PIXEL ALCHEMIST —RAPID IMAGE SYNTHESIS AND EDITING WITH GENERATIVE DEEP LEARNING

TITUS EBBECKE

MAY 2022

UNIVERSITY OF CALIFORNIA, BERKELEY MDES

TABLE OF CONTENTS

ABSTRACT

To get the most out of computers, human-machine interfaces should ideally use a language that human designers are already familiar with. Unfortunately, most traditional software deploys a highly abstract and low-level syntax that is designed to be easily understood by a computer but not by a person. This makes computer-aided creation inaccessible and inefficient.

Moving through disentangled latent vector spaces of generative deep learning networks has recently allowed us to independently manipulate a vast amount of fuzzy high-level parameters that can describe an equally large amount of design decisions the way humans naturally do.

This thesis combines a graphical user interface with a tactile hardware controller to simultaneously control an arbitrary amount of user-defined parameters for semantic image editing with generative adversarial networks. The result is a system that can precisely edit synthetic images from any visual domain with fuzzy text prompts, allowing people with limited technical knowledge to utilize their rich natural language abilities to iteratively design complex imagery with unprecedented speed in a non-destructive manner.

1 MOTIVATION

The complex nature of how humans experience reality allows them to precisely communicate highly nuanced concepts. A musician may describe a composition as *wild and ambitious, young and free* or a cymbal as *dry* or *wet*. Likewise, an art critic could claim a painting to be *misogynist* or *unsettling*, and a fashion designer may view a dress as either *progressive* or *nostalgic*. The sheer amount of combined emotions, tastes, and shared experiences that can serve as easily interpretable analogies allow people to articulate ideas with tremendous precision. Due to shared human experiences, many cultures naturally associate *warm* tones, colors, or personalities as comforting and pleasant, while *cold* things are perceived as disheartening, distant, and provocative, regardless of whether the object in question can actually have a temperature. Such terminologies are here considered *fuzzy* and *high-level* as they are representational phrases that cannot be objectively connected to a single property.

The ability to communicate with *fuzzy* language is crucial for creating complex artifacts with other humans and machines. This can be illustrated with a hypothetical interaction between an architect and a client planning a modern museum at a port. The client has a specific vision for the style of the desired building but lacks the ability to express it in a concrete way with objective low-level parameters like exact measurements. Instead of presenting the architect with a floor plan or 3D model, the client would rather vocally describe their idea as a "*flowing silhouette with a facade shining like the waves of the surrounding water*". This single sentence directly informs a large range of decisions the architect has to make, from the appearance of the building, over material choice, to the layout of the building even though the client did not explicitly specify any of these factors.

This highly efficient way of communication can also benefit people who browse artifacts rather than create them. An obscure yet successful example can be found in a recent trend on the video-sharing platform YouTube. Here, some users are compiling music lists where the chosen tracks are not aligned by artist or genre but by a specific emotion they evoke. To communicate the content of the playlist, the creators give them highly-abstracted titles such as "*this playlist will make you feel like a greek goddess in a ruin garden*" – a collection of dreamy piano pieces and enchanting lullabies. A playlist titled "*You're in love with the villain of the story*" consists of tragic classics full of gloomy strings while "*random burst of energy at 2am*" is a compilation of bright, noisy, and intense electronic hyperpop.

Although these metaphorical titles are disconnected from any objective properties of the chosen music, they still manage to be highly descriptive. When combined with a thumbnail that visually encompasses the selected atmosphere, such forms of describing music can convey the acoustic content significantly better than the titles of the music pieces or the names of the artists.

Traditional design software, in contrast, cannot interpret such language as it instead operates with low-level syntaxes that can take years to master. This comes at great expense and inaccessibility. The previously mentioned fictional architect, for example, would have to work together with skilled 3D artists that tenaciously translate the client's fuzzy description into cartesian coordinates of digital vertices, normal maps of virtual materials, or integers that determine the appearance of some mesh. The ability to express an idea through artificial media with a computer is thus reserved for an elite that has the resources required to learn and practice computer-aided creation.

2. SEMANTIC IMAGE GENERATION WITH DEEP LEARNING

Recent leaps in generative deep learning, however, have revealed that even self-supervised models can automatically learn representations of high-level parameters. Such models are trained on large datasets, often images, which they can reproduce after a long training process. Generative adversarial networks, or GANs (Goodfellow et al. 2014), variational autoencoders (VAEs) (Kingma and Welling 2013), and diffusion models (Sohl-Dickstein et al. 07--09 Jul 2015; Dhariwal and Nichol 2021; Ramesh et al. 2022) are among the most successful types of generative frameworks, capable of creating completely novel images of faces, cars, and landscapes that are virtually indistinguishable from real ones.

Despite revolutionizing the way computers generate imagery, the initially random nature of synthesis with deep learning made these models unsuitable for controlled design processes where the machine has to follow the intentions of a human designer.

Recently, Radford et al. (18--24 Jul 2021) introduced the highly influential "Contrastive Language-Image Pretraining" (CLIP) model, a system trained on 400 million image-text pairs to predict how related a text description is to an image.

CLIP greatly accelerated research on semantic image synthesis. When combined with a generative framework, it can allow users to generate an image from a text prompt describing the desired image (Ramesh et al. 18--24 Jul 2021). By using CLIP as a discriminator rating generated images against the text prompt, a system can gradually optimize for this rating until a satisfactory image is found.

Unfortunately, there is no single image that uniquely satisfies a text description unless the prompt is exceedingly long and specific. Even if this would be the case, designers naturally want to adjust a result to explore multiple design directions or to incorporate changes that a collaborator or the designer itself desires. The limitations of communication and interactive tools already prevent human designers from perfectly interpreting the intentions of another collaborating human. In some cases, a person may even fail to be satisfied with a creation of their own, even if it suits their original intention, as the human imagination often lacks the resolution of a concrete artifact. Assuming that a machine can simply generate a perfect image from a single input is unrealistic.

Digital design tools must therefore allow for iterative adjustments. The randomness and inability to gradually edit parts of a resulting image make simple text-to-image approaches insufficient no matter their ability to interpret a target description.

3 DISENTANGLEMENT OF LATENT VECTOR SPACES

A breakthrough came with the observation of disentangled characteristics of the high-dimensional latent vector spaces of GANs (Chen et al. 2016; Esser, Rombach, and Ommer 2020; Lin et al. 2019; Kazemi, Iranmanesh, and Nasrabadi 2019). Disentanglement refers to the independence of semantically meaningful parameters inside bespoke latent space. If such a space is disentangled, it tends to have spatial directions that exclusively affect individual properties of the generated output. Moving along such a direction can correspond to the change in the hair length of a generated face or the color of an artificial car. Further observations revealed that this also includes fuzzy parameters like the *sportiness* of a car, the *happiness* of a face, or even the degree to which an artificial person looks like Taylor Swift. This opened up the possibility to independently edit a vast amount of parameters by simply moving along their respective dimensions in the latent vector space.

4 STYLECLIP GLOBAL DIRECTIONS

Finding and labeling semantically meaningful directions remained a challenge. Initially, this process either required manual inspection and annotation or pre-trained classifiers. Both severely limit the number of accessible parameters. In order to leverage the unique advantages of self-supervised deep learning, a system must be able to automatically scale with more parameters and datasets.

In their paper "StyleCLIP', Patashnik et al. (2021) demonstrate that directions in the high-dimensional vector space of CLIP's pre-trained neural networks could be transposed to the vector space of a GAN. This idea is spawned from the assumption that despite inherent differences between a vector space of text embeddings and a vector space of image embeddings, both can share semantically meaningful global directions that have high cosine similarity. The manifolds of CLIP's text embeddings and image embeddings could thus be normalized to allow a user to map the direction Δt, representing a direction from a neutral to a target text prompt, from CLIP's language-image embedding space to StyleGAN's (Karras, Laine, and Aila 2019; Karras, Laine, et al. 2020) stylespace S. In theory, this finding allows a user to visually manipulate a GAN-based image with every semantic parameter that CLIP encodes.

The unique advantage of these *global* directions is the ability to universally change images in real-time with a nearly unlimited amount of text prompts. In contrast, loss-based optimization usually requires gradient descent over many generated images until a satisfactory one is found, essentially brute-forcing the way to a target. This can take minutes and is therefore unsuitable for a rapid iterative design process. More importantly, most previous works on semantic image editing with generative deep learning are limited to a preset number of parameters. This inhibits the advantage of using deep learning for media editing, as conventional software systems already offer a range of precise parameters by manually engineering a desired programmatic function.

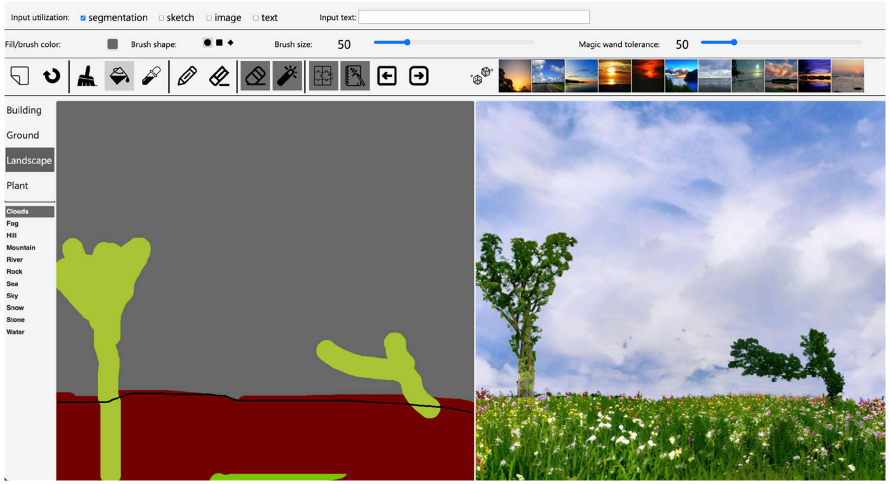
Allowing users to spontaneously utilize thousands of fuzzy semantic parameters with minimal preprocessing has been previously unheard of in the world of computational image processing. Discoveries like global directions for semantic image editing mark a milestone in the pursuit of aligning the interfaces of computational design systems with traditional human-human communication.

5 A GRAPHICAL USER INTERFACE FOR RAPID SEMANTIC IMAGE GENERATION AND EDITING

To make the advantages of StyleCLIP accessible to designers and artists, I developed a dynamic graphical user interface (GUI). Similar to the GUI that Patashnik et al. (2021) propose, users can set a neutral prompt and a target prompt. The former should describe the visual content of the GAN-generated base image while the latter describes the desired change in the target image. Users can determine the intensity a of this change, as well as how many aspects of the image should be changed through the threshold parameter β. A high β threshold will only touch the most relevant parts of the base image. For example, when manipulating the target prompt *brown eyes* on an artificial face with a high β value, only the color of the iris will change while a low β value may also affect the nose, wrinkles, and even the gaze of the face. The broad nature of some target prompts may require a low β threshold by default. When a face is supposed to look more *surprised*, for example, StyleCLIP would ideally change a range of visual elements, including raising eyebrows, adding wrinkles on the forehead, and opening the mouth.

6 CONVENTIONAL INTERFACES LIMIT DEEP LEARNING BASED IMAGE EDITING

Having access to a large number of arbitrary variables renders traditional controllers like mice and cursors insufficient. The use of such traditional interfaces makes image synthesis with systems like GauGAN (Park et al. 2019), DALL-E (Ramesh et al. 18--24 Jul 2021), or even StyleCLIP's GUI hard to use and inefficient. GauGAN requires the user to draw semantic masks that get translated into a photorealistic style. This process is tedious with an inaccurate touchpad or mouse and only works well with domains that do not require precise masks such as landscapes. While it replicates the



GauGAN's graphical user interface. A user has to paint a simple semantic segmentation on the left side and can translate the composition into a range of photorealistic styles as seen on the right.

widespread use of sketches as a form of communication, it does not allow to iterate the result with fuzzy parameters that apply to the overall appearance rather than clearly defined parts of the image. Systems like DALL-E are outstanding at generating images that fit even highly specific text prompts. However, editing is limited to the binary presence of objects in bespoke prompts. A gradual control over the intensity of parameters does not exist. Additionally, each change requires the user to type in a new text, making an iterative design process mentally demanding due to the effort associated with typing on a keyboard. StyleCLIP already offers a GUI to change a single prompt with one pair of α and β controllers. The interface consists of just one slider which may change a parameter that conveys a multitude of visual aspects but is still limiting as it requires a prompt to contain every desired change. Controlling the individual intensity of each of these changes is therefore not supported. Lastly, most interfaces rely on a cursor that can only be at one place at a time. This slows the design process down further by making simultaneous control of multiple parameters impossible.

These limitations are detrimental to the user experience because of the uncertain nature of how artificial intelligence creates output. Due to the novelty of writing to a computer what it should design, users often feel overwhelmed as they expect a clear syntax, the way they are used to from conventional software.

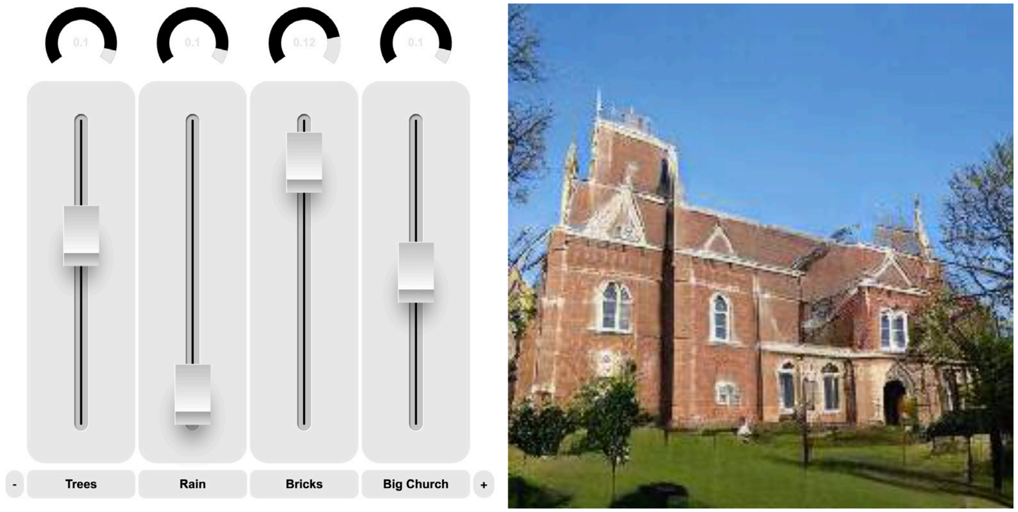
More importantly, it is often not given that a system actually interprets input the way the user intends to. Targeted image generation with deep learning, therefore, requires speculation and a lot of iteration. In fact, many semantic interfaces for generative models have unpredictable preferences for specific phrases and grammar while the models themselves have significant limits to what they are able to generate in the first place. Combined, these factors create a need for systems that allow designers to quickly experiment with multiple target prompts and their intensity to achieve an intuitive, rapid, and precise design process.

7 A HIGH DOF CONTROLLER FOR SEMANTIC IMAGE EDITING

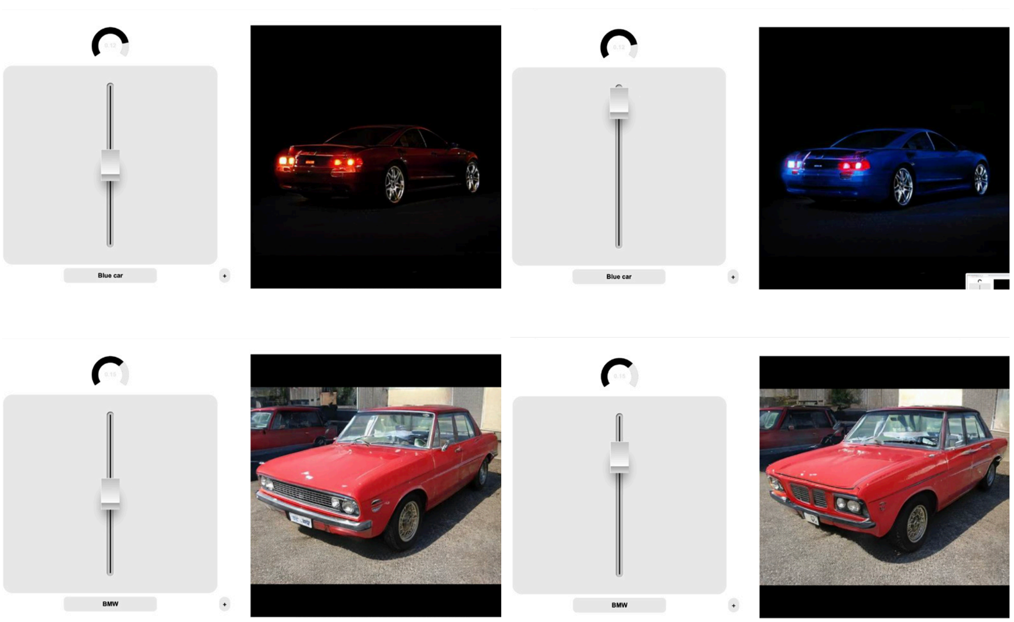
Given the sheer amount of possible parameters that can be manipulated with generative models, interfaces must therefore allow users to quickly specify and edit a large set of parameters in parallel. I propose the use of input controllers with high degrees of freedom (DOF) as a solution to this problem. A high-DOF system allows its user to independently control multiple inputs simultaneously. This project proposes a system that accepts MIDI sliders, traditionally used in music production, to simultaneously control the α and β values of multiple user-defined semantic parameters. The MIDI controller used in this project contains five mechanical sliders and five knobs and is connected to a computer via USB. A graphical user interface, visually resembling that slider, is displayed on a monitor where a user can enter a target prompt for each slider. The same GUI can be used to add and remove parameters dynamically.

Once a prompt has been entered, the corresponding hardware slider will control how prominent that parameter is in the edited image by adjusting StyleCLIP's α value. A knob above each slider determines the β threshold. If multiple parameters are used, each global direction matrix is computed, added together, and normalized. A change in a slider will trigger this computation, generate an image, encode it, and send it to the web frontend in real-time.

The visual domain of the generated image can be selected by choosing from a list of pre-trained StyleGAN 2 (Karras, Laine, et al. 2020) models. A new random image is generated each time a user clicks on one of the models.



The Pixel Alchemist GUI in use. Four sliders are actively controlling the amount of trees and the weather in the generated image, as well as the materiality and size of the church.



StyleCLIP global directions can target very specific properties of an image, like the color of a car. The prompt "BMW" may add a kidney grille to an artificial car as it is a distinctive visual feature of that brand.

8 PRE-TRAINED MODELS

These models include faces, cars, landscapes, portrait paintings, and modern graphic design posters. Each dataset requires a few hours of preprocessing on a GPU such as an NVIDIA P100. The custom design poster dataset has been trained on 7445 hand-selected posters from professional designers and studios[1] using two NVIDIA A100 GPUs and a StyleGAN 2 architecture with adaptive discriminator augmentation (Karras, Aittala, et al. 2020).

The visual designs of the real posters are distinctively experimental and feature complex compositions with high degrees of asymmetry and diverse color palettes. I choose this type of data for two reasons. First, to foreshadow possible applications for a conventional designer, and second, to showcase the unique ability of deep neural networks to identify and reproduce latent patterns that usually require human mastery. Designing a poster according to the high standards of professional artists requires years of developing the right aesthetic intuition as it goes well beyond aligning shapes on primary axes.

Due to the high visual complexity of the already limited training data, the generated posters have low fidelity and mostly generate abstract shapes. As generative models begin to be able to reproduce typography, this quality issue should be partly resolved in the future.

For more robust datasets, StyleCLIP is capable of adjusting a large range of parameters. Using the LSUN churches dataset, one can change how prominent *gothic* architecture is in the generated building, how many *trees* are visible, or whether the facade is made of *bricks*. A user can either only change the *color* of a generated car by utilizing a high β threshold while leaving the base image untouched, or make a whole design look more like a *sports car* or a *van*. Interestingly, the system is able to distill complex design patterns. When the prompt "*BMW*" is used with a high β threshold on a car, the output only splits the front grill in two. This is called a "kidney grille", a detail that serves as the signature design feature of BMW cars.

[1] *Courtesy of: typo/graphic posters, André Felipe Menezes, www.typographicposters.com*

P. 12–13: A grid of artificial posters generated with a StyleGAN2 ADA model trained on graphic design posters.

Noteworthy is also the ability to move in the opposite direction of a text prompt by applying a negative a value. This applies the opposite effect of whatever prompt the user entered and thus offers an even more nuanced way of instructing the generator. Imagine a car designer who has gotten the feedback that their design looks too *sporty*. The designer could type in the prompt “*sports car*” and simply apply a negative value. The unique advantage I propose with Pixel Alchemist is the use of multiple parameters in parallel. A designer can dynamically add and remove any of the previously mentioned parameters and iteratively adjust each individual slider. This creates a new design process where a designer has to first, think of text prompts that serve as semantic building blocks describing the output and second, gradually adjust the impact of each of these building blocks to their liking.

9 SYSTEM INFRASTRUCTURE

Making the system accessible to a wide audience was a primary goal of this project. The software runs on a self-contained Google Colaboratory notebook which provides the code, a ready-to-use software environment, and the computer hardware necessary to run it. Part of this code is a flask server that delivers a dynamically generated HTML file to a public webserver where a user can access the graphical web interface and see generated results update in realtime.

10 BIAS

The generative models utilized in this project are biased. The neural networks used in CLIP and StyleGAN can only find and reproduce rules that were present in the training data. The more a pattern is present in the training set, the more the program will recognize it as a rule and follow it during inference. A dataset that is taken from a biased world will therefore always produce biased results unless exhaustive measurements are taken to remove these biases from the data or counter them in the learning architecture. This bias is especially visible when generating human faces. Simply entering certain races as a prompt, for example, can add glasses to a face or affect their facial expression with a low β threshold.

The potential for harmful effects caused by amplifying biases with machine learning systems also applies to this project. I attempt to address this issue by showing a highlighted note explaining this bias whenever a user selects a face-generating dataset. While this does not remove all potential for harm, it does clearly hinder users from claiming that the generated results would provide any fundamental truths about society.

The hardware setup of PixelAlchemist. A MIDI slider array connects to a laptop running the Google Colaboratory notebook.

11 DISCUSSION AND LIMITATIONS

GANs require huge amounts of data and cannot produce accurate results unless that data has a strong commonality. This is an important reason why GANs are traditionally benchmarked with human faces as they are symmetric with relatively little variance. In contrast, the LSUN car dataset cannot reliably generate realistic images of cars. It regularly creates distorted vehicles or even unrecognizable objects. The GAN model trained on posters suffered from a very limited amount of training images and an extraordinarily high degree of visual diversity. This resulted in an unstable training process and abstract results. This abstraction renders many global CLIP directions ineffective.

It is important to highlight that these limitations arise solely from the availability of data and the ability of a chosen GAN to handle diverse yet small datasets. As long as research on generative frameworks continues to improve, the same approach to global directions can likely be used to achieve dramatically better results. Very large diffusion models like DALL-E 2 have recently suggested that such issues on visual diversity and quality may have already been solved.

These problems also explain the lack of models that synthesize media that is actually useful to designers. Human faces, while very suitable to showcase the ability of StyleGAN, are hardly relevant for most design processes. However, once generative frameworks become good at producing high-quality results with small amounts of diverse data, we can directly integrate them into systems like Pixel Alchemist to study the full potential for designers with datasets that generate floor plans, 3D models, 2D graphic layouts, and even fonts.

12 OUTLOOK

Allowing people to fluently communicate the visual concepts they imagine to a computer using natural language will significantly alter the way we design with computers. Most adult humans are already capable of verbally describing ideas in great detail with a large range of rhetorical methods. Utilizing this skill for human-computer interaction could therefore drastically lower the complexity and effort needed to generate any desired image, and even text, video, audio, 3D model, or trajectory of a physical robot. <u>Traditional software has been limited either by its inability to translate fuzzy input into meaningful output or by its inability to scale as every additional generative function requires manual engineering</u>. Considering that the deep learning systems mentioned here have already solved both issues, it becomes easy to imagine that they can have a major impact on the way we design with computers.

The historic success of skeuomorphic principles in human-computer interaction suggests that future software will fluently interface with the immensely complex system we call natural human language. Eventually, we will instruct computers by referencing analogies, metaphors, and anecdotes to communicate the highly nuanced visions in our heads. Pixel Alchemist is designed to be an evolving platform that not just outlines this future but one that offers primitive yet direct access to it today.

Special thanks to André Felipe and Flávia Menezes from typo/graphic posters for supporting this project with data, my collaborator Xinyi Zhu, my advisers Prof. Kyle Steinfeld and Prof. Eric Paulos, my father, mother, and sisters for their continued support, and my beloved friends from Berkeley who have been my most valuable resource over the past two years.

REFERENCES

Chen, Xi, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets." Advances in Neural Information Processing Systems 29. https://proceedings.neurips.cc/paper/2016/hash/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Abstract.html.

Dhariwal, Prafulla, and Alexander Nichol. 2021. "Diffusion Models Beat GANs on Image Synthesis." Advances in Neural Information Processing Systems 34 (December): 8780–94.

Esser, Patrick, Robin Rombach, and Bjorn Ommer. 2020. "A Disentangling Invertible Interpretation Network for Explaining Latent Representations." 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr42600.2020.00924.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Nets." Advances in Neural Information Processing Systems 27. https://proceedings.neurips.cc/paper/5423-generative-adversarial-nets.

Karras, Tero, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. "Training Generative Adversarial Networks with Limited Data." Advances in Neural Information Processing Systems 33: 12104–14.

Karras, Tero, Samuli Laine, and Timo Aila. 2019. "A Style-Based Generator Architecture for Generative Adversarial Networks." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4401–10.

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. "Analyzing and Improving the Image Quality of Stylegan." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8110–19.

Kazemi, Hadi, Seyed Mehdi Iranmanesh, and Nasser Nasrabadi. 2019. "Style and Content Disentanglement in Generative Adversarial Networks." In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), 848–56.

Kingma, Diederik P., and Max Welling. 2013. "Auto-Encoding Variational Bayes." arXiv [stat.ML]. arXiv. http://arxiv.org/abs/1312.6114v10.

Lin, Zinan, Kiran Koshy Thekumparampil, Giulia C. Fanti, and Sewoong Oh. 2019. "InfoGAN-CR: Disentangling Generative Adversarial Networks with Contrastive Regularizers." https://openreview.net/forum?id=oo384_Owakk.

Park, Taesung, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. "Semantic Image Synthesis with Spatially-Adaptive Normalization." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2337–46.

Patashnik, Or, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. "Styleclip: Text-Driven Manipulation of Stylegan Imagery." In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2085–94.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 18--24 Jul 2021. "Learning Transferable Visual Models From Natural Language Supervision." In Proceedings of the 38th International Conference on Machine Learning, edited by Marina Meila and Tong Zhang, 139:8748–63. Proceedings of Machine Learning Research. PMLR.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. "Hierarchical Text-Conditional Image Generation with CLIP Latents." arXiv [cs.CV]. arXiv. http://arxiv.org/abs/2204.06125.

Ramesh, Aditya, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 18--24 Jul 2021. "Zero-Shot Text-to-Image Generation." In Proceedings of the 38th International Conference on Machine Learning, edited by Marina Meila and Tong Zhang, 139:8821–31. Proceedings of Machine Learning Research. PMLR.

Sohl-Dickstein, Jascha, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 07--09 Jul 2015. "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics." In Proceedings of the 32nd International Conference on Machine Learning, edited by Francis Bach and David Blei, 37:2256–65. Proceedings of Machine Learning Research. Lille, France: PMLR.

CODE

The code is stored and hosted at GitHub and can be found at: https://github.com/titusss/Pixel Alchemist.git