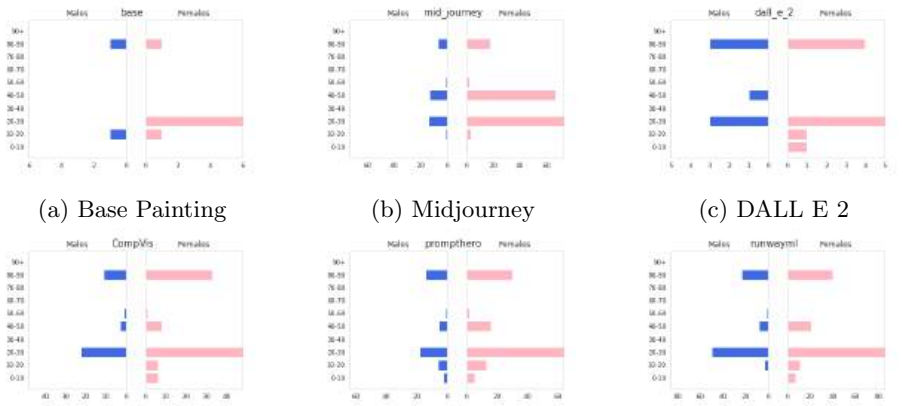
(a) Model 1



(b) Model 2

Figure 5.1: Number of people detected per generator prompt 0

The population pyramid in Figure 5.2 shows the distribution of males and females and their respective age categories. It indicates that there is not an equal disparity of Male and Female characters detected, while the age groups seem to focus on the 20-30-year-olds for all the images. The age of 'Adam and Eve' is presumed to be around the same age, this shows that the age prediction is accurate. However, the gender distribution does not represent 2 entities of a male and a female but seems to be predominantly female. The reasons for this will be discussed further in the discussion. This issue arises in all of the other prompts as well.



(a) Base Painting (b) Midjourney (c) DALL E 2

(d) CompVis (e) Prompthero (f) Runwayml

Figure 5.2: Population Pyramids of the Generators

In Figure 5.3, we can see the sentimental value score for various images. The majority of images have a positive sentiment, indicated by the brighter pixels. Overall, the sentiment of the images is positive, with the Stable Diffusion models having a higher mean score than the base paintings. This suggests that for prompt 0, the AI generates more positive sentiment values. However, the Mid-Journey and DALL E have a similar mean score to the base paintings. When we refer back to Table 5.1, we can see that DALL E still performs poorly in sentimental comparison to the base paintings. This indicates that the 8 x 8 map of sentimental value does not align with the same locations as the base paintings.

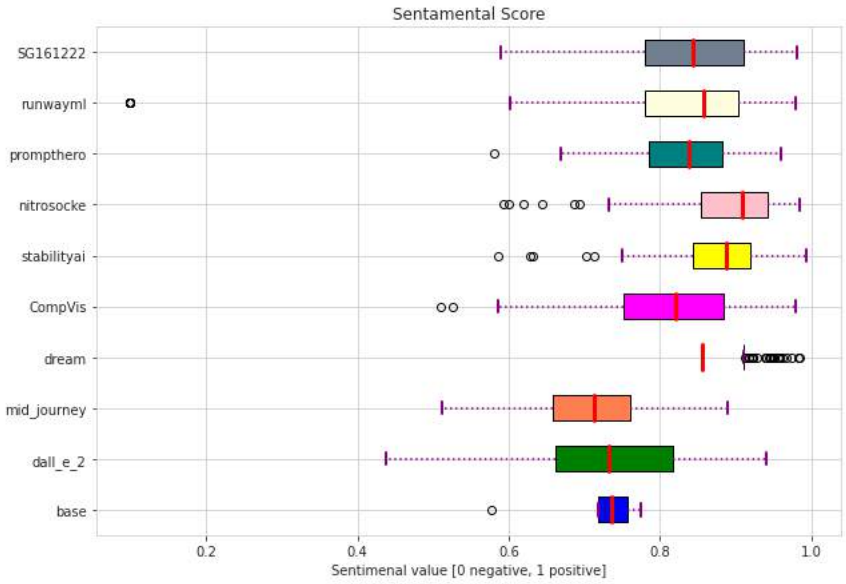


Figure 5.3: Mean Sentimental score from each generator prompt 0

5.1.2 Prompt 1 Tower of Babel

Table 5.2: Prompt 1 score

Generator	Model 1 Score	Model 3 Score	Pipeline Final Score
Midjourney	0.086986	**0.152223**	**0.119604**
SD[dream]	**0.085516**	0.175068	0.130292
SD[prompthero]	0.090185	0.171775	0.130980
SD[SG161222]	0.126137	0.164494	0.132222
SD[CompVis]	0.090471	0.177714	0.134092
SD[stabilityai]	0.103747	0.165785	0.134766
SD[nitrosocke]	0.088247	0.183435	0.135841
SD[runwayml]	0.087092	0.189718	0.138405
DALL-E	0.086078	0.190783	0.138430

Table 5.2 shows the pipeline results for prompt 1. Midjourney has the smallest difference overall with the base paintings than any other AI generators. Prompt 1 'The Tower of Babel' is a prompt that focuses mainly on the sentiment since it does not make a direct reference to how many characters there are. This can be seen in 5.4, the range of humans detected varies and shows low mean score with multiple outliers. This shows the prompt is not ideal for human detection.

The sentimental mean score is similar to the base paintings as shown in figure

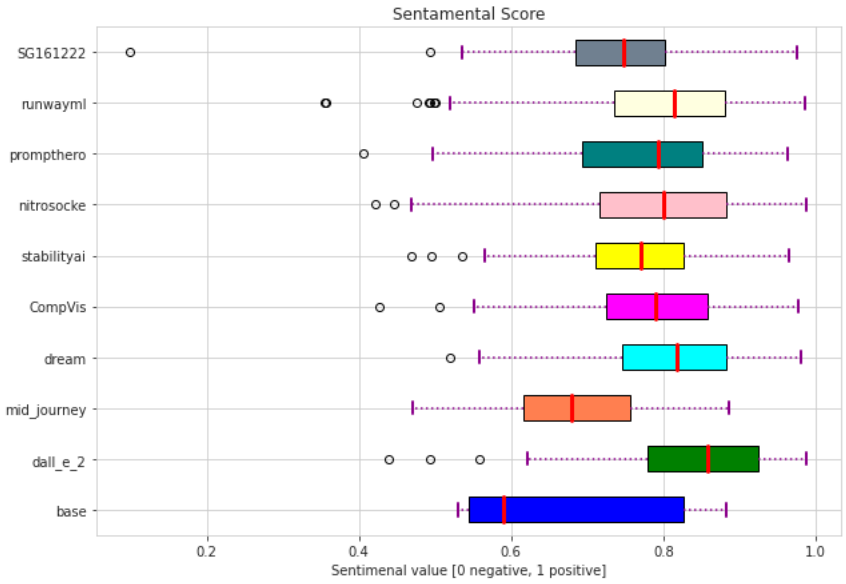


Figure 5.4: Mean Sentimental score from each generator prompt 1

5.4. However, the base paintings seem to have a more natural sentimental mean score than the generators, with a higher positive value for the sentiment. Here, Midjourney shows a similar score to the base paintings. The fact that all the images have a higher mean in sentimental score compared to the base paintings seems that the AI does not interpret the prompt as negatively as the humans. The context of the more positive images shows a limitation from the prompt interpretation of the text and will be discussed further.

5.1.3 Prompt 2 Binding of Isaac

Table 5.3: prompt 2 score

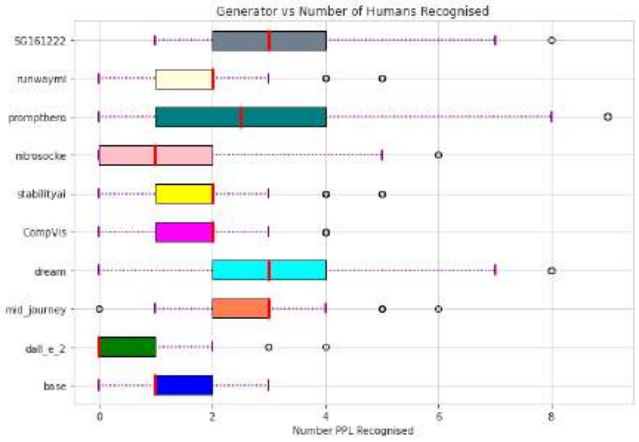
Generator	Model 1 Score	Model 3 Score	Pipeline Final Score
SD[runwayml]	0.074278	0.169080	**0.121679**
SD[CompVis]	**0.069459**	0.174341	0.121900
SD[stabilityai]	0.072664	0.172033	0.122349
SD[nitrosocke]	0.074587	0.174468	0.124528
Midjourney	0.086868	**0.167156**	0.127012
DALL-E	0.085009	0.181598	0.133303
SD[SG161222]	0.096199	0.176700	0.136450
SD[prompthero]	0.109132	0.172435	0.140784
SD[dream]	0.112952	0.190198	0.151575

Prompt 2 focuses on the story of Abraham sacrificing his son. The Prompt was used to detect the age categorization of Abraham and his son. The sentimental score was also considered necessary in this prompt.

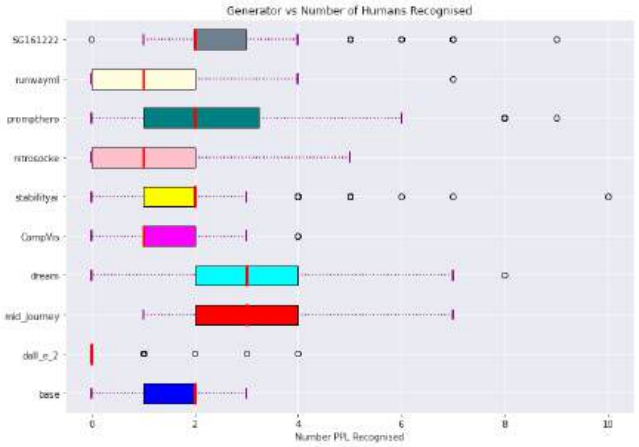
Table 5.3 shows that the Stable Diffusion model has the lowest overall score, especially in human detection difference against the base images. Understanding the result from model 1 for this prompt, we can look at 5.5b. The box plot shows that the mean score for the base paintings is much lower than AI generators other than DALL E, whose mean score was close to 0 shows almost no people were detected in the images. The base painting scoring a mean of 1.5 people shows that in most paintings, roughly 2 were detected, showing some accuracy for what we are looking for. For Stable Diffusion, even though some versions, like Nitrosocke, have the same mean score as the base, the mean for most models tends to have a larger number of humans detected showing more than 2 people were detected per image. The AI incorporates more characters than humans interpreted from the prompt.

The age group detected in model 1 can be seen in the 5.6

The population pyramid shows that in the base paintings, there is evidence that two age categories are shown in the prompt. This is also somewhat present in the AI generators with them having the 80-90-year-old category and the 20-30 category being the most populace. This shows that 2 categories can fill the Father-son role. However, the AI does not separate the characters in age the same as the base paintings. The gender classification also shows more females than males detected in the images which is the same limitation as for prompt 0. The age categorization of the two characters Abraham and his son shows that the AI did not interpret the age difference.

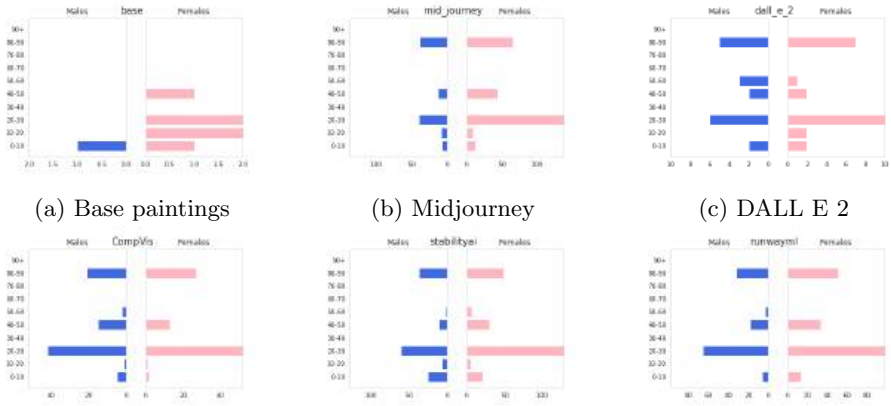


(a) Number of people detected per each Generator for Model 1



(b) Number of people detected per each Generator for Model 2

Figure 5.5: Number of people detected per generator



(a) Base paintings (b) Midjourney (c) DALL E 2

(d) CompVis (e) StabilityAI (f) Runwayml

Figure 5.6: Population Pyramids of the Generators

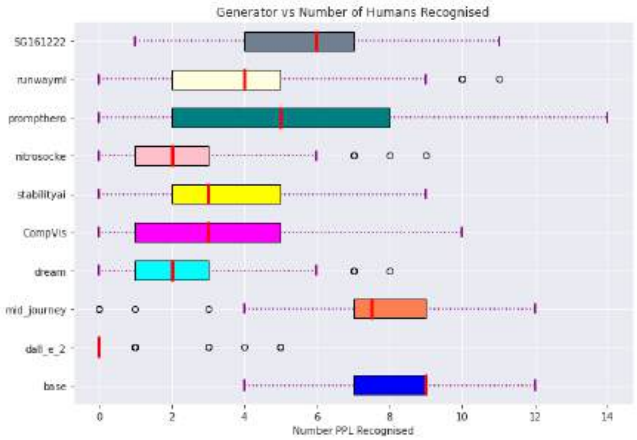
5.1.4 Prompt 3 The Last Supper

Prompt 3 is the most popular prompt, "The Last Supper". The prompt evaluated the number of people detected since it refers to 13 specific male individuals. The score of the pipeline and its analysis will show the effectiveness of large groups of characters presented in a prompt. Table 5.4 shows the scores of each generator, and we can see that mid-journey has the lowest score in each category.

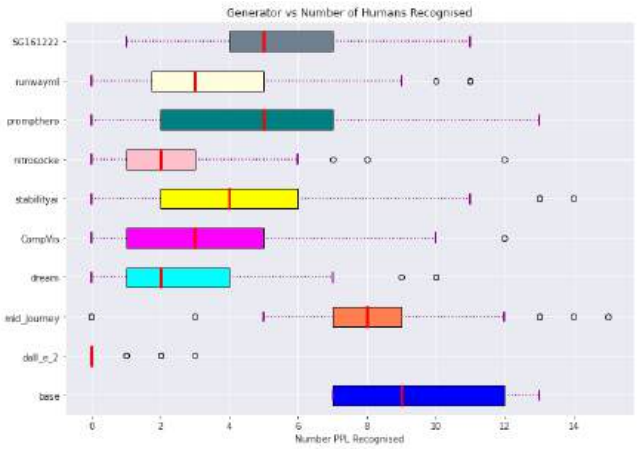
Table 5.4: Prompt 3 score

Generator	Model 1 Score	Model 3 Score	Pipeline Final Score
Midjourney	**0.108251**	**0.149140**	**0.128696**
SD[SG161222]	0.152054	0.149589	0.150822
SD[prompthero]	0.180073	0.152455	0.166264
SD[stabilityai]	0.197527	0.151170	0.174348
SD[runwayml]	0.206087	0.151606	0.178847
SD[CompVis]	0.226405	0.158818	0.192612
SD[dream]	0.240988	0.162855	0.201921
SD[nitrosocke]	0.248754	0.168720	0.208737
DALL-E	0.324126	0.173894	0.249010

5.7 shows the number of people detected in the images as a box plot. We can see for both model 1 and model 2 that the AI images' mean score is less than the base paintings, indicating that fewer characters are created in the images. Midjourney performed best with the most similar mean score to the base, while DALL E produced a mean score of 0 detected people. This shows that even though Stable Diffusion and Midjourney create few outliers that detect



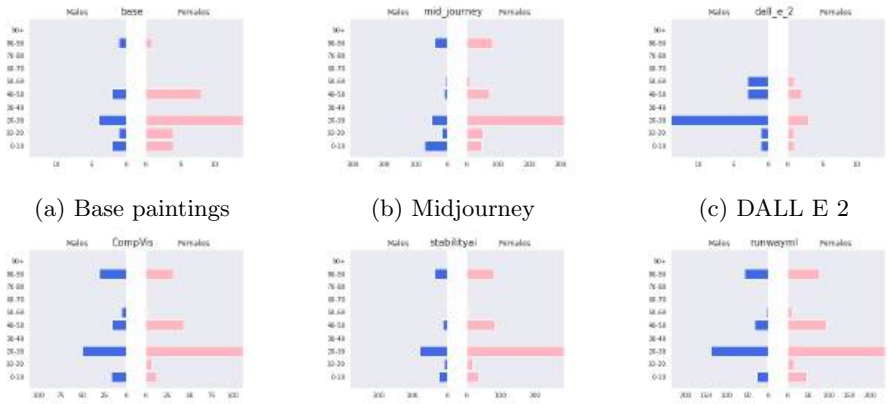
(a) Model 1



(b) Model 2

Figure 5.7: Number of people detected per generator prompt 3

numerous people, the generators do not produce the same amount of people as an artist has incorporated in their artworks.



(a) Base paintings (b) Midjourney (c) DALL E 2

(d) CompVis (e) StabilityAI (f) Runwayml

Figure 5.8: Population Pyramids of the Generators prompt 3

The population pyramid 5.8 shows exciting results. Prompt 3 refers only to male characters. However, the pyramid indicates that the pipeline mainly detects female characters. This is a limitation within the age and gender detection mentioned before.

The sentimental mean score can be seen in Fig 5.9, where it can be seen that the images were mainly labelled as positive. Which shows that bright pixels were mainly used in the images.

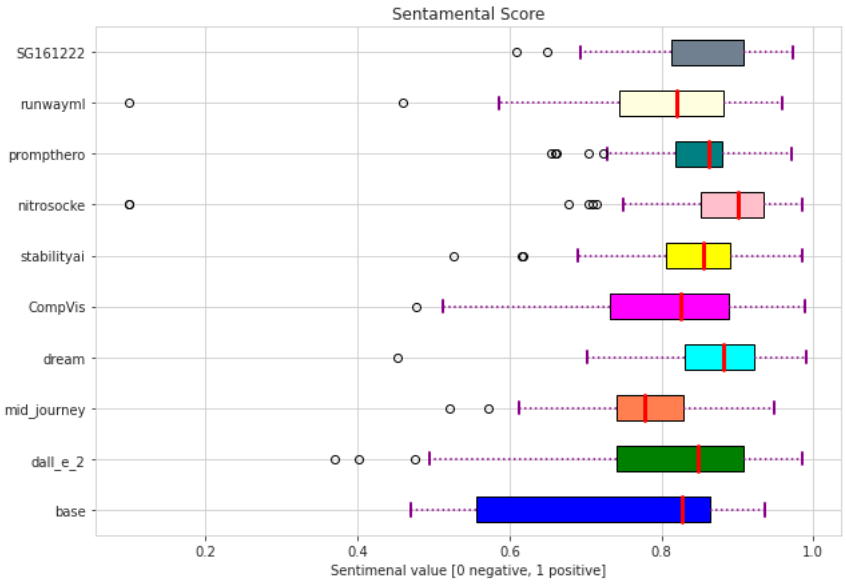


Figure 5.9: Population Pyramid Prompt 3

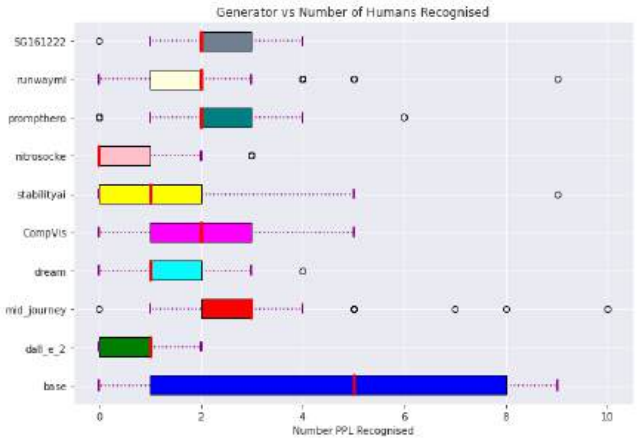
5.1.5 Prompt 4 Moses Found

Prompt 4 focused on female characters and age categorization. Table 5.5 shows the results. Mid-journey has the best results for both model 1 and model 3.

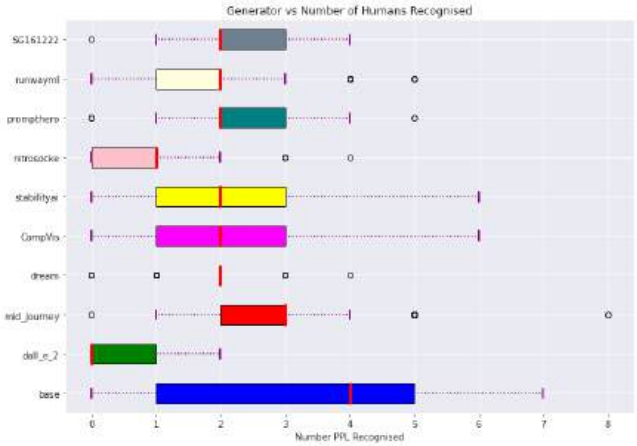
Table 5.5: Prompt 4 score

Generator	Model 1 Score	Model 3 Score	Pipeline Final Score
Midjourney	**0.108251**	**0.149140**	**0.144841**
SD[SG161222]	0.152054	0.149589	0.159754
SD[prompthero]	0.180073	0.152455	0.161308
SD[stabilityai]	0.197527	0.151170	0.165702
SD[runwayml]	0.206087	0.151606	0.166535
SD[CompVis]	0.226405	0.158818	0.169044
SD[dream]	0.240988	0.162855	0.169835
SD[nitrosocke]	0.199292	0.168720	0.184006
DALL-E	0.202878	0.173894	0.188386

Figure 5.10 shows that other than the mid-journey, all of the other generators fall within the range of the base paintings in the number of people detected. However, the mean number of people detected is far lower than the base paintings. This indicates that for prompt 4, the Renaissance painters interpreted the text with multiple individuals present. The prompt indicates multiple individuals being present in the text with "daughter", "maidens" and "baby". This



(a) Model 1

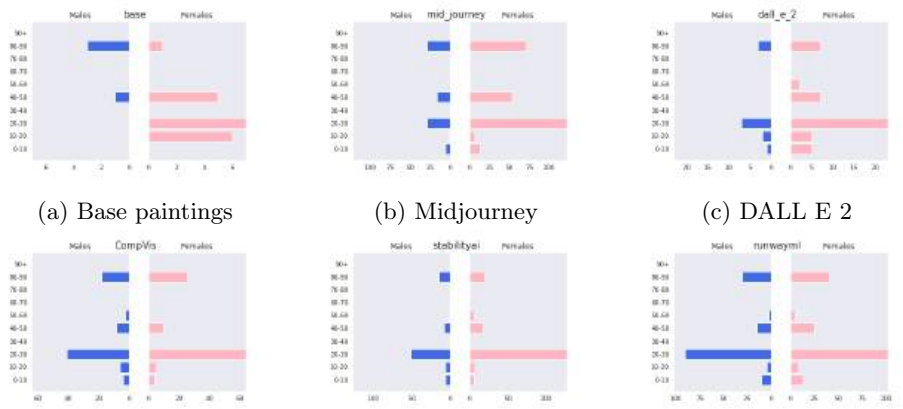


(b) Model 2

Figure 5.10: Number of people detected per generator prompt 4

implies at least a minimum of 4 characters being present in the prompt. The mean score of all generators falls below this threshold. This implies all the text generators fall short of capturing these processes from the text. On the other hand, the human-made Base painting does fall above this criteria with a mean score of 4 and 5 for Cameralyze and model 1, respectively.

Figure 5.10 shows that other than the mid-journey all of the other generators fall within the range of the base paintings in the number of people detected. however, the mean number of people detected is far lower than the base paintings. This indicates that for prompt 4 the Renaissance painters interpreted from the text multiple individuals present. The prompt indicates multiple individuals being present in the text with "daughter", "maidens" and "baby". This implies at least a minimum of 4 characters being present in the prompt. The mean score of all generators falls below this threshold. This implies all the text generators fall short of capturing these processes from the text. On the other hand, the human-made renaissance painting does fall above this criteria with a mean score of 4 and 5 for Cameralyze and custom Detectron2 and AlexNet model, respectively.



(a) Base paintings (b) Midjourney (c) DALL E 2

(d) CompVis (e) StabilityAI (f) Runwayml

Figure 5.11: Population Pyramids of the Generators prompt 4

Figure 5.11 shows the age and gender detection of all the images processed through model 1. We can state that the text-to-image generators generate a majority of female characters. In the prompt, "daughter" and "maidens", which are female characters, are the focal point, and we can state that the generator incorporates these attributes into the generated images. We can also see that the age group mainly targeted is from 20 -30 years which is an exciting interpretation. This is also visible in the Renaissance paintings where the detector has the same age and gender relationship to the extent that 20-30-year-olds are the majority category as well as females. Therefore the process of identifying human characteristics from age to gender is similar, while the generated images for this prompt fail to incorporate the right number of people.

5.2 Result Summary

Table 5.6 shows the score of all the prompts, Midjourney being the best performing AI regarding similarities to the base paintings. From all the previous tables, we can see that mid-Journey excels in the sentimental score, having the smallest difference from the base paintings. Reasons for this will be discussed further in chapter 7 and 6. Stable Diffusion models are ranked from 2nd to 8th, the model version chosen has different solid points for different points. In addition, prompt 3 outperforms Midjourney. In addition, human detection outperforms mid-journey in prompts 0, 1, and 2.

The table also shows that Promphero is the best version of stable diffusion with the most similarities to the base paintings. The difference in each version of Stable Diffusion has solid points for the different prompts. This is an aspect which will be discussed later on.

Overall we can state that SRQ2A and SRQ2B can be evaluated in this pipeline. The scores show the comparison of human art to AI art. Even though each Generator scores differently per prompt, there are trends like Midjourney having the most similarities and DALL E least to the base paintings.

		Base	Midjourney	Dall E 2	CV	PH	SAI	DA	NS	SG	RW
Prompt 0	Male STD	0.547723	0.636430	0.196946	0.742369	0.867831	0.678159	0.763498	0.825717	0.814921	0.76078
	Male Mean	1.4000	0.621212	0.040000	0.880000	1.120000	0.680000	1.230000	0.960000	1.065000	0.71000
	Female STD	0.836660	1.250000	0.261116	0.673525	0.715979	0.841377	0.841115	0.670240	0.829361	0.65737
	Female mean	0.8000	0.568182	0.050000	0.530000	0.950000	0.825000	0.860000	0.555000	0.840000	0.49500
	Number STD	0.836660	0.841711	0.378594	0.853927	0.867540	0.924254	0.900000	0.966668	0.943624	0.98887
	Number mean	2.2000	1.871212	0.090000	1.410000	2.070000	1.505000	2.090000	1.515000	1.905000	1.20500
Prompt 1	Male STD	0.894427	1.991822	0.140705	1.557128	1.478841	2.233679	0.464497	0.942519	1.403907	1.34230
	Male Mean	0.600000	2.537190	0.020000	0.860000	0.570000	1.025000	0.080000	0.310000	0.670000	0.41500
	Female STD	2.915476	0.926277	0.000000	1.243001	1.233006	1.782551	0.171447	1.042791	2.141619	0.61112
	Female mean	2.000000	0.603306	0.000000	0.480000	0.430000	0.780000	0.030000	0.305000	0.920000	0.22000
	Number STD	2.701851	2.324886	0.140705	2.523506	2.399495	3.718347	0.548552	1.798038	3.220834	1.80501
	Number mean	3.600000	3.140496	0.020000	1.340000	1.000000	1.805000	0.110000	0.615000	1.590000	0.63500
Prompt 2	Male STD	0.836660	1.111379	0.469149	0.957427	1.575828	1.020592	1.506183	0.887646	1.270302	0.98041
	Male Mean	0.800000	2.695652	0.110000	1.050000	2.040000	1.440000	2.290000	0.855000	2.120000	0.94000
	Female STD	0.447214	0.484227	0.345096	0.580752	0.868936	0.684061	0.877093	0.638603	0.899120	0.52569
	Female mean	0.800000	0.253623	0.110000	0.310000	0.550000	0.380000	0.720000	0.315000	0.575000	0.24500
	Number STD	1.140175	1.116057	0.612661	1.039814	2.015571	1.275041	1.696669	1.103261	1.601499	1.16082
	Number mean	1.600000	2.949275	0.220000	1.360000	2.590000	1.820000	3.010000	1.170000	2.695000	1.18500
Prompt 3	Male STD	2.302173	2.188727	0.325825	2.022000	2.022000	2.282642	2.108281	1.597352	1.884284	2.08270
	Male Mean	7.600000	6.960784	0.070000	2.180000	2.180000	3.340000	2.140000	1.465000	4.585000	2.65500
	Female STD	0.707107	1.015125	0.242878	0.895499	0.895499	1.067755	0.659047	0.889639	0.976662	1.02960
	Female mean	2.000000	1.137255	0.040000	0.810000	0.810000	0.840000	0.500000	0.750000	1.030000	0.98500
	Number STD	2.792848	2.390202	0.469149	2.430841	2.430841	2.724835	2.199725	1.899160	2.188773	2.52433
	Number mean	9.600000	8.098039	0.110000	2.990000	2.990000	3.340000	2.640000	2.215000	5.615000	3.64000
Prompt 4	Male STD	2.121320	0.864319	0.277798	0.716614	0.688726	0.749941	0.522233	0.389691	0.742111	0.62443
	Male Mean	2.000000	0.343750	0.060000	0.540000	0.480000	0.520000	0.300000	0.170000	0.545000	0.45500
	Female STD	0.894427	0.963629	0.577350	0.951925	0.910100	1.107262	0.717741	0.808498	0.803510	0.96646
	Female mean	1.400000	2.476562	0.500000	1.270000	2.000000	1.510000	1.500000	0.640000	1.760000	1.27500
	Number STD	2.880972	0.983595	0.640707	1.116407	0.846621	1.223349	0.603023	0.870453	0.559949	1.07838
	Number mean	3.400000	2.820312	0.560000	1.810000	2.480000	2.030000	1.800000	0.810000	2.305000	1.73000

Table 5.6: Summary of all prompt results

rank	Generator	Pipeline Overall Score
1	Midjourney	**0.1270022**
2	SD[prompthero]	0.1446072
3	SD[stabilityai]	0.144655
4	SD[SG161222]	0.1460142
5	SD[CompVis]	0.1477002
6	SD[runwaym]	0.1512578
7	SD[dream]	0.1562376
8	SD[nitrosocke]	0.1563282
9	DALL-E	0.1788744

Chapter 6

Religious and Aesthetic Analysis

In order to answer SRQ2c, we examine the images in the database and explore the importance of religious accuracy and aesthetic quality in the images that are generated. This research question is vital given the interdisciplinary nature of the task. While the pipeline and evaluation compare the images to Renaissance paintings, they do not address the accuracy of religious context and aesthetic features in the images. These aspects will be evaluated in this chapter.

6.1 DALL E

DALL E 2 has produced the least text-accurate images relating to the prompt. The images generated usually fall short of incorporating elements or features described in the text. For example, in prompt 3, the generator produces mainly incomprehensible text as seen in fig 6.1, 6.2 and 6.3. The images have no link to the text and do not contain any reference or attributes to the last supper.

While prompt 0 and prompt 2 have biblical references with images incorporating angel wings, halos and medieval glass that are commonly seen in churches. This shows some biblical reference in the images for some prompts, but it still lacks in connecting the image to the actual story referred to in the prompt. On the other hand, prompt 1 mainly generates arid landscapes or forests, which may relate to phrases from the prompt such as "the whole earth", "a plain", "the face of the whole earth", and "abroad over the face of the whole earth". The images from prompt 1 do not show any accuracy to the story of the Tower of Babel, only the incomprehensible landscape, which has no relationship to the prompt as seen in fig 6.4, 6.5 and 6.6.

Prompt 4 depicts mainly a photographic aesthetic image of a female character. This is evident in the prompt. However, the character's attire and physical appearance do not match the biblical reference, with images of Native Americans, Victorian era, modern-day era clothes and characters as seen in fig 6.7,

Figure 6.1: DALL E text generation prompt 3 example

Figure 6.2: DALL E text generation prompt 3 example

Figure 6.3: DALL E text generation prompt 3 example

Figure 6.4: DALL E prompt 1 landscape forest example

Figure 6.5: DALL E prompt 1 landscape arid example

Figure 6.6: DALL E prompt 1 landscape arid example

6.8 and 6.9.

Figure 6.7: DALL E prompt 4 Modern Asian examples

Figure 6.8: DALL E prompt 4 Native American

Figure 6.9: DALL E prompt 4 Victorian era

DALL E is the least accurate generator for producing these biblical prompts

compared to the others. With a lack of accuracy but for some prompts, there are biblical references in the images even though the references are taken out of context of the prompt.

6.2 Midjourney

The first Aesthetic feature identified in Midjourney is the attention to detail. This can be seen in Fig 6.10, 6.12 and 6.12. Midjourney is fine-tuned towards generating realistic hands. However, it's still facing problems. Some show more than five fingers, some show the overlapping of multiple hands. This flaw does not however affect the overall performance in generating the most realistic hands compared to the other generators.

Figure 6.10: Midjourney prompt 1 hands detail

Figure 6.11: Midjourney prompt 1 hands detail

Figure 6.12: Midjourney prompt 1 hands detail

Midjourney also has consistency in generating detailed faces, This leads to characters having detailed emotions visible on their face. Fig 6.13, 6.14 and 6.15 show the detail in the faces of all the characters depicted. This emotion reflects an uneasy mood this aspect which can be seen in the detail can not be captured in the pipeline's sentimental model since it only looks at pixels. The aesthetic detail thus carries weight in analyzing the image.

Figure 6.13: Midjourney prompt 3 emotion

Figure 6.14: Midjourney prompt 3 emotion

Figure 6.15: Midjourney prompt 3 emotion

One issue with Midjourney is the limited range of images it produces. The images tend to focus on characters and lack diversity, as seen in both DALL E and Stable Diffusion. As a result, the images often share similarities in terms of environment, weather, objects, and art style, as shown in the accompanying figures 6.16, 6.17 and 6.18.

Figure 6.16: Midjourney prompt 0 same pattern

Figure 6.17: Midjourney prompt 0 same pattern

Figure 6.18: Midjourney prompt 0 same pattern

When it comes to religious accuracy in Midjourney, the content produced provides the most realistic religious perspective for the majority of the prompts' generated images. These images resemble storybook illustrations, with historical attire, environments, and characters from the prompts incorporated. However, prompt 0 has some shortcomings in this regard. The religious accuracy is not as strong, and the characters depicted seem to be from Western countries seen in figure 6.16, and the attire is not historically accurate, which shows the limitation and Western bias for prompt 0 image generation. Overall Midjourney shows a high level of aesthetic details in human anatomy and every aspect of the image, but the trade-off comes from a lack of variety in the images.

6.3 Stable Diffusion

It appears that Stable Diffusion combines the styles of both DALL-E and Midjourney, resulting in unique images without a specific theme like those in Midjourney. This allows for the generation of science fiction and fantasy themes. For instance, Prompt 1 showcases these themes as depicted in the accompanying figures 6.19, 6.20 and 6.21.

The images produced exhibit a diverse range of aesthetic styles, lacking any fixed standard. While some images accurately acknowledge religious prompts with halos and crosses, others fall short in their religious accuracy. It is worth noting that there are inaccuracies in certain images produced by Stable Diffusion, such as the use of the cross from a prompt of the old testament example seen in figure 6.22. Which indicates biblical inaccuracy.

One of the drawbacks of Stable Diffusion is that it can create difficulty in accurately depicting human autonomy, such as eyes and hands. Additionally,

Figure 6.19: Stable Diffusion Africa theme prompt 1

Figure 6.20: Stable Diffusion Science fiction theme prompt 1

Figure 6.21: Stable Diffusion Middle ages theme prompt 1

Figure 6.22: Stable Diffusion Cross present in Old Testament prompt 2

some of the resulting images may be nearly identical to the reference image, showing little difference. An example of this can be seen in Figures 6.24 and 6.23. which shows Tower of Babel painting was taken as a reference.

Overall Stable Diffusion gives the most comprehensible versatile results for the prompt. It does not have an aesthetic, artistic style or level of detail as Midjourney, but it creates images that are more detailed than DALL E and more versatile than Midjourney. However, it has less detail than Midjourney.

Figure 6.23: fig:Stable Diffusion generated image

Figure 6.24: fig:Midjourney prompt 3 emotion example 2

Chapter 7

Discussion

7.1 Survey

According to the survey results, there was no unanimous agreement among the participants. This suggests that achieving 100% accuracy in basic tasks like identifying age and gender through object detection depends on the person doing it. It also highlights the complexity of the task. This explains why model 1 and model 2 yielded different outcomes, as each model takes a unique approach and has varying interpretations of the detected human, reflected in the survey participants' responses. The survey's primary value to the thesis is its potential for future research, and it will be revisited in chapter 8.

7.2 Models limitation

Model 1 and model 2 do answer SRQA and RQ2B, showing that it is possible to use object detection and sentimental models to compare AI images to human-made artwork. However, there are limitations to the models that have affected the model's accuracy.

In Figure 7.1, Detectron2 was unable to identify all 13 humans in the picture, only recognizing 12. This limitation may be due to the fact that most of the images used were art pieces. Detectron2 is primarily trained on photograph image types such as COCO, LVIS, and cityscapes, which could explain the reduced accuracy in artistic work. This limitation becomes more obvious as the realism of the art images decreases, making it more difficult for the model to identify human characters. To increase accuracy, one solution would be to use labelled artistic data to train the Detectron2 model.

One of the limitations of the model is its tendency to predict female characters even when they are actually male. This was observed in prompts 0, 2, and 3. An example of this limitation can be seen in Figure7.1, where the characters have male gender traits but are depicted in biblical attire and theme, which the model associates with female attributes. The physical attributes of long

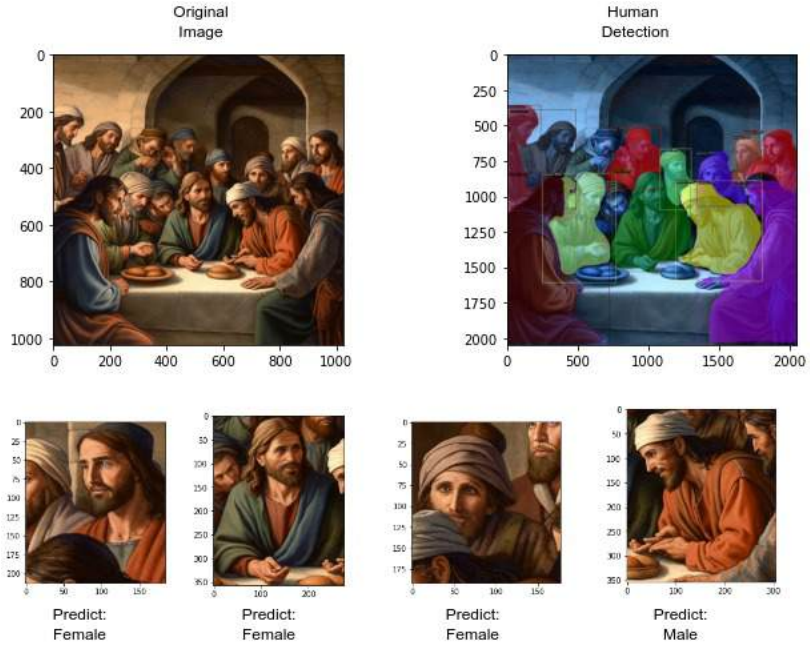


Figure 7.1: Detecting humans and classifying them process example

hair, robes, and head scarves, commonly seen in biblical times, are generalized by the model to be associated with the female gender. This may be due to the training data from ImageNet, which associates long hair with females. This limitation affects the accuracy of gender classification in biblical characters, leading to misclassification in the detector despite their accurate representation in the images.

The sentimental model also had its limitation even though it looks at the pixel brightness as a factor to depict the sentiment of the image. Images that have negative sentiment toward people such as 7.2 scored to be neutral. The score reflects the weakness of bright spots in the image, skewing the score to positive. Therefore for a more accurate sentiment score, other aspects of the image have to be considered other than pixel brightness.

7.3 Aesthetic Link to Pipeline

Some religious accuracy and aesthetic features of specific models explain why a generator might perform better. Midjourney's attention to detail increases the overall realism. This increases the object detection performance as mentioned before. Some versions of Diffusion have issues with human detection since some of the images produced were more abstract.

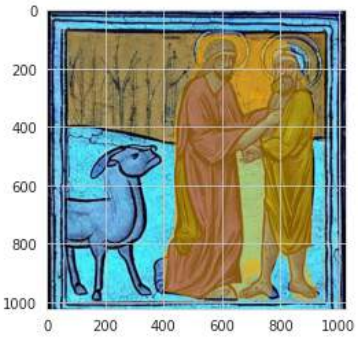
Based on the aesthetic style of the base painting, it appears that the artist

(a) painting: Carvaggio(1598)

(b) Heat Map sentiment Carvaggio(1598)

Figure 7.2: Abstract Image limitation in detection

(a) SD-RUN-P2-191.jpg



(b) Prediction Finds 1 Character

Figure 7.3: Sentiment bright spot weakness

prioritized realism in their imagery. The focus of the painting was mainly on the characters, resulting in a sentimental tone reminiscent of the Midjourney. The pipeline results support this observation, as the sentiment of the Midjourney was most similar to that of the base paintings. This suggests that the art style of the Renaissance period and the Midjourney have similarities in their use of bright pixels.

Chapter 8

Conclusion and Future Work

The thesis provides a data set of biblical images produced by AI generators. Creating this database answers RQ1 and contributes to the interdisciplinary study of computer science, theology and art. The survey also labels the base paintings and can be utilized in future work to understand the classification of the given paintings.

The RQ2 was tackled with model 1 and model 2. The result of the models answers SR2A and SR2B hover with some limitations to both models. The results show how machine learning techniques can be used to evaluate art., proving that the pipeline effectively compares humans to AI biblical by comparing objects and sentimental value and scoring them.

Through this study, it has become apparent that there are still various areas worth researching. Every perspective thus far has brought forth more information on the matter. The database and images are valuable contributions that can be utilized in various ways. The pipeline is an area that has room for growth, with the potential to incorporate additional machine-learning models to improve the accuracy and descriptive nature of art comparisons. Furthermore, the pipeline can be applied to aspects beyond biblical image studies and could be integrated into art institutes. The thesis also presents the opportunity to understand AI from a theological perspective, analyzing how each image generated can differ based on context, accuracy, art style, theme, and other interpretative features. The pipeline aims to incorporate models that classify landscape, facial emotion, and weather, which are not commonly compared in artworks and provide a new context for evaluation. Another approach would be to compare different versions of the bible and evaluate the images produced when using the New King James Version versus the New American Standard Bible. Additionally, some prompts had truncated text so that the study could assess images produced based on different truncated versions. In sum, this thesis offers numerous opportunities for future research and uniquely explores the

intersection of Art, Theology, and Computer Science.

Appendix A

Prompt used for generation of images King James Version (KJV)

Prompt 0

therefore the Lord God sent [Adam and Eve] forth from the garden of Eden, to till the ground from which he was taken. He drove out the man; and at the east of the garden of Eden he placed the cherubim, and a flaming sword which turned every way, to guard the way to the Tree of life.

Prompt 1

Now the whole earth had one language and few words. And as men migrated from the east, they found a plain in the land of Shinar and settled there. And they said to one another, "Come, let us make bricks, and burn them thoroughly." And they had brick for stone, and bitumen for mortar. Then they said, "Come, let us build ourselves a city, and a tower with its top in the heavens, and let us make a name for ourselves, lest we be scattered abroad upon the face of the whole earth." And the Lord came down to see the city and the tower, which the sons of men had built. And the Lord said, "Behold, they are one people, and they have all one language; and this is only the beginning of what they will do; and nothing that they propose to do will now be impossible for them. Come, let us go down, and there confuse their language, that they may not understand one another's speech." So the Lord scattered them abroad from there over the face of all the earth, and they left off building the city. Therefore its name was called Babel, because there the Lord confused[a] the language of all the earth; and from there the Lord scattered them abroad over the face of all the earth.

Prompt 2

When they came to the place of which God had told him, Abraham built an altar there, and laid the wood in order, and bound Isaac his son, and laid him on the altar, upon the wood. Then Abraham put forth his hand and took the knife to slay his son. But the angel of the Lord called to him from heaven, and said, "Abraham, Abraham!" And he said, "Here am I." He said, "Do not lay your hand on the lad or do anything to him; for now I know that you fear God, seeing you have not withheld your son, your only son, from me." And Abraham lifted up his eyes and looked, and behold, behind him was a ram, caught in a thicket by his horns; and Abraham went and took the ram, and offered it up as a burnt offering instead of his son. So Abraham called the name of that place The Lord will provide;[a] as it is said to this day, "On the mount of the Lord it shall be provided."

Prompt 3

And when it was evening he came with the twelve. And as they were at the table eating, Jesus said, Jesus said, "Truly, I say to you, one of you will betray me, one who is eating with me." They began to be sorrowful, and to say to him one after another, "Is it I?" He said to them, "It is one of the twelve, one who is dipping bread into the dish with me. For the Son of man goes as it is written of him, but woe to that man by whom the Son of man is betrayed! It would have been better for that man if he had not been born." And as they were eating, he took bread, and blessed, and broke it, and gave it to them, and said, "Take; this is my body." And he took a cup, and when he had given thanks he got gave it to them, and they all drank of it. And he said to them, "This is my blood of the[b] covenant, which is poured out for many. Truly, I say to you, I shall not drink again of the fruit of the vine until that day when I drink it new in the kingdom of God."

Prompt 4

Now the daughter of Pharaoh came down to bathe at the river, and her maidens walked beside the river; she saw the basket among the reeds and sent her maid to fetch it. When she opened it she saw the child; and lo, the babe was crying. She took pity on him and said, "This is one of the Hebrews' children." Then his sister said to Pharaoh's daughter, "Shall I go and call you a nurse from the Hebrew women to nurse the child for you?" And Pharaoh's daughter said to her, "Go." So the girl went and called the child's mother. And Pharaoh's daughter said to her, "Take this child away, and nurse him for me, and I will give you your wages." So the woman took the child and nursed him.

Appendix B

Human Art

Table B.1: Table of Paintings chosen for Thesis with Artist

Prompt 0	Prompt 1	Prompt 2	Prompt 3	Prompt 4
Michelangelo (1512)	Bartolomeo Cavarozzi (1598)	Lucas van Valckenborch (1594)	Juan de Juanes (1560)	Lucas Van Valckenborch (1635)
Jan Bruehel (1624)	Lucas Gassel (1539)	Pieter Breugel (1563)	Peter Paul Rubens (1632)	Toussaint Gelton (1645)
Benjamin west(1760)	Carvaggio(1598)	Grimmer(1604)	Il Tintoretto(1592)	Jan Kosten(1650)
Izaak van Oosten(1628)	Titiaan(1542)	Hendrick van Cleve(1570)	Hans Holbein de Jonge(1527)	Paolo Veronse(1570)
Cornelis van Poelenburg(1652)	Rembrandt(1635)	Frederik van Valckenborch(1600)	Leonardo da Vinci(1495)	Bartholomeus Breenbergh(1622)

Appendix C

Survey

multiple screenshots [1]

Figure C.1: survey example people detection

[1] Survey results can be found here `https://drive.google.com/drive/folders/15RMSd6QnSt6VU6byMDmaO73iPbbwH4Fz?usp=drive_link`

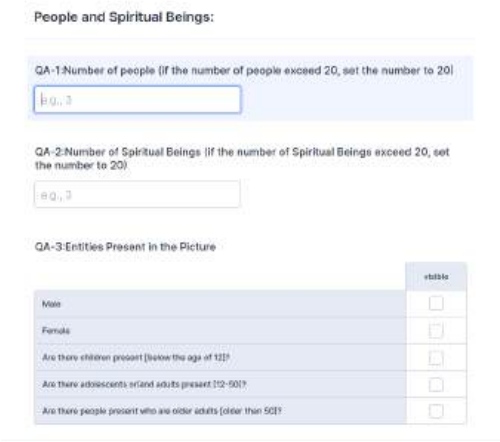


Figure C.2: survey example weather for future work

Appendix D

formula example

$$\| \frac{TargetImagePeople}{NumberOfPeopleMax} - \frac{TargetBasePeople}{NumberOfPeopleMax} \|$$

Figure D.1: number of people comparison formula

$$\| \frac{TargetImageNumberFemale}{NumberOfPeopleMax} - \frac{TargetBaseNumberFemale}{NumberOfPeopleMax} \|$$

Figure D.2: number of gender comparison formula

$$\| \frac{(TargetImageNumberAgeGroup(0-10) - TargetBaseNumberAgeGroup(0-10)) + \ldots + (TargetImageNumberAgeGroup(90-100) - TargetBaseNumberAgeGroup(90-100))}{NumberOfPeopleMax} \|$$

Figure D.3: age array group comparison formula

$$\| \frac{TargetIMAGESentimentalValue - TargetBASESentimentalValue)}{2} \|$$

Figure D.4: sentimental full image comparison formula

$$\| \frac{(Patch[0]TargetIMAGESentimentalSum() - Patch[0]TargetBASESentimentalSum()) + \ldots + (Patch[63]TargetIMAGESentimentalSum() - Patch[63]TargetBASESentimentalSum())}{64} \|$$

Figure D.5: sentimental patch comparison formula

Appendix E

VR Exhibition

The output of this research includes a virtual reality exhibition in the NU building and online at `https://shuai.ai/art/seeing/`.....using ArtSteps[1]....

The exhibition was a part of the Network Institute's end-of-year celebration on 11th July.

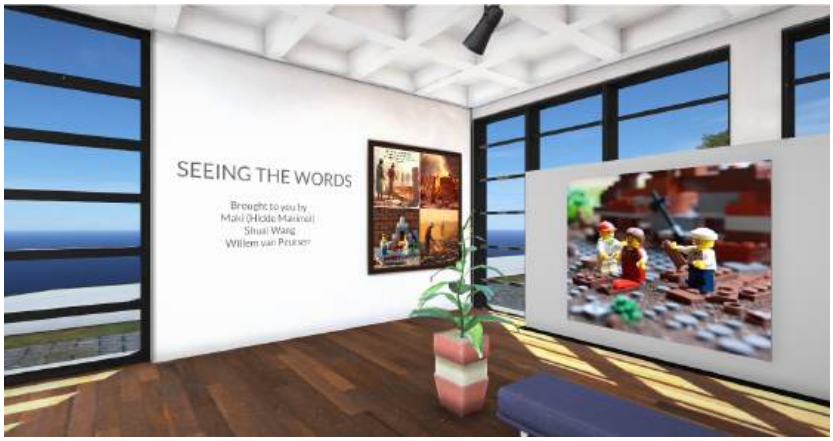


Figure E.1: A really Awesome Image

[1] `https://www.artsteps.com/`

Figure E.2: A really Awesome Image

Figure E.3: A really Awesome Image

Bibliography

[1] Abdullah M Abu Nada, Eman Alajrami, Ahmed A Al-Saqqa, and Samy S Abu-Naser. Age and gender prediction and validation through single user images using cnn. 2020.

[2] Siddharth Agarwal, Harish Karnick, Nirmal Pant, and Urvesh Patel. Genre and style based painting classification. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 588–594. IEEE, 2015.

[3] R Anand, T Shanthi, MS Nithish, and S Lakshman. Face recognition and classification using googlenet architecture. In *Soft Computing for Problem Solving: SocProS 2018, Volume 1*, pages 261–269. Springer, 2020.

[4] Yaniv Bar, Noga Levy, and Lior Wolf. Classification of artistic styles using binarized features derived from a deep neural network. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13*, pages 71–84. Springer, 2015.

[5] Eva Cetinic and Sonja Grgic. Automated painter recognition based on image feature extraction. In *Proceedings ELMAR-2013*, pages 19–22. IEEE, 2013.

[6] Eva Cetinic, Tomislav Lipic, and Sonja Grgic. A deep learning perspective on beauty, sentiment, and remembrance of art. *IEEE Access*, 7:73694–73710, 2019.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[8] Elliot J Crowley and Andrew Zisserman. In search of art. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13*, pages 54–70. Springer, 2015.

[9] Omid E David and Nathan S Netanyahu. Deeppainter: Painter classification using deep convolutional autoencoders. In *Artificial Neural Networks*

and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25, pages 20–28. Springer, 2016.

[10] Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, and Iti Mathur. *Natural language processing: python and NLTK*. Packt Publishing Ltd, 2016.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Joo-Wha Hong and Nathaniel Ming Curran. Artificial intelligence, artists, and art: attitudes toward artwork produced by humans vs. artificial intelligence. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(2s):1–16, 2019.

[13] Nejc Ilenič. *Globoki modeli avtorstva umetniških slik*. PhD thesis, Univerza v Ljubljani, 2017.

[14] Koichi Ito, Hiroya Kawai, Takehisa Okano, and Takafumi Aoki. Age and gender prediction from face images using convolutional neural network. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 7–11, 2018.

[15] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014.

[16] S Karayev, M Trentacoste, H Han, A Agarwala, T Darrell, A Hertzmann, and H Winnemoeller. Recognizing image style. bmvc. 2014.

[17] Daniel Keren. Painter identification using local features and naive bayes. In *2002 International Conference on Pattern Recognition*, volume 2, pages 474–477. IEEE, 2002.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[19] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[20] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops*, June 2015.

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[23] Joanne Morra. Utopia lost: Allegory, ruins and pieter bruegel's towers of babel. *Art History*, 30(2):198–216, 2007.

[24] Ramona Cristina Popa, Nicolae Goga, and Maria Goga. Extracting knowledge from the bible: A comparison between the old and the new testament. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, pages 505–510, 2019.

[25] Insha Rafique, Awais Hamid, Sheraz Naseer, Muhammad Asad, Muhammad Awais, and Talha Yasir. Age and gender prediction using deep convolutional neural networks. In *2019 International conference on innovative computing (ICIC)*, pages 1–6. IEEE, 2019.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[27] Benoit Seguin, Carlotta Striolo, Isabella diLenardo, and Frédéric Kaplan. Visual link retrieval in a database of paintings. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I 14*, pages 753–767. Springer, 2016.

[28] Lior Shamir and Jane A Tarakhovsky. Computer analysis of art. *Journal on Computing and Cultural Heritage (JOCCH)*, 5(2):1–11, 2012.

[29] Gjorgji Strezoski and Marcel Worring. Omniart: a large-scale artistic benchmark. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–21, 2018.

[30] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. 2019.

[31] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 29, 2015.