

MSC INDIVIDUAL PROJECT

UNIVERSITY OF THE ARTS LONDON

CREATIVE COMPUTING INSTITUTE

Old Sights, New Visions

Controlled Uses of Diffusion Based Image-to-Image
Translation for generative video

Author:

Adam Cole

Supervisor:

Prof. Mick Grierson

November 21, 2022

Abstract

This study investigates controlled uses of diffusion-based image-to-image translation for generative video within the context of an arts practice. Image translation has proven to be a powerful tool in creating video art, but achieving both temporally coherent and visually diverse results remains a challenge with little formal research into the subject. Through a sequence of targeted studies of increasing complexity, this paper outlines the control techniques that were necessary to achieve quality results for images, animation, and video. I used my findings in the creation of two original art pieces, *Kiss/Crash* and *Crash Me, Gently*, which use these techniques to reflect on the nature of images in the age of AI.

Acknowledgements

It has been an absolute privilege to spend the last year at the Creative Computing Institute among such a welcoming group of tutors and students. I'd like to thank our course leader, Phoenix Perry, for showing the deep relationship between art and play. I'd like to thank my professors Rebecca Fiebrink and Terence Broad, whose thoughtful introduction to the world of creative AI gave me the technical and theoretical foundation for the work in this project. And finally, I'd like to thank my advisor Mick Grierson whose sound advice helped guide this project at every turn. Hanging out in your office to talk freely about art, culture, technology, or whatever else happened to be on my mind that week has been a highlight of my academic career.

Contents

1	Introduction	7
1.1	Motivation	7
1.2	Research Goals	8
2	Background	9
2.1	Technical Context	9
2.1.1	Diffusion Models	9
2.1.2	Generative Process	9
	Forward Diffusion	9
	Reverse Diffusion	10
2.1.3	Text Guidance	11
2.1.4	Image Translation	11
2.2	Brief History of Diffusion Models in Practice	12
2.2.1	The Big Corporate Players	12
2.2.2	Rise of the Open Source Community	12
2.2.3	New Tools, New Culture	13
2.3	Image to Image Translation	13
2.3.1	Using Diffusion Models	13
	Experimental Video Art	13
	Cinema and SFX	14
	3D Worlds	14
2.3.2	Beyond Diffusion Models	14
3	Methodology	17
4	Results	18
4.1	Tools	18
4.2	Basic Image Translation	18
4.2.1	Control Techniques	18
	Standard Text-to-Image Settings	18
	Image Strength	19
4.2.2	Results	21
4.3	Basic Animation Translation	22
4.3.1	Control Techniques	22
	Frame Blending	22
	Consistent Seed	24
	Keyframe Extraction and Frame Interpolation	24
4.3.2	Results	24
4.4	Complex Animation Translation	26

4.4.1	Control Techniques	26
	Frame Blending Modulation	26
	Seed Selection	26
4.4.2	Results	27
4.5	Organic Animation Translation	29
4.5.1	Control Techniques	29
	Video Masking	29
	Color Profile Coherence	30
4.5.2	Results	30
4.6	Camera Footage Translation	32
4.6.1	Control Techniques	32
	Strength Scheduling	32
	Noise Blending	33
	DreamBooth	34
	Stable WarpFusion	34
4.6.2	Results	35
5	Final Works	37
5.1	<i>Kiss/Crash</i>	37
5.2	<i>Crash Me, Gently</i>	37
6	Evaluation	39
6.1	Technical Evaluation	39
6.1.1	Still Image	39
6.1.2	Animation Studies	39
6.1.3	Final Works	40
6.2	Aesthetic Evaluation	41
6.2.1	<i>The Last Supper, 1981</i> – Image Study	41
6.2.2	Animation Studies	42
6.2.3	Final Works	42
	<i>Kiss/Crash</i>	42
	<i>Crash Me, Gently</i>	43
7	Discussion	44
7.1	Technical Contribution	44
7.2	Aesthetic Contribution	44
7.3	Closing Remarks	45
8	Conclusion	46
A	Links to Video Materials	47

List of Figures

1.1	From left to right, an example of Dada [17], Pop [60], and Postmodern [26] art which translates existing imagery into self-reflective, critical art.	8
2.1	Forward diffusion process [36]	10
2.2	Reverse diffusion process [36]	10
2.3	Example of text prompted results using classifier-free guidance. [35] .	11
2.4	Example of image-to-image translation with the prompt, "A fantasy landscape, trending on artstation". The Left is the original, and the right is the result. [44]	12
2.5	Examples of images generated with diffusion tools. From left to right: DALLE-2 [37], Imagen [1], Midjourney [2]	12
2.6	Examples of image-to-image translation using warp fusion. Left is Temporary by Roope Ranisto [41], right is Blue Crystal Fire by Aiplague [3]	14
2.7	Examples of image translation used for narrative film. Left is a horror film [28], and right is a Spider-Man-inspired action film [10].	14
2.8	Frame from real-time VR experience applying Stable Diffusion translation to surrounding living room [54]	15
2.9	Still frame from <i>Blade Runner—Autoencoded</i> , generated by passing an original frame from the film through the neural network [8]	15
2.10	Still frame from <i>Learning To See</i> . Left is the live camera feed and right is the translated output of the neural network	16
4.1	Input image \rightarrow noised input image \rightarrow diffusion output after denoising	19
4.2	The Last Supper, 1981	21
4.3	Input video for basic animation study	22
4.4	Visualization of image processing steps for frame blending. Note that the visual difference between the previously generated frame and outputted frame is purposefully small to create smooth results when played back.	23
4.5	Grid of video results from the basic animation study	25
4.6	Input video for complex animation study	26
4.7	Example of seed images considered for the complex animation results.	27
4.8	Grid of video results from the complex animation study	28
4.9	Input video of squash and stretched sphere for organic animation study	29
4.10	The result when the video mask was very strong creates very strong contours.	29
4.11	Result when video mask was less strong allowed for more image diversity.	30

4.12	Effect of color coherence on results. The top sample did not use color coherence and the colors became increasingly saturated. The bottom sample did use color coherence resulting in cleaner colors and a more consistent feel from beginning to end.	30
4.13	Grid of video results from the organic animation study.	31
4.14	<i>Crash Me, Gently</i> input frames	32
4.15	<i>Kiss/Crash</i> input frames	32
4.16	Example of strength scheduling. The result looks more like input at the start (left) and increasingly matches the prompt by the end (right).	33
4.17	Noise blending increasing from top 0 (top row) to 1 (bottom row). Less noise blending results in less diverse images over time. More noise blending results in more variation over time.	33
4.18	Training samples of iconic Hollywood kisses used to fine-tune the diffusion model used for these studies.	34
4.19	Images created with fine-tuned DreamBooth model	34
4.20	Strength scheduling examples used in <i>Crash Me, Gently</i>	36
6.1	Detail from the organic animation study showing a side-by-side comparison of the input and its translation. Formal features like color and composition clearly carried over, but the result is still stylistically unique from the original.	40
6.2	Left: the Mona Lisa by Leonardo da Vinci, Right: L.H.O.O.Q., Marcel Duchamp's transgressive "translation: of the original [15]	41
6.3	Left: the original Last Supper by Leonardo da Vinci, Right: <i>The Last Supper, 1981</i> , transgressive "translation: of the original	41

List of Tables

4.1	Impact of image strength on the output of image-to-image diffusion	20
-----	--	----

Chapter 1

Introduction

The morality of art consists in the perfect use of an imperfect medium.

— Oscar Wilde [61]

1.1 Motivation

In a world already drowning in images, the oncoming tidal wave of AI generated “art” is reason for pause if not outright concern. This is not just because the technology has the potential to upend the creative industries and impact existing artists and spectators in unpredictable ways [27]; But because images are incredibly powerful: their ability to entertain, delight, and inspire is matched by their potential to misinform, control, and incite [30].

The debate on the benefits and dangers of images goes back thousands of years, most notably to Plato’s allegory of the cave which connects the idea of images to misinformation and incarceration [39]. However, it’s in the 20th century when the production, distribution and consumption of images became deeply tied to technological advancements like photography, mass media, and instant visual communications that a renewed discourse on images became more vital [30].

This discourse can easily be extended into the 21st century, especially in relation to AI generated images. For example, it is not a difficult mental leap to imagine how an image generating machine fits neatly within the logic of late capitalism as outlined by Fredrich Jameson [23], exaggerates the experience of hyperreality introduced by Jean Baudrillard [6], or amounts to another evolution of spectacle as discussed by Guy Debord [12]. But perhaps the most nuanced and interesting point of view for me personally is Susan Sontag’s *On Photography* where she describes our modern relationship to images as follows:

A society becomes modern when one of its chief activities is producing and consuming images, when images that have extraordinary powers to determine our demands upon reality, and are themselves coveted substitutes for firsthand experience, become indispensable to the health of the economy, the stability of the polity, and the pursuit of private happiness. [56]

If we are to accept that images have the extraordinary powers to determine our demands upon reality, how will a technology which can generate endless images in a fraction of a second impact the health of our economy, the stability of our polity,

and our pursuit of private happiness? And, how then, are artists expected to use this technology responsibly?

While there is no right answer, we'd benefit greatly by recognizing that this technology is revolutionary in its capabilities, but familiar in the questions it raises about the nature of images, representation and authenticity. As such, we can look to past art movements which responded to these same questions for strategies in confronting the core issue. Relevant examples include Dada collage, Pop Art appropriation, and Postmodern pastiche. In all these cases, artists used the prevalent image-making technology of their time and repurposed it in some way which both spoke to societal issues and revealed something unspoken about the way the technology worked.

In the context of diffusion based generative art, I'm particularly interested in the ability of these models to apply image-to-image translation. Like the work of Marcel Duchamp and Barbra Kreugar, this allows artists to use existing images as material in the art-making process. I wonder, then, how can we augment, invert, and negate existing images in composition and meaning using image-to-image translation in the same spirit of avant-garde artists from the past 100 years.



Figure 1.1: From left to right, an example of Dada [17], Pop [60], and Postmodern [26] art which translates existing imagery into self-reflective, critical art.

1.2 Research Goals

Using these artists as inspiration, my research question is how the diffusion based image-to-image process can be twisted on itself to reveal something true about AI generated imagery and our digital culture in general. Additionally, how can this technique be used to address the artistic themes of particular interest to me, specifically the way intimacy is mediated through technology and interrogating the gap between the real experience of love and its artificial representations.

My goal is that through my research, I will develop the necessary control techniques to generate quality image translations and ultimately use these skills in the creation of original audio-visual compositions that speak to the nature of our image world in the age of AI.

Chapter 2

Background

In this section, I will give the technical overview necessary to understand how diffusion based image-to-image translation works. I will then give a brief history of the tools which fueled the proliferation of text-to-image content and the open-source community which has driven much of the experimental innovations in this space, especially for image-to-image content. Finally, I will expand our focus to look at AI artworks that don't use diffusion models. I will show how these artists used their research into different image translation techniques to create poetic art pieces and how that serves as a model for my own research methodology.

2.1 Technical Context

2.1.1 Diffusion Models

Diffusion-based models are a class of generative AI models which recently gained popularity due to their impressive ability to create diverse, high-quality content in many domains. For example, they are being used to generate high-resolution images [44], audio [38], video [20] and 3D assets [40] as well as targeted editing tools like in-painting [29], super-resolution [51] and colorization [49]. Inspired by research into thermodynamics [55], diffusion models were greatly improved for the task of image generation [18] eventually beating out GANs to achieve state-of-the-art results [14]. They continue to be a major research focus for many generative media types with improvements in speed of sampling [34, 25, 52], quality of results [21], and novel applications like molecule design [63].

2.1.2 Generative Process

The diffusion process as defined in [18] works by incrementally destroying the data through small additions of Gaussian noise until the sample becomes pure noise itself and then training a model to learn how to reverse that noising process. The result is a model able to progressively move from complete noise toward a clean sample representative of the dataset. For images, this works in two key steps outlined below.

Forward Diffusion

The forward diffusion process is defined as a Markov chain of T steps where at each step a small amount of Gaussian noise is added to the sample until the sample

becomes pure noise. The corollary of the Markov chain is that each step depends only on the one proceeding it.

For a given sample \mathbf{x}_0 in a real data distribution, $q(\mathbf{x})$ ($\mathbf{x}_0 \sim q(x)$), the forward diffusion process of T steps can be formulated as follows:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

where β_1, \dots, β_T is a variance schedule which ensures that \mathbf{x}_T is a nearly isotropic Gaussian for sufficiently large T .

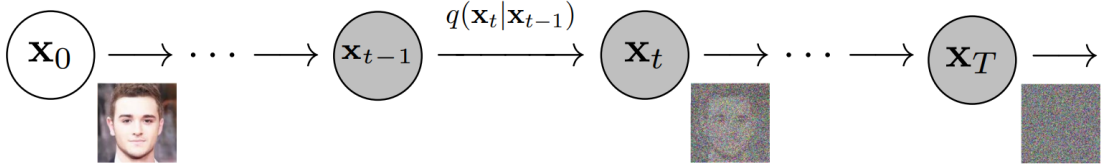


Figure 2.1: Forward diffusion process [36]

Reverse Diffusion

The reverse diffusion process moves in the opposite direction: from pure noise to a sample representative of the data distribution. The process begins with a pure Gaussian sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We can then approximate the steps $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ with a parameterized model p_θ (i.e. a neural network) as follows:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

where the mean parameterized by $\boldsymbol{\mu}_\theta$ and the variance parameterized by $\boldsymbol{\Sigma}_\theta$ are conditioned on noise level t .

We can reformulate stepping through this trajectory from pure noise at \mathbf{x}_T to a sample \mathbf{x}_0 from the distribution as:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$

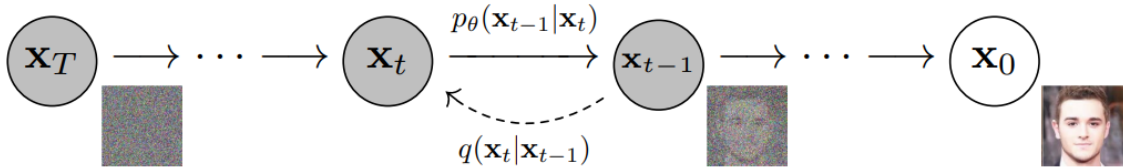


Figure 2.2: Reverse diffusion process [36]

The details of training such a model are beyond the scope of this paper but can be further investigated in [18].

2.1.3 Text Guidance

In addition to generating high-quality samples from pure noise, we can also guide the content of the sample with a text prompt. This was initially implemented using classifier guidance, where a classifier model was trained alongside the diffusion model for the purpose of text guidance [14]. Since then, this technique was improved by classifier-free guidance where a single generative model (trained jointly on a conditional and unconditional diffusion model) is able to handle text guidance without a secondary classifier model both simplifying training and improving the quality of results [19, 35]. With a text prompt, we are able to control the scale of the classifier-free guidance where higher values adhere more closely to the prompt at the cost of sample diversity.



Figure 2.3: Example of text prompted results using classifier-free guidance. [35]

2.1.4 Image Translation

One more method we have to guide the output of a text-to-image diffusion model is via image translation. There are many implementations and applications of image translation [49, 53, 62], but within the context of this paper we will focus on the one introduced in [32] and further developed by [44].

With this image-to-image translation method, the key technical detail to remember is that in the Markov chain of diffusion steps, each step is only dependent on the one preceding it. As a result, we can think of image translation as starting somewhere in the middle of the chain as opposed to starting from pure noise. Instead, we start with a partially noised version of the input image by applying the forward diffusion process on the input image for a given number of steps. We then run the reverse diffusion process from that point for the same number of steps resulting in a new image from the sample distribution. The similarity of the generated image to the input depends on the amount of noise applied at the start of the process and the corresponding number of reverse diffusion steps. At each reverse diffusion step, classifier-free guidance is able to push the result toward the text prompt. This is the key mechanism we will use for controlled image-to-image translation.



Figure 2.4: Example of image-to-image translation with the prompt, "A fantasy landscape, trending on artstation". The Left is the original, and the right is the result. [44]

2.2 Brief History of Diffusion Models in Practice

2.2.1 The Big Corporate Players

Recently a suite of creative tools based on advances in diffusion model research has emerged. The first to make a cultural splash was the release of DALLÉ-2 by OpenAI in April 2022 [37] which publicized impressive results with their text-to-image diffusion model [62]. This model was trained on a massive dataset of more than 500 million captioned images scraped from the internet [62]. DALLÉ-2 began as a closed beta which offered an easy UI to try text-to-image generation. Soon after Google announced Imagen, their iteration of a text-to-image diffusion model which benefited from further integrating large transformer language models into the generative process [50]. At the same time, smaller companies like Midjourney released their own versions of text-to-image models and interfaces to compete with OpenAI [47].



Figure 2.5: Examples of images generated with diffusion tools. From left to right: DALLÉ-2 [37], Imagen [1], Midjourney [2]

2.2.2 Rise of the Open Source Community

While these tools were either invite-only or behind a paywall, a collection of open-source tools were developed, released, and continually improved upon by a community of artists and developers. An especially influential one was Disco Diffusion [11], a CLIP-guided diffusion model and corresponding notebook created by Katherine

Crowson which later collected additional features from open-source contributions. Self-described as “a frankensteinian amalgamation of notebooks, models and techniques for the generation of AI Art and Animations”, Disco Diffusion introduced several experimental features like generative 2D animation and video translation [5]. The community often shared their experiments like [11] via Python notebooks on Google Colab, a site that allows users to easily run GPU-based machine learning code in the cloud.

In August 2022, Stability AI publicly released Stable Diffusion, an open-sourced diffusion model intended to compete in quality and performance with OpenAI and Google’s models [33]. Similar to the community around Disco Diffusion, a collection of tools and interfaces quickly arose around the release. For example, the collective Deforum released and actively maintain their Stable Diffusion notebook which adopted many of the features from Disco Diffusion including image and video translation [13].

2.2.3 New Tools, New Culture

The ease of use of these tools from DALLÉ to Stable Diffusion has led to an explosion of AI-generated content [31]. The impact of these images has begun to enter mainstream consciousness as they proliferate across social media [31], cause controversy in traditional art settings [7, 43, 46], and get picked up in major media outlets [9, 45, 59].

2.3 Image to Image Translation

2.3.1 Using Diffusion Models

Existing examples of image-to-image translation using diffusion models have come primarily from an online community of engineers and visual artists who often share their discoveries through Twitter and Discord. Their work has used open-sourced tools like Stable Diffusion to explore new aesthetics, techniques, and applications.

Experimental Video Art

One of the most technically advanced contributions has been WarpFusion by Alex Spirin [57] who has been exploring solutions for the problem of consistency in video translation. Visual artists such as Roope Ranisto [41] and AIPlague are using WarpFusion for experimental video art and music video production [3].

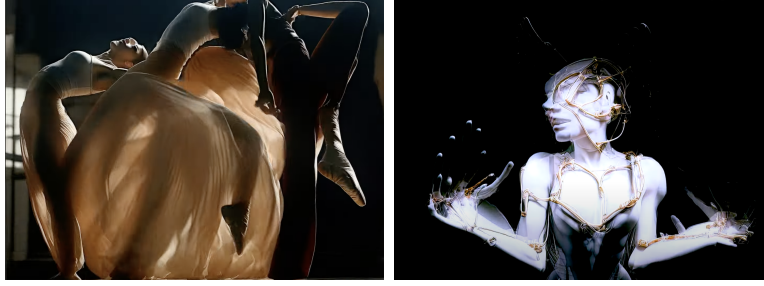


Figure 2.6: Examples of image-to-image translation using warp fusion. Left is Temporary by Roope Ranisto [41], right is Blue Crystal Fire by Aiplague [3]

Cinema and SFX

Some artists with a background in cinema and visual effects are using image-to-image translation in combination with a style transfer tool known as EBSynth [24] as part of their special effects toolkit. A couple of examples are Scott Lightiser [28] and Corridor Digital [10] who are using image translation to create original short video narrative content. In both these cases, the artists are bringing together multiple AI tools and techniques into a larger VFX production ecosystem.

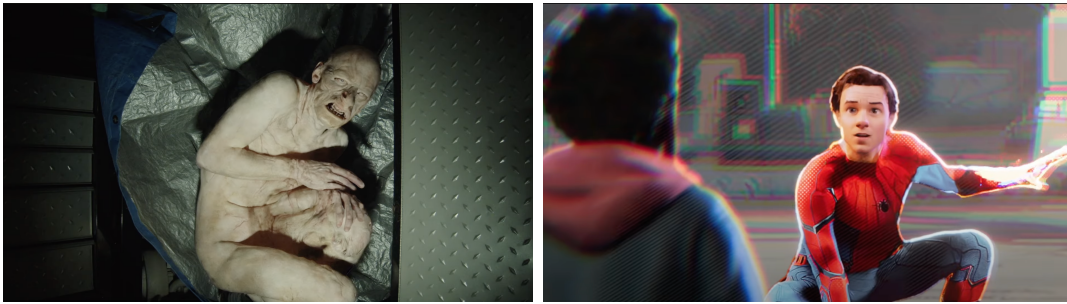


Figure 2.7: Examples of image translation used for narrative film. Left is a horror film [28], and right is a Spider-Man-inspired action film [10].

3D Worlds

Finally, image-to-image translation is being used for the creation of 3D worlds in both AR and VR. For example, @ScottieFoxTV [54] is creating real-time AR/VR experiences by piping in images from his surroundings, translating them through Stable Diffusion, and stitching them back together in VR as a 360 environment.

2.3.2 Beyond Diffusion Models

Using image-to-image translation to control the outputs of generative models has been explored in the sphere of AI art and generative machine learning long before diffusion models. While this has often been done in the sphere of technical research such as style transfer [16], pix2pix [22], and cycleGAN [64], there are some examples of artists specifically exploring this technique worth noting.

In AutoEncoding Blade Runner [8], the authors train a Variational Auto Encoder on the film Blade Runner. After training the model, each individual frame of the



Figure 2.8: Frame from real-time VR experience applying Stable Diffusion translation to surrounding living room [54]

film was then passed back through the network, which reconstructed each image “from memory” resulting in a totally reimagined film with a dreamy relationship to the original. The authors explain that the resulting artifact “seeks to emphasize the ambiguous boundary” between the original and its “replicant”. It also exposes the imperfect way these generative systems encode and represent “reality” and makes a metaphor for the way human cognition may operate in a similarly flawed manner.



Figure 2.9: Still frame from *Blade Runner—Autoencoded*, generated by passing an original frame from the film through the neural network [8]

Another example is Learning to See [4], an interactive installation where a camera feed of household objects like wires or keys is translated in real-time into imagery of a specific domain like clouds, waves, or stars. Changes to the arrangement of real objects have a clear relationship to the composition of the outputted video. The authors explain that this project exposes the learned bias of the underlying networks and how that reflects our own self-affirming cognitive biases: “[when] the trained network looks out into the world via the camera, it can only see what it already knows.” Similar to AutoEncoding Blade Runner, this project exposes the internal mechanism of how these images are constructed while also making a powerful statement on the way humans construct meaning.

These two projects are by no means an exhaustive review of critical art using AI image translation. Rather, these projects provide a good blueprint for merging technical research with the creation of a visual art that reflects on the very nature of the images generated. I will use a similar method in my research into diffusion models and their artistic potential.

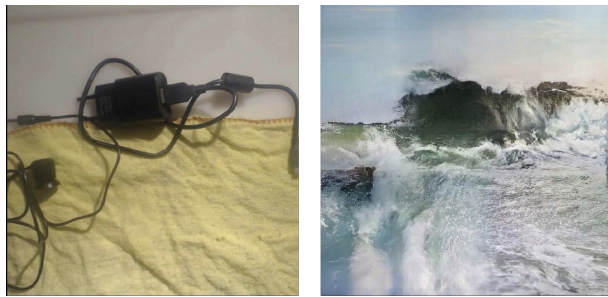


Figure 2.10: Still frame from *Learning To See*. Left is the live camera feed and right is the translated output of the neural network

Chapter 3

Methodology

To investigate controlled uses of image-to-image translation using diffusion models, I will use a practice-based auto-ethnographic study using a sequence of targeted experiments. I'll start with the most basic example, the translation of a single image, and incrementally add complexity with each study by moving through various animation styles and finally recorded video. At each step, I'll explore what techniques of control are necessary to generate a quality translation.

As a self-reflective study, the definitions of “control” and “quality translation” are subjective, but I will outline them as follows:

Control: Control in relation to image translation is the levers or techniques we can use to navigate the spectrum of possible outputs for a given input. For individual images or frames, this will be expressed as adding controls to balance the space of possibilities between a totally random result and returning the original image. For video, this will also apply to controlling visual coherence across time.

Quality Translation: For a successful image translation, the result should have a compositional or thematic relationship to the original. This can include a relationship in form, color, texture, or content. For video, this can also include capturing the sense of motion from the original input. Additionally, a quality translation should also have aesthetic merit, both formally and conceptually.

The result at each stage will be a sequence of images or video clips that can be evaluated in terms of control and quality. In the end, I will employ these experimental techniques in the creation of two audio-visual compositions *Kiss/Crash* and *Crash Me, Gently* which relate to my themes of interest.

Chapter 4

Results

In the following section, I will outline the various studies I completed over the course of my research. For each study, I will detail the techniques which were necessary to develop a quality translation and then share a sample of the results.

4.1 Tools

To create the works below, I relied heavily on the open-source tools described in the introduction, specifically notebooks for Disco Diffusion [11] and Stable Diffusion [13]. It should be noted there is not a one-to-one relationship between models. For example, using a specific selection of settings with one model will lead to a totally different result in another. However, the *instincts* around how changing these settings influence the output are consistent across the models I experimented with.

4.2 Basic Image Translation

The first study was the translation of a single image: specifically an image of the Last Supper by Leonardo Da Vinci. The goal was to translate the original from a somber catholic image into a celebratory queer one thereby inverting the meaning and composition of the original.

Starting my research with a basic still image allowed me to experiment with the simplest levers available when working with image-to-image translation. These include the seed, prompt, total diffusion steps, sample algorithm, guidance scale, and most importantly the image strength.

4.2.1 Control Techniques

Standard Text-to-Image Settings

The following are the most basic settings for controlling many text-to-image frameworks. While these are all incredibly influential, they are not specifically related to the task of image translation, and delving into the details of these techniques is beyond the scope of this paper. Instead, I will provide a brief description of each.

Prompt: This essential component of any text-to-image process defines the content and style of the desired image generation. Small changes to the prompt will significantly impact the resulting image.

Classifier free guidance scale: The degree to which the generated image should be guided toward the text prompt. Higher values lead to a stronger representation of the text prompt, but too high a value will lead to over-saturated or otherwise deformed results.

Total steps: The total number of diffusion steps to generate the image. Higher values may lead to images with more fine-grain detail, but returns diminish after a certain threshold.

Sampler algorithm: The algorithm used when sampling the model from noise to image. Different samplers can lead to different end results.

Seed: The seed of the randomly generated Gaussian noise is used to begin the diffusion process. Changing the seed will change the overall image composition of the output.

Image Strength

Image strength is one of the most effective controls we have in determining the amount the output image will resemble the input image. Image strength can be thought of as a value between zero to one and works by controlling where we begin the diffusion process. For example, if the total number of steps is 100 and the image strength is 70%, the number of diffusion steps will be 30. Our starting image will be the original image with 30 steps of Gaussian noise added, and we then denoise 30 steps from there resulting in a different, but related image to the original.

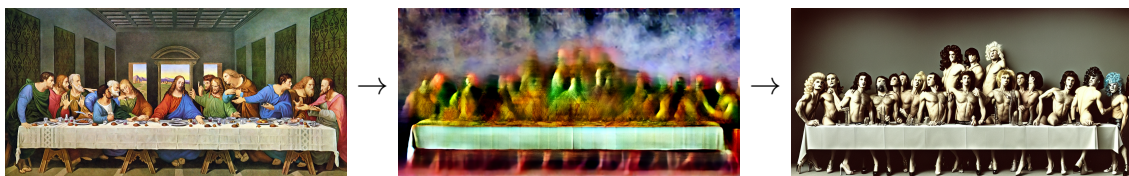


Figure 4.1: Input image \rightarrow noised input image \rightarrow diffusion output after denoising

A higher image strength results in images more similar to the input, while a lower input strength results in less similar results. Table 4.1 outlines the effect of this setting in more detail.





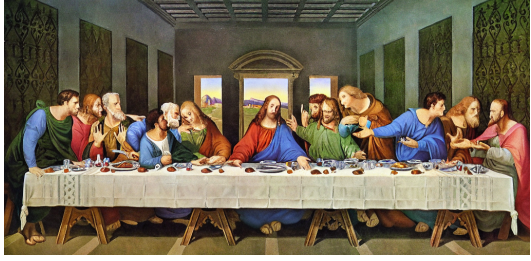
Image strength	Effect on output	Image
Zero	Output unrelated to input, totally prompt driven.	
Low	Large compositional and thematic elements remain in the final output, but the style and content may change drastically compared to the input.	
Medium	A balance between low and high outcomes.	
High	Small details, textures, and features of the image may be affected, but will greatly resemble the input.	
One	Matches the input exactly.	

Table 4.1: Impact of image strength on the output of image-to-image diffusion

4.2.2 Results

Gaining an intuition for these basic settings was essential before moving on to more complex inputs. While each lever has a clear impact on the control of the output, achieving a quality translation required small iterative changes to the prompt, strength, seed, etc. until a result of sufficient quality was achieved. This process is the starting point for any successful image translation. Below is the final piece created at the end of that process titled *The Last Supper, 1981*.



Figure 4.2: The Last Supper, 1981

4.3 Basic Animation Translation

The second study involved the translation of a basic moving image. For this study, I chose the most simple animation example: a white sphere moving up and down on a black background. Video translation works by breaking the initial video down into individual frames, running each frame through the image-to-image translation pipeline, and then rejoining the translated images into a new video. The biggest challenge here compared to the single image-to-image translation from the first example is balancing a sense of temporal coherence across the video without sacrificing image diversity as the sequence progresses. I achieved this using a host of techniques broken down below.

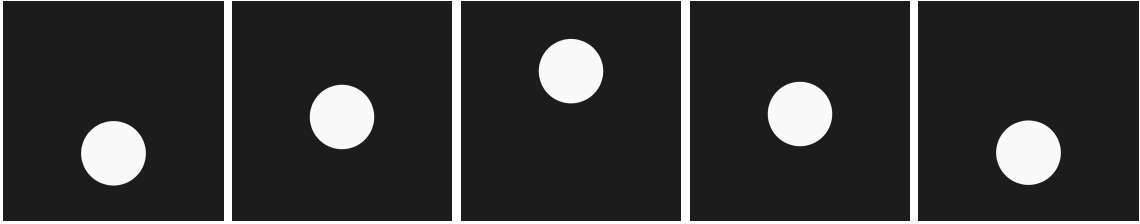


Figure 4.3: Input video for basic animation study

4.3.1 Control Techniques

Frame Blending

Every diffusion step produces a new image, and we must look for tricks to relate the new image not just to the input frame but to the frame generated before it. As a result, we can consider the major inputs for every frame generated (besides the first) as:

1. The input frame from the video.
2. The previous video frame generated.

The simplest strategy which was sufficiently effective for this study was to overlay the previous frame on top of the input frame and pass that modified blended image through the diffusion process. However, if the image strength is not strong enough, the composition from the modified image won't sufficiently transfer to the generated output and there will be no gain in terms of temporal coherence. To account for this, we can break the diffusion process into two key steps:

1. For the first input frame, run the diffusion process at a low image strength meaning the result will resemble the prompt more strongly than the input image.
2. For each subsequent frame, run the diffusion process on the modified image (aka the input image blended with the previous frame) at a low image strength. This allows elements from the first generated image which most resembles the text prompt to be carried through the entire generated video, while also capturing the compositional details from the input video.

This technique was based on the way Disco Diffusion handles 2D animations [11].

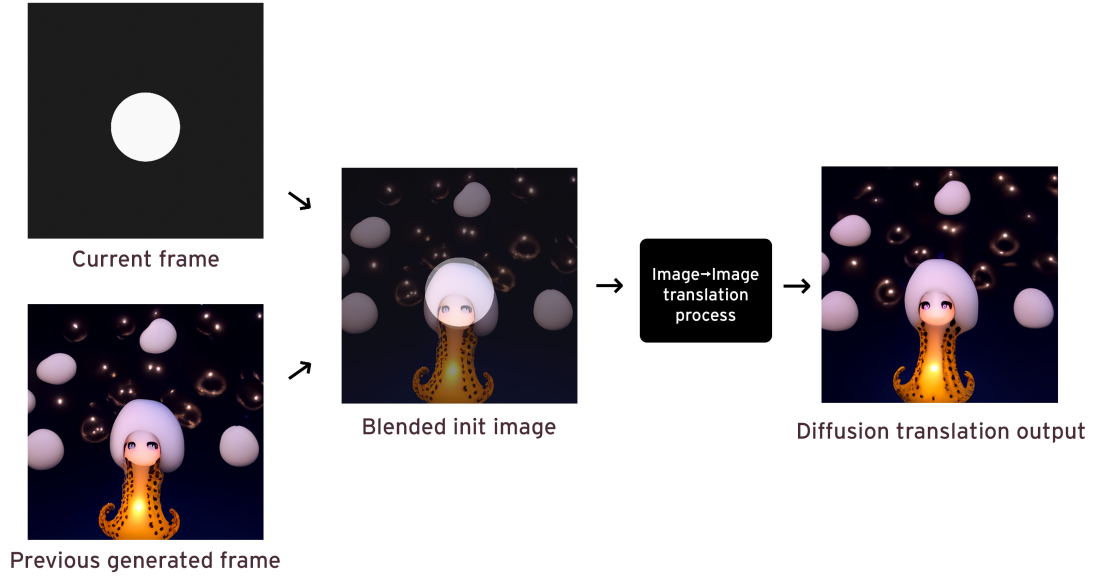


Figure 4.4: Visualization of image processing steps for frame blending. Note that the visual difference between the previously generated frame and outputted frame is purposefully small to create smooth results when played back.

Algorithm 1 Frame blending pseudocode

```

DiffusionArgs  $\leftarrow$  standard diffusion params like prompt, sampler, etc.

for frameNum  $\leftarrow$  0 to TotalFrameCount do
  inputFrame  $\leftarrow$  GETINPUTVIDEOFRAME(frameNum)

  if frameNum is 0 then
    imageStrength  $\leftarrow$  LowImageStrength
    initImage  $\leftarrow$  inputFrame
  else
    imageStrength  $\leftarrow$  HighImageStrength
    initImage  $\leftarrow$  inputFrame * BlendStrength + prevGeneratedFrame * (1 - BlendStrength)
  end if

  outputImage  $\leftarrow$  RUNDIFFUSIONTRANSLATION(initImage, imageStrength, DiffusionArgs)
  prevGeneratedFrame  $\leftarrow$  outputImage
end for

```

Consistent Seed

To assist temporal coherence across the animation, we want to keep the input conditions as consistent as possible. One important way we can do this is by maintaining a consistent seed during the generation of the animation. By keeping the underlying noise the same across the translation process, similar frames with only small variations have a greater likelihood of outputting related compositions. For smooth animations, this means consecutive translated frames are more likely to flow together neatly.

Keyframe Extraction and Frame Interpolation

When running image-to-image generation for a video, every frame runs the risk of derailing the animation with totally new, unrelated content. The above techniques outline some strategies to mitigate that risk, but another option is to simply use fewer frames. Instead, we can extract every second or fourth frame from the initial video, and run those frames through the diffusion process. We can then run the generated images through an interpolation software like FILM (Frame Interpolation for Large Motion) [42] resulting in a “smooth video” in spite of the fact that we only used half or a quarter of the input frames.

4.3.2 Results

Using the above techniques, I generated four variations of the input video documented below in 4.5. These results can also be seen as part of the *Animation Studies* video in Appendix A.

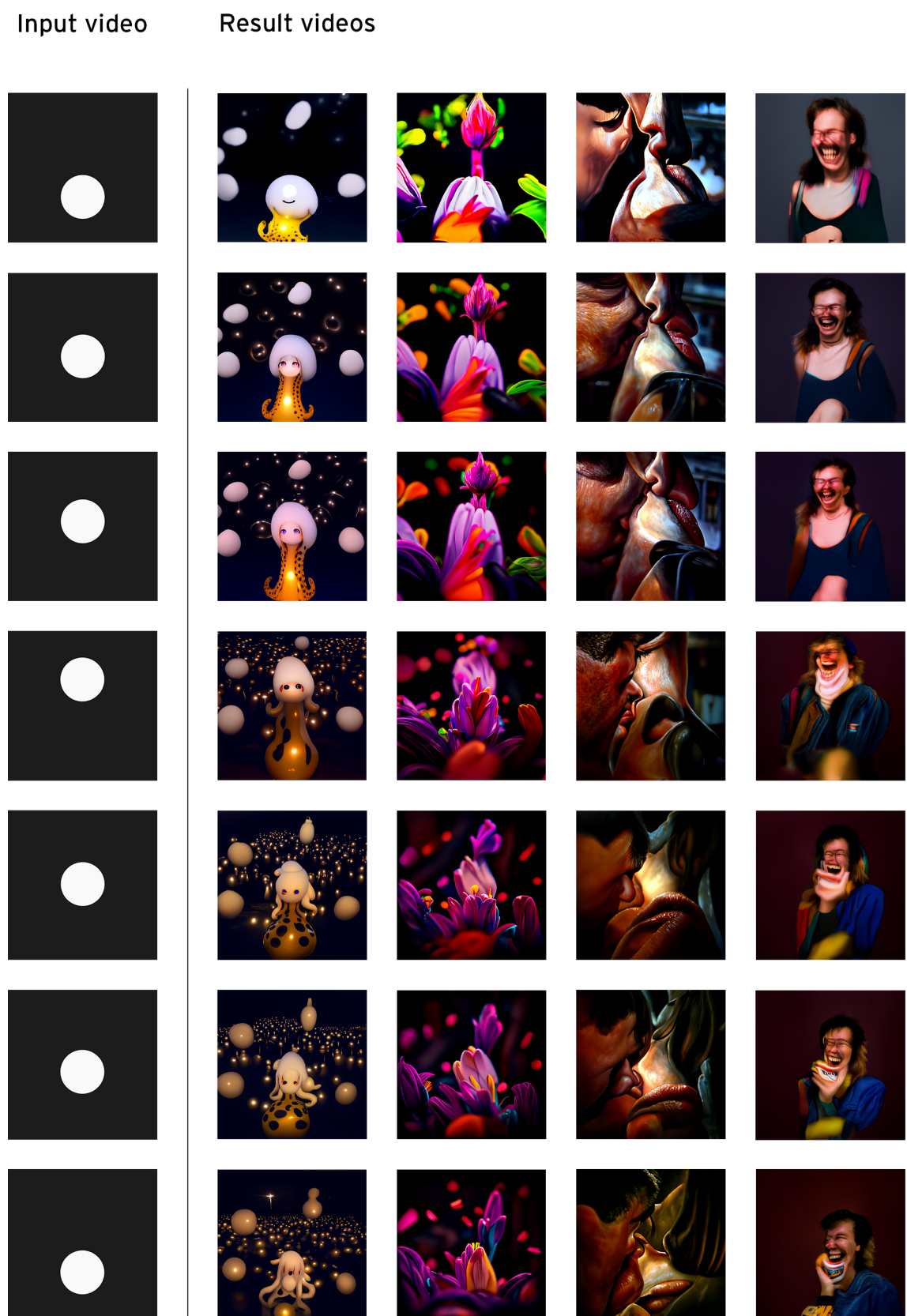


Figure 4.5: Grid of video results from the basic animation study

4.4 Complex Animation Translation

For the next study, I wanted to increase the degree of complexity of the source animation. Instead of a single moving shape, I used Processing to generate a geometric animation. Compared to the previous study, this one introduced multiple points of motion, color, texture, and composition. The addition of these features required an updated set of control techniques to achieve a quality translation.



Figure 4.6: Input video for complex animation study

4.4.1 Control Techniques

Frame Blending Modulation

One risk of the frame blending technique is that elements that were generated early in the video might get stuck in the blended overlay and consistently feature in the animation even if it no longer has a relationship to the input video.

One effective strategy here was to modulate the degree to which I blended the original input frame with the previous frame. This had the effect of clearing the history captured by the overlay at given intervals and kept the resulting frames tied to the progression of the animation. The risk here is that running the modulation at too short an interval would lead to choppy temporal coherence because the benefits of frame blending are lost too frequently. Finding the right interval was the result of trial and error. I also considered experimenting with a degree of randomness here since the rhythm of the interval can sometimes be seen in the resulting video.

Seed Selection

In the input animation, I start with an empty canvas and slowly build up the composition. However, in practice, I found running the diffusion process on the video in reverse, starting with the most complex frame and moving backward to a blank screen, to be a more effective strategy in terms of controlled results. Doing so has two benefits:

1. By starting with the final animation frame, we can spend some time exploring seeds that will determine the final composition of the completed animation. When played forward, the animation will build to the final frame we hand-picked.
2. I found the frame blending modulation technique to be more effective when removing content from the canvas as opposed to building up the final composition.



Figure 4.7: Example of seed images considered for the complex animation results.

4.4.2 Results

Three key results were generated from this starting video outlined in figure [4.8](#). These videos are also documented within the *Animation Studies* in Appendix [A](#).

Input video

Result videos

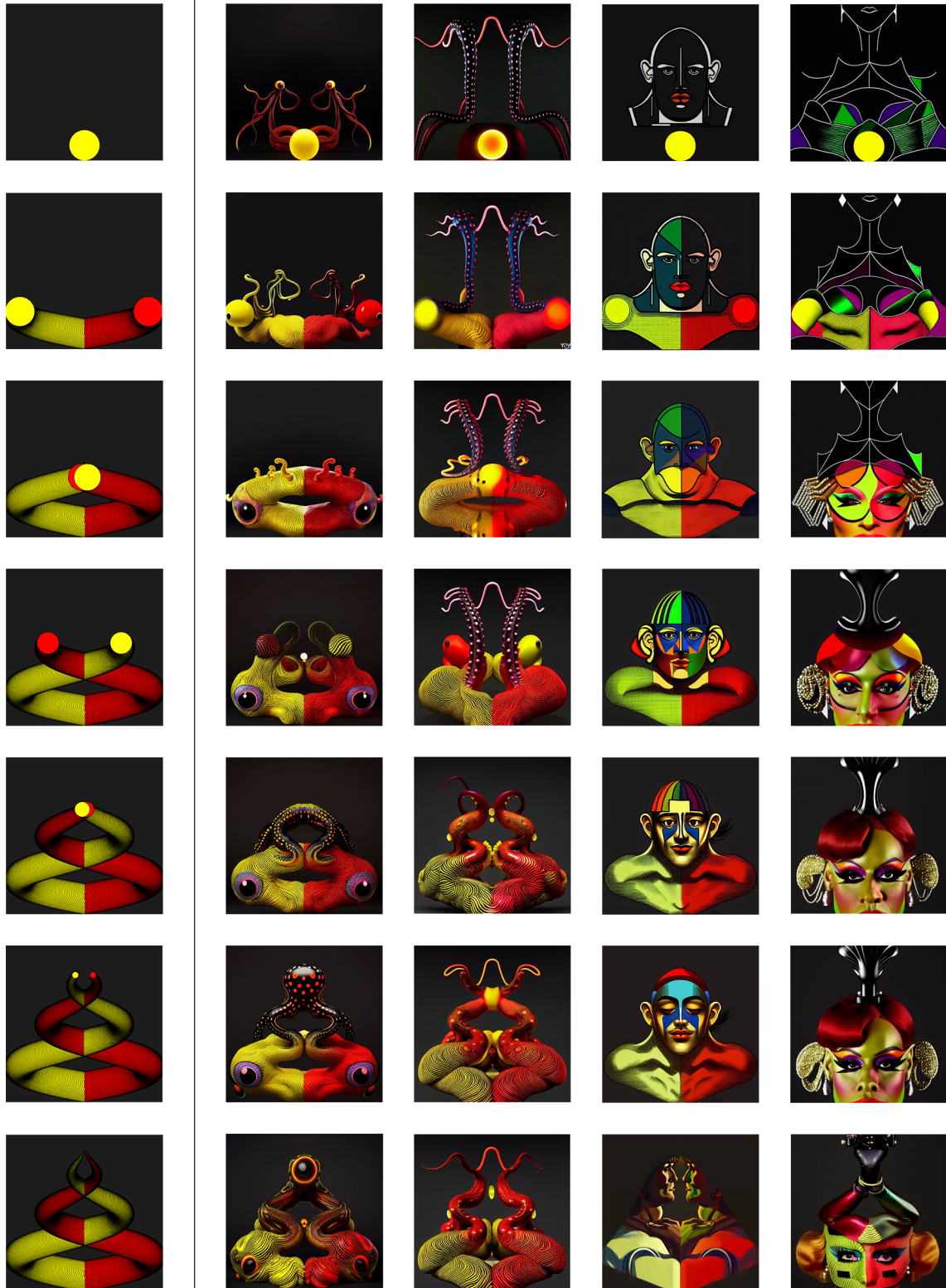


Figure 4.8: Grid of video results from the complex animation study

4.5 Organic Animation Translation

Inspired by Disney’s 12 principles of animation, my next study increased the complexity by working with an organic life-like squash and stretch animation. The difference here is that the underlying motion needs a stronger sense of consistency and conformity compared to the last two studies where a totally abstract result would be acceptable.

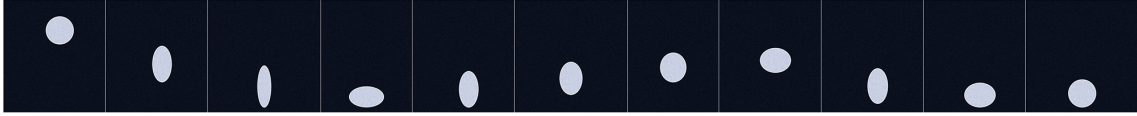


Figure 4.9: Input video of squash and stretched sphere for organic animation study

4.5.1 Control Techniques

Video Masking

One of the primary methods of maintaining the shape and timing of the input animation was through video masking. Masking in the context of diffusion models is the process of controlling where the diffusion is run on a given image->image translation process. A mask can be defined with a grayscale image where the brighter the pixel, the more diffusion steps will be run at that location. For video, we can create a video mask that conforms to the input video and allows us to reinforce the contours of the central action. This was essential for the squash and stretch animation where the organic fluctuations of the shape needed to be captured in the translation result.

When the mask values were very strong, I was able to generate fairly clear squash and stretch results.

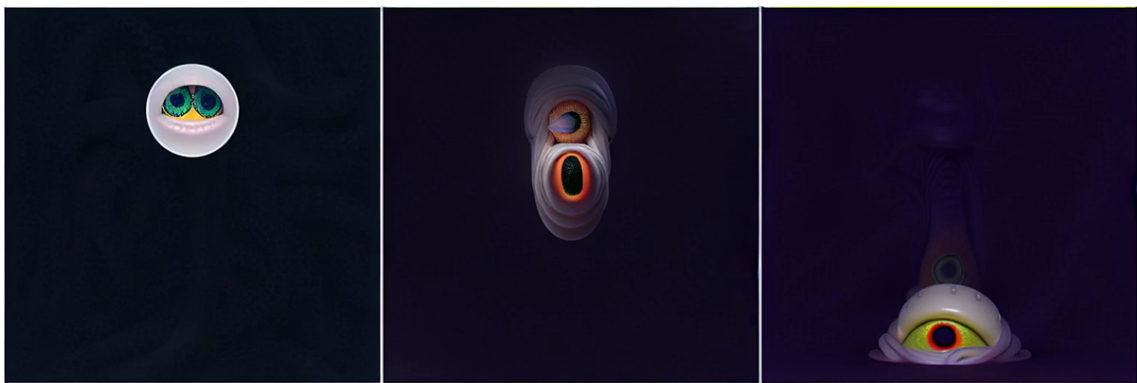


Figure 4.10: The result when the video mask was very strong creates very strong contours.

However, more interesting results were generated by lowering the strength of the mask meaning the translated image had more flexibility. This allowed for more interesting compositions which filled in the background while also capturing the organic sense of motion from the input video.

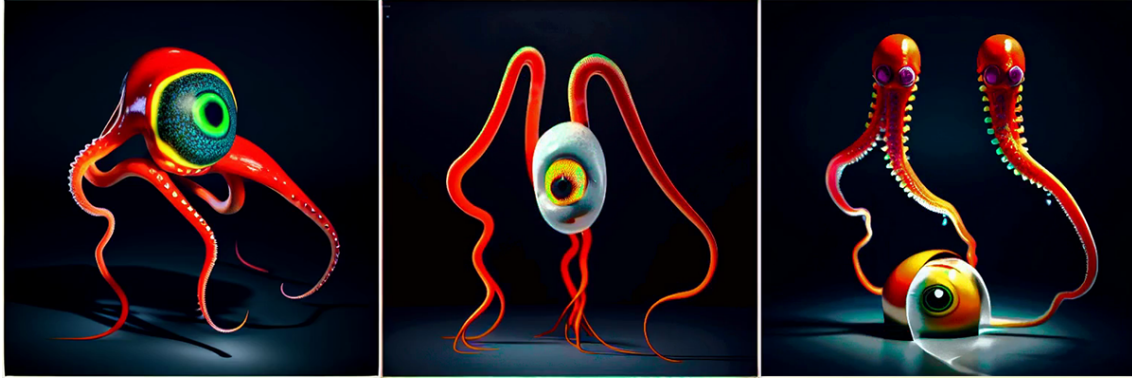


Figure 4.11: Result when video mask was less strong allowed for more image diversity.

Color Profile Coherence

Due to the nature of frame blending, it's possible for the colors of the resulting animation to become increasingly saturated as aspects of the previous frame are cycled back through the diffusion process. To keep the color profile of the translated video consistent, I took the color profile of the initial diffused frame and applied that profile to every diffused frame as a post-processing step before moving on to the next frame. This technique was inspired by the way color consistency is handled for 2D and 3D animation generation within the Deform notebook but repurposed for video translation [13].



Figure 4.12: Effect of color coherence on results. The top sample did not use color coherence and the colors became increasingly saturated. The bottom sample did use color coherence resulting in cleaner colors and a more consistent feel from beginning to end.

4.5.2 Results

Several video results were generated from this input animation, four of which are represented in figure 4.13. The videos are also available within the *Animations Studies* video link in Appendix A.

Input video

Result videos

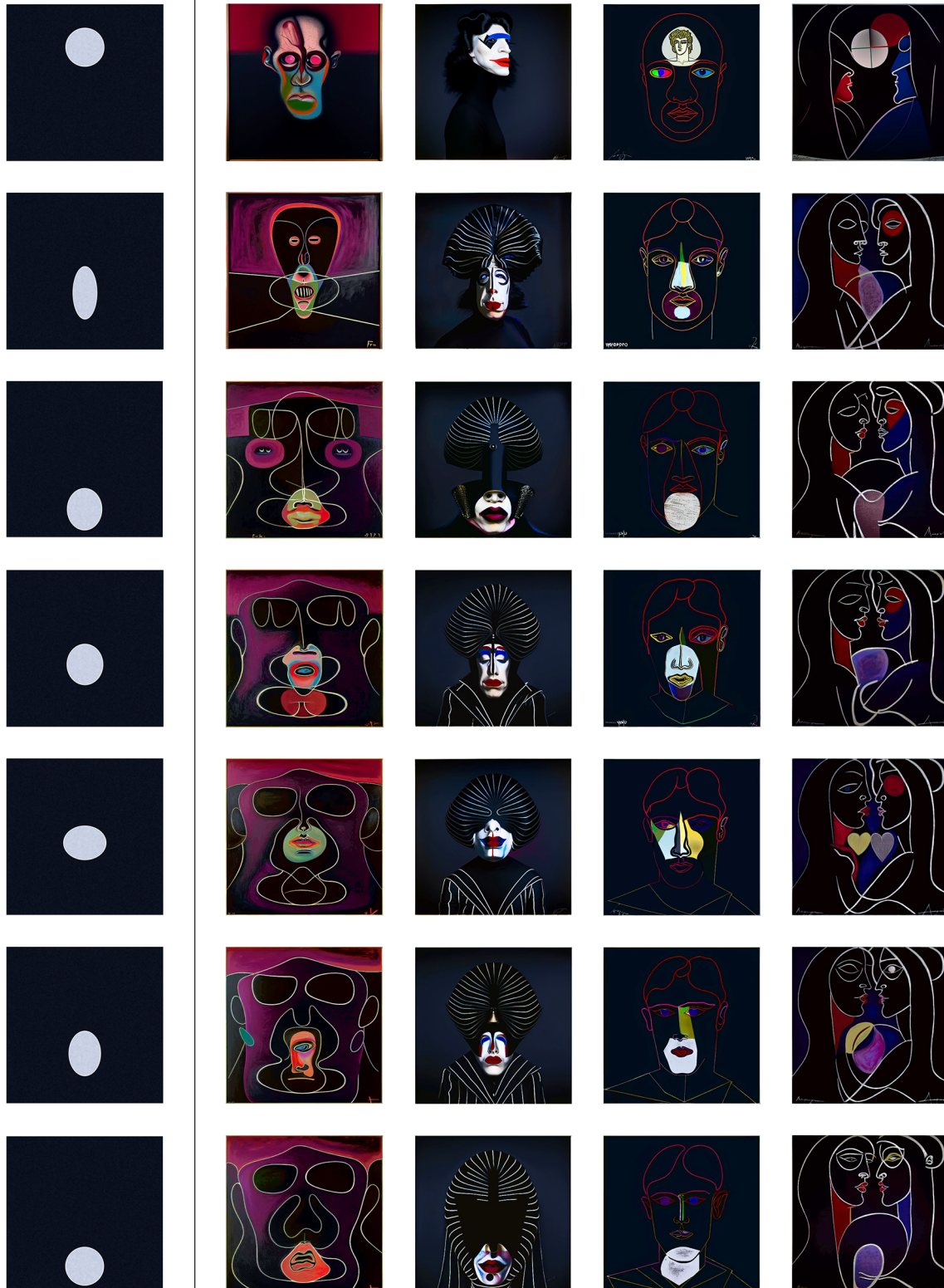


Figure 4.13: Grid of video results from the organic animation study.

4.6 Camera Footage Translation

The final study used real-life footage as the input video. Live video is more complex than the animation studies above because, unlike animation, we are not in control of every element in the frame. Furthermore, the photo-realism of the input might necessitate a photo-realistic translation which is less forgiving than the abstract results from the animation studies. For this study, I used two primary input videos: one is a video of a car crash, and the second is a video I composited together of a kiss between two versions of myself.



Figure 4.14: *Crash Me, Gently* input frames



Figure 4.15: *Kiss/Crash* input frames

In addition to the control techniques described above, I also experimented with the following techniques to generate a quality translation.

4.6.1 Control Techniques

Strength Scheduling

Strength scheduling is a technique where instead of designating an image strength for the whole video, I interpolate the image strength from beginning to end. As the video progresses, the diffusion process gets stronger and the resulting frames deviate more and more from the original input. For this to work effectively, I needed to interpolate not just the image strength but any other factor that impacts the image translation like the degree of frame blending and color coherence. Similar to the seed selection technique described above, I found that running this process in reverse by starting with the highest level of diffusion and interpolating downwards to generate the best results since the first frame has a large impact on the result of the animation both in terms of content, composition, and color profile. One last

control was adding ease functions for the strength schedule which allowed finer grain control over the effect.

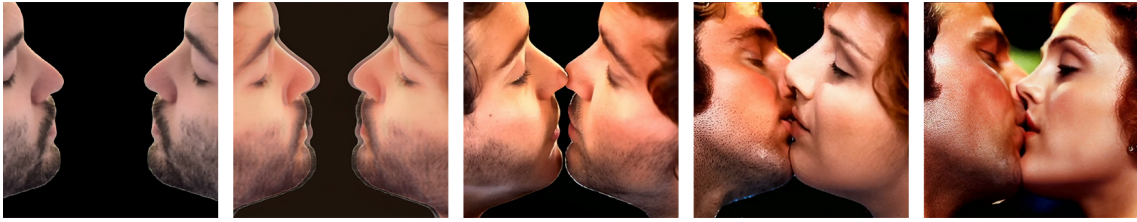


Figure 4.16: Example of strength scheduling. The result looks more like input at the start (left) and increasingly matches the prompt by the end (right).

Noise Blending

In previous examples, I almost always kept the seed consistent throughout the animation, but with these specific videos, I felt the need to add more noise variation over time. However, changing the seed in every frame was too dramatic. To control these extremes I added the ability to blend the underlying noise of successive frames meaning I could control how similar or different successive frames should be. This has a greater impact when the image strength is low and the generated image is driven more by the seed value.

When the blend value is high, the noise values used across the animation remain fairly stable and the resulting video has less variation. When the blend value is low, there are greater differences in the underlying noise of each frame and the resulting video has more variation

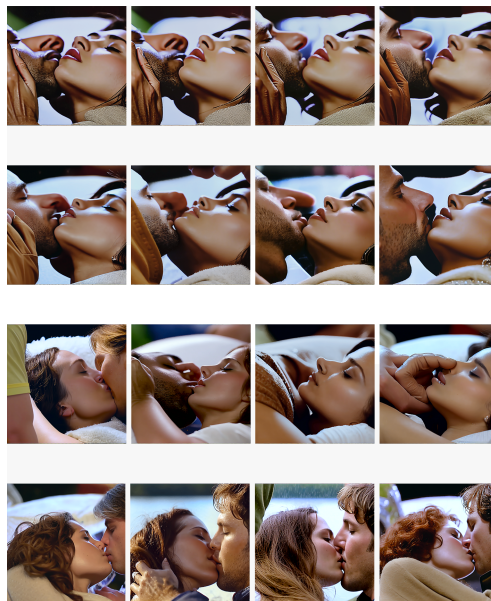


Figure 4.17: Noise blending increasing from top 0 (top row) to 1 (bottom row). Less noise blending results in less diverse images over time. More noise blending results in more variation over time.

DreamBooth

DreamBooth [48] is an incredibly powerful way to control the results of text-to-image generation using diffusion models. It allows us to fine-tune a diffusion model using textual inversion with just a handful of image samples. In this case, I was interested in generating images reflective of Hollywood cinematic kisses so I fine-tuned a stable diffusion model on a small dataset of 30 Hollywood kisses.

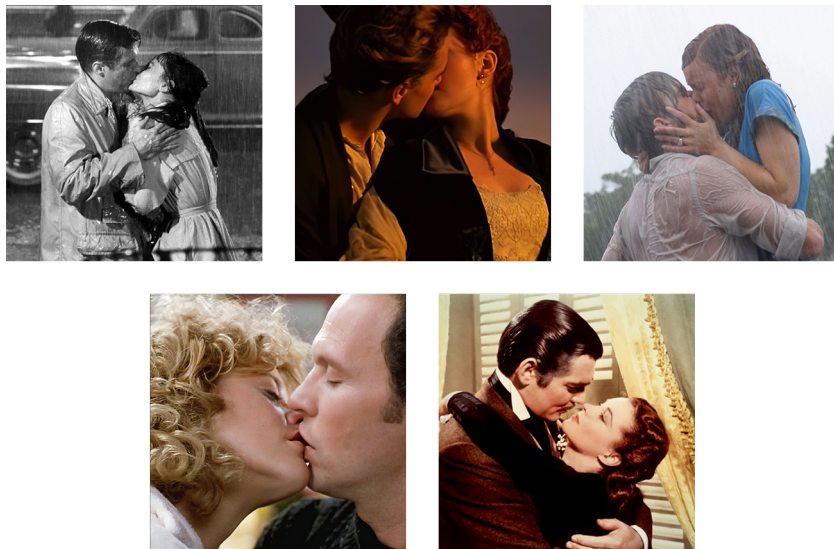


Figure 4.18: Training samples of iconic Hollywood kisses used to fine-tune the diffusion model used for these studies.

This fine-tuned model surprised me with how well it was able to capture the composition and style of the classic Hollywood kiss in various contexts. In relation to controlling image-to-image translation, using DreamBooth to fine-tune a model can focus the stylistic and compositional makeup of the diffusion process leading to more temporal and thematic coherence.

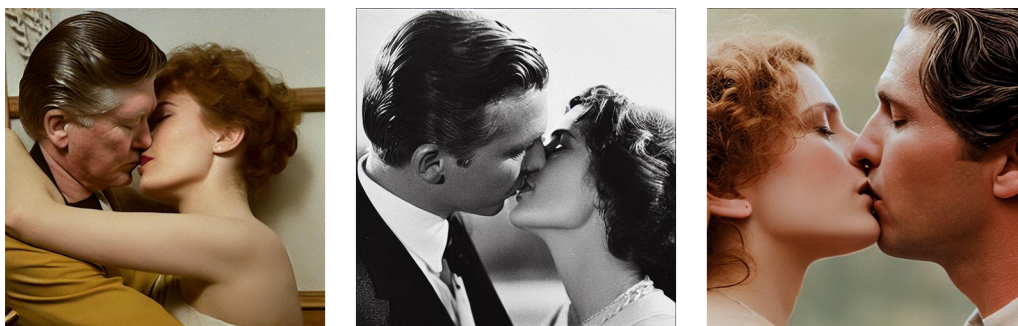


Figure 4.19: Images created with fine-tuned DreamBooth model

Stable WarpFusion

WarpFusion [57] is a technique developed by Alex Spirin specifically to reduce noise and improve consistency for video translation. Specifically, it uses optical flow maps to warp the previously generated frame onto the input frame of the next translation

step. This can be thought of as a slightly more advanced version of the frame-blending technique described above, but instead of considering just the previous generated frame and current input frame it also considers the change in motion of the input frame in relation to the previous frame. This technique is worth highlighting in spite of the fact that it's not my personal work due to its effectiveness for live video and the fact that it's under active development and improving continuously. It works neatly with many of the control techniques outlined above like video masks. As of the writing of this paper, the version of the notebook compatible with Stable Diffusion models and used for this study is only available through a paid Patreon subscription [58].

4.6.2 Results

The results generated here were used in the production of the final works described in more detail in the following section.

Input video



Result variations

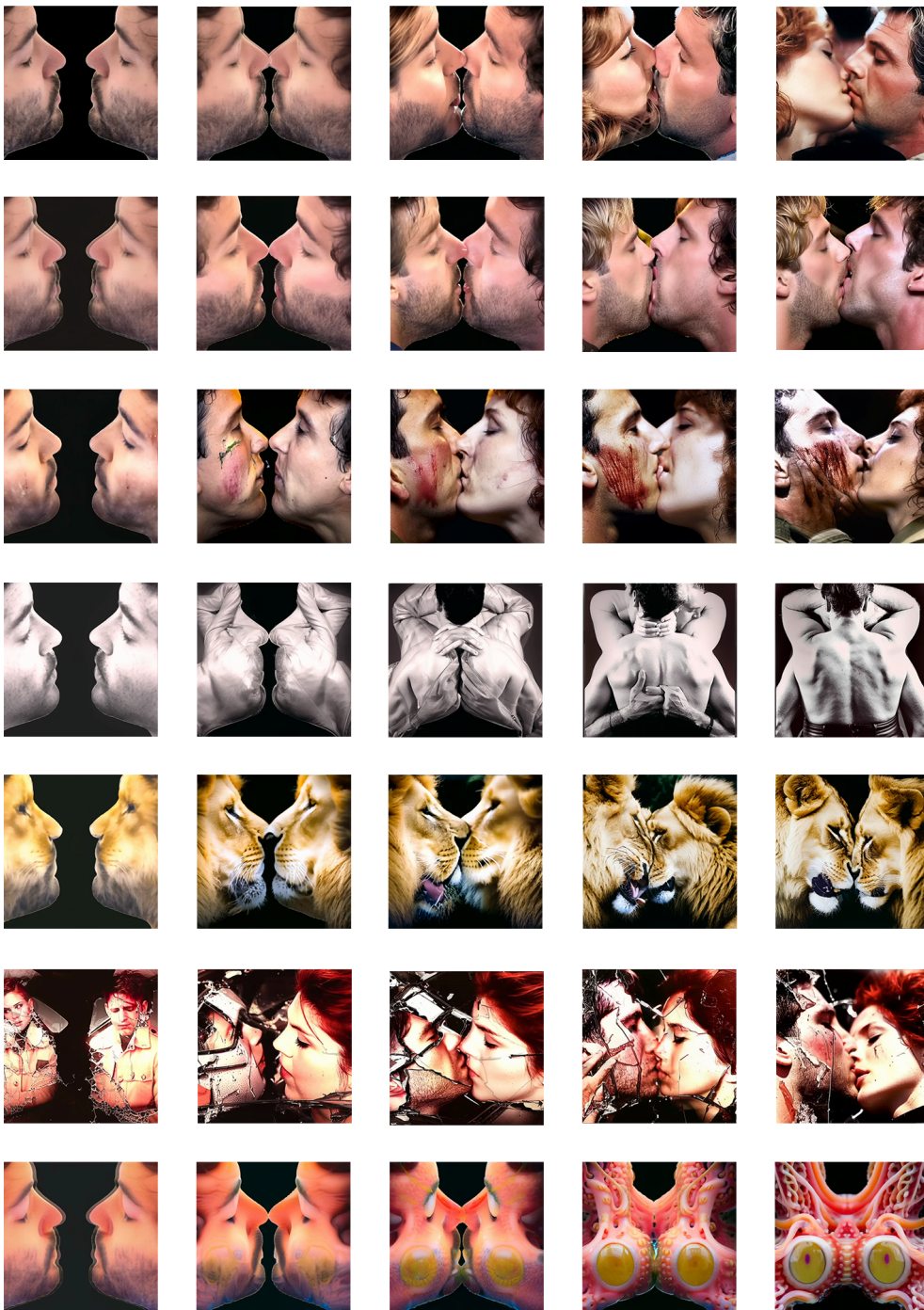


Figure 4.20: Strength scheduling examples used in *Crash Me, Gently*

Chapter 5

Final Works

The goal at this stage was to use the individual techniques explored throughout this paper to create thoughtful audio-visual compositions which spoke to the themes of intimacy, desire and loneliness in our image-saturated culture. Both pieces described below use mechanisms of control for the image-to-image translation that guide the aesthetic quality of the generated footage such as frame blending, strength scheduling, and models fine-tuned with DreamBooth. Video documentation of both works can be found in [Appendix A](#).

5.1 *Kiss/Crash*

In this piece, a car crash is turned into a kiss. Technically, we use the test crash footage of a classic 1950s Hollywood car as the starting video and use strength scheduling to morph the car crash smoothly into a romantic Hollywood kiss. Over about a minute in length, the cars repeatedly crash into one another. The colliding metal transforms into faces pressed deeply against each other. Shattered glass turns into romantic raindrops surrounding the central lovers. The sound of the crash smoothly transitions into a romantic song rich in 1950s Hollywood nostalgia. As the video progresses, the impacts crescendo in speed and intensity, and the image translations become more erotic and unhinged. The growing tension is released in the end by watching the full slow-motion video of the crash with no image translation applied. The association between these images becomes so ingrained that the car crash unedited takes on an eerily romantic tone on its own.

5.2 *Crash Me, Gently*

If the last piece turned a car crash into a kiss, this piece turns a kiss into a destructive car crash. *Crash Me, Gently* is an interactive installation using a diffusion model trained on iconic Hollywood kisses. A vintage CRT TV plays a video of me kissing myself. As the faces move closer together, the image is transformed into a romantic Hollywood kiss. There is a foot pedal the user can step on and as they do the pacing of the images accelerate. The kisses become more intense and cinematic, but they also become increasingly disjointed, pornographic, and violent. Contexts start to slip into one another: celebrities, creatures, and images of destruction begin to appear in the role of the central lovers. The tone of the audio also accelerates, with

sounds of metal crashing and shattered glass whenever the couple's lips meet. The image becomes increasingly destabilized as the viewer accelerates the pedal until finally released after which the TV returns to its calm looping kiss.

The starting point for this piece was the short video clip of my kiss in figure 4.14. I then used the techniques above such as strength scheduling, DreamBooth, and warp fusion to generate a large variety of clips: some of which were tamer and more similar to the input and some of which were more intense both in their distance from the input and strangeness of their content. Using these pre-rendered AI-generated images as a starting point, I was able to use OpenFrameworks to build an interactive installation incorporating them. Since they are all translations of the same input video, I was able to play with seamlessly cutting between them and gradually increasing the intensity of the images. Here control exists not just in the creation of the images themselves but also in the user experience of them.

Chapter 6

Evaluation

These results will be evaluated on two fronts: their technical proficiency and aesthetic value. Here, technical proficiency refers to how successfully formal qualities of the original were translated to outputs in terms of composition, color, and motion as well as visual and temporal coherence. The aesthetic evaluation will consider the works more holistically to measure their critical and poetic qualities.

6.1 Technical Evaluation

6.1.1 Still Image

The final image, Last Supper 1981, shows technical control over the final output by being sufficiently similar to the original but bringing out a completely new style and content. The main composition of the piece strongly resembles the input, specifically in the placement of the table, the row of guests behind it, and dramatized poses of the characters. However, the style is completely changed from a Renaissance painting to a modern black-and-white photograph and instead of apostles in robes, we have semi-nude queer men in drag and high heels. The level of control achieved here mainly through image strength and prompting is demonstrated by the ability to capture sufficient compositional details of the original while dramatically altering its content and meaning.

A reasonable criticism is that unlike the original, aspects of the translation are “imperfect”: faces are incomplete and bits of the scenery are unrealistic. In a later section, I explain how this can be read as part of its aesthetic value.

6.1.2 Animation Studies

In addition to static qualities like composition, the *animation studies* needed to capture the sense of motion from the original. They also had to be temporally coherent so that they play back as smooth, stand-alone animations.

In the study of the basic moving circle, some outputs like the floating squid capture the color and form of the original video quite closely while others like the fleshy abstract kiss simply capture the sense of motion suggested by the input video. All of these examples achieved a smooth temporal coherence, though, primarily through a mixture of frame blending and interpolation.

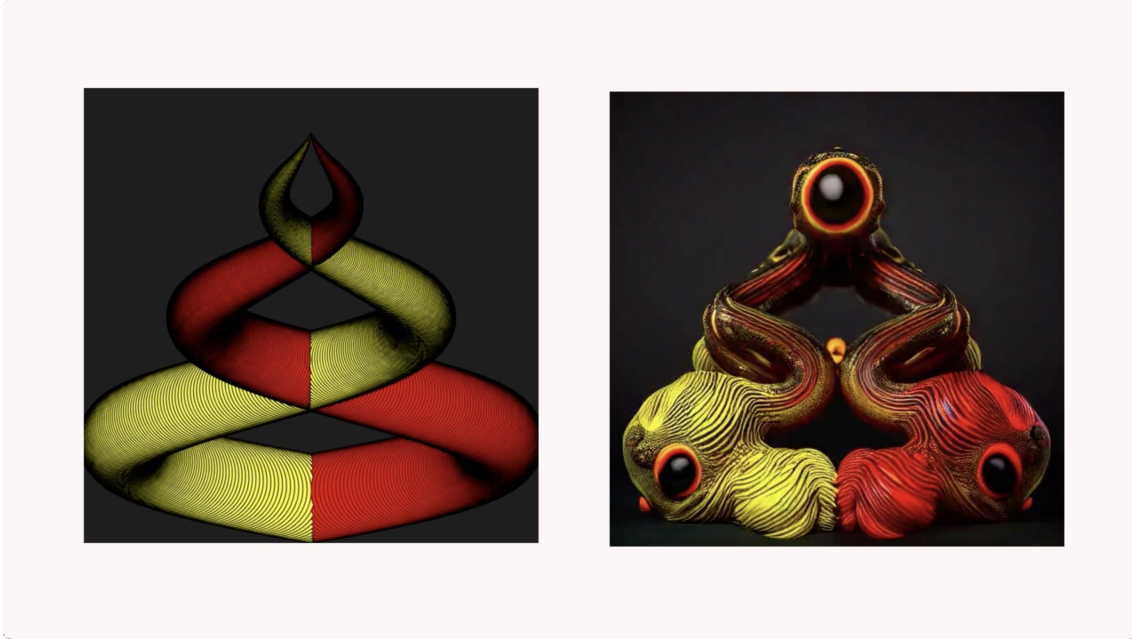


Figure 6.1: Detail from the organic animation study showing a side-by-side comparison of the input and its translation. Formal features like color and composition clearly carried over, but the result is still stylistically unique from the original.

These general qualities can be seen across the other two animation studies. In the complex animation outputs, elements of the original geometric animation can be seen in each result in terms of the red and yellow color scheme, pyramid-like composition, and motion which starts from the bottom and builds upward. The results sufficiently diverge from the original and each other, using styles that span 3D renders, graphic design, and photo-collage. The same can be said of the organic studies which are each coherent animations on their own but also capture the squash and stretch motion of the input in creatively diverse ways. For example, in one video the falling shape is captured by a face that seems to melt and reform while in another it plays a narrative role by turning into a pair of hearts at its peak before falling again.

In all these results, the balance achieved between alluding to the original input and outputting diverse, smooth results points to the technical success of the control techniques developed throughout this study.

6.1.3 Final Works

The technical quality of the video translations which are used in the final works can be seen in the photo-realistic results and the control over the intensity of the images. The results in 4.20 show examples of Hollywood-style kisses that match the composition of the input video but also look like they could be real frames from a film. In the later results farther down, we see how the intensity of content could be increased, introducing scenes of violence or sexuality, while still maintaining the main compositional features of the original. The same control over intensity is demonstrated in the *Kiss/Crash* translation where strength scheduling is used to visualize that crescendo of intensity.

6.2 Aesthetic Evaluation

In addition to achieving technically proficient results, this paper's stated goal is to explore the artistic potential of image translation in a way that reflects the power of representation and the nature of AI. Claiming the success of these results aesthetically is a more challenging task, but by evaluating the results closely, I may show what is meant by a "meaningful translation".

6.2.1 *The Last Supper, 1981* – Image Study



Figure 6.2: Left: the Mona Lisa by Leonardo da Vinci, Right: L.H.O.O.Q., Marcel Duchamp's transgressive "translation: of the original [15]

The first completed result was the translation of *The Last Supper*, inspired partially by the transgressive and satirical "image translation" of the Mona Lisa by Marcel Duchamp.

As discussed in the previous section, *The Last Supper, 1981* is clearly an allusion to the iconic painting. However, what makes it a meaningful translation is its stylistic qualities and political connotations.

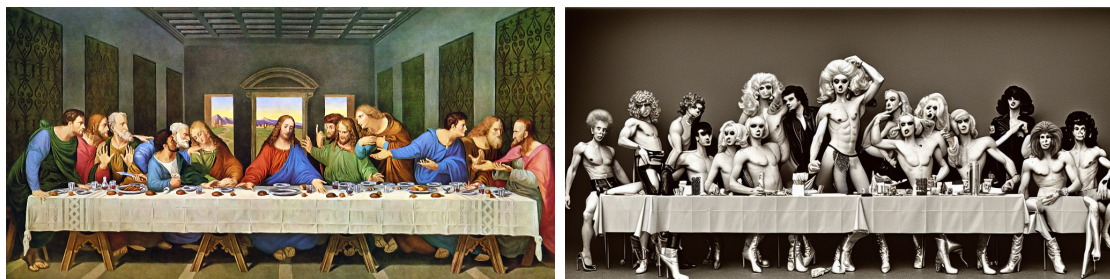


Figure 6.3: Left: the original Last Supper by Leonardo da Vinci, Right: *The Last Supper, 1981*, transgressive "translation: of the original

This image borrows from a particular 1980s gay aesthetic popular in New York City at a time when AIDS was just about to ravage the community and be largely

ignored by the Christian right wing of the country who believed AIDS to be a religious plague. Repurposing the Christian iconography reimagines the Last Supper as a moment of queer joy and points to their imminent betrayal by the government. Here the role of Jesus is satirically replaced by a buff gay man in drag. The fact that an AI clearly generated this image, bringing to life placeholders of individuals who might not have survived this era, adds another level of meaning.

A close reading of this image reveals it as more than a technically proficient translation of the original but also as a meaningful one. Here the meaning is derived from the cultural and political implications of the result in relation to the original, and the transgressive nature of its religious context.

6.2.2 Animation Studies

Next, I'd like to consider the animation studies collectively. Individually, the animation studies are visually compelling. While some can stand on their own as art pieces, what I found most interesting was looking at the studies in aggregate.

By seeing all the results for a given study side by side, you get a sense of the unlimited number of variations that can be produced from a single input. It makes the process of image translation visible to an audience who may not be familiar with this technology.

Doing so speaks to the nature of these AI tools and the way they are able to easily manipulate digital surfaces. These augmentations have the potential to be incredibly beautiful but can also feel quite shallow, like a thin veneer placed between us and the original. In that sense, they become illustrative of our increasingly mediated relationship to reality. One gets a sense of the way AI translation will be used to mediate between the reality of the original and our perception of it, especially with the specter of the metaverse on the horizon.

Here the aesthetic quality of the results is both in the visual abundance of the individual animations, as well as how collectively they reveal a truth about the nature of AI imagery.

6.2.3 Final Works

To evaluate the final works, I will describe what makes them meaningful both in terms of their themes as well as the role AI imagery plays within the logic of the artwork. Both of these works deal with the theme of desire and the gap between experience and representation in the digital age.

Kiss/Crash

Kiss/Crash creates meaning through the suggestive metaphor conjured between two seemingly unrelated images: the car crash and the kiss. These are images common in Western media, from the destructive nature of action movies to the passionate romance scenes in dramas, but are also prevalent daily across advertisements and social media news feeds. Here, the juxtaposition of these two mirrored poses creates an uneasy equivalence questioning the relationship between the two. What destruction is triggered by romance? What pleasure is there in a car crash? The metaphor in the work is suggestive but not moralistic allowing the viewer to create provocative connections between the two.

One possible connection to be made is between the mixture of Thanatos and Eros, the *death drive* and the *love-drive*. The primary experience of the viewer is one of arousal as they watch the crash repeatedly turned into an intimate scene, an experience reinforced by the acceleration of the video's rhythm and sexualized content. The connection between death and love points to our innate human desire for intensity, intimacy, and authentic feeling. But that desire, both in this video and in our culture in general, is sublimated by representations sometimes of extreme destruction and other times through exaggerated romance. The desire is real. The images are not. As an audience, we are left feeling the gap between the representation of that intensity and our actual desire for it.

The role of AI imagery here is twofold. First is in the formal ability to smoothly translate from one iconic image to the other, thereby creating the metaphor between the two in a visually stimulating way. The second is in the exaggerated spectacle of the images which highlights their artificiality and the degree to which the experience is simulated.

Crash Me, Gently

In *Crash Me, Gently*, meaning is derived from the contrast between the individual experience of loneliness and the desire for real, sensuous experiences with others. The isolated central figure, who is in a cold mirrored kiss with himself, is repeatedly transformed into a warm sensuous Hollywood kiss. In this piece the self is quite literally projected into the romanticized representation of love, visually demonstrating the ways we attempt to see ourselves in these stylized images. The physicality of the television display mimics the medium by which we often experience these scenes and reminds us of their status as artificial images.

Central to this piece is its interactivity which invites the viewer to press the foot pedal and accelerate the intensity of the experience. The viewer becomes complicit in the hunt for more exaggerated representations of "love". The frenetic changes of context as the pedal is pressed, including the fast-paced mixture of violence, sex, celebrity, and disaster, mimics the increasing schizophrenia of our digital world. The immediate result of pressing the pedal and being bombarded with these sensual scenes is perversely satisfying, but it also questions the way desire can be aroused by representations while never truly being satisfied.

The AI-generated imagery plays a similar role here as in *Kiss/Crash* by allowing us to translate the self into representations of cultural myths and suggesting the way artificial representations may increasingly stand in for real sensations in the future.

Chapter 7

Discussion

7.1 Technical Contribution

In this paper, I outlined a series of techniques and control mechanisms to achieve quality results for image and video translations. However, this is not to say the system has been perfected or that the problem of image translation using diffusion models is solved. Many of the visual outputs shared here were the result of trial and error and experimentation. Creating a quality translation still takes a lot of manual labor, and there remains much that can be improved in the realm of temporal coherence. Additionally, the techniques outlined here are not comprehensive. The subset of methods explored in this paper were solutions to specific challenges that arose while working on these particular studies. As such, there are numerous additional avenues to explore within this framework.

Rather, this paper’s key technical contribution is providing an instinctual understanding of how to accomplish quality image translations more than any singular method described. As noted above, not every technique mentioned here was totally original. Some like DreamBooth and WarpFusion were open-source tools pulled together to address my research goal. Some techniques like video masks were original at the time of my implementation but have since shown up independently in various open-source notebooks.

By showing which control techniques were most effective, how they work internally, and when they should be used, I’ve given a good overview of the problem space and outlined the available routes to take when navigating it. Artists and engineers who engage with this topic will be able to pick and choose which techniques might be most beneficial for their specific goals and achieve similar results. Finally, it provides a solid foundation of research into the topic of diffusion-based image translation which can be further built upon in academic settings or by the open-source community pushing this medium forward.

7.2 Aesthetic Contribution

One of the goals of this paper was to repurpose an image-making technique with a critical eye in the same spirit as former avant-garde artists. The key contribution in this space was developing an art that makes the process of AI translation visible and hints at its potential dangers.

Additionally, I found the method used here to be extremely effective and one worth pursuing further by other artists and myself. Specifically, I began with an open discomfort with diffusion models and the effect they might have on art and society. But by methodically learning how they work, both technically and aesthetically, I was able to engage with the implications of the medium more thoughtfully. This is the same method used by many artists of the past who contended with their uneasy relationship to technology as well as more recently in the AI community by [8] and [4]. The technical research outlined here provided a clear pathway to the creation of meaningful art.

7.3 Closing Remarks

Susan Sontag, writing 50 years before the arrival of AI imagery, made the following call to action which seems just as relevant today, if not more so:

Images are more real than anyone could have supposed. And just because they are an unlimited resource, one that cannot be exhausted by consumerist waste, there is all the more reason to apply the conservationist remedy. If there can be a better way for the real world to include the one of images, it will require an ecology not only of real things but of images as well. [56]

I opened this essay with concern for the oncoming flood of artificial images, but my ultimate goal with this project was to consider the ways we can use artificial images in a self-reflective way that heightens our sense of the real. My hope is that instead of creating more visual waste, I've added to the worthwhile ecology of images: an addition that doesn't further drown us in the flood, but instead helps us float to the surface where we can see the sun and feel the light on our skin.

Chapter 8

Conclusion

I began this essay by questioning how artists could use prompt-driven generative art in a self-reflective critical way. I hypothesized that I could use diffusion-based image-to-image translation to meaningfully augment, invert, and negate existing imagery in a way that revealed a truth about the nature of artificial images. I tested this hypothesis through a sequence of increasingly complex studies where I developed the necessary levers of control to develop quality translations. I then used those skills in the development of two new audio-visual works which reflected on the nature of images in the age of AI. Finally, I reflected on how these findings may be used technically by other artists and considered what made these AI projects aesthetically meaningful. In doing so, I called on artists to reflect on how the AI tools we use in our practice connect to the past, can be critical of the present, and might imagine a worthwhile future.

Appendix A

Links to Video Materials

Animation Studies Collected video work from the animation studies in chapters 4.3 – 4.5 <https://vimeo.com/773245117>

Kiss/Crash Video of completed work <https://vimeo.com/773240946>

Crash Me, Gently Documentation video of interactive installation <https://vimeo.com/773388839>

Bibliography

- [1] URL: <https://imagen.research.google/>.
- [2] Midjourney showcase. URL: <https://www.midjourney.com/showcase/>.
- [3] aiplague. Aiplague - blue crystal fire (4k) ai video / stable diffusion, Sep 2022. URL: <https://www.youtube.com/watch?v=W4A-oGlkSh4>.
- [4] Memo Akten, Rebecca Fiebrink, and Mick Grierson. Learning to see: You are what you see. In *ACM SIGGRAPH 2019 Art Gallery*, SIGGRAPH '19, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3306211.3320143.
- [5] Alembics. Alembics/disco-diffusion. URL: <https://github.com/alembics/disco-diffusion>.
- [6] Jean Baudrillard. *Simulacra and simulation*. University of Michigan Press, 1994.
- [7] Vittoria Benzine. In pictures: Dall-e makes its gallery debut in a show where all the works were created with an assist from a.i., Nov 2022. URL: <https://news.artnet.com/art-world/in-pictures-dall-e-makes-its-gallery-debut-in-a-show-where-all-the-works-were->
- [8] Terence Broad and Mick Grierson. Autoencoding blade runner: Reconstructing films with artificial neural networks. In *ACM SIGGRAPH 2017 Art Gallery*, SIGGRAPH '17, page 376–383, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3072940.3072964.
- [9] Laurie Clarke. Is generative ai really a threat to creative professionals?, Nov 2022. URL: <https://www.theguardian.com/technology/2022/nov/12/when-ai-can-make-art-what-does-it-mean-for-creativity-dall-e-midjourney>.
- [10] @CorridorDigital. The spider-verse joins the mcu, Nov 2022. URL: <https://twitter.com/CorridorDigital/status/1591958349612224514>.
- [11] Katherine Crowson and et al. Disco diffusion, 2021. URL: <http://discodiffusion.com/>.
- [12] Guy Debord. *The society of the spectacle*. Rebel Press, 2005.
- [13] Deforum. Deforum stable diffusion. URL: https://colab.research.google.com/github/deforum-art/deforum-stable-diffusion/blob/main/Deforum_Stable_Diffusion.ipynb.

- [14] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. [arXiv:2105.05233](#).
- [15] Marcel Duchamp. L.h.o.o.q., 1919. URL: https://en.wikipedia.org/wiki/File:Marcel_Duchamp,_1919,_L.H.O.O.Q.jpg.
- [16] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015. [arXiv:1508.06576](#).
- [17] Raoul Hausmann. The art critic, 1919. URL: <https://www.tate.org.uk/art/artworks/hausmann-the-art-critic-t01918>.
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. [arXiv:2006.11239](#).
- [19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. [arXiv:2207.12598](#).
- [20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. [arXiv:2204.03458](#).
- [21] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance, 2022. [arXiv:2210.00939](#).
- [22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2016. [arXiv:1611.07004](#).
- [23] Frederic Jameson. *Postmodernism or the cultural logic of late capitalism*. Duke Univ. Press, 1992.
- [24] Ondrej Jamriska. Ebsynth: Fast example-based image synthesis and style transfer, 2018. URL: <https://github.com/jamriska/ebsynth>.
- [25] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. [arXiv:2206.00364](#).
- [26] Barbara Kruger. Untitled (i shop therefore i am), 1987. URL: <https://www.artsy.net/artwork/barbara-kruger-untitled-i-shop-therefore-i-am>.
- [27] Hye-Kyung Lee. Rethinking creativity: creative industries, ai and everyday creativity. *Media, Culture & Society*, 44(3):601–612, 2022. [arXiv:https://doi.org/10.1177/01634437221077009](#), [doi:10.1177/01634437221077009](#).
- [28] Scott Lighthiser. Stable diffusion | vfx / story context test, Nov 2022. URL: <https://www.youtube.com/watch?v=2CmRj8TdR8w>.
- [29] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models, 2022. [arXiv:2201.09865](#).
- [30] Sunil Manghani, Arthur Piper, and Jon Simons. *Images: A Reader*. SAGE, 2006.

- [31] Stephen Marche. We’re witnessing the birth of a new artistic medium, Sep 2022. URL: <https://www.theatlantic.com/technology/archive/2022/09/ai-art-generators-future/671568/>.
- [32] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sedit: Guided image synthesis and editing with stochastic differential equations, 2021. [arXiv:2108.01073](https://arxiv.org/abs/2108.01073).
- [33] Emad Mostaque. Stable diffusion public release, Sep 2022. URL: <https://stability.ai/blog/stable-diffusion-public-release>.
- [34] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. [arXiv:2102.09672](https://arxiv.org/abs/2102.09672).
- [35] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photo-realistic image generation and editing with text-guided diffusion models, 2022. [arXiv:2112.10741](https://arxiv.org/abs/2112.10741).
- [36] Ryan O’Connor. Introduction to diffusion models for machine learning, Sep 2022. URL: <https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/>.
- [37] OpenAI. Dall-e 2, Apr 2022. URL: <https://openai.com/blog/dall-e-2/>.
- [38] Santiago Pascual, Gautam Bhattacharya, Chunghsin Yeh, Jordi Pons, and Joan Serra. Full-band general audio synthesis with score-based diffusion, 2022. [arXiv:2210.14661](https://arxiv.org/abs/2210.14661).
- [39] Plato. *The republic*. Penguin books, 2007.
- [40] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022. [arXiv:2209.14988](https://arxiv.org/abs/2209.14988).
- [41] Roope Rainisto. Temporary, Oct 2022. URL: <https://www.youtube.com/watch?v=Ev2ddH5PJR0>.
- [42] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion, 2022. [arXiv:2202.04901](https://arxiv.org/abs/2202.04901).
- [43] Adi Robertson. How deviantart is navigating the ai art minefield, Nov 2022. URL: <https://www.theverge.com/2022/11/15/23449036/deviantart-ai-art-dreamup-training-data-controversy>.
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. [arXiv:2112.10752](https://arxiv.org/abs/2112.10752).
- [45] Kevin Roose. A.i.-generated art is already transforming creative work, Oct 2022. URL: <https://www.nytimes.com/2022/10/21/technology/ai-generated-art-jobs-dall-e-2.html>.

- [46] Kevin Roose. An a.i.-generated picture won an art prize. artists aren't happy., Sep 2022. URL: <https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html>.
- [47] Janus Rose. Inside midjourney, the generative art ai that rivals dall-e, Jul 2022. URL: <https://www.vice.com/en/article/wxn5wn/inside-midjourney-the-generative-art-ai-that-rivals-dall-e>.
- [48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2022. [arXiv:2202.04901](https://arxiv.org/abs/2202.04901).
- [49] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models, 2021. [arXiv:2111.05826](https://arxiv.org/abs/2111.05826).
- [50] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. [arXiv:2205.11487](https://arxiv.org/abs/2205.11487).
- [51] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement, 2021. [arXiv:2104.07636](https://arxiv.org/abs/2104.07636).
- [52] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models, 2022. [arXiv:2202.00512](https://arxiv.org/abs/2202.00512).
- [53] Hiroshi Sasaki, Chris G. Willcocks, and Toby P. Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models, 2021. [arXiv:2104.05358](https://arxiv.org/abs/2104.05358).
- [54] @ScottieFoxTTV. Stable diffusion in vr + touchdesigner = realtime immersive latent space, Oct 2022. URL: <https://twitter.com/ScottieFoxTTV/status/1578387866572525570>.
- [55] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. [arXiv:1503.03585](https://arxiv.org/abs/1503.03585).
- [56] Susan Sontag. *On Photography*. Penguin books, 2019.
- [57] Alex Spirin. Disco diffusion v5.2 - warp by alex spirin. URL: https://colab.research.google.com/github/Sxela/DiscoDiffusion-Warp/blob/main/Disco_Diffusion_v5_2_Warp.ipynb.
- [58] Alex Spirin. Sxela is creating stuff using ai in an unintended way | stable warp fusion. URL: <https://www.patreon.com/sxela>.
- [59] Nitasha Tiku. Ai can now create any image in seconds, bringing wonder and danger, Sep 2022. URL: <https://www.washingtonpost.com/technology/interactive/2022/artificial-intelligence-images-dall-e/>.

- [60] Andy Warhol. Marilyn diptych, 1962. URL: <https://www.tate.org.uk/art/artworks/warhol-marilyn-diptych-t03093>.
- [61] Oscar Wilde. *The Picture of Dorian Gray*. Penguin Books, 2000.
- [62] Julia Wolleb, Robin Sandkühler, Florentin Bieder, and Philippe C. Cattin. The swiss army knife for image-to-image translation: Multi-task diffusion models, 2022. [arXiv:2204.02641](https://arxiv.org/abs/2204.02641).
- [63] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications, 2022. [arXiv:2209.00796](https://arxiv.org/abs/2209.00796).
- [64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017. [arXiv:1703.10593](https://arxiv.org/abs/1703.10593).