Human Involvement Can Improve Current Image Synthesis Methods within the Domain of Art

Zachary Upstone

Bachelor of Science in Computer Science
The University of Bath
2023/2024

Human Involvement Can Improve Current Image Synthesis Methods within the Domain of Art

Submitted by: Zachary Upstone

Copyright

Attention is drawn to the fact that copyright of this dissertation rests with its author. The Intellectual Property Rights of the products produced as part of the project belong to the author unless otherwise specified below, in accordance with the University of Bath's policy on intellectual property (see `https://www.bath.ac.uk/publications/university-ordinances/attachments/Ordinances_1_October_2020.pdf`).
This copy of the dissertation has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the dissertation and no information derived from it may be published without the prior written consent of the author.

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of Bachelor of Science in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Abstract

Image Synthesis within the domain of art remains a difficult task due to its complex and human nature. Many current models either suffer from insufficient user control or inadequate output quality. This dissertation employs Reinforcement Learning Human Feedback (RLHF) as a solution to resolve both these problems simultaneously. A "feedback loop" model is presented that overlays an existing model. This allows for the refinement of output images by a user, thus improving control. This method also allows for the collection of losses from the user's choices to allow for further updates to the underlying network. This effectively enables the model to learn from a user. The results produced by this dissertation show a quantitative improvement in user control over baseline models. They also show the qualitative success of the RLHF implementation. However, further analysis is required to confirm whether this RLHF implementation improves output quality.

Contents

List of Figures

List of Tables

Acknowledgements

Thanks to Professor Peter Hall, for all his help and guidance during the supervision of this project.

This research made use of Hex, the GPU Cloud in the Department of Computer Science at the University of Bath.

Chapter 1

Introduction

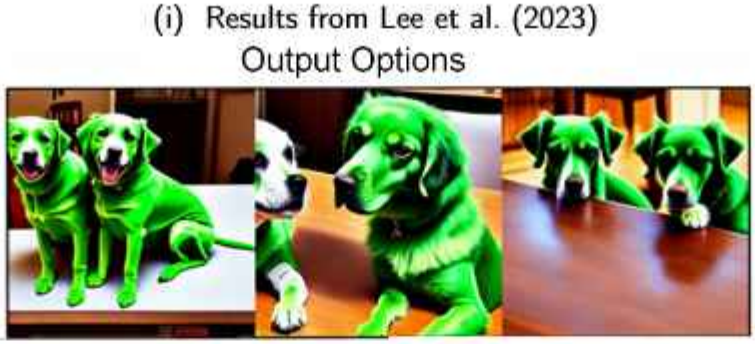
Image synthesis is the generation of images; in our case from inputs such as text prompts or sketches (Baraheem, Le and Nguyen, 2023). Various image synthesis models have experienced huge growth in recent years (Xu et al., 2023). However, the state-of-the-art remains subpar within the realm of art, both in terms of output quality and user control. This inadequacy is due to the complex and human nature of art (Leder et al., 2012). As many current image synthesis models look to move away from human involvement, this dissertation investigates the converse. Instead, it investigates whether reintroducing more human involvement in this process can resolve this inadequacy.

1.1 Dissertation map

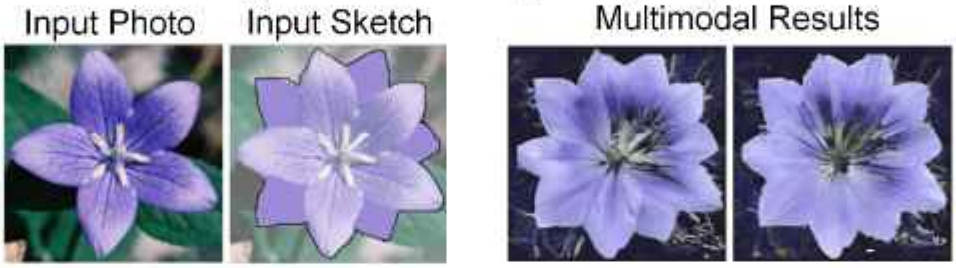
This section maps out the general structure of this dissertation.

- Investigation of the current literature surrounding image synthesis.
- Explanation of the gaps identified in the problem description.
- Discussion of research around this dissertation's suggested solution to the identified problems.
- In-depth analysis of this dissertation's design of a possible solution for these problems, incorporating the aforementioned research.
- More detailed discussion regarding the specifics of development and testing as well as an outline of a user study plan.
- Examination of testing results and user study results.
- Conclusion of the research and this investigation's discoveries including an evaluation of the solution's success.

(i) Results from Lee et al. (2023)



(ii) Results from Cheng et al. (2022)



(iii) Results from Liu et al. (2020b)

(iv) Results from this dissertation's solution

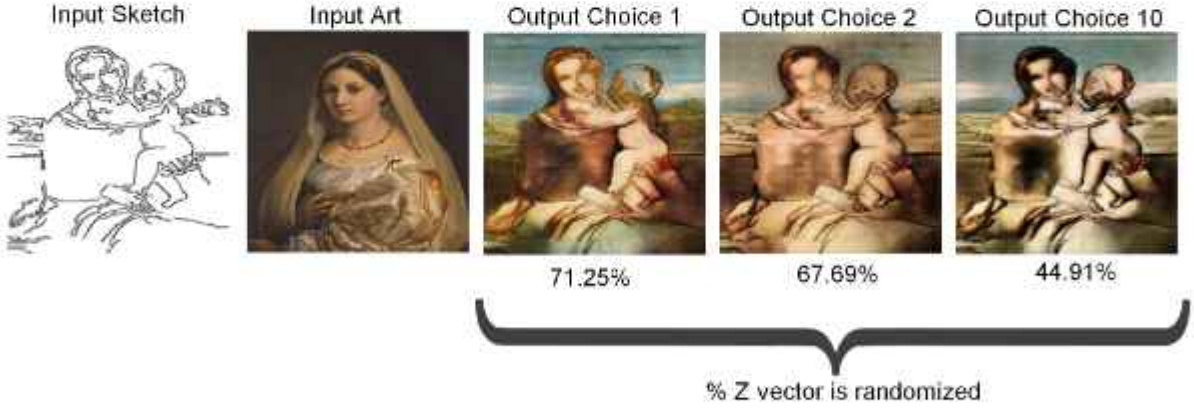


Table 1.1: Table showing a comparison of different state-of-the-art image synthesis models. This table can be used to compare different models' inputs and outputs for quality and control. The first row (i) shows the output of the work of Lee et al. (2023), the only input is a text prompt, however, it allows for multiple output options; in this case, it took "two green dogs on the table". The second model (ii) by Cheng et al. (2022) takes an input photo and an input sketch giving more control. The third model (iii) by Liu et al. (2020b) is specific to the domain of art and allows input sketch and input style. Finally at the bottom is this dissertation's model (iv), it takes a sketch and style input, allows output options at each stage, and allows the output to be progressed.

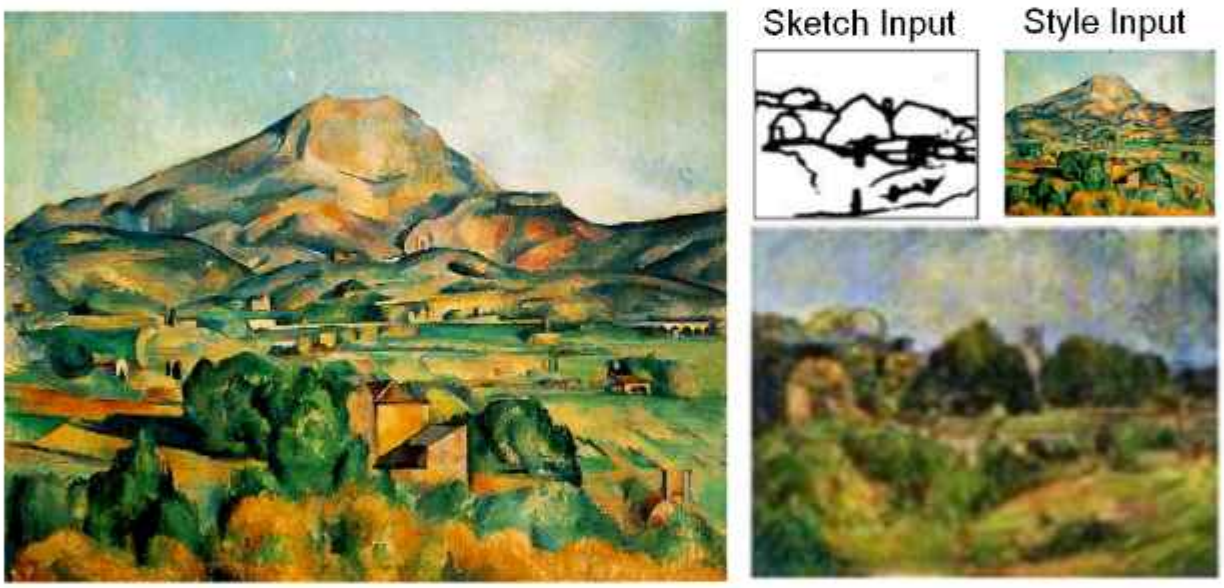
1.2 Problem description

The inadequacies with state-of-the-art image synthesis models are set out below as two distinct problems.

1.2.1 Problem A: Poor quality of images synthesised for art

An example of this is displayed in Figure 1.1 (below), on the left is a real Post-Impressionist painting by Cezanne (Cezanne, 1895), and on the right is the image produced by one of the state-of-the-art image synthesis tools (Liu et al., 2020b) in a Post-Impressionist style. The synthesized images lack the details and in particular to this style, the sharp edges and use of lines that are present in the original painting.

Furthermore, the generated image lacks the semantic detail of the genuine article. The manner in which the "paint" is applied is not reflective of the real painting in any way. For example, the Cezanne painting can be seen to be made of small planes, whereas, the generated image has no recognizable form.



(a) (b)

Figure 1.1: Figure for comparison of real art and image produced by state-of-the-art image synthesis methods. (a): Real post-impressionist painting by Cezanne (Cezanne, 1895), (b): Image in the style of "Paul Cezanne" produced by start-of-the-art (Liu et al., 2020b), including taken sketch and style inputs.

1.2.2 Problem B: Lack of control users have in synthesising images

Many state-of-the-art image synthesis models have interactive tools (such as Playform (Liu et al., 2020a) and Nvidia-Canvas (Nvidia Canvas, n.d.)) to assist with user control over image output. Currently, the majority of the existing solutions focus on taking a text input (Xu et al., 2023). However, this leads to a lot of variance within the final image produced due to the randomness of diffusion; a popular image synthesis method (Hertz et al., 2022).

Furthermore, other approaches take a sketch as an additional input. However, these continue to be inadequate, for example in Figure 1.2, where the text input was "volcano in the style of Turner" the output image does not align with the intended style of Turner as well as containing artefacts. Thus demonstrating that the State-of-the-Art still faces the same problem; users lack control over outputs.

Figure 1.2: Figure showing input and output of a Playform (Liu et al., 2020a) usage. Left: Sketch inputted into Playform with prompt "Volcano erupting", Middle: Resulting generated image, Right: Real painting "The Slave Ship" by Turner (1840)

Throughout the artistic process, the trajectory of a piece often evolves as it is created (Mace and Ward, 2010), resulting in a dynamic "feedback loop".

The majority of existing solutions do not allow users to modify the generated image during or after its synthesis, thus limiting control. During this investigation, this dissertation found only three papers (Zhu et al., 2018; Ghosh et al., 2019; Kazemi, Taherkhani and Nasrabadi, 2020) that emulate this "feedback loop", only one of which (Kazemi, Taherkhani and Nasrabadi, 2020) incorporated Reinforcement Learning Human Feedback (RLHF). Despite this limited research scope, the emergence of other papers exploring this area suggests established precedent and literature support for its potential success.

1.3 Ideal solution

A perfect system would be able to project an imagined image from a user's mind and produce it as a reality for others, filling in any gaps in their imagination. This would evolve as the user sees the model's output and changes their artistic direction. It is this adaptation that this dissertation attempts to include in the image synthesis process to improve current models.

1.4 Project contributions

This project has approached the aforementioned issues by increasing user control and allowing the network to improve based on user feedback. Contributions include:

- Implementation of a "feedback loop" of images for users to choose from to fine-tune results and specify individual fluctuations, to improve user control.
- Implementation of Reinforcement Learning Human Feedback (RLHF) on a multiple choice system, to improve the underlying network and therefore potentially improve output quality.

- Evaluation of the success of the proposed solution.

Before moving on to related work, it is important to explain the two key areas this dissertation brings together briefly, these are explained in Section 1.5 and Section 1.6 below.

1.5 What is Sketch to Image?

Sketch to Image is a form of generative image rendering where the input is a sketch and the output is a fully rendered image (Ramy and hosny Barakat, 2022). This form of image synthesis utilises lines from the sketch input to designate the semantic information of the output; this is often achieved through the use of a neural network. This model allows far more control than text-to-image models such as Xu et al. (2023). It also has variants where the style can be an input image (such as Liu et al. (2020b); Zhu et al. (2017b)) which some papers conject allows more control (Richardson et al., 2020).

1.6 What is Reinforcement Learning Human Feedback (RLHF) and why should it be included?

Reinforcement Learning Human Feedback is a powerful technique for training models (specifically neural networks) for hard-to-quantify problems (Daniels-Koch and Freedman, 2022). It introduces the human into the loop, giving feedback on how an agent is performing and thus allowing for purpose-built agents. This method is suitable for this dissertation's aims as humans remain superior to current state-of-the-art models in this domain as shown by Figure 1.1. Evidently, this is due to art being opinionated and therefore difficult to quantify numerically for normal learning techniques.

Chapter 2

Literature and Technology Survey

2.1 Introduction

Recently, AI-generated art has gained significant attention and has become a popular point of interest and concern in the art world. This is mainly due to the progress at which improvements are being made (Liu et al., 2020b) and its increased accessibility (available in tools like Playfrom (Liu et al., 2020a) and Nvidia-Canvas (Nvidia Canvas, n.d.)). Most research has been focused on going from prompts to images, however, "A picture is worth a 1000 words" (Brisbane, 1913), and a sketch can be used as an alternate input.

Firstly, Section 2.2 highlights several gaps within current models. Leading on from this, Section 2.3 details current models for synthesising images and evaluates their shortcomings. Section 2.4 discusses this dissertation's suggested solution to these identified shortcomings. Finally, Sections 2.5.1 and 2.5.2 discuss current models that already employ similar solutions for the image synthesis of art.

2.2 Gaps in the literature

As mentioned in the problem description 1.2, this dissertation has identified several gaps in the relevant literature. To summarize these include:

- Style being poorly captured from inputs (demonstrated by Liu et al. (2020b)).
- Output image quality being lacklustre.
- Lack of user control over generated images (a few examples allow control over the seed e.g. Playform (Liu et al., 2020a) but this means little to users).
- Lack of a good dataset containing sketches and corresponding art pieces.

Many others exist but fall out of the scope of this project and are not addressed within. However, this dissertation takes a key step in giving users more control over the art they generate, while also enabling the underlying network to undergo further learning.

2.3 Current methods of Image Synthesis for the domain of art

Currently, there are many solutions for going from sketch and prompt to a completed artwork including Playform (Liu et al., 2020a), Nvidia-Canvas (Nvidia Canvas, n.d.), and Fotor (Fotor AI, n.d.). However, based on research and this dissertation's aims, BiCycleGan (Zhu et al., 2017b) was chosen to be the base model for this dissertation's solution. This was decided as it can produce results using a sketch and style input. Once implemented this dissertation adds its modifications to address the problems discussed in Section 1.2. The background for choosing BiCycleGAN (Zhu et al., 2017b) as well as other related work is discussed in this section.

2.3.1 Diffusion models

Diffusion Models have become more popular recently as, unlike Generational Adversarial Networks (GANs, another popular image synthesis technique) which are discussed in Section 2.3.4, they are more stable and scalable (Cheng et al., 2022). Diffusion models are based on real-life diffusion (Cheng et al., 2022). To train, they take a complete image with structure and areas of focus which diffuse out until it is random noise. Once trained, the agent can reverse this and produce images from just noise.

Cheng et al. (2022) have produced a diffusion model that also takes sketches as shown in Figure 2.1. This allows a lot more control over what is output by the diffusion and in a sense takes a second style input (in the form of the image you wish to correct). However, their model also only focuses on producing realistic images. This limits the artist to producing photo-like images instead of art. A lot of current solutions behave in a similar manner opting to focus on producing only realistic outputs. This is probably because the domain of art holds more nuance than the domain of realism (Goldman, 2020).

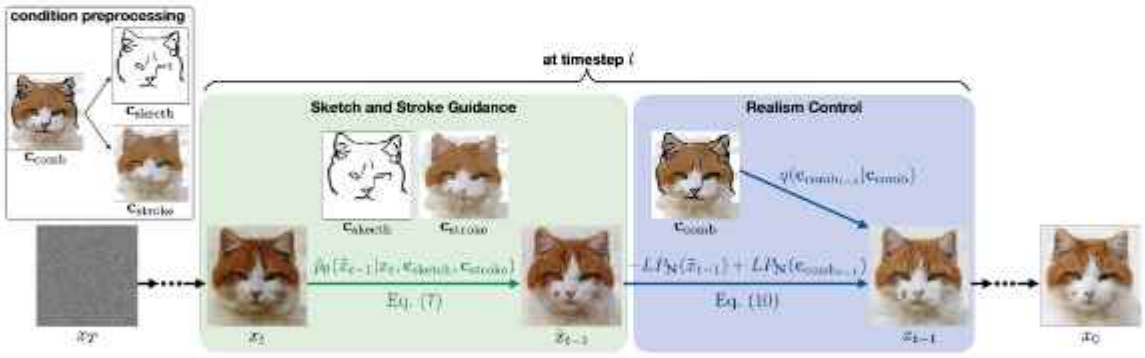


Figure 2.1: Representation of how a diffusion models work in Cheng et al. (2022) work. Inputs are a sketch and a photo, output is the photo adjusted to the sketch.

Wang et al. (2023) use diffusion models to clean up sketches, instead opting to use a non-Markovian diffusion process. The idea of correcting sketches could be used within an assistive art tool. However, Wang et al. (2023) state that the level of abstraction is not controlled within their implementation which an artist would care about deeply. This could instead be partially conveyed by a style image. However, as will be shown later, this dissertation has opted down another path to investigate improving image synthesis output quality.

2.3.2 Multi-Density Translation Networks (MDTN)

Another approach seen in the literature is that of Huang et al. (2020). Their work employs a Multi-Density Sketch Generator (MDSG) that takes parameters to the continuous sketch representation space. They then utilise a Multi-Density Translation Network that generates images conditioned on different levels of sketches outputted by the MDSG (Huang et al., 2020). This shows good results, especially for colouring images, and can achieve different styles with anime and realistic face generation. Huang et al. (2020) work also shows the ability to return different images from different levels of sketches as shown in Figure 2.2. They achieve this by training on different "densities" of sketches. This is something this dissertation could implement in the future, as some style is conveyed in the user's sketch input. Consequently, a better understanding of input sketches could lead to better results. However, this falls out of the direct scope of this project, so was deemed additional for this investigation.

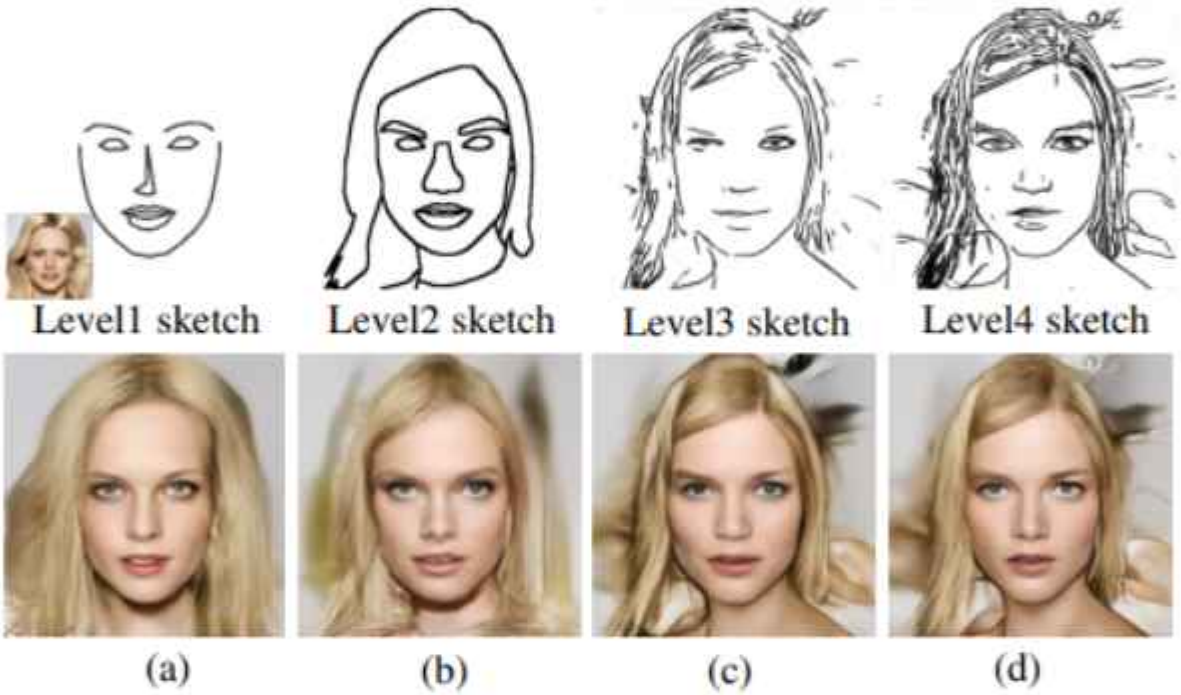


Figure 2.2: Results from Huang et al. (2020). In their model, they can produce final images from different levels of detail sketches due to their training strategy.

Huang et al. (2020) also use a unique dataset collection method proposed by Kang, Lee and Chui (2007). They use Coherent Line Drawing (CLD), as they claim other methods produce "short isolated edge fragments" (Huang et al., 2020). Once again this work has yet to be applied to art, whereas other works have and the output quality is low. In Section 2.3.4 and Section 3.2 this dissertation discusses its dataset preparation methods in more depth.

2.3.3 Hierarchical Network Architecture

Collomosse et al. (2017) implemented a hierarchical triplet network to learn how to produce images utilising both style and structure constraints. Firstly, Collomosse et al. (2017) trained a style network and then completely separately a structured network. These are "normalized and concatenated" (Collomosse et al., 2017) into input vectors for the main network. In

their model, the style image is an input painting and the structure input is an input sketch. This produces some impressive results as it attempts to disentangle content and aesthetics (Collomosse et al., 2017). Moreover, they allow additional control by changing the weights of categories on images. This work is a key example of keeping the human involved in the process. This dissertation takes user involvement further as to improve underlying networks based on human interactions.

Collomosse et al. (2017) work uses the Behance Artistic Media! (BAM!) dataset (Wilber et al., 2017). This dataset contains "65 million contemporary artworks" which allows Collomosse et al. (2017) the success they see. They also used Mechanical Turk (MTurk) (Mechanical Turk, n.d.) to evaluate retrieval accuracy (Collomosse et al., 2017). However, their work is based on image retrieval not generation. Despite this, it shows an interesting use of other tools that are relevant to this dissertation's contributions and allows users more control over structure and style through the use of multiple input images.

Figure 2.3: Figure showing work by Collomosse et al. (2017). Their model can take two inputs: a sketch input and a style input, then its output takes into account this style input. This Figure shows it taking a sketch of a bird plus a watercolour-style input and outputs the images on the right.

2.3.4 Generational Adversarial Networks (GAN)

GANs take up a large space of the current solutions for assistive AI art tools, so this dissertation has investigated them extensively. GANs use a discriminator instead of optimizing per pixel reconstruction error (Chen and Hays, 2018). This discriminator critiques the image and attempts to distinguish the real and fake, thus leading the generator to produce sharper images (Chen and Hays, 2018). This is the basic concept of an adversarial network. Pix2Pix (Isola et al., 2016) is a key example of a GAN and many solutions are built on top of it including BiCycleGAN (Zhu et al., 2017b) which this dissertations solution utilises.

Advances in quality of synthesised images

Many recent works have attempted to improve output image quality by modifying this basic GAN model. CoGS (Ham et al., 2022) follows a similar framework to this dissertation's solution's final design. They produce images from sketches and style images similar to the work of Collomosse et al. (2017) to decouple control the user has over the output structure and its appearance (Ham et al., 2022).

Their work importantly uses VAE's (variational autoencoder) (Kingma and Welling, 2013) which optionally modify the loss at the end. This refinement process is missing in a lot of current solutions. It explores local latent space as visualized in Figure 2.4. As discussed in Section 3.3, this dissertation's ideas draw from this to produce multiple similar images by

exploring local latent space. This also links back to the fine-tuning nature that real artists follow as they adapt their work during its creation. However, Ham et al. (2022) do not give the user more control over the produced outputs. Many solutions solve either one of these two problems of control or quality. This dissertation addresses both in parallel.

Figure 2.4: Latent space representation of songbird images from Ham et al. (2022) work. Different images correspond to different style vector inputs.

Xia, Yang and Xue (2019) suggest Stroke Calibration Networks (SCN) and Image Synthesis Networks (ISN) as another approach to improving output quality by modifying the GAN structure. They extend GAN structure to generate images that are faithful to a constraining input sketch (Xia, Yang and Xue, 2019). The SCN converts the sketch to be more like an edge map so the ISN can then convert it to an image. This allows the ISN to be trained on edge maps showing one approach to dealing with the dataset problem (the dataset problem is discussed further in Section 2.3.4). However, this is fundamentally flawed, as shown by their results, by using the SCN, outputs lose all style. This works in Xia, Yang and Xue (2019) case as they are trying to produce photorealistic images with little guidance. However, it would not work for art, as if user inputs are abstracted in any way, they will lose some of the meaning they wish to convey.

This dissertation's work follows a very similar design to Liu et al. (2020b) work as they go from sketches and style images to outputs in the domain of art. However, their results are lacklustre and do not represent the style they claim to. Liu et al. (2020b) show significant qualitative and quantitative improvements by introducing three new features, Dual Mask Injection (DMI),

Feature Map Transfer (FMT), and Instance De-Normalization (IDM) of which this dissertation incorporates FMT. This is further discussed in Section 4.1.1.

Dataset Preparation

Many current GANs struggle due to a lack of a coherent dataset of real sketches linked with real art pieces. Such a dataset is hard to produce as sketches are art in themselves. This means that they fall on both sides of the set therefore making it hard to sketch an art piece without the sketch displaying a unique style itself. This dissertation refers to this as "the dataset problem". The following sections talk broadly about how existing literature deals with this dataset problem.

Before moving on, edgemaps must be discussed. Edge maps are a black-and-white output image of where edges lay in an input image. They are often deduced by an algorithm. Many solutions opt to use real paintings and edge maps found from such paintings as these are easier to procure than real sketches.

Scribbler (Sangkloy et al., 2016b) is a GAN that uses edge maps and small colour scribbles to produce complete images. This work is flawed due to the quality of the output images. However, they have initiated an interesting approach to dataset preparation. Instead of using typical edge maps, they produce different sketch styles using xDog (Winnemoller, Kyprianidis and Olsen, 2012). Sangkloy et al. (2016b) then go on to show how their model works with different input sketch styles effectively. This is very important as it keeps the user in mind and could be applied to this dissertation's solution in the future. If applied it would allow the model to infer more style information from the input sketch and therefore produce a more representational output. Liu et al. (2017) work also use xDog (Winnemoller, Kyprianidis and Olsen, 2012) to obtain different levels of sketches.

Another approach that utilises dataset modifications for quality improvements is from Lu et al. (2017). They explore the opposite direction and use the sketch as a weak constraint so it is not a hard lock-on style. For this, they used the sketchy database (Sangkloy et al., 2016a), however, there is no equivalent for art pieces. In their research, they produce sketches using xDog (Winnemoller, Kyprianidis and Olsen, 2012) and they also fine-tune them using the PC (photocopy) (Photocopy Effect, 2016) effect from Photoshop and the FDoG filter (Kang, Lee and Chui, 2007). This dissertation does not recreate this level of dataset production, however, it could be a good future direction.

SketchyCOCO (Gao et al., 2020) shows better results through the use of freehand sketches as training data instead of edge maps, however, this method of collection is onerous and arduous. This is the ideal solution but was impractical for the scope of this project.

Finally, CycleGAN Zhu et al. (2017a) use unpaired training data in the absence of paired data and shows good results from it. Their work goes both from photos to paintings, but, also works in reverse to further improve itself by utilising cycle consistency. Unpaired training data could be a useful solution to the edge map problem, the BAM! (Wilber et al., 2017) dataset has both sketches and paintings within it. Originally we planned to utilise this, however, the output quality was poor. Instead, this dissertation's solution is built using BiCycleGan (Zhu et al., 2017b). Unlike CycleGAN Zhu et al. (2017a), this allows for the additional input of a style image which as shown by the work of Collomosse et al. (2017) gives the user more control. However, BiCycleGAN is unable to process unaligned data, this dissertation's solution

initially intended to re-implement this feature of CycleGAN within BiCycleGAN but with little success. Therefore, an aligned dataset was used instead.

Work that gives Humans more Control

Finally, this investigation into GANs will move on to work that brings the Human-in-the-loop (HITL), particularly for agent refinement. This also helps with the problem of users lacking control.

Elgammal et al. (2017) work focuses on "arousal" (Elgammal et al., 2017). They modify the GAN structure by making the critic element also check that art is not "too novel" and marking images produced by this as "not art". They achieved this by adding a style classification and style ambiguity losses. This identifies a major issue with existing solutions, defining the creativity of machine-synthesized images is an open and hard question (Elgammal et al., 2017). Colton (2008) came up with three criteria that a creative solution should cover:

- the ability to produce novel artefacts.
- the ability to generate quality artefacts.
- the ability to assess its creation.

Elgammal et al. (2017) work uses novel loss functions to try and build closer to human-like performance. However, they still opt to keep the human out of the creation process. It could be argued, that this is counter-productive as humans will verify the quality of the final product during testing.

Ghosh et al. (2019) keep the Human-in-the-loop (HITL) by use of an agent that suggests possible completions for a drawing. To achieve this they use an alternate approach to class conditioning (Ghosh et al., 2019) similar to that used by Lee, Zitnick and Cohen (2011). In this work, they attempt to allow for multiple solutions from a single sketch input which would greatly improve control. This work looks promising, however, they achieve these different outputs by having multiple models for each category of output. This dissertation opts to allow for additional style image inputs for more control. Ghosh et al. (2019) have yet to extend their work to art and allow the user to physically choose from the recommendations. This is limiting as it still relies on the users' skills. Overall this work fails to address the control issue.

Zhu et al. (2018) also keeps the Human-in-the-loop (HITL) by allowing users to make edits in real-time to generated images (Zhu et al., 2018). This creates an evolving image like Ghosh et al. (2019). However, their approach has a narrow field of view allowing only one alternate option and the quality is quite low. Although this editing system keeps the human involved, it is still very dependent on an individual's skill in the case of art. This project's solution is similar but suggests a novel method for such updates.

Work throughout this section suggests improvements can be made by utilising human feedback.

2.3.5 Other techniques

Bui et al. (2018) have implemented a Convolutional Neural Network (CNN) with multi-stage regression to measure similarities between sketches and images. This could be used to go from sketches to images in a manner similar to that of Internet Image Montage (Chen et al., 2009). Bui et al. (2018) used web search to utilise a large bank of data as well as a triplet

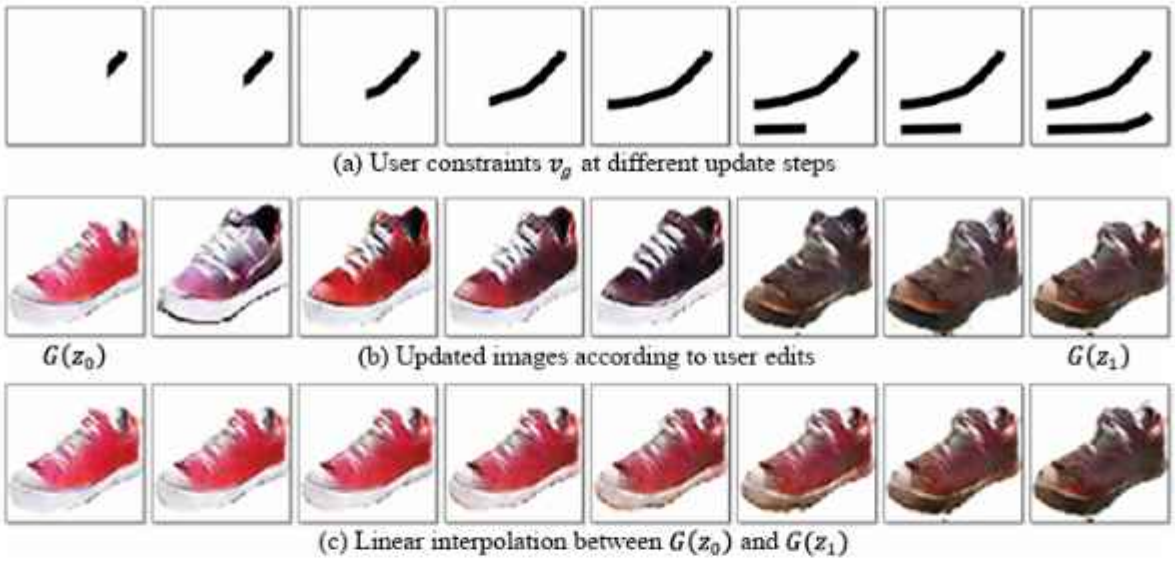


Figure 2.5: This Figure shows work by Zhu et al. (2018). As a user draws a sketch input the output image changes. It demonstrates live updates to outputs as the user input changes, showing a "feedback loop".

architecture which they found to have "superior performance" (Bui et al., 2018). However, Internet Image Montage (Chen et al., 2009) shows how this translation is hard. In Chen et al. (2009) work they found image backgrounds caused issues when trying to display indoor scenes and that having multiple tagged objects caused artefacts to occur (Chen et al., 2009). By shifting the focus away from this labelling, and focusing instead on a continuous scene with one sketch as a structural focus, no intermediate data is lost. This is important for art as it often can not be fragmented into discrete pieces.

2.4 Reinforcement Learning Human Feedback (RLHF)

2.4.1 The latent variables problem

The problem that all of the current solutions try to answer is producing a model that can produce high-quality and accurate (to a user's desire) outputs based on inputs. GANs use the discriminator and generator model to refine several networks to produce their results. However, none have come close to producing good-looking art. These networks have latent variables that training tries to uncover. Art has many profound latent variables that can not just be determined by pixel value, as shown by Mace and Ward (2010) analysis of the artistic process. As agents are improved and produce better art, it must be remembered that humans fundamentally decide what is and what is not art. As touched on by Bendel (2023) if an image synthesis model managed to produce work that was considered novel, it may just reflect the designer of the model's creativity.

2.4.2 Why RLHF resolves this

RLHF has been defined in Section 1.6. As seen, many current solutions only include the human at the beginning (providing inputs), and at the end (in testing and evaluating images)

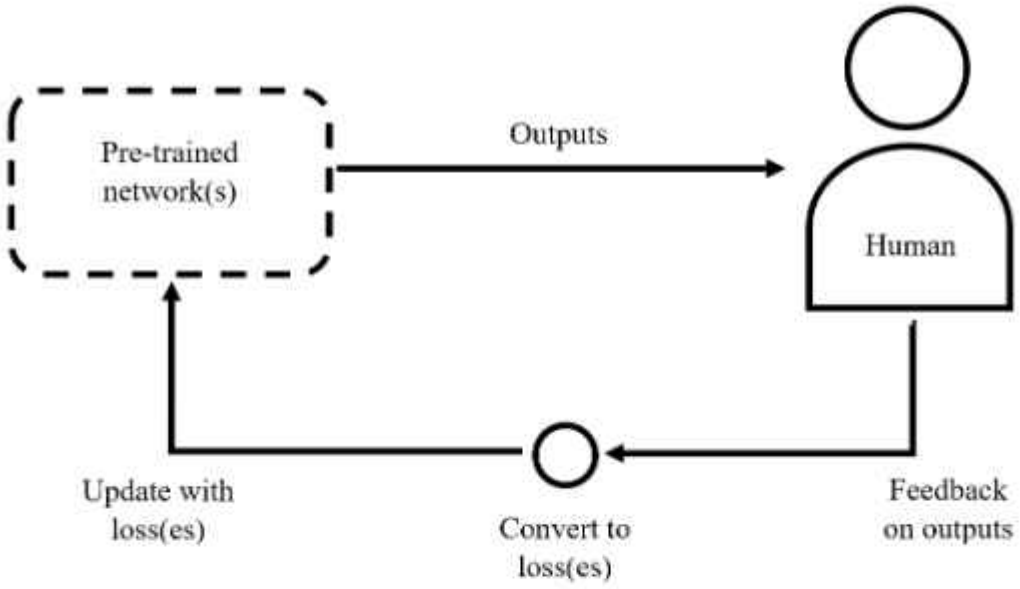


Figure 2.6: Figure showing how RLHF works, outputs by a network are somehow rated by Humans, this rating is in turn somehow converted to a loss for the network.

of the image synthesis process. RLHF is the converse of this. RLHF has gained a lot of interest recently due to its success with ChatGPT (Ramponi, 2023) and other works (Li, Yang and Wang, 2023). By having feedback on the quality of an image one can readjust the aforementioned latent variables to achieve a more optimal policy. RLHF is relevant for the image synthesis of art, as humans are still superior to current solutions. This is shown by Figure 1.1. However, ranking art is a subjective matter so to be able to achieve a feedback system, there needs to be a way for a human to objectively rank the quality of outputs. By casting the problem to how accurately a model produces a piece of a certain style, this becomes less subjective. Furthermore, by giving multiple options and allowing the user to progress from a chosen option, we allow for the evolution of the image and the underlying model. This returns art to the form it is most easily judged, visually. Like The Blind Watchmaker (Dawkins, 1986), suggested by Richard Dawkins, artists will move in small steps to build up something very complicated.

Since this is a human problem of identification, this dissertation employs RLHF and aims "to learn the human's underlying reward" (Li, Yang and Wang, 2023). The RLHF phase of the project is defined and discussed in Section 3.3.3.

2.4.3 Making the solution human

There is little research on producing multiple solutions from a single input as mentioned earlier when talking about Ghosh et al. (2019). By including random noise in the style vector produced in BiCycleGAN (Zhu et al., 2017b) this leads to multiple similar outputs, allowing the user different choices. However, this does not allow users much more control as they are still unable to specify sufficient detail. To answer this, in this dissertation, a method similar to VAE's (Kingma and Welling, 2013) like CoGs (Ham et al., 2022) is used. By adding random noise to each of the choices, local latent space is explored. Then by creating a "feedback loop", where the chosen style vector (with its random noise) is input again, this can "evolve" the image based on user choices. This can be visualised as taking little steps in the local latent space.

Overall, this solution is simpler than VAE's, however, their concept is emulated by creating a network that learns from this "feedback loop" idea.

2.4.4 Limitations of RLHF

Next, it is important to talk about some limitations of RLHF. Firstly, humans may be misaligned (Casper et al., 2023), i.e. they may think the wrong answer is correct. Although art is subjective, an expert's opinion is more respected. Furthermore, not every individual represents the overall consensus. This is why this dissertation focuses on a network being trained on how close it is to a style.

Another problem is that good oversight is difficult (Casper et al., 2023). Bowman et al. (2022) perform an in-depth analysis of this problem. They summarise that within RLHF systems it is difficult to identify when machines are outperforming humans. In this investigation, to counteract this the size of the user study is deliberately kept small. As this dissertation has a small user study, testing the effectiveness of the proposed RLHF employs a unique method, discussed in Section 4.3.1.

Finally, RLHF often suffers from "reward hacking" (Casper et al., 2023), there are many ways to fit the human feedback dataset. Sometimes this will be done incorrectly leading "to causal confusion and poor out-of-distribution generalization" (Casper et al., 2023; Tien et al., 2022). Overall this dissertation does not address many of these issues, as they are outside the scope, but they are important to note.

2.5 Work that incorporates RLHF within Image Synthesis

2.5.1 Reinforcement Learning (RL) and Art Synthesis

Many existing works already incorporate RL into image synthesis. Krishna et al. (2021) uses RL in tandem with image synthesis, they use image synthesis to accommodate for the lack of a data set for their problem. They train an RL agent to reconstruct CT scan images using synthesised data for training. This is the converse of what this dissertation implements, but, shows the potential for success of this pairing of techniques. Khurana et al. (2023) presents another combination of RL techniques with image synthesis. They present a diffusion model where images are ranked at the end. This dissertation opts to not use a ranking system to implement RLHF, however, it is suggested as potential future work in Section 6.3. Interestingly, in their work, they rank images by their usefulness for a task. This is not suitable for the domain of art but it shows one approach to aligning users' subjective opinions for an agent's learning.

Finally, Kirstain et al. (2023) presents PickScores a large dataset of image prompts and human preferences, an example of a preference method. This dissertation opts to implement this technique. Importantly, they talk about how such data could be used for RLHF, showing the potential for this project in the literature and bringing this discussion to the next section.

Figure 2.7: Figure showing work by Kirstain et al. (2023). Their work consists of a large dataset of synthesised images and user preferences. Darkened images are the ones not chosen by real human feedback, this could be incorporated into a diffusion model.

2.5.2 RLHF and Art Synthesis

RLHF has been incorporated into image synthesis already. Firstly, Liang et al. (2023) trains a model on 18k pieces of human feedback on artefacts and misrepresentation of text prompts. Unfortunately, their solution still gives the user insufficient control. Importantly, this focuses more on semantic information, (i.e. what is present in the image). This highlights a flaw of this dissertation's approach, there being no way to definitely "know" what the sketch input is. However, as stated previously, sketch inputs still allow for a lot more user control than text inputs. Future work here to resolve this could be to include not only a sketch and style input but also a text prompt input. Importantly, they identify aesthetic quality as one of the human feedback rankings.

Figure 2.8: Figure showing synthesized image of pandas riding a motorcycle by Liang et al. (2023). This work incorporated RLHF via humans ranking images on different scoring metrics. These are shown on the right-hand side. Humans also gave feedback by highlighting any words from the prompt that were misaligned in the image. Liang et al. (2023) came up with this unique scoring system allowing users to rate an image on different factors including plausibility (how likely an image is), alignment (if there are any out-of-place artefacts in an image), Aesthetics (how clean an image looks), and an overall score.

Lee et al. (2023) shows good results through the use of RLHF. By using a large dataset of human feedback to train a scorer for generated images they can then improve the underlying network further. Unfortunately, their results are only suitable for improving semantic representation which as mentioned (in Section 2.3.5) is not as important in art.

Xu et al. (2023) has a similar approach, they use a large dataset of 137k human responses to train a scorer for a diffusion-based model. They then use ReFL (Xu et al., 2023) to use said scorer to improve future synthesis. Importantly they use a ranking system instead of each image being given a score. This is to normalise data between individuals, which as seen in Section 3.3 is the path this dissertation's solution follows. However, they base this on user preference which this dissertation avoids, as it makes the issue more subjective. As seen in Section 4.3.2, this dissertation gets users to choose from output options based on how well a chosen style is represented. This is done to be more consistent and improve the agent more objectively.

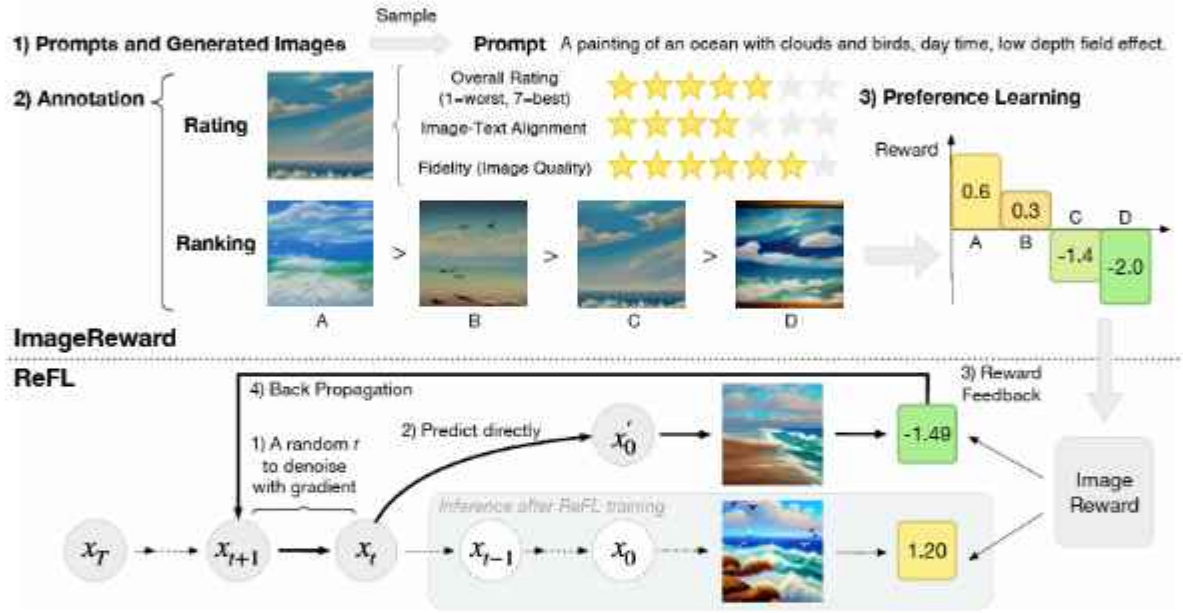


Figure 2.9: Diagram showing how Xu et al. (2023) work uses ranking of images to incorporate RLHF. Images ranked lower are given negative rewards and those ranked higher positive rewards. Images are also rated on alignment and fidelity.

Finally, Kazemi, Taherkhani and Nasrabadi (2020) work shows a great case of putting the human back in control and focusing on realising the user's mental image. Their work produces images quickly, which is realistic for human use and does not need a large dataset of human feedback. This dissertation's work is heavily influenced by their work. They also discuss their work being adaptable to GANs showing further evidence in the literature that this approach is a suitable solution.

Overall, RLHF is a powerful tool for improving model quality. Many current solutions fail to realise that RLHF can also be used alongside allowing users more control. By incorporating both users can more accurately go from a mental image to a produced image. This dissertation takes parts from many of the models discussed here to try and resolve the issues of user control and output quality highlighted in Section 1.2.

2.6 Ethical consideration for Image Synthesis

It is important to talk about the ethical considerations of art synthesis. Art synthesis raises many questions as identified by Bendel (2023), these include but are not limited to:

- Do people compare themselves to digital art?
- Will digital art change the standard of beauty?
- Is digital art deceptive and tricking people into what is real art?
- Are digital art models discriminatory?
- How do digital art models handle censorship and bias?

It is important to note this dissertation builds this tool and investigates this problem area with the express desire to aid artists with producing what they intend to. Many of the normal issues faced by art synthesis are addressed by the belief that art is a human endeavour and thus any art produced by such a tool is only as innovative as its creator. By focusing on giving users more control, this dissertation does not raise many of the questions normal art synthesis methods face. Bendel (2023) work discusses this topic in more detail.

2.7 Summary of research

To summarise, this dissertation has investigated many of the existing solutions for the image synthesis of art. It has identified gaps in the form of lacklustre quality and control in the domain of image synthesis for art. It then went on to discuss RLHF as a powerful tool that could resolve this.

Chapter 3

Method

This chapter describes the method employed to design a system that can be used to discover whether human feedback improves current art synthesis methods. The proposed solution involves adding a new user control method and RLHF on top of an existing model. Both of these additions were made to address the gaps in the literature identified in Section 2.2.

This dissertation's model involves giving the user multiple output options. Users choose from between these outputs, "evolving" the image similar to the work of Zhu et al. (2018); Chen and Hays (2018). RLHF then improves the underlying network based on these choices.

This work is built upon the work in Zhu et al. (2017b), an implementation of which can be found in the repository here BicycleGan.

This chapter explains how this dissertation's solution was designed including:

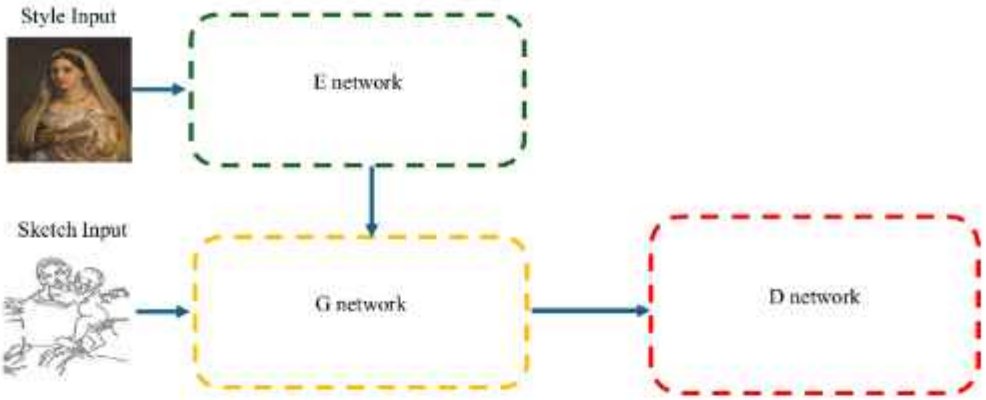
- The underlying network, shown by Figure 3.1b
- The dataset, discussed in Section 3.2
- The user flow, which is shown by Figure 3.2
- The UI, shown by Figure 3.4
- The RLHF add-on, shown by Figure 3.5

3.1 Initial training of the underlying model

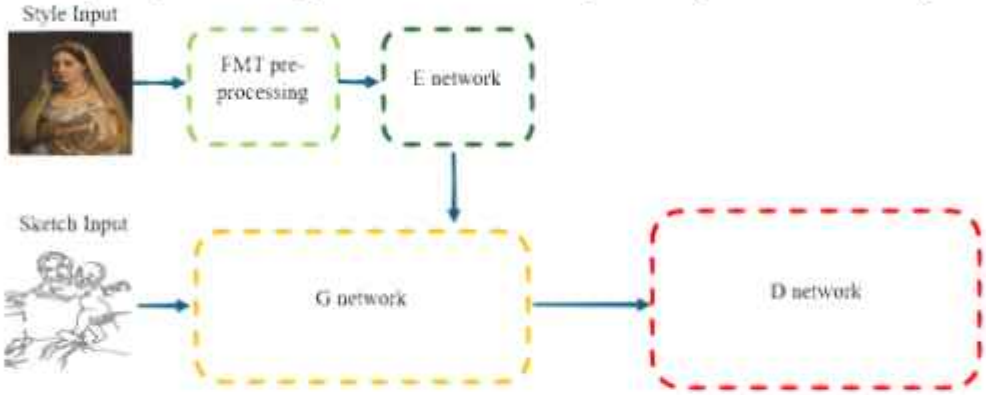
During development, several underlying models to work with RLHF were considered. BiCycleGAN (Zhu et al., 2017b) was chosen, due to its ability to take a style input which allows more control. This style input is utilised by this dissertation's solution to generate multiple images from one sketch.

BiCycleGAN consists of three networks:

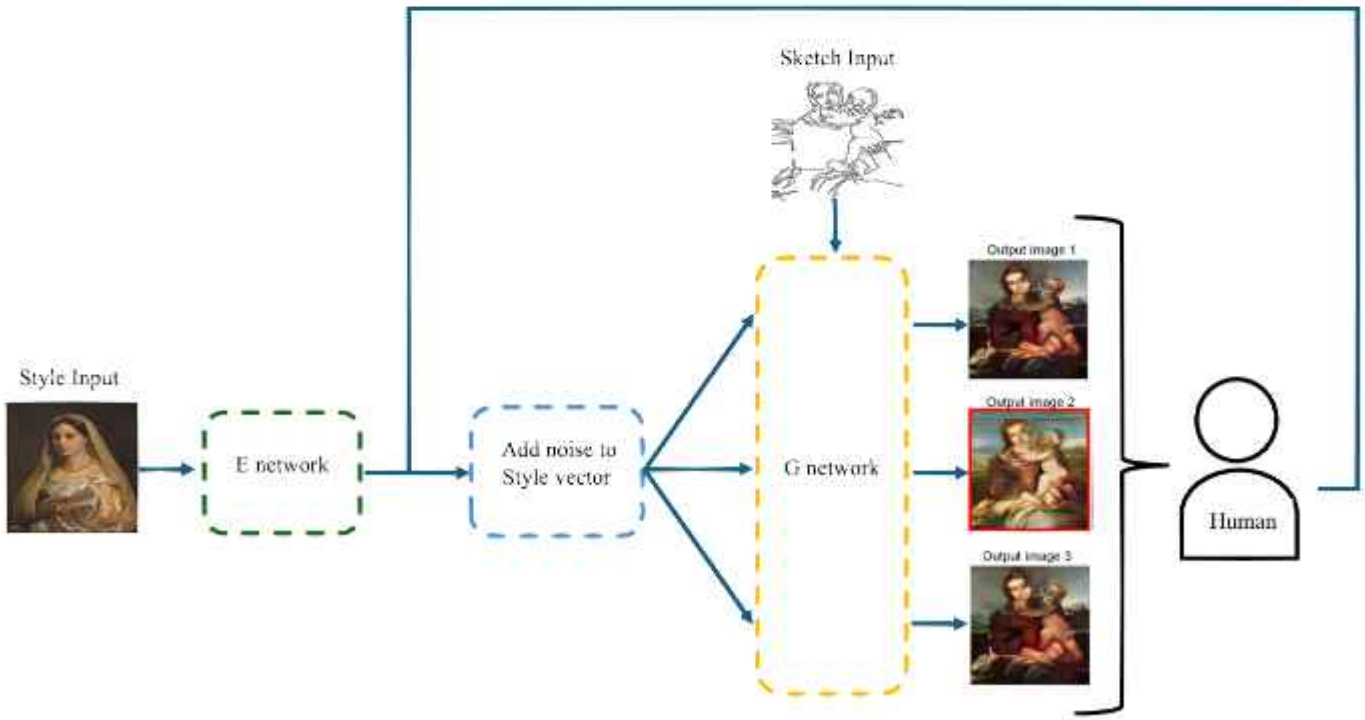
- Extractor (E) - represented by the green dotted border.
- Generator (G) - represented by the orange dotted border.
- Discriminator (D) - represented by the red dotted border.



(a) Diagram showing workflow of basic BiCycleGAN (Zhu et al., 2017b)



(b) Diagram showing workflow of modified BiCycleGAN. This dissertation utilised this training method to train the initial underlying networks. FMT pre-processing was added from the work of Liu et al. (2020b) as it improved results significantly.



(c) Diagram of the workflow during the RLHF stage. Users are given a choice of images in each iteration. Initially, a user inputs a sketch and style image, but, in all following iterations, their choice is used instead of the style input.

A diagram of their setup can be seen in Figure 3.1a. The E network takes a style image and tries to summarise this numerically by outputting a latent style vector. The G network takes a sketch image and this style vector as inputs and outputs a new image. The D network takes this generated image as its input and "reviews" it. Following this, it outputs whether the image is real or fake as well as other details. The D network's "review" allows for additional updates to be performed on the other networks.

This dissertation modified this framework by adding Feature Map Transfer (FMT) from Liu et al. (2020b) work as this dramatically improved the quality of the underlying trained network; this is discussed further in Section 4.1.1. A basic diagram of this dissertation's updated underlying model used to train the base networks can be seen in Figure 3.1b.

A facet of BiCycleGAN discovered during the initial training is that the Generator (G) network starts to overfit after training for too long. This overfitting is evident as the Generator(G) begins to disregard the input style vector and instead produces what it believes is a good representation of a general learned style. This disregarding could be due to the E, G and D networks not updating at the same rates, as their losses are calculated differently. For this reason, the number of epochs the networks were trained for was not fixed, instead, they were trained until they began to exhibit signs of this overfitting.

In theory, this overfitting issue could be resolved with the correct dataset, however, this is not the focus of the investigation and would be a time-consuming process. More exact details about training for the user study are discussed in Section 4.3.2.

3.2 Dataset choice and preparation

Next, a dataset to work with had to be chosen. The dataset this dissertation used was decided on after researching and testing what would work optimally with the underlying networks and RLHF. Networks were trained on datasets of fixed styles with roughly 1000 paired images. Art pieces were collected from WikiArt (WikiArt dataset, n.d.), and then edge maps were created of each art piece and paired up to create a paired dataset. For testing, it was decided that High Renaissance paintings would be used as they included many portraits. This in turn allowed more variables to be controlled during the user study by getting participants to draw people instead of anything for the sketch input.

3.2.1 Edge Map creation

Ideally, real sketches would have been used as some style is often conveyed within the sketch input. This dissertation started by utilising them and experimented with them extensively. However, due to limited success with such datasets (shown by Figure A.2 in Appendix A) and time constraints, edge maps were used instead. This is further discussed in Section 4.1.1. Many different edge map creation methods were experimented with. Canny (Canny, 1986) was used in the end due to its simplicity as this area was not the investigation's focus.

When using Canny it was important to decide what minimum and maximum thresholds to use. After investigations, which can be seen in Appendix A in Figure's A.3a and A.3b, this dissertation's implementation settled on using a minimum threshold of 75 and a maximum of 200 for the collected High Renaissance art pieces.

3.3 This dissertation's model

This section discusses the design of this dissertation's additions of a new user control method and RLHF.

3.3.1 User flow

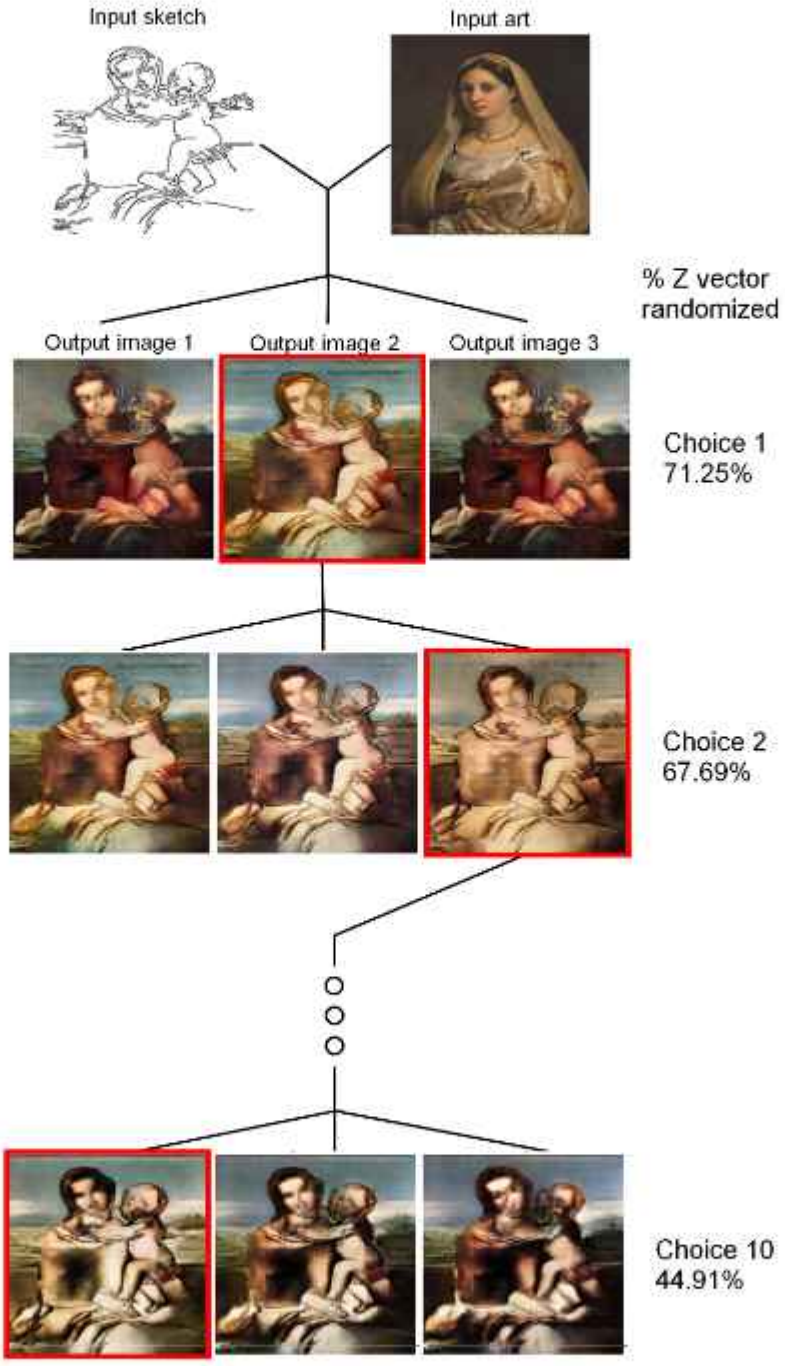


Figure 3.2: Diagram showing use case of this dissertation's tool. Sketch and style images are input three choices are output, and the chosen output is used with the sketch input again to produce three more outputs. A full use case of this can be found in Appendix A Figure A.1

Firstly, to create a new user control method a user flow (of how a user would interact with the model) was designed. Figure 3.2 presents how a user interacts with this new user flow. Users input a style and sketch image similar to BiCycleGAN (Zhu et al., 2017b), after this, they are presented with three options. Users then choose from these options and are presented with three more options based on their choice. This can continue indefinitely, however, the variance of the options is designed to decrease with each iteration. The decaying variance enables the user to hone in on a particular style.

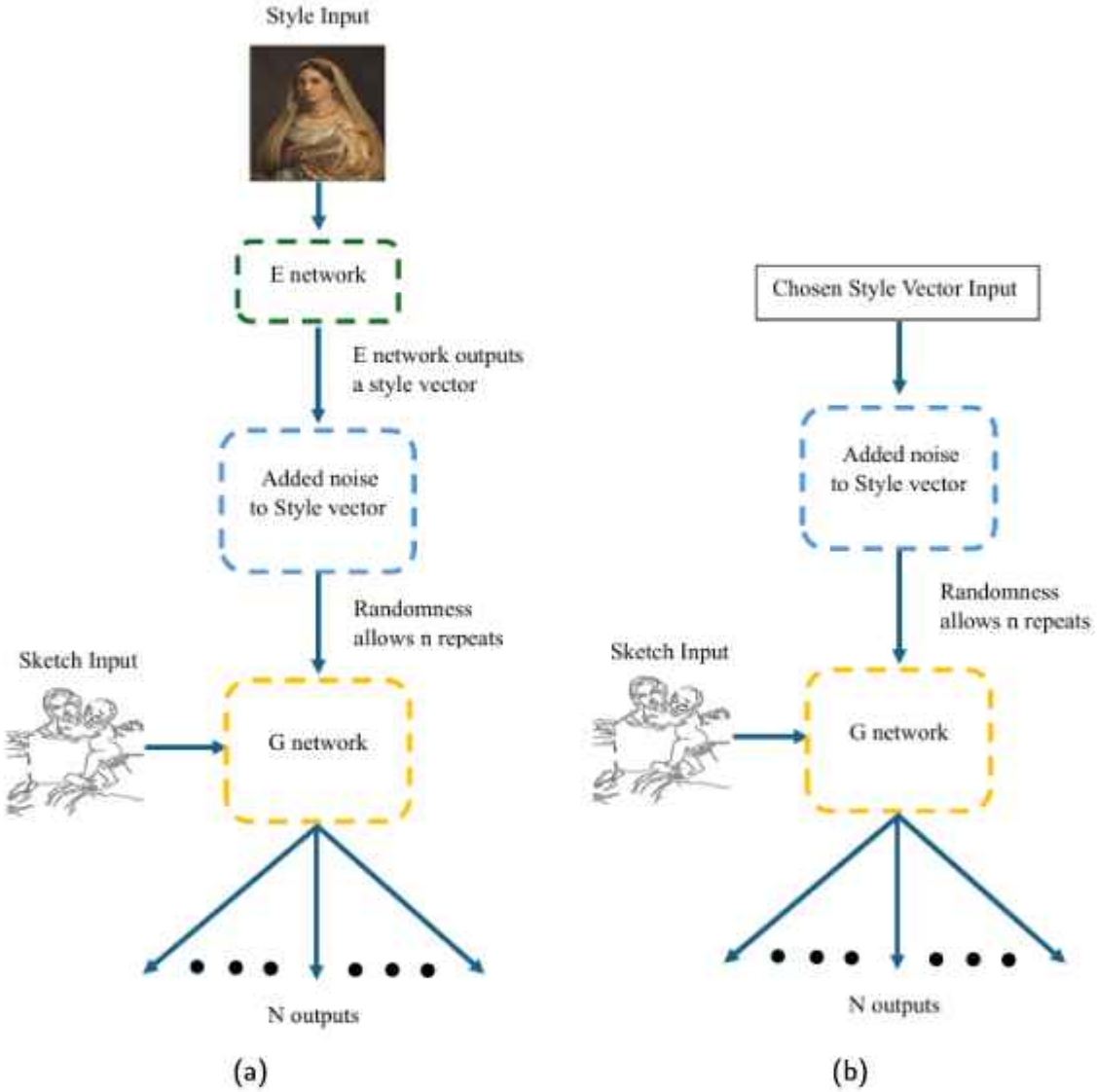


Figure 3.3: Diagram showing how this dissertation added randomness to produce multiple outputs. In both (a) and (b) it adds random noise to the latent style vector, this allows the production of multiple outputs from fixed inputs. Figure (a) is utilised during the first iteration of the tool, for all following iterations Figure (b) is followed. The style vector is kept after each choice to provide this new input. This is also shown in Figure 3.1c.

To achieve this flow, all images had to be given some variance when generated. This also incited learning for the RLHF and made the difference between the choices explicit for users. This randomness was added to the images after the Extractor (E) network stage. The repository by Zhu et al. (2017c) follows a similar approach, however, they only perform this during initial training. This dissertation's solution adds this randomness by following Figure 3.3a for the first

iteration where the user inputs the style and sketch image and Figure 3.3b for all following iterations.

Throughout this flow process, feedback is collected for the RLHF updates.

3.3.2 User Interface

A user interface was added as it was necessary to aid with usability for the user study. It was decided it would include input style, input sketch, last picked image, and future image choices. This gave users more control and understanding by organizing data that was relevant to them. The approach for the UI was decided on for simplicity, for example, a drawing app feature was not included as many existing apps serve this purpose (such as FireAlpaca (2023)). Also, by allowing for any input the user could produce the sketch by any means they desired. A figure of what this UI looks like while in use can be seen in Figure 3.4. It is simplistic and could be improved in the future. However, it should be noted after a quick explanation no participants struggled to use it.

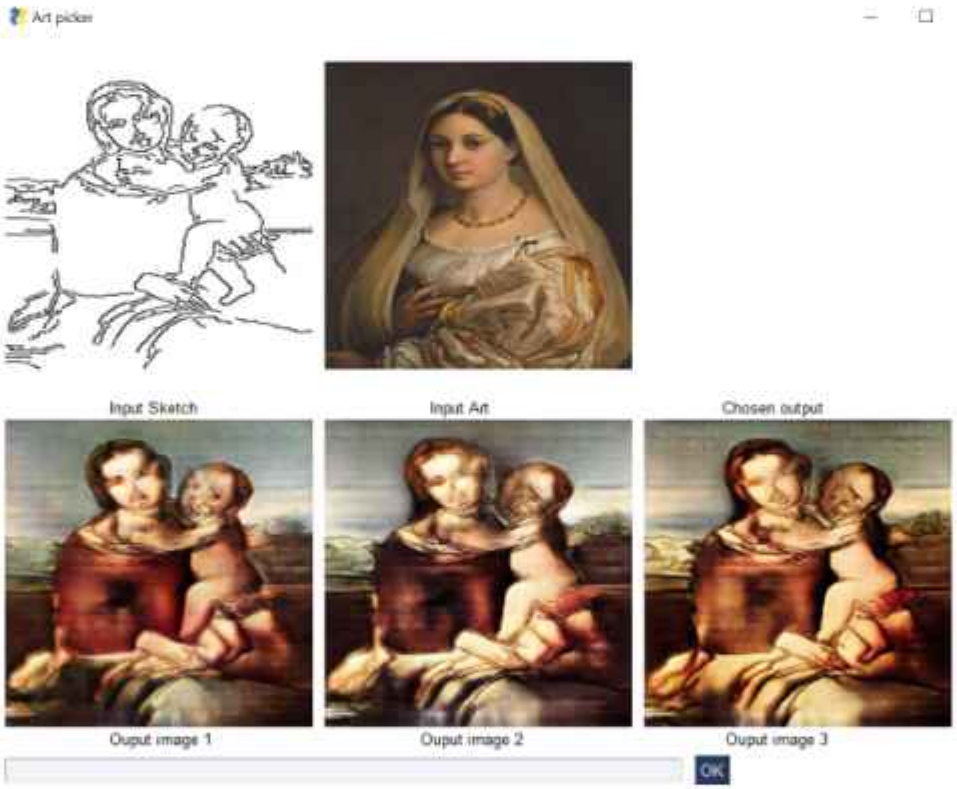


Figure 3.4: Screenshot of the User Interface of this dissertation's tool. This was used for the user study.

3.3.3 RLHF network updates

Finally, RLHF was added underneath the new user control flow. It updated the underlying model's networks based on the losses that **could** be gathered from the new user interactions. The learning already implemented by BiCycleGAN (Zhu et al., 2017b) was also replicated alongside these new losses. Since the losses calculated by the RLHF add-on are much larger, the model undergoes increased learning from less training.

This dissertation's design used three distinct losses. Figure 3.5 shows how these were calculated.

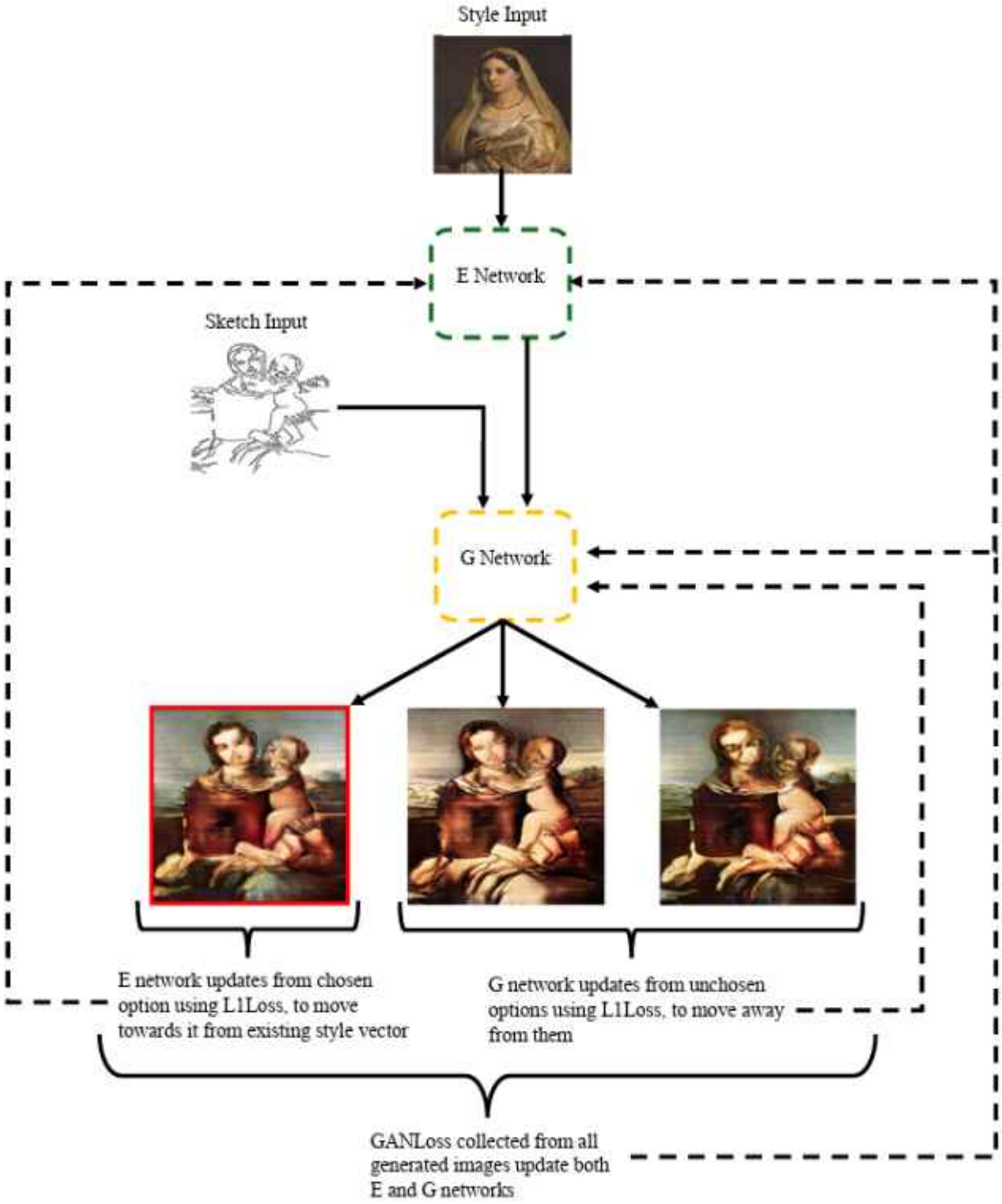


Figure 3.5: Diagram showing the 3 different losses collected from a single choice of the use case. At the top it shows the two inputs feeding into the networks, then the outputs produced. The chosen image contributes to a new loss for the Extractor (E) network and the unchosen images contribute to a new loss for the Generator (G) network. All images also contribute to an existing GANLoss for both networks.

Explicitly, these are:

- **E loss**: Difference between the chosen image's style vector and the style vector with no added noise.
- **G loss**: Pixel difference between the chosen image and the images not chosen by the user (as a sort of negative loss).
- **GE loss**: GANLoss from all images, this is a loss replicated from the model's normal training.

Importantly, as the human was a substitute for the Discriminator (D) network and it was not used beyond initial training this was not updated.

Generator (G) Network Updates

For updating the Generator (G) network two losses were calculated. Firstly, the difference between the chosen image and the unchosen images, which acted as a "negative loss" that would aim for higher loss values until a set cap was met (this was set to 20). Secondly, the Generator (G) network is also updated using the GE loss, this loss also updates the Extractor (E) network. It was calculated using BiCycleGan's (Zhu et al., 2017b) already existing GANLoss implementation.

Extractor (E) Network Updates

The other new loss was the difference between the unaltered style vector and the chosen image's style vector. This and the aforementioned GE loss updated the Extractor (E) network.

3.4 Other design possibilities

To reach this final design many alternate approaches were also considered:

3.4.1 Initial model training

Alternates to this dissertation's base model training strategy that were considered include:

- Training the model for an optimal length. With more time and computation a better underlying model could be produced. However, due to time constraints, and this not being this dissertation's key focus, models were trained until overfitting or ~2000 epochs.
- Training multiple underlying models and comparing results. This would allow for a further understanding of which base models benefit from RLHF the best, as the most optimal standalone model may not perform optimally when combined with RLHF. However, again this would be very time-consuming, so has been left to future work.
- Using an unaligned dataset. Originally this project started out using an unaligned dataset but achieved lacklustre results. With additional time this could be investigated more thoroughly, this is further discussed in Section 4.1.1.

3.4.2 User interaction system

Substitute design choices for the interactive system that were considered include:

- Adding an internal drawing system. This was not included as it would be limiting, such software already exists for this explicit purpose with a plethora of additional features such as FireAlpaca (2023). If this was added it would consume a lot of time to develop and end up limiting how users utilise the tool. By taking an existing image users can create sketch input images using any tool they favour.
- Adding the ability to specify how many options are available at each iteration in the tool. Initially, this was intended to be implemented. However, it felt unnecessary for the research's investigation, but, would be an interesting future addition as the system would be able to perform more learning after each iteration.
- Altering the choice system to follow a ranking image system. This would follow the work of Xu et al. (2023) who use this method to incorporate RLHF into image synthesis. This provides more data that can be used when calculating losses and therefore further learning from a single iteration. This could be incorporated within the system by taking the ranked one image as the input for the "feedback loop" and would be an interesting future addition.

3.4.3 RLHF network updates

As for alternate RLHF network updates, the possibilities are endless. The chosen method was decided based on what was possible and logical on top of the chosen framework (BiCycleGAN (Zhu et al., 2017b)). An interesting alternative would be that mentioned above with list ranking data. However, due to the difficulty with collecting data for RLHF, it is hard to gauge what is best without considerable time input into testing with humans. With the testing performed on the code in Section 4.3 this solution was found to be the best in terms of update speed and accuracy.

Chapter 4

Implementation and Testing

4.1 Implementation

This section will go into further detail about the implementation of this dissertation's solution, importantly, it discusses the specific challenges the proposed solution faced. Chiefly, for RLHF to be implemented a neural network had to be used, which influenced many other implementation choices.

4.1.1 Changes during development

This subsection discusses changes that were made to improve this dissertation's model during development. However, firstly, it is important to note, that during development, many other paper's claims of quality such as sketch2art (Liu et al., 2020b) and CycleGAN (Zhu et al., 2017a) could not be recreated.

Training - Adding Feature Map Transfer (FMT)

As discussed in Section 3.1, BiCycleGAN (Zhu et al., 2017b) was used for the underlying network as it incorporated an Extractor (E) network. This was necessary for the project's design as this is where the noise was planned to be added (into the latent vector produced by the Extractor (E) network).

During development, the Extractor(E) network was identified as a weak point in this architecture. For this reason, Feature Map Transfer (FMT) from Liu et al. (2020b) work was used as it dramatically improved results from the model. This is likely due to it helping remove semantic information from style images for training. Due to time constraints, other parts of Liu et al. (2020b) work were excluded, these could be added in a future extension to improve the underlying agent further.

Dataset - Changing from real sketches to Edge Maps

During the development of the system, the dataset underwent many changes and faced much scrutiny. It became a repeated roadblock as it severely limited the quality of outputs. To start Behance Artistic Media (BAM!) (Wilber et al., 2017) a collection of human-verified art data and WikiArt (WikiArt dataset, n.d.) a collection of genre-sorted famous artworks were used. Sketches for the dataset were collected from BAM! (Wilber et al., 2017), these could then

be matched with paintings collected from both WikiArt and BAM! (Wilber et al., 2017) to create an unpaired dataset. The choice of BiCycleGAN (Zhu et al., 2017b) was partly made due to it being an extension to CycleGAN (Zhu et al., 2017a), a different model which allowed unpaired training. However, due to the low quality of results that were collected from this training method, unpaired data was not used. A preliminary result of this method is shown in Figure A.2 in Appendix A. With more time, these issues could be resolved, but this has been left to future work.

Instead, a dataset consisting of paintings gathered from WikiArt (WikiArt dataset, n.d.) and edge maps was used. This was because it achieved a better quality of output. Originally, it was intended that the dataset would cover multiple genres, however, this led to mixed results which could be attributed to a lacklustre Extractor (E) network. In the future, this could be improved by expanding the latent vector size and including more parts of Liu et al. (2020b) work.

RLHF - Altering when updates are performed

Many of the final decisions for the RLHF networks design are based on the results of the test discussed in Section 4.3.1. The Extractor (E) network caused repeated problems. This was due to the "feedback loop" creating a circular dependency as the style vector was reused. To resolve this the Extractor (E) network was only updated at the end of a user's interaction with the tool.

During testing, it was also found that the Generator (G) network improved significantly when only updated at the end of each use case. This could be due to it being connected to the Extractor (E) network (which was only updated at the end of a use case). As a result of this, both networks were updated in this manner. This saves a lot of time during user use as network updates are not performed during operation, instead, they are performed on the closure of the application. It also allows for a more fixed user experience as altering networks during generation could lead to image deviations becoming more drastic.

4.2 Resources

4.2.1 Hardware

This dissertation made use of the Computer Science department's Hex Cluster at the University of Bath. The Garlick cluster (six GPUs NVIDIA GeForce RTX 2080 — 8.00 GB, 120Gb ram) was used to train initial models. The final tool was lightweight because the RLHF implementation does not require much compute, this is because it relies on human responses. As a result, it can be run on a laptop (Intel(R) Core(TM) i5-1035G1 CPU @ 1.00GHz (8 CPUs), ~1.2GHz and 8192MB RAM) which was ideal for the user study.

4.2.2 Software

A Torch implementation of BiCycleGAN (Zhu et al., 2017c) was used as it is very lightweight. The code implemented for this dissertation with all its additions is available on request, email: zu213@bath.ac.uk.

The key repositories used can be found below:

- WikiArt.
- CycleGan.
- BicycleGan.
- Sketch2art-pytorch.

4.3 Testing

The test plan was twofold. Firstly, self-evaluations were performed both during and after development. Secondly, a user study was performed at the end of development.

4.3.1 Self review and testing

Throughout the project, all modifications to the code were tested by observation of results. This was most important for the RLHF additions as testing of the UI was simple and performed during development. Furthermore, as shown later in Section 5.2 the user study also proved that the UI improved user control.

The most important test on the code was to ensure the suggested RLHF additions worked, i.e. the network would change correctly in accordance with feedback. Collecting enough human feedback to test the RLHF implemented by this project was impossible with the time limit of the project. So instead of this, a simple "human" that always picks the darkest images was programmed to use the tool. After running this on the same image roughly 1000 times, the results can be seen below in Figure 4.2. It had moved towards a darker depiction as the "human" intended. In Figure 4.1 there is a diagram showing how the model changed as the "human" interacted with the system. It updates quickly and therefore the RLHF is efficient.

Figure 4.1: Diagram showing the progression the network makes if darker images are always chosen. This is a visual representation of the results of this dissertation's test for the network updates. As can be seen, the images get darker as the iterations go on, as the "human" always picked the darker images this can be seen to be working.

It is also important to note in Figure 4.2 that the three options still show variance after the "human" has used the tool for 1000 iterations. This shows that the latent vector still remains important to the Generator (G) network after this dissertation's updates. This is contrary to what happened with normal training, where at this point the latent style vector would start to be disregarded. Since the RLHF implementation was successful for this use case, it can be conjected that, provided a human repeatedly picked the correct style, the network's understanding of such a style would improve.

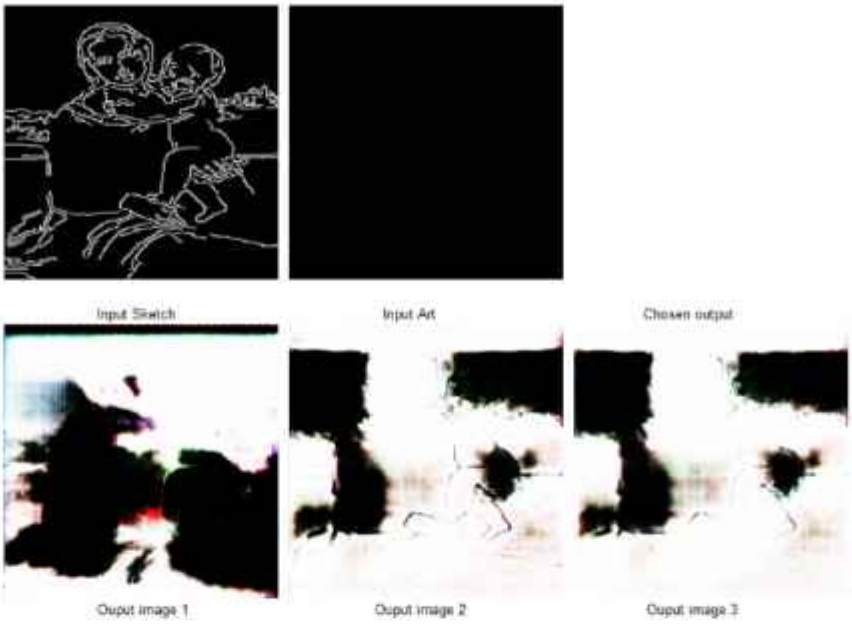


Figure 4.2: Diagram showing tools output after 1000 iterations of darkest output being chosen. Significantly, the variety in the three options remains, this is a problem that networks trained through the normal method faced as they overfitted.

4.3.2 User study

For the user study, BiCycleGAN (Zhu et al., 2017b) was trained using paintings from the WikiArt dataset with the created edgemaps discussed in Section 3.2. It was trained for 1675 epochs using this data set with a learning rate of 0.0002. The user study was then split into two further parts:

Part 1: Users using the tool

In the first part, the tool created in this dissertation was given to users to produce images of their own. The instructions were given to try and produce an image of High Renaissance style with the input sketch being a person and to keep going until they were satisfied. Users could produce the sketch using any drawing tool they wanted. The randomness was set to decay to ~0.034% by 30 iterations. By tweaking the randomness decay variable this can be modified, for these tests decay was set to 0.95.

Some of the images produced by users can be seen in Figure 5.1. Results are further discussed in Section 5.

Part 2: Participants judging improvements

During Part 1 initial and final images were produced representing each use case. For part 2 these two images from each user were presented to new participants for them to decide which more closely reflected the style input. Users were given the style inputs that the user from part 1 used. This was included to improve clarity and prevent a lack of knowledge about High Renaissance paintings from affecting results. The format of the document participants were given can be seen in Appendix A Figure A.4. Results can be seen in Appendix B, and are discussed further in Section 5.

4.3.3 Issues faced with implementation

A lot of challenges were faced during the development of the solution. Firstly, the Extractor (E) network could not be updated while the tool was being used. This was due to only one style vector being used from the network yet multiple updates being performed on it.

Secondly, early during development, the model would learn to maximise the "negative loss" and disregard minimising other losses. To counteract this, a cap was added to this "negative loss" which meant the model would still focus on minimising other losses instead of just maximising this.

4.3.4 Test failures

There were three key test failures during implementation:

Firstly, as this dissertation's tool is limited to one genre it does not work for other styles. This could be resolved provided enough training and a large enough dataset along with changes to the code to incorporate more of Liu et al. (2020b) contributions. However, this was not recreated successfully.

This dissertation's tool also struggles with less detailed sketch inputs. This could be resolved by incorporating work from Huang et al. (2020).

Finally, outside of other's work, the tool struggled with variety. Due to the randomness of the output options, they can often be similar to each other. This could be resolved with a further understanding of the style vector's latent space. With this, the output options could be better distributed, however, this would take further work.

Chapter 5

Results

This section discusses the results gathered, including identifying the successes and limitations of this dissertation's solution. Overall, the system successfully granted users more control. Importantly, the user study was split into two parts, with participants playing the role of users in the first part and judges in the second part. This is further explained in Section 4.3.2.

5.1 Qualitative results

5.1.1 User study

In the first part of the study, images were collected from users utilising this dissertation's tool. Users aimed to produce an image that most closely reflected the input style image. This was achieved as users made choices between the offered solutions. Some of the images from their first and last choices can be seen in Figure 5.1.

As stated previously, in this study users aimed to capture style. Style for visual arts is defined as a "distinctive manner which permits the grouping of works into related categories" (Fernie, 1995). Although the created agent is trained to work with High Renaissance paintings, which in itself is a genre or style, this can be split into many sub-styles. Many users tend towards more colourful final images when using the tool; the resulting images reflect this. The final images represent the sub-styles more accurately but are still imperfect. Therefore, it can be argued that the tool captured these sub-styles more accurately than existing works.

Although there are improvements it is hard to see if they are great or small. The network in its current state does not understand what to take from the style image. Instead, it outputs a style it has learned from training. However, with a human assisting this often falls closer to the intended style than it would without them.

As discussed in Section 3.1, as the network was trained further by normal methods, it became less varied as the style vector was discounted. This was evidenced by the tool not showing the same amount of variation across the three images. The alternate RLHF network updates implemented by this dissertation did not show this same issue. Users picked more colourful (if not more stylish) options which led to the converse, the networks outputted a wider breadth of images. This implies these changes have overcome this issue. However, as discussed this is impossible to verify without copious amounts of real human input.

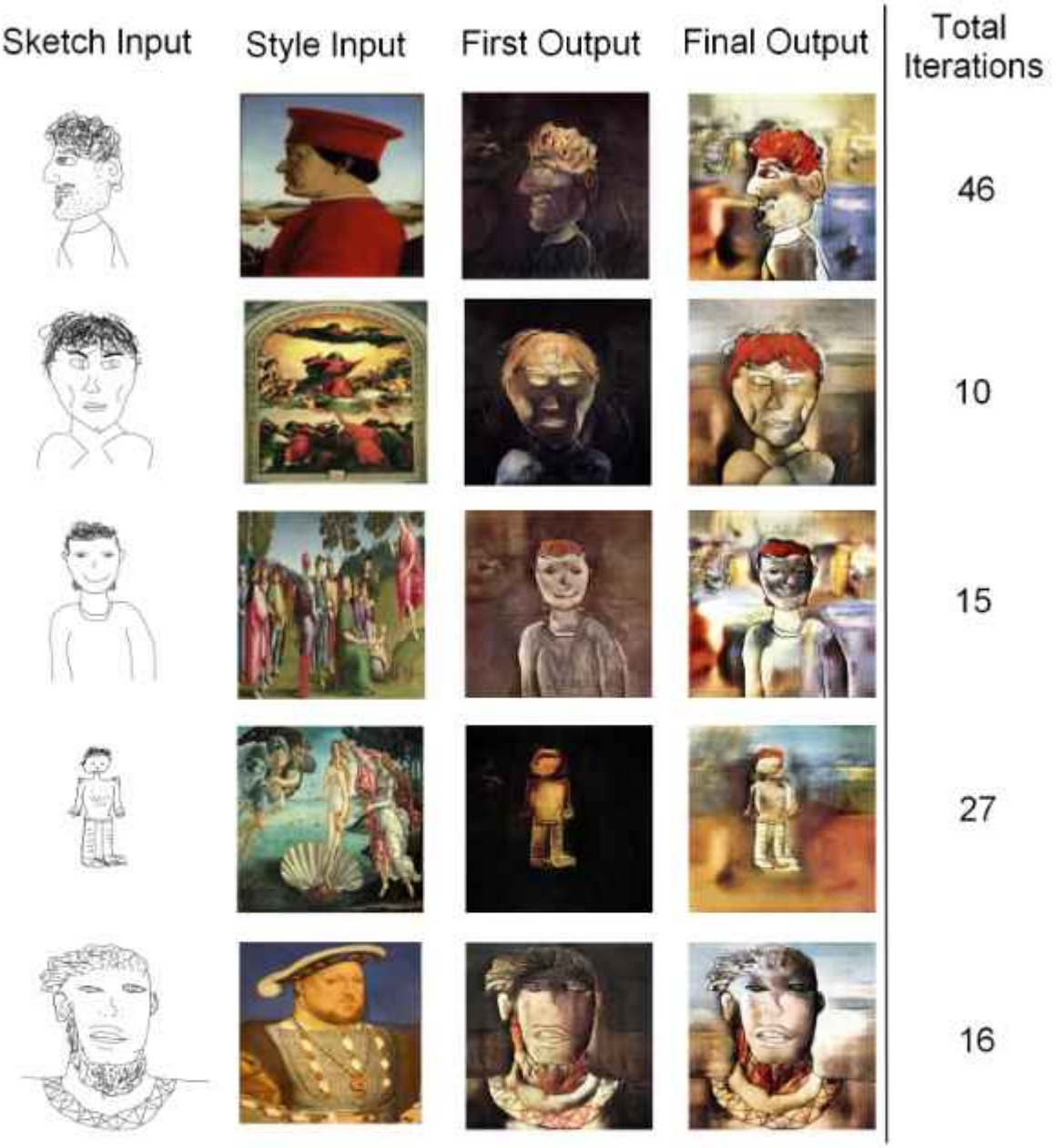


Figure 5.1: Results of the first part of the user study. First column: Input sketches of people drawn by users. Second column: Input style images, chosen by users to be a High Renaissance painting. Third Column: Initial output chosen by users to match style. Fourth Column: Final output chosen by users to match style. Fifth Column: Number of iterations between first and final images undertaken by users.

5.1.2 Self study

Some results from the self-study can seen in Figure 5.2 and Figure 5.3. Throughout development, the tool was experimented with, however, as can be seen from Figure 5.2, outputs still lack accuracy. One of the key issues with current image synthesis models is disentangling style and semantic information. This dissertation's model faces the same problems. Although overall colour can be improved through human feedback, its accuracy about placement (semantic information) is still low. For example in Figure 5.2 the tool was steered towards darker and

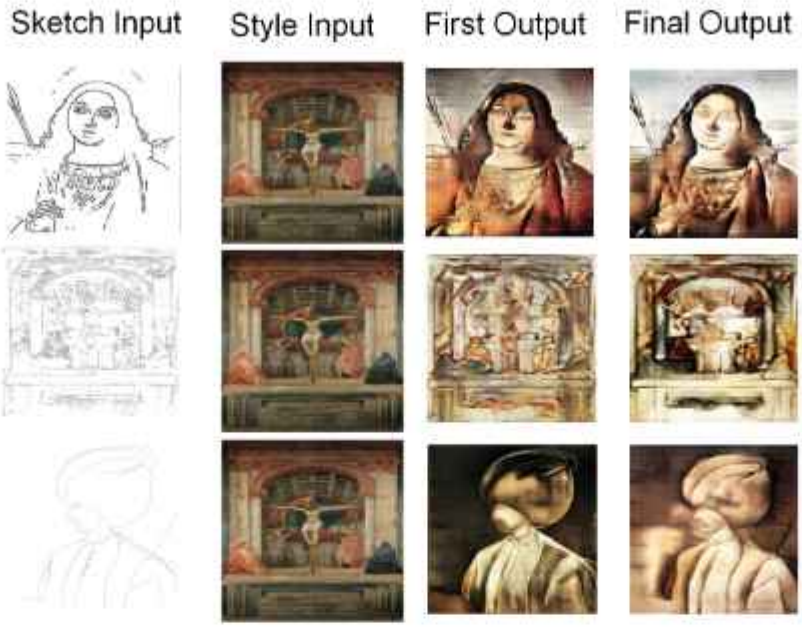


Figure 5.2: Figure showing some first and final results produced by this dissertation's tool. Each use case was run for ~20 iterations. All these cases used a fixed style image for clarity. We were able to guide the network towards sharper and darker images which we felt better represented the style image.

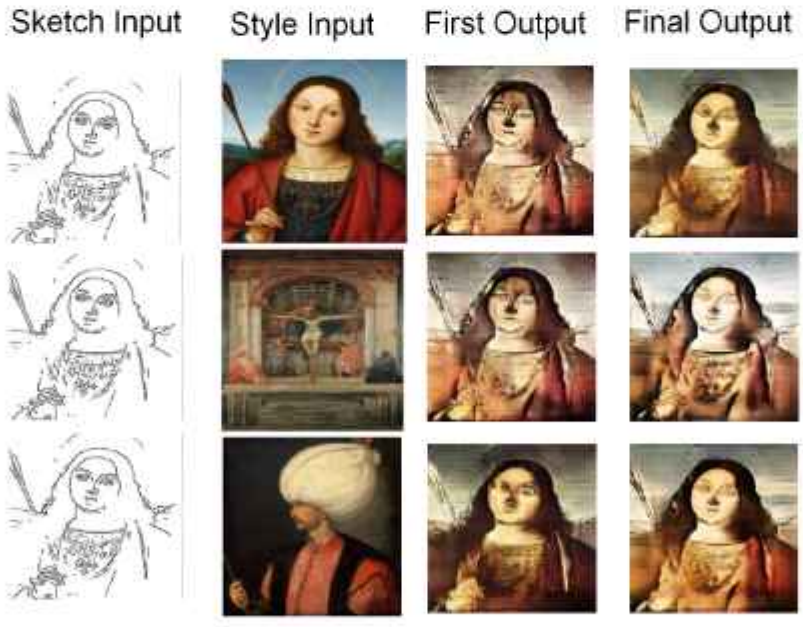


Figure 5.3: Figure showing some first and final results produced by this dissertation's tool. Each use case was run for ~20 iterations. All these cases used a fixed sketch input for clarity. This allowed the network to be guided to more colourful images in most cases. This was done as this normally better reflected each style input. The top and bottom rows become smoother images, whereas the middle row leads to a sharper image based on our interpretation of the style input.

sharper colours which could be argued better represented the fixed style image. However, the faces present in the output images do not match the skin colours of the style inputs.

Despite all this, it is important to look beyond the quality, as it is still not close to the stage of real art. It is also missing the human feedback necessary to perfect it. With this in mind, it is clear that this dissertation's solution has made improvements over its underlying model.

5.2 Quantitative results

Table 5.1 shows the outcomes of the 1-tailed test that was performed on the collected data. The test's result suggests this tool improved user control when directed towards a specific style. This test was completed on the results of the second part of the study (see Section 4.3.2 for further explanation). If this dissertation's solution showed no improvement in user control, the results of this test would be close to random. However, for all the images it is statistically significant to be a value >1.5 (the split between option 1 or 2) up to a p-value of 0.004.

Also, included are the statistical study results when the worst-performing images from the first part of the study were removed. This is labelled as "All-" in Table 5.1. Throughout the study, it could be argued that some users were more successful at trying to replicate a style than others. Removing two user's output images from the analysis reduced the p-value to < 0.001. This implies that it is almost guaranteed that the tool enabled a user to have more control over style than just using the underlying network.

However, this second result is not as credible as the first, since data has been removed. Despite this, the first result is significant enough to show the quantitative success of this endeavour.

One Sample T-Test

	t	df	p
All	2.748	109	0.004
All-	4.466	87	< .001

Note. For the Student t-test, the alternative hypothesis specifies that the mean is greater than 1.5.
Note. Student's t-test.

Table 5.1: Table showing results of statistical study. "All" is the total results when all works produced by users in part 1 of the study are included. "All-" is the results when a select few results from the first part of the study are removed.

The results of this two-part study have been collated into two graphs below. The first graph in Figure 5.4, clearly shows an overall favouring for the final option produced, that is the image produced after the modifications of this dissertations tool. This graph also displays how the results of a few use cases were not as well defined, with images one and nine being key examples of contradiction to the norm. This shows that the tool still relies on a user's skill.

This dependence on skill could either be attributed to the quality of sketch input or the quality of choices made within the tool. However, the sketch input did not have much impact on the study's results as it was the style that was being tested. If the sketch is the factor being limited by user skill it could be resolved by incorporating the work of Huang et al. (2020).

Alternatively, if the quality of choices made within the tool was limited by skill, this suggests that users lack clarity for the choices they make. This could be attributed to foresight. With the proposed solution, users are unable to understand the next generated images clearly before they make a choice. A future addition to resolve this would be the ability to backtrack.

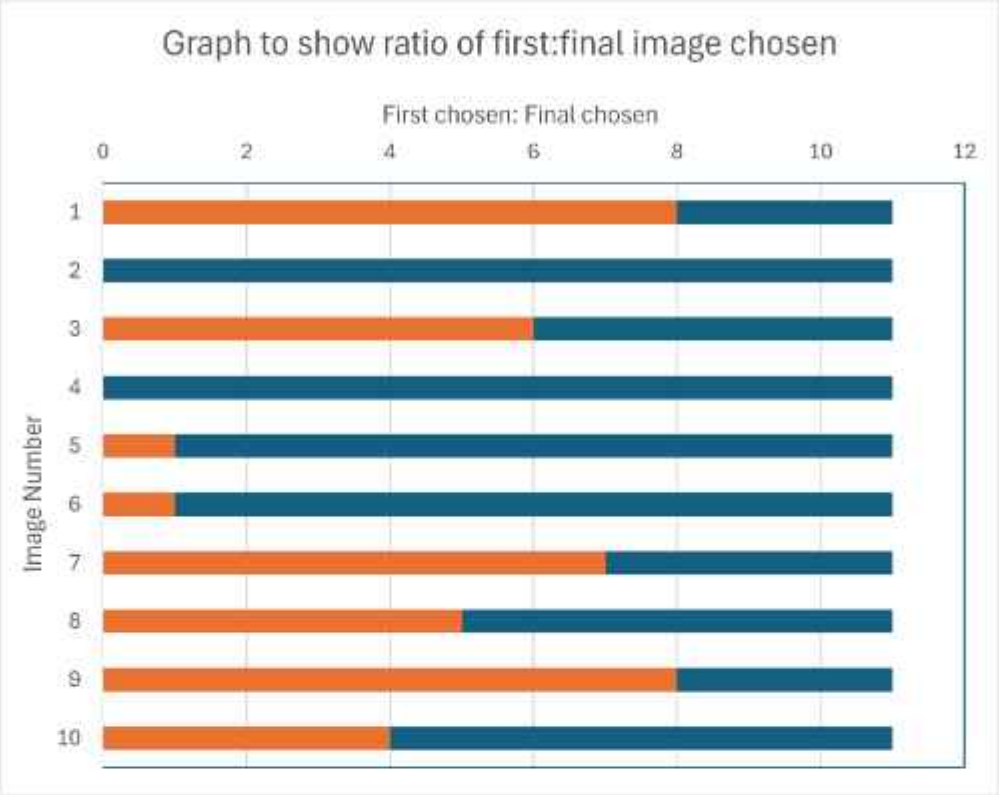


Figure 5.4: Graph showing ratio of first output:final output chosen by the second part of the user study. The orange parts of the column represent the number of times the first output was chosen for each use case, the blue part represents the converse. As can be seen in some images the improvement is more clear cut whereas in others it is more ambiguous.

This second graph in Figure 5.5 shows how many option 2's each user picked. The average is above 5 and as shown statistically significant. However, as can be seen, many users were close to random in their choices. During the study it became apparent that participants favoured returning an even distribution of 1's and 2's, this suggests that the results were not always clear-cut enough. The way to resolve this issue could be to increase the size of the style vector to achieve a better latent space representation. Alternatively, the entire underlying network could be overhauled, to more closely replicate the work of Liu et al. (2020b). Despite all this, this graph proves that there were no outlier participants in the second part of the user study.

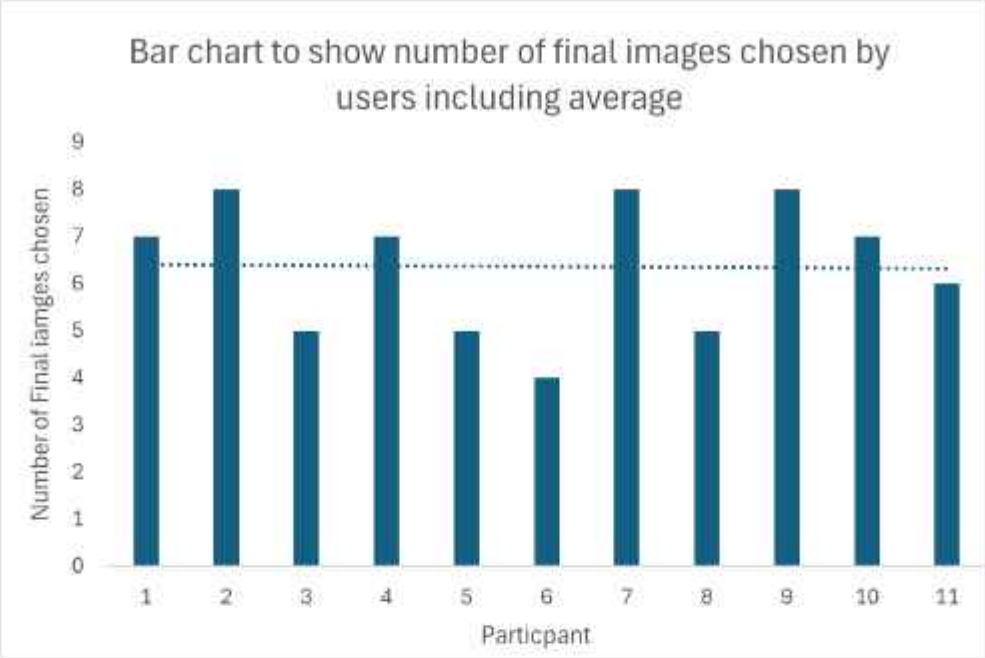


Figure 5.5: Graph showing the number of final outputs each participant chose. This was put together to see if some participant's results were outliers to the pattern in the second part of the user study. However, it can be seen that none were outliers.

5.3 Successes

The model succeeded in:

Firstly, all users found it intuitive to use. This is likely due to the simplistic design. As mentioned in Section 4.3.2, users could use whichever art tool they desired to produce the sketch and procure the style image (of the High Renaissance) style. This enabled a lot of freedom which made the task easier for each user to approach the task however they desired. During the user study, purposely vague explanations of what to do were given to observe if the tool was intuitive. The tool also seems to be more straightforward than existing online tools like Playform (Liu et al., 2020a) and Nvidia-Canvas (Nvidia Canvas, n.d.) from the user's reactions in the user study. This is despite the tool not having a simple way to input images.

Secondly, the tool improved the output quality of the individual use cases as shown by the results of the user study in Section 5.2. During the development of the tool, many images were produced for evaluation of the success of the tool. An example can be seen in Figure 5.6. As can be seen, the final choice represents the colours of the style image a lot more closely, whereas the first choice contains a lot of colours that are not present in the style. Currently, the network has learned to add colours based on its general understanding. The final image, however, has a better colour representation, as it could be steered to a desired style. Once the tool is closed, the network updates. These updates lead to the network moving away from the images it originally produced, and moving towards producing images similar to the user's final choice. With appropriate choices, the quality of future outputs will improve.

Finally, the tool gave users more control. As highlighted in Section 2.2, one of the key gaps in the literature was the lack of control current image synthesis methods allowed. This, in turn, would limit the system's capacity for RLHF. The approach taken, although simplistic, did not

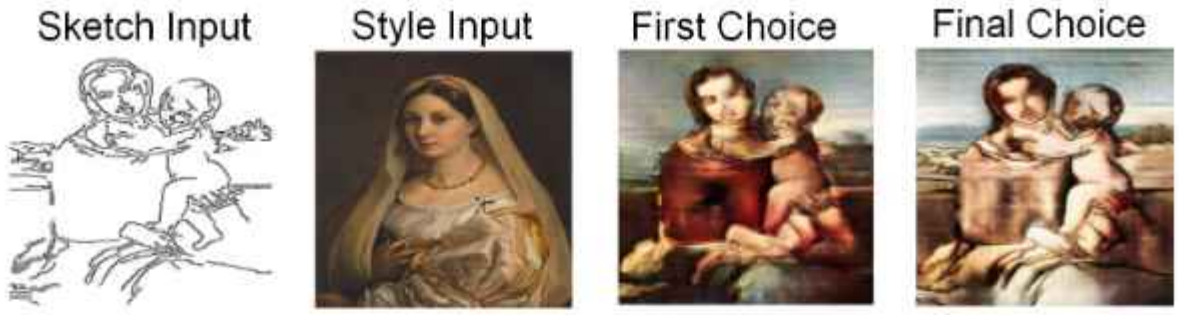


Figure 5.6: Display of successful use of this dissertation's tool to steer close to a desired style. The final choice has colours and textures that are closer to the style inputs than that of the first choice. This took 17 iterations to complete.

confuse users with abstract inputs, (such as seed numbers Playform (Liu et al., 2020a) use). Instead, with image choices a user can use visual information to steer towards the output they desire. This was proved to give more control than the standalone underlying network by the statistical study.

5.4 Limitations

However, despite successes, it is clear this dissertation's solution is far from producing real art. For example, this dissertation's implementation of RLHF does not improve the network's understanding of semantic information. This is shown in Figure 5.6 where there are still structural issues that surround detailed areas such as faces. Overall, the quality is still low despite improvements. This can be seen when comparing images synthesised by this dissertation's solution with real artwork.

Another failure is related to Figure 5.7, as seen below, the input style image was just a black square. However, despite using the tool for over 30 iterations, the "Final Choice" was the darkest image produced. Although in the RLHF testing in Section 4.3.1 darker images were produced, this is over multiple iterations. Ideally, this dissertation's solution would be able to produce a much darker image within a single use case.

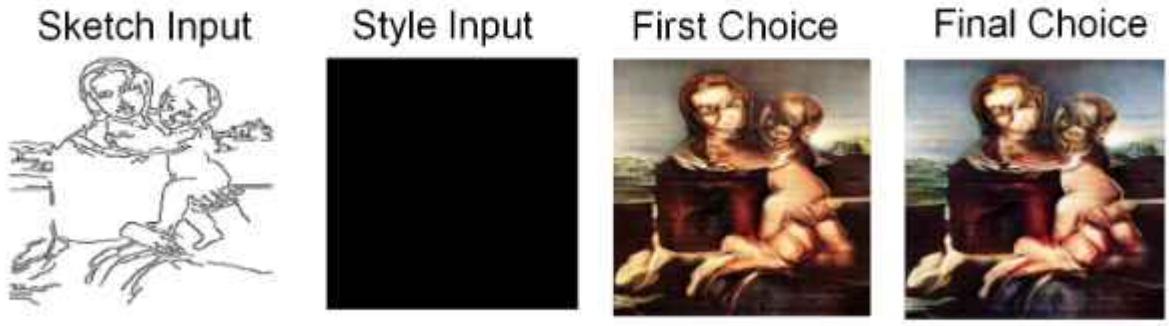


Figure 5.7: Display of shortcoming of this tool. As can be seen, the tool does not enable the user to produce a much darker image, which should be possible within 30 iterations.

Finally and most importantly, the investigation was also limited by the ability to collect human feedback. With more feedback, it could be seen if the RLHF updates significantly improved

the network. This would in turn more clearly justify this dissertation's proposed solution and whether it improves output quality.

5.5 Results summary

Overall, the control users have over current image generation was improved by this dissertation's model, as shown by the statistical study. By incorporating RLHF, this in turn will, improve the quality of the outputs. However, the latter is difficult to prove without more data collection. Despite these advances, user control and image quality in the realm of art are still subpar.

Chapter 6

Conclusion

This dissertation has presented a new "feedback loop" system to further current image synthesis tools for the domain of art. It achieved this by incorporating RLHF to enhance existing models based on human feedback. This has elicited final results, for individual use cases, better aligning with the mental image a user set out to realise. This advancement is attributed to improved user control.

6.1 Achievements

In summary, this work has reached the following achievements:

- Implementation of a "feedback loop" of images for users to select and more accurately realise their mental image through fine-tuning. Thus, allowing users more control.
- Invention of a novel technique that has enabled further learning, by utilising the aforementioned "feedback loop". This model has the potential to improve image quality. However, further analysis through additional data collection is required to confirm this.
- Concluded that RLHF can successfully improve the current state-of-the-art models in terms of user control.

6.2 Problems faced

During its development, this project also faced issues including:

- RLHF networks updating to maximise the new "negative loss" for the Generator (G) network and disregarding other losses collated for the Generator (G) network.
- Underlying networks of the current state-of-the-art being inadequate, despite extensive training. Due to time constraints, this dissertation was unable to recreate other papers' work in their entirety.
- Variance not being appropriate. Due to the use of a latent style vector, the style is abstracted and therefore data is lost. This led to issues with the output options, where their distribution was sub-optimal. With more extensive training, larger latent variable size and more fine-tuning of variables, this could be resolved.

As a result of these issues and time constraints, the model had to be limited to one genre for the user study and self-testing. This does not show an adequate representation of the ideal solution mentioned in Section 1.3. In this ideal solution, the latent vector represents any style and the Generator (G) network can translate this to any image. The limitations of this dissertation's solution are discussed in Section 5.4.

6.3 Future work

Future work that could improve this dissertation's solution includes:

- Implementation of the choice system as a ranking system and modifying the RLHF to take advantage of this, similar to the works of Xu et al. (2023); Kirstain et al. (2023). This could further improve output quality.
- Further optimisation of the underlying networks to fit the dataset for best performance. This would allow a more accurately quantifiable evaluation of the success of this dissertation's contributions. This would include incorporating the excluded parts of the work of Liu et al. (2020b).
- Further improvement and experimentation with the dataset, including incorporating Huang et al. (2020) work. This would allow the underlying networks to take sketches of different qualities. This could be combined with the work of Lu et al. (2017) to produce a more varied sketch to art dataset.
- Experimentation with a pessimistic approach to the RLHF system inspired by the work of Li, Yang and Wang (2023). This has the potential to improve the amount of learning from a single-user interaction.
- Modifying the user tool to allow the ability to backtrack or alter the number of choices available to users. Thus improving user control.
- Overhauling this dissertation's model to combine stable-diffusion and GANs to allow for the input of not only a sketch and style image but also a semantic text prompt.

Overall, this dissertation has found that RLHF in the form of a "feedback loop" does improve user control and with more appropriate feedback may improve output quality. Thus, contributing to the two key gaps earlier identified in the literature.

Word count: 10461

Word Count

```
File: Dissertation.tex
Encoding: utf8
Sum count: 12430
Words in text: 10671
Words in headers: 275
Words outside text (captions, etc.): 1479
Number of headers: 77
Number of floats/tables/figures: 35
Number of math inlines: 5
Number of math displayed: 0
Subcounts:
  text+headers+captions (#headers/#floats/#inlines/#displayed)
  24+13+0 (1/0/0/0) _top_
  145+0+0 (0/0/0/0) Abstract
  38+1+0 (1/0/0/0) Preamble
  869+46+192 (9/3/0/0) Introduction
  4023+93+323 (22/9/0/0) Literature and Technology Survey
  1803+44+282 (14/5/1/0) Method
  1498+54+92 (16/2/2/0) Implementation and Testing
  1700+13+414 (8/8/2/0) Results
  571+11+176 (6/8/0/0) Conclusion
```

Bibliography

Baraheem, S.S., Le, T.N. and Nguyen, T.V., 2023. Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook [Online]. Available from: https://link.springer.com/article/10.1007/s10462-023-10434-2.

Bendel, O., 2023. Image synthesis from an ethical perspective [Online]. Available from: https://link.springer.com/article/10.1007/s00146-023-01780-4.

Bowman, S.R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Kernion, J., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lovitt, L., Elhage, N., Schiefer, N., Joseph, N., Mercado, N., DasSarma, N., Larson, R., McCandlish, S., Kundu, S., Johnston, S., Kravec, S., Showk, S.E., Fort, S., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Mann, B. and Kaplan, J., 2022. Measuring progress on scalable oversight for large language models [Online]. Available from: https://arxiv.org/abs/2211.03540.

Brisbane, A., 1913. The post standard [Online]. Available from: https://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words.

Bui, T., Ribeiro, L., Ponti, M. and Collomossea, J., 2018. Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression [Online]. Available from: [Online].\Availablefrom: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiO39CTz9-CAxUvW0EAHR9sDfEQFnoECBUQAQ&url=https%3A%2F%2Fopenreview.net%2Fpdf%3Fid%3Dc8CW1RwdGJ&usg=AOvVaw3fB62w0VL-N3-uSZeS6e2L&opi=89978449.

Canny, J., 1986. A computational approach to edge detection [Online]. Available from: https://ieeexplore.ieee.org/document/4767851.

Casper, S., Davies, X., Shi, C., Gilbert, T.K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E.J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D. and Hadfield-Menell, D., 2023. Open problems and fundamental limitations of reinforcement learning from human feedback [Online]. Available from: https://arxiv.org/abs/2307.15217.

Cezanne, P., 1895. Mont sainte-victoire [Online]. Available from: https://www.pinterest.co.uk/pin/mont-saintevictoire-paul-cezanne--614248836663824952/.

Chen, T., Cheng, M.M., Tan, P., Shamir, A. and Hu, S.M., 2009. Sketch2photo: Internet

image montage [Online]. Available from: https://cg.cs.tsinghua.edu.cn/papers/SiggraphAsia_2009_sketch2photo.pdf.

Chen, W. and Hays, J., 2018. Sketchygan: Towards diverse and realistic sketch to image synthesis [Online]. Available from: https://arxiv.org/abs/1801.02753.

Cheng, S.I., Chen, Y.J., Chiu, W.C., Tseng, H.Y. and Lee, H.Y., 2022. Adaptively-realistic image generation from stroke and sketch with diffusion model [Online]. Available from: https://arxiv.org/abs/2208.12675.

Collomosse, J., Bui, T., Wilber, M., Fang, C. and Jin, H., 2017. Sketching with style: Visual search with sketches and aesthetic context [Online]. Available from: https://openaccess.thecvf.com/content_ICCV_2017/papers/Collomosse_Sketching_With_Style_ICCV_2017_paper.pdf/.

Colton, S., 2008. Creativity versus the perception of creativity in computational systems [Online]. Available from: https://cdn.aaai.org/Symposia/Spring/2008/SS-08-03/SS08-03-003.pdf.

Daniels-Koch, O. and Freedman, R., 2022. The expertise problem: Learning from specialized feedback [Online]. Available from: https://arxiv.org/abs/2211.06519.

Dawkins, R., 1986. The blind watchmaker [Online]. Available from: [Online].\Articleavailablefrom:https://en.wikipedia.org/wiki/The_Blind_Watchmaker.

Elgammal, A., Liu, B., Elhoseiny, M. and Mazzone, M., 2017. Can: Creative adversarial networks generating "art" by learning about styles and deviating from style norms [Online]. Available from: https://arxiv.org/abs/1706.07068.

Fernie, E., 1995. Art history and its methods: A critical anthology [Online]. Available from: https://books.google.co.uk/books?vid=ISBN9780714829913&redir_esc=y.

Firealpaca [Online], 2023. Available from: https://firealpaca.com/.

Fotor ai [Online], n.d. Available from: https://www.fotor.com/.

Gao, C., Liu, Q., Xu, Q., Wang, L., Liu, J. and Zou, C., 2020. Sketchycoco: Image generation from freehand scene sketches [Online]. Available from: https://arxiv.org/abs/2003.02683.

Ghosh, A., Zhang, R., Dokania, P.K., Wang, O., Efros, A.A., Torr, P.H. and Shechtman, E., 2019. Interactive sketch fill: Multiclass sketch-to-image translation [Online]. Available from: https://arxiv.org/abs/1909.11081.

Goldman, A.H., 2020. Realism about aesthetic properties [Online]. Available from: https://www.jstor.org/stable/431968.

Ham, C., Tarres, G.C., Bui, T., Hays, J., Lin, Z. and Collomosse, J., 2022. Cogs: Controllable generation and search from sketch and style [Online]. Available from: [Online].\https://arxiv.org/abs/2203.09554.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y. and Cohen-Or, D., 2022. Prompt-to-prompt image editing with cross attention control [Online]. Available from: https://arxiv.org/abs/2208.01626.

Huang, J., Liao, J., Tan, Z. and Kwong, S., 2020. Multi-density sketch-to-image translation network [Online]. Available from: https://arxiv.org/abs/2006.10649/.

Isola, P., Zhu, J.Y., Zhou, T. and Efros, A.A., 2016. Image-to-image translation with conditional adversarial networks [Online]. Available from: https://arxiv.org/abs/1611.07004.

Kang, H., Lee, S. and Chui, C.K., 2007. Coherent line drawing [Online]. Available from: http://cg.postech.ac.kr/papers/kang_npar07_hi.pdf.

Kazemi, H., Taherkhani, F. and Nasrabadi, N.M., 2020. Preference-based image generation [Online]. Available from: https://openaccess.thecvf.com/content_WACV_2020/papers/Kazemi_Preference-Based_Image_Generation_WACV_2020_paper.pdf.

Khurana, V., Singla, Y.K., Subramanian, J., Shah, R.R., Chen, C., Xu, Z. and Krishnamurthy, B., 2023. Behavior optimized image generation [Online]. Available from: https://arxiv.org/abs/2305.01569.

Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes [Online]. Available from: https://arxiv.org/abs/1312.6114/.

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J. and Levy, O., 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation [Online]. Available from: https://arxiv.org/abs/2305.01569.

Krishna, A., Bartake, K., Niu, C., Wang, G., Lai, Y., Jia, X. and Mueller, K., 2021. Image synthesis for data augmentation in medical ct using deep reinforcement learning [Online]. Available from: https://arxiv.org/abs/2103.10493.

Leder, H., Gerger, G., Dressler, S.G. and Schabmann, A., 2012. How art is appreciated. [Online]. Available from: https://psycnet.apa.org/record/2011-28761-001.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M. and Gu, S.S., 2023. Aligning text-to-image models using human feedback [Online]. Available from: https://arxiv.org/abs/2302.12192.

Lee, Y.J., Zitnick, C.L. and Cohen, M.F., 2011. Shadowdraw: Real-time user guidance for freehand drawing [Online]. Available from: https://dl.acm.org/doi/10.1145/2010324.1964922.

Li, Z., Yang, Z. and Wang, M., 2023. Reinforcement learning with human feedback: Learning dynamic choices via pessimism [Online]. Available from: https://arxiv.org/abs/2305.18438.

Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., Sun, J., Pont-Tuset, J., Young, S., Yang, F., Ke, J., Dvijotham, K.D., Collins, K., Luo, Y., Li, Y., Kohlhoff, K.J. and Deepak Ramachandran, V.N., 2023. Rich human feedback for text-to-image generation [Online]. Available from: https://arxiv.org/abs/2312.10240.

Liu, B., Song, K., Zhu, Y. and Elgammal, A., 2020a. Sketch-to-art: Synthesizing stylized art images from sketches [Online]. Available from: https://www.playform.io/.

Liu, B., Song, K., Zhu, Y. and Elgammal, A., 2020b. Sketch-to-art: Synthesizing stylized art images from sketches [Online]. Available from: https://arxiv.org/abs/2002.12888.

Liu, Y., Qin, Z., Luo, Z. and Wang, H., 2017. Auto-painter: Cartoon image generation from sketch by using conditional generative adversarial networks [Online]. Available from: https://arxiv.org/abs/1705.01908.

Lu, Y., Wu, S., Tai, Y.W. and Tang, C.K., 2017. Image generation from sketch constraint using contextual gan [Online]. Available from: https://arxiv.org/abs/1711.08972.

Mace, M.A. and Ward, T., 2010. Modeling the creative process: A grounded theory analysis of creativity in the domain of art making [Online]. Available from: https://www.tandfonline.com/doi/abs/10.1207/S15326934CRJ1402_5.

Mechanical turk [Online], n.d. Available from: https://www.mturk.com/.

Nvidia canvas [Online], n.d. Available from: https://www.nvidia.com/en-gb/studio/canvas/.

Photocopy effect [Online], 2016. Available from: https://www.youtube.com/watch?v=QNmniB_5Nz0.

Ramponi, M., 2023. The full story of large language models and rlhf [Online]. Available from: https://www.assemblyai.com/blog/the-full-story-of-large-language-models-and-rlhf/.

Ramy, A. and Barakat, N. hosny, 2022. Sketch to image using generative adversarial networks (gan) [Online]. Available from: https://www.researchgate.net/publication/362481944_Sketch_to_Image_Using_Generative_Adversarial_Networks_GAN.

Richardson, E., Alaluf, Y., Azar, Y., Patashnik, O., Shapiro, S., Nitzan, Y. and Cohen-Or, D., 2020. Encoding in style: a stylegan encoder for image-to-image translation [Online]. Available from: https://arxiv.org/abs/2008.00951.

Sangkloy, P., Burnell, N., Ham, C. and Hays, J., 2016a. The sketchy database: Learning to retrieve badly drawn bunnies [Online]. Available from: https://faculty.cc.gatech.edu/~hays/tmp/sketchy-database.pdf.

Sangkloy, P., Lu, J., Fang, C., Yu, F. and Hays, J., 2016b. Scribbler: Controlling deep image synthesis with sketch and color [Online]. Available from: https://arxiv.org/abs/1612.00835.

Tien, J., He, J.Z.Y., Erickson, Z., Dragan, A.D. and Brown, D.S., 2022. Causal confusion and reward misidentification in preference-based reward learning [Online]. Available from: https://arxiv.org/abs/2204.06601.

Turner, J., 1840. The slave ship [Online]. Available from: https://en.wikipedia.org/wiki/The_Slave_Ship.

Wang, Q., Deng, H., Qi, Y., Li, D. and Song, Y.Z., 2023. Sketchknitter: Vectorized sketch generation with diffusion models [Online]. Available from: https://openreview.net/pdf?id=4eJ43EN2g6l.

Wikiart dataset [Online], n.d. Available from: https://www.wikiart.org/.

Wilber, M.J., Fang, C., Jin, H., Hertzmann, A., Collomosse, J. and Belongie, S., 2017. Bam! the behance artistic media dataset for recognition beyond photography [Online]. Available from: https://arxiv.org/abs/1704.08614.

Winnemoller, H., Kyprianidis, J.E. and Olsen, S., 2012. Xdog: an extended difference-of-gaussians compendium including advanced image stylization [Online]. Available from: https://users.cs.northwestern.edu/~sco590/winnemoeller-cag2012.pdf.

Xia, W., Yang, Y. and Xue, J.H., 2019. Cali-sketch: Stroke calibration and completion for high-quality face image generation from human-like sketches [Online]. Available from: https://arxiv.org/abs/1911.00426.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J. and Dong, Y., 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation [Online]. Available from: https://arxiv.org/abs/2304.05977.

Zhu, J.Y., Krähenbühl, P., Shechtman, E. and Efros, A.A., 2018. Generative visual manipulation on the natural image manifold [Online]. Available from: https://arxiv.org/abs/1609.03552.

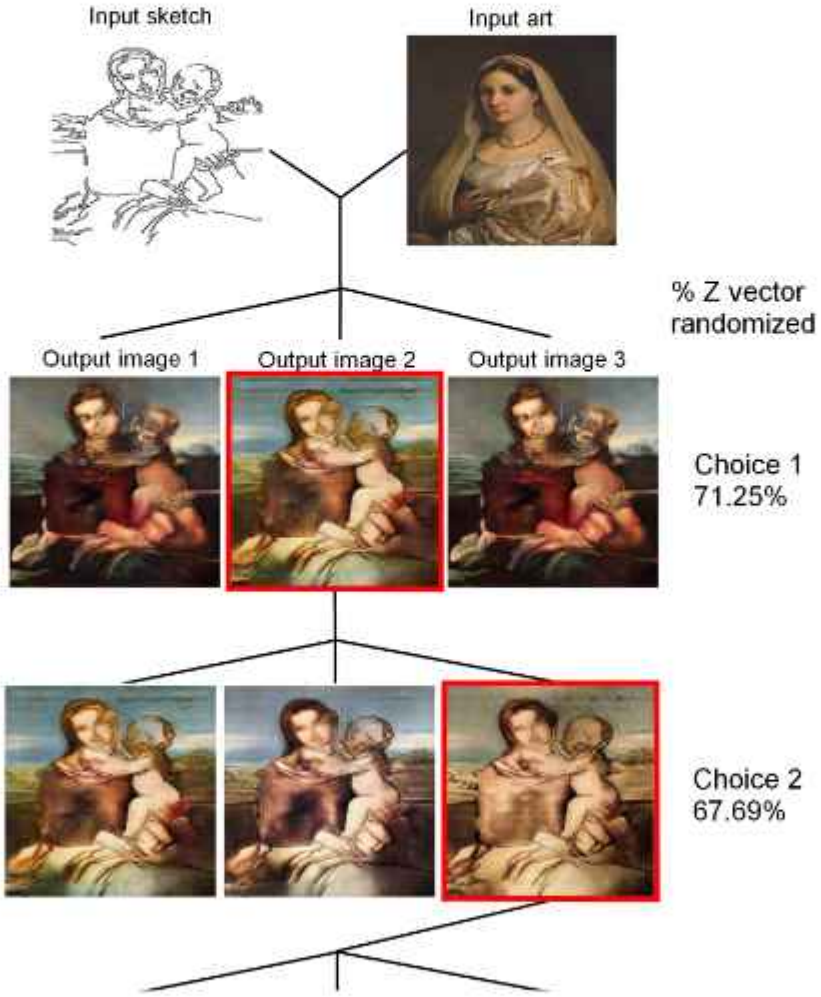
Zhu, J.Y., Park, T., Isola, P. and Efros, A.A., 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks [Online]. Available from: https://arxiv.org/abs/1703.10593.

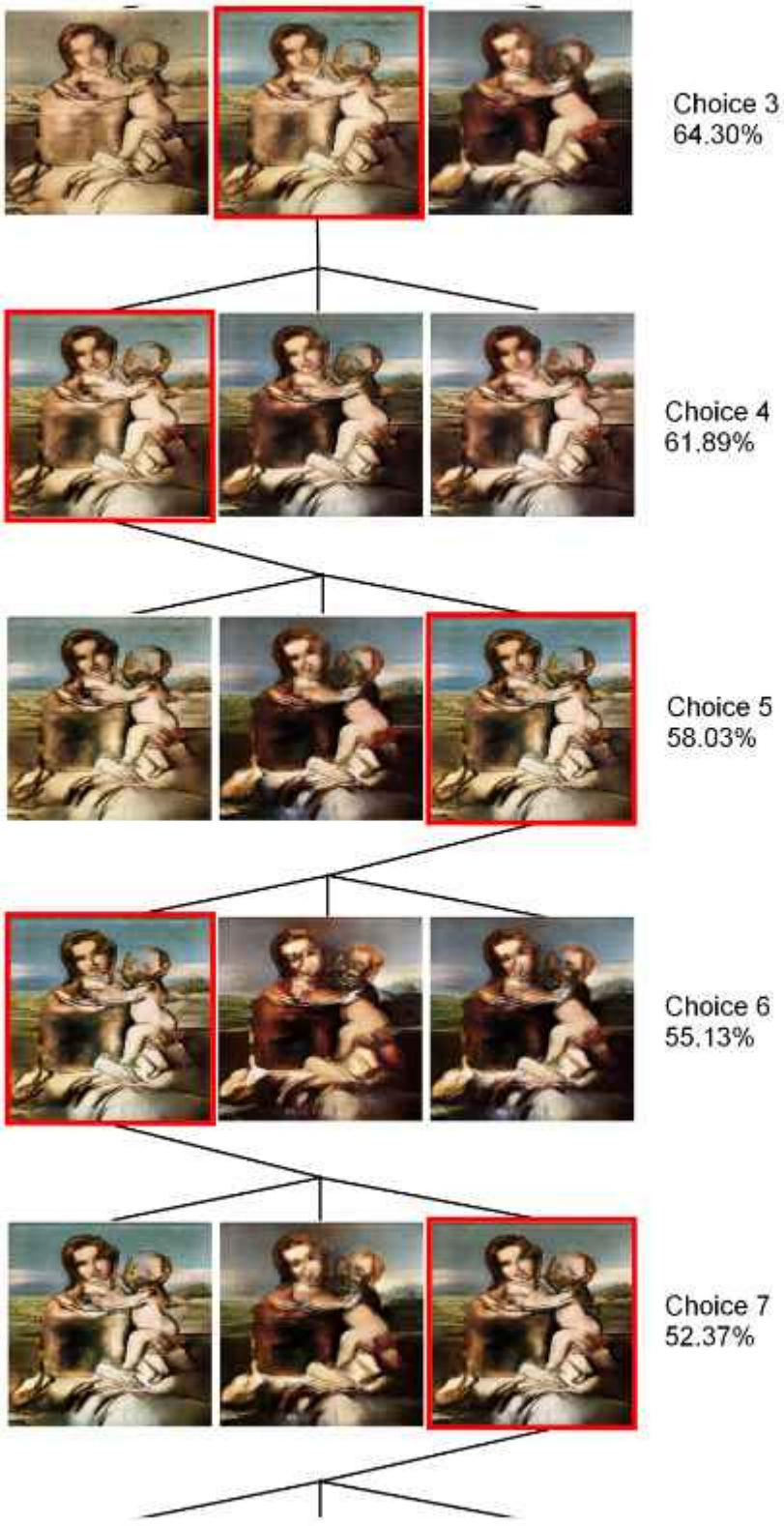
Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O. and Shechtman, E., 2017b. Toward multimodal image-to-image translation [Online]. Available from: https://arxiv.org/abs/1711.11586.

Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O. and Shechtman, E., 2017c. Toward multimodal image-to-image translation [Online]. Available from: https://github.com/junyanz/BicycleGAN.

Appendix A

Additional Diagrams




Choice 3
64.30%
Choice 4
61.89%
Choice 5
58.03%
Choice 6
55.13%
Choice 7
52.37%

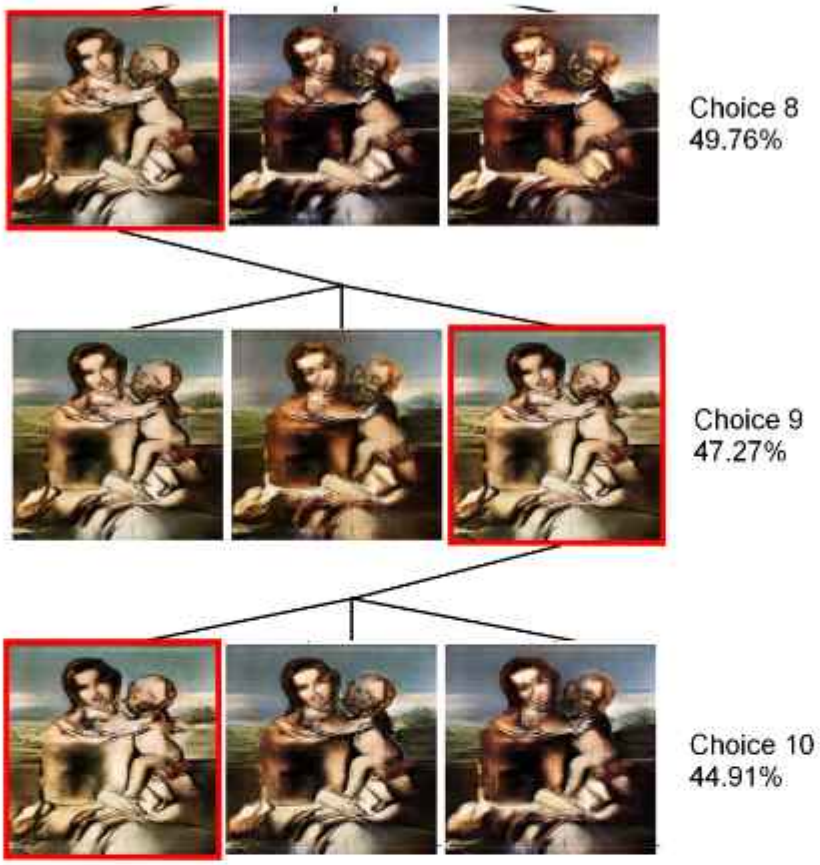


Figure A.1: Full use case diagram of this dissertation's tool, mentioned in Figure 3.2. The z-vector's variance decreases with each iteration.

Figure A.2: Diagram showing results from this dissertation's previous model, this model was trained on unpaired data of real sketches and real art, as can be seen the results were lacklustre.

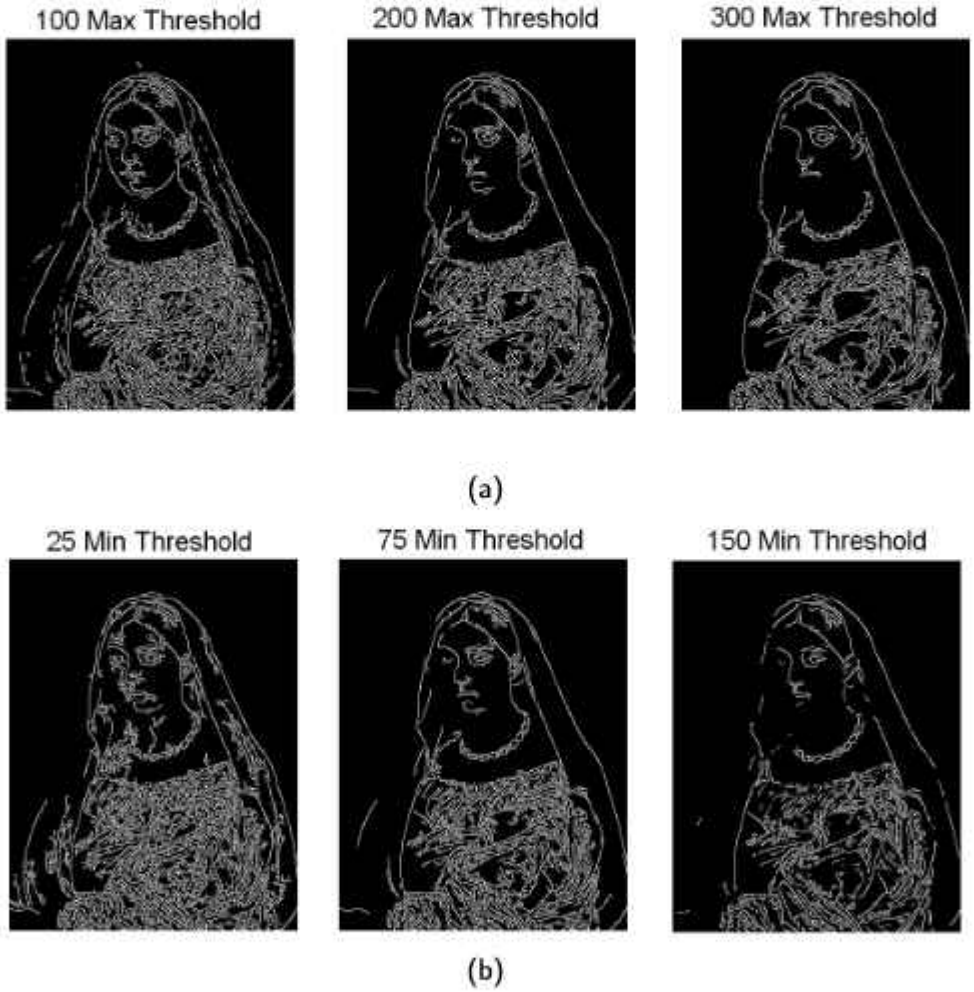


Figure A.3: Diagram showing canny edge map detection experiments. Figure (a) shows experimentation with the maximum threshold modified. Figure (b) shows experimentation with the minimum threshold modified. For both cases, too low a threshold leads to too many edges being picked up, but with too high a threshold too few edges are picked up.

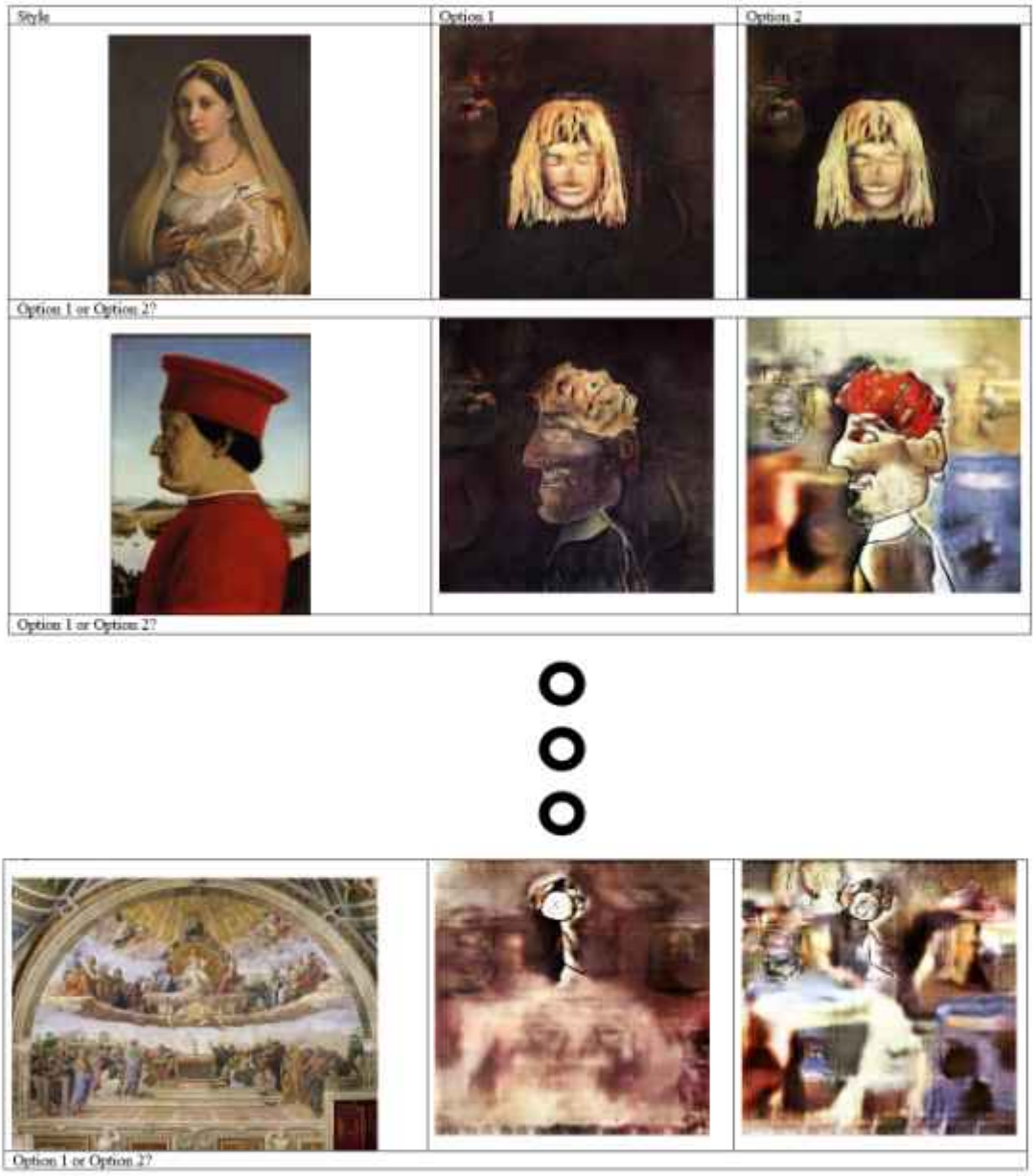


Figure A.4: Diagram of a sample of questions provided to participants in the second part of the user study. Participants were provided with the style image and first and final images produced from parts of the first part of the study. They chose between these two options as to what most closely matched the style of the style input.

Appendix B

Raw Results Output

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2	Question	User 2 - 1	User 2 - 2	User 2 - 3	User 2 - 4	User 2 - 5	User 2 - 6	User 2 - 7	User 2 - 8	User 2 - 9	User 2 - 10	User 2 - 11		Average
3	1	2	2	1	1	1	1	1	1	2	1	1		1.272727
4	2	2	2	2	2	2	2	2	2	2	2	2		2
5	3	2	1	1	2	1	1	2	1	2	2	1		1.454545
6	4	2	2	2	2	2	2	2	2	2	2	2		2
7	5	1	2	2	2	2	2	2	2	2	2	2		1.909091
8	6	2	2	2	2	2	1	2	2	2	2	2		1.909091
9	7	1	2	1	2	1	1	2	1	1	2	1		1.363636
10	8	2	1	1	1	2	2	2	1	2	1	2		1.545455
11	9	1	2	1	2	1	1	1	1	2	1	1		1.272727
12	10	1	2	2	1	1	2	2	2	1	2	2		1.636364
13	no of 2's	7	8	5	7	5	4	8	5	8	7	6		
14														
15		Average overall:		1.636364		Average without 1 or 9:		1.72727		Average no of two:		6.363636		

Table B.1: Raw results of part 2 of the user study.

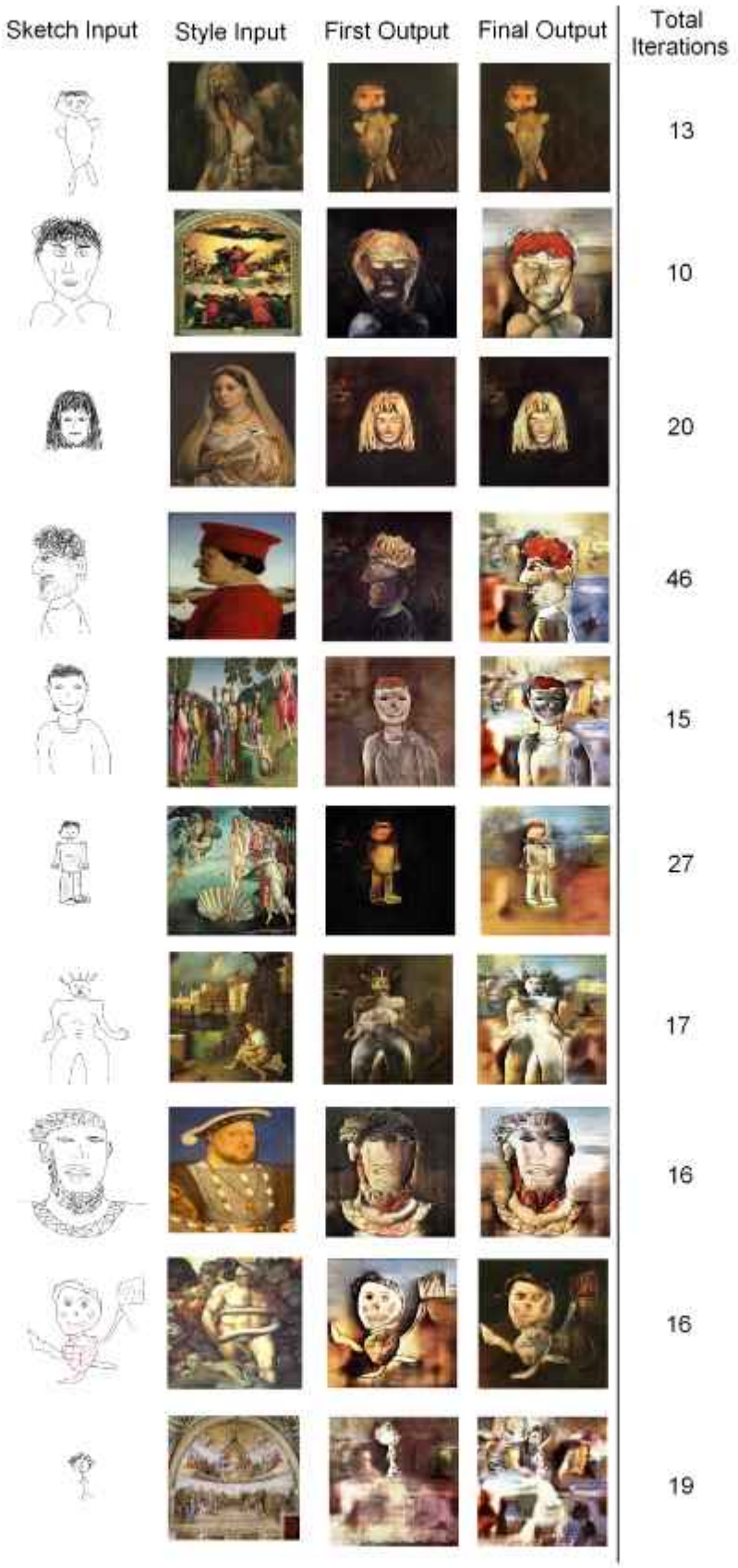


Figure B.1: First and final outputs of part 1 of the user study.