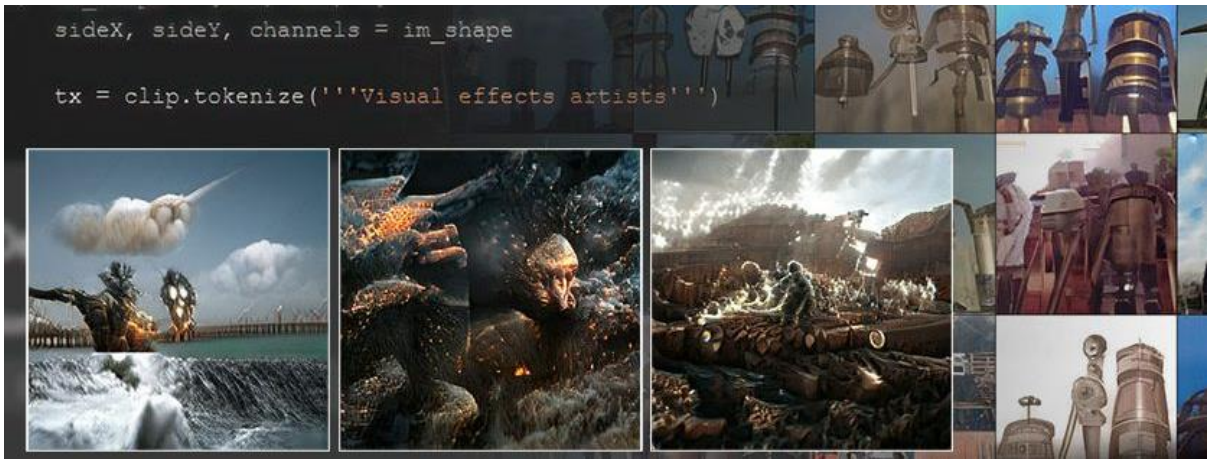
The Future of AI Image Synthesis

By Martin Anderson



First published March 1st, 2021 at:
https://rossdawson.com/futurist/implications-of-ai/future-of-ai-image-synthesis/
Web-archived version

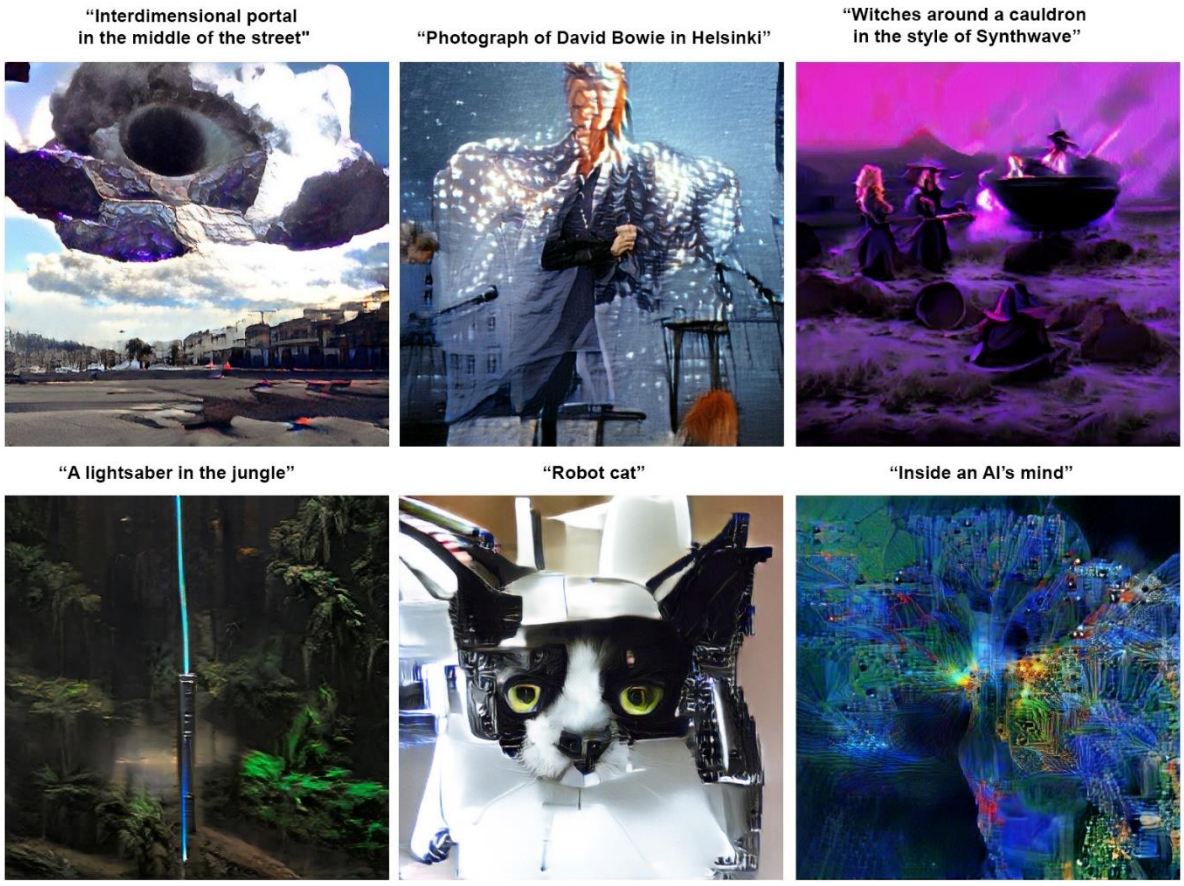
Shortly after the new year 2021, the Media Synthesis community[1] at Reddit began to become more than usually psychedelic.

The board became saturated[2] with unearthly images depicting rivers of blood[3], Picasso's King Kong[4], a Pikachu chasing Mark Zuckerberg[5], Synthwave witches[6], acid-induced kittens[7], an inter-dimensional portal[8], the industrial revolution[9] and the possible child of Barack Obama and Donald Trump[10].

These bizarre images were generated by inputting short phrases into Google Colab notebooks (web pages from which a user can access the formidable machine learning resources of the search giant), and letting the trained algorithms compute possible images based on that text.

In most cases, the optimal results were obtained in minutes. Various attempts at the same phrase would usually produce wildly different results.

In the image synthesis field, this free-ranging facility of invention is something new; not just a bridge between the text and image domains, but an early look at comprehensive AI-driven image generation systems that don't need hyper-specific training in very limited domains (i.e. NVIDIA's landscape generation framework GauGAN [on which, more later], which can turn sketches into landscapes, but only into landscapes; or the various sketch>face Pix2Pix projects, that are likewise 'specialized'[11]).



Example images generated with the Big Sleep Colab notebook[12]. Above, the input text, below, the result. (https://old.reddit.com/r/MediaSynthesis/)

The quality is rudimentary at best, like a half-formed remembrance of dreams, and the technology clearly nascent; but the long-term implications for VFX are significant.

So where did these technologies come from?

A New Wave of Multimodal Neural Networks

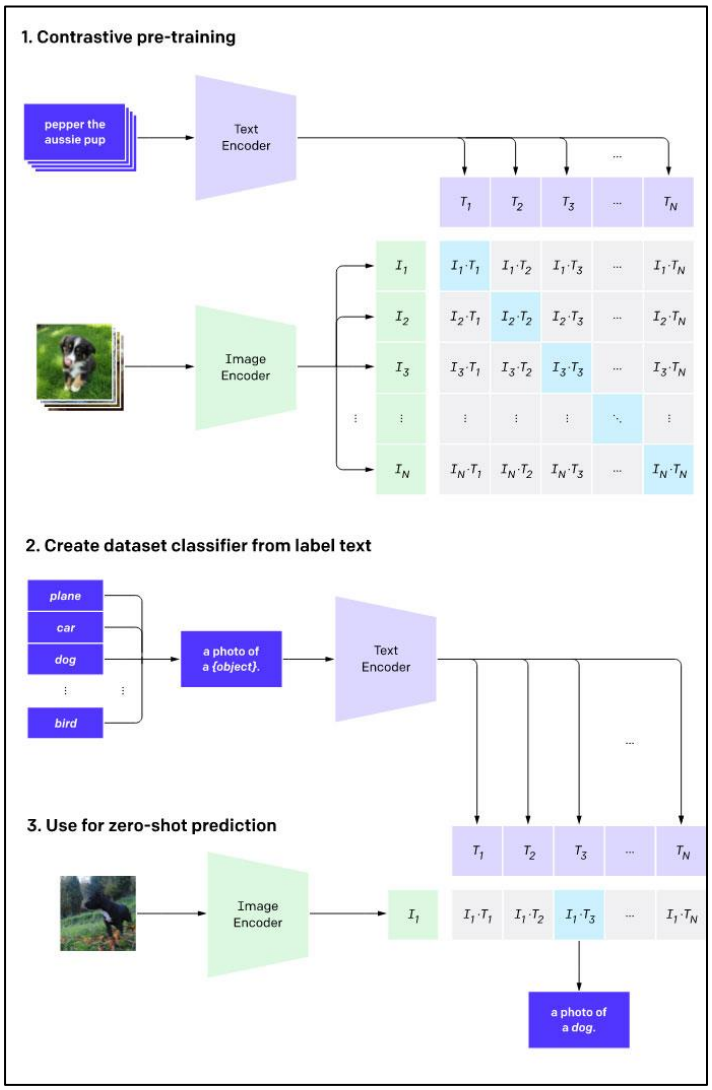
In early 2021 artificial intelligence research laboratory OpenAI, fresh from the previous year's press sensation[13] regarding the power of its GPT-3 autoregressive language model , released details of two new multimodal[14] neural networks capable of straddling two diverse computing domains: text and imagery.

OpenAI: CLIP

CLIP[15] (Contrastive Language-Image Pre-Training) is a zero-shot[16] neural network that evaluates the relevance of a text snippet for any given image without the same optimization that preceding networks had to undertake in order to achieve this.

CLIP is trained not on a fixed and pre-labeled dataset, but on internet-obtained images (image content), which it attempts to pair with the most apposite of 32,768 randomly-sampled text snippets.

With this approach, CLIP breaks away from previous systems which – the company claims – are typically over-optimized for benchmark tests, and have inferior general applicability to 'unseen' image datasets.

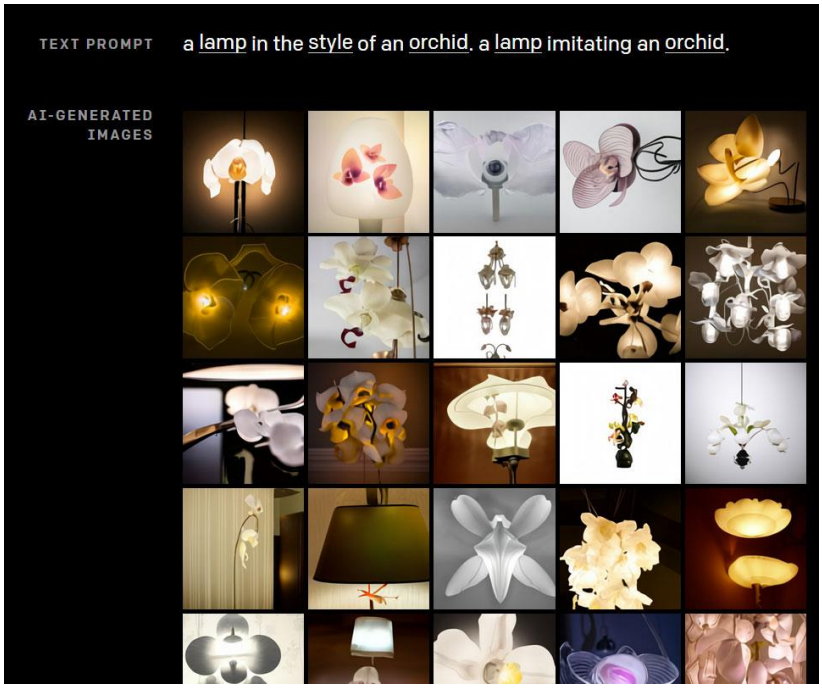


CLIP's architecture trains a text encoder and an image encoder to find matching image/caption pairs in its source dataset. As the loss values decrease (i.e. it gets better at doing this), the model gains the ability to perform 'zero shot' classification.

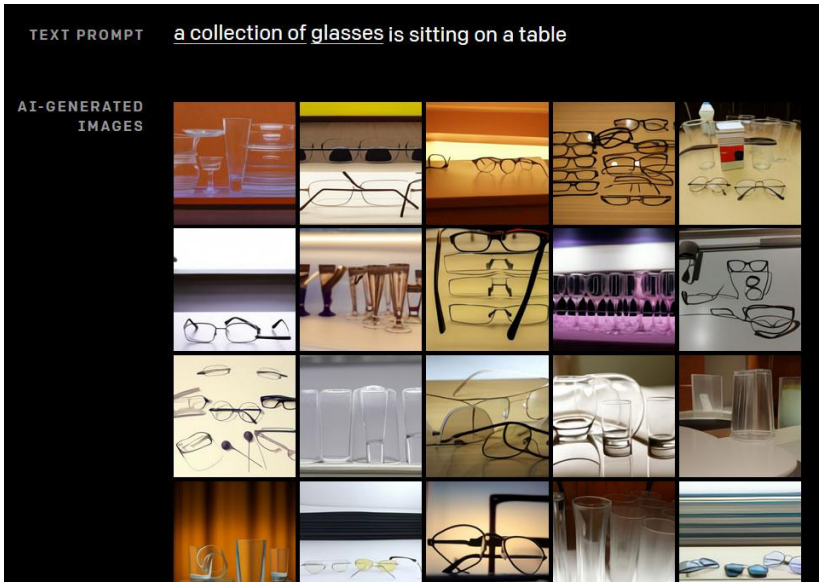
In tests, CLIP was able to match the performance of ImageNet's benchmark ResNet50[17] challenge without the benefit of referring to any of the dataset's 1.28 million labeled images – an unprecedented advance. Subsequently OpenAI made CLIP accessible via a Colab.

OpenAI: DALL-E

DALL-E[18] uses 12 billion out of the 175 billion parameters of the GPT-3 dataset to generate text-image pairings capable of producing relatively photorealistic images — depending on the availability of image source material that will be summoned up from the data set by the text prompt:



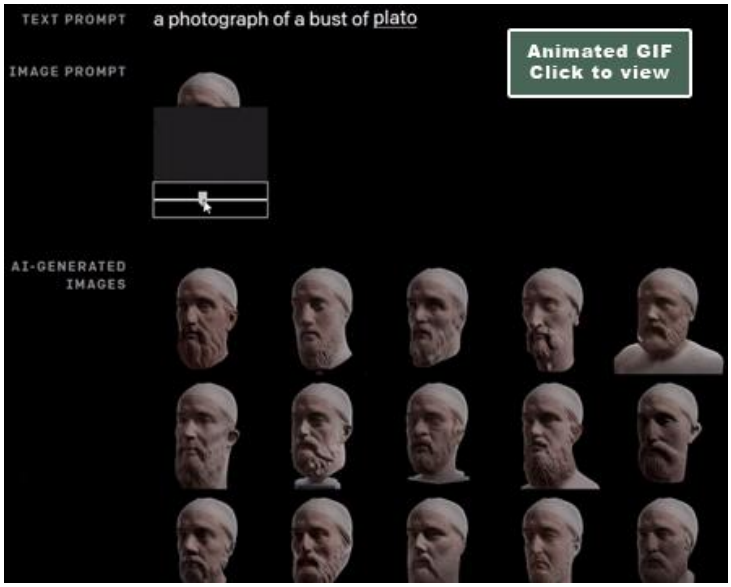
The system is prone to the occasional semantic glitch:



The ambiguity of the term 'glasses' shows up in DALL-E's output. (https://openai.com/blog/dall-e/)

As the creators note, DALL-E is disposed to confuse colors and objects when the number of requested objects in the text prompt are increased. Also, rephrasing the prompt into grammatically complex or unfamiliar forms can kill the model's interpretive capabilities (for instance, where the object of the possessive is unclear).

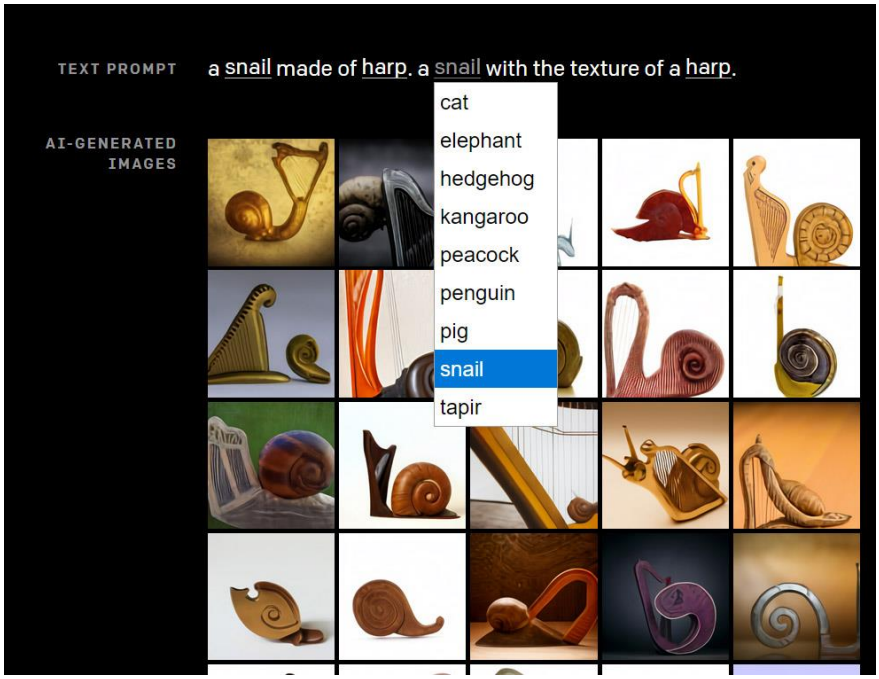
If the web-derived image data for a text prompt is frequent and detailed[19], it's possible to create a genuine sense of instrumentality, in the style of a 3D modelling interface, albeit not remotely in real time:



*Dall-E has high instrumentality, even if consistency remains a problem. Here the AI reconstructs a bust of Plato using a partial image at various rotations, and a text prompt. (*https://openai.com/blog/dall-e/*)*

Presumably the inclusion of the previously-generated image in a sequence of this nature would improve continuity across the frames; but in the above example, as far as we know, DALL-E is 'starting fresh' with each image, based on the text prompt, and on what it can see of the top of Plato's head.

DALL-E's transformative and interpretive capabilities are currently capturing the imagination of image synthesis enthusiasts, not least because the systems have been quickly commoditized by researchers and enthusiasts (see 'Derivative Works' below):



Within the limits of the most productive subjects in the contributing datasets, DALL-E can weld some pretty incongruous elements together.

DALL-E builds on OpenAI's Image GPT generator[20], a proof-of-concept transformer model capable of finishing incomplete images (albeit with a wide range of possible responses).

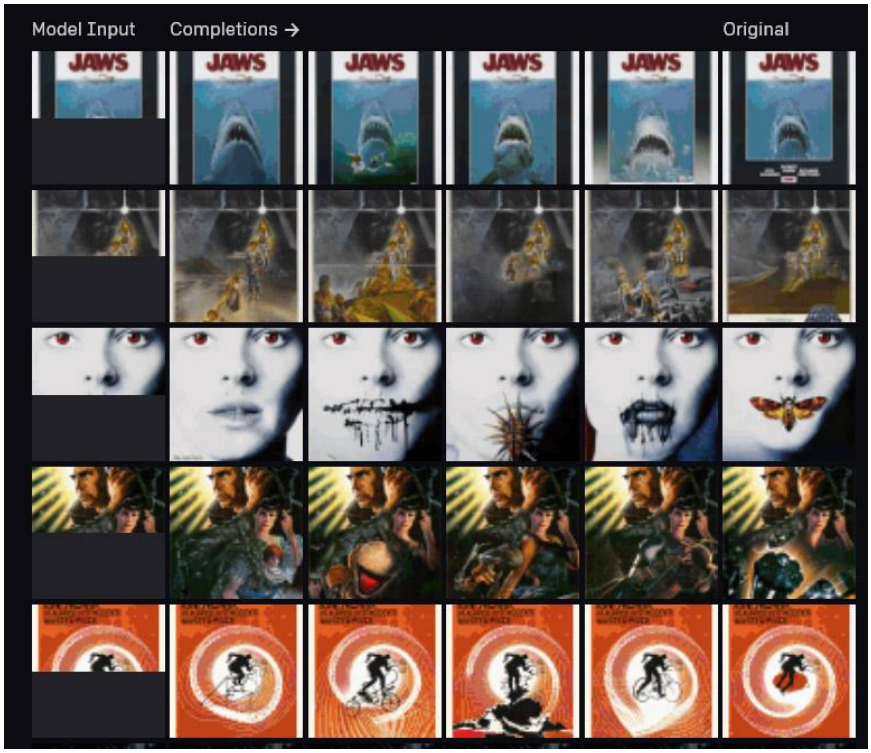


Image GPT attempts to 'complete' famous movie posters. The model has a high computational cost, and is trained at low resolutions of 32×32, 48×48, and 64×64. To further improve performance, the system uses a pared-down 9-bit color palette.(https://openai.com/blog/image-gpt/)

A simplified third-party adaptation of DALL-E's source[21] is available at GitHub[22], with an accompanying Colab[23].

Derivative Works

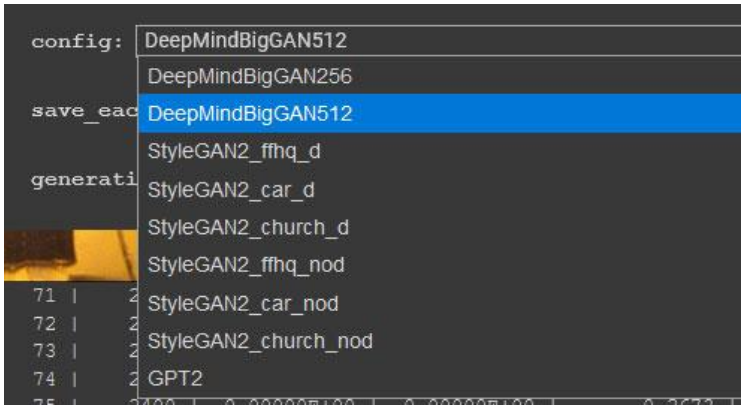
Sadly, there is no general access to these extraordinary transformative capabilities; OpenAI will only release the dVAE decoder and convolutional neural network (CNN) encoder from DALL-E, and there is no official plan to release the transformer model that powers the higher quality images seen in the 'official' posts (see above). However, a comment from an OpenAI developer on GitHub in February states that the company is 'considering' API access to DALL-E, which would extend the interpretive power of third-party image synthesis systems.

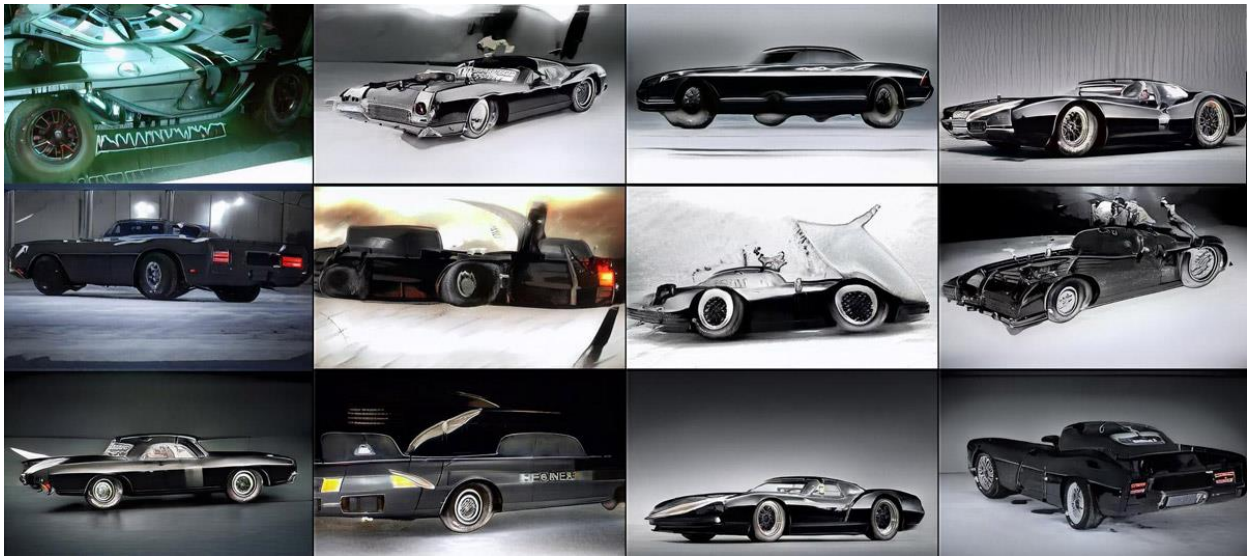
In the meantime, developers have attempted to recreate the functionality of DALL-E with such APIs as are available. Though the results are not as delineated or cohesive as those produced by the cloistered-off DALL-E, they hint at what can be achieved by leveraging text-to-image technologies that are powered by wide-ranging image datasets.

Clip Glass

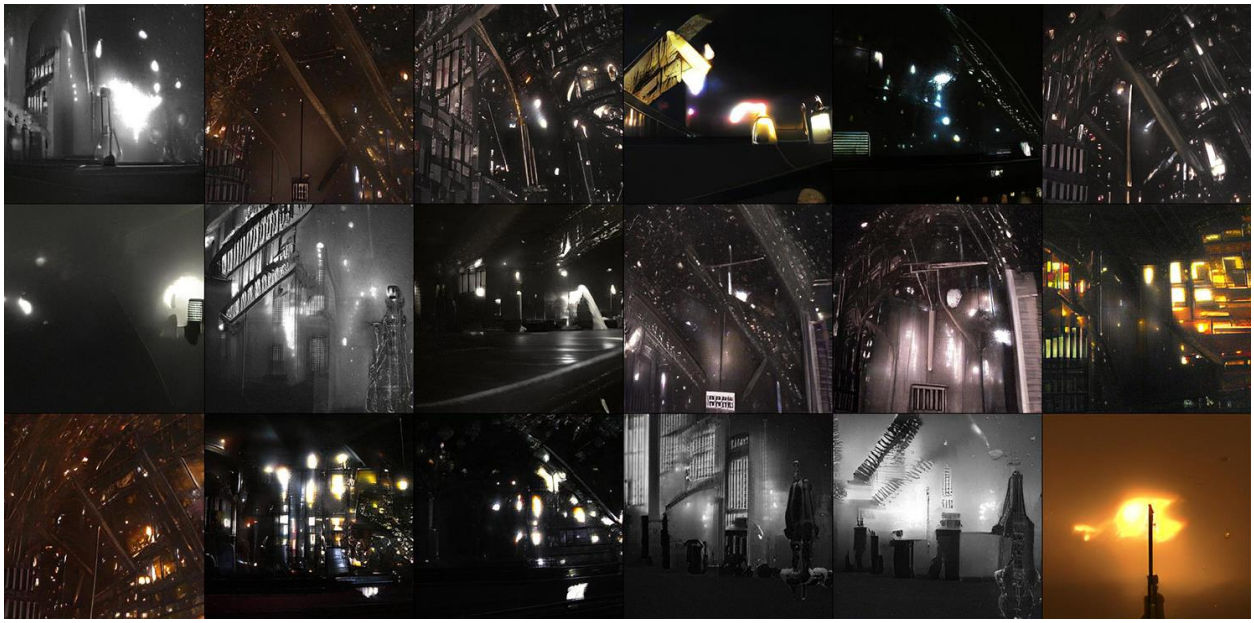
Research out of the University of Pisa recently produced CLIP-GlaSS, which builds on OpenAI's framework to incorporate BigGAN[24], StyleGAN2[25] and GPT-2[26] as the catalysts for text-prompted image generation.

CLIP-GlaSS operates across a set of distinct GAN frameworks, of which the most generalized is DeepMind's BigGAN. The other dedicated GANs offer domains for face generation, a car, a church, and for text generation via GPT-2, as well as higher quality variants (for the image-based GANs).





CLIP-GlaSS attempts to recreate 'the batmobile' with the StyleGAN2 'car' set.



A cathedral at night in the style of Blade Runner' – *CLIP-GlaSS uses StyleGAN2's 'church' setting to approximate scenery similar to the cult 1982 sci-fi film. The 'church' and 'car' sets will occasionally feature people, whereas the StyleGan2 ffhq (facial feature) set is sharply focused on faces, with minimal transformations outside of the portrait area.*

The overnight popularity of CLIP-GlaSS may have something to do with the very accessible Colab notebook that the Italian researchers have made available.

Unlike Big Sleep (see below), the quality of CLIP-GlaSS images usually improves when allowed to train a little longer. But since these are 'zero shot' generators, they tend to reach their quality ceiling (a perceived 'minimum loss') quickly.

Big Sleep and Aleph2Image

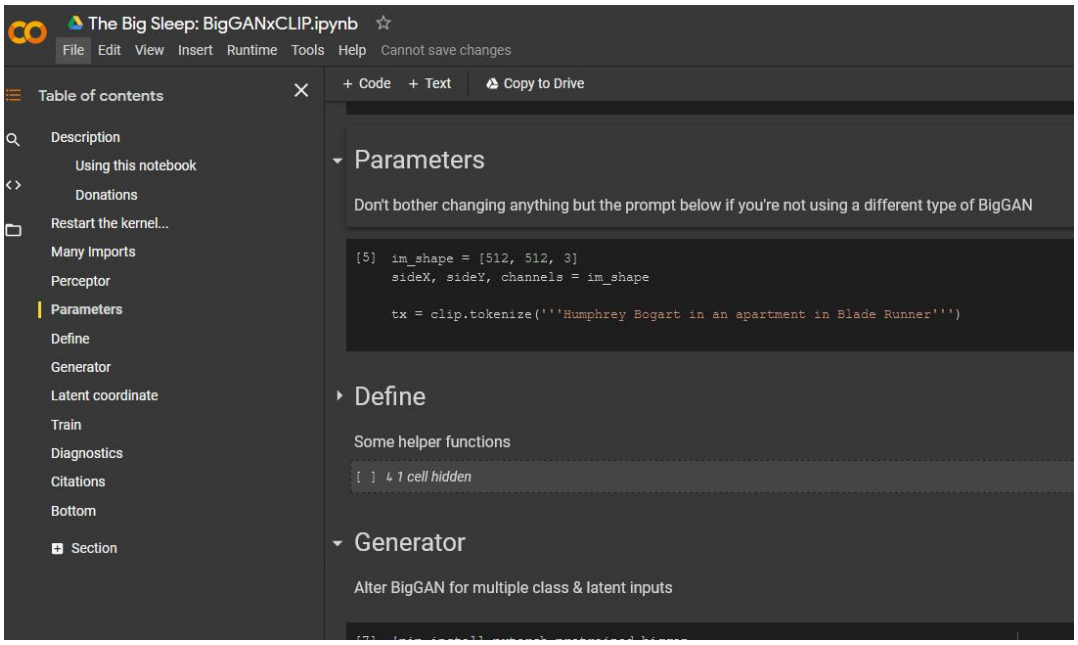
Programmer[27] and University of Utah PhD student[28] Ryan Murdock maintains a number of CLIP-derived projects, including Aleph2Image and BigSleep – the latter of which has so entranced the Reddit media synthesis group.

Images derived in the BigSleep Colab from the text prompt 'Visual Effects Artists'. Besides the CGI sand-storm and the studio lights in the third image, are all these pictures so pastoral and painterly because the prompt contains the word 'artist'?

'I created BigSleep and Aleph because I love making neural art,' Murdock told me. 'and I wanted to see if it was possible to get a backdoor around DALL-E not being released.'

BigSleep interprets 'A cathedral in the style of Blade Runner'. *At this stage of assimilation, and at this resolution, it is difficult to say whether the BigGAN powering BigSleep is chiefly drawing on production sketches or actual stills or footage, since the results in themselves are similar in quality to production concept art.*

BigSleep has the most user-friendly Colab notebook out of any of the current offerings:



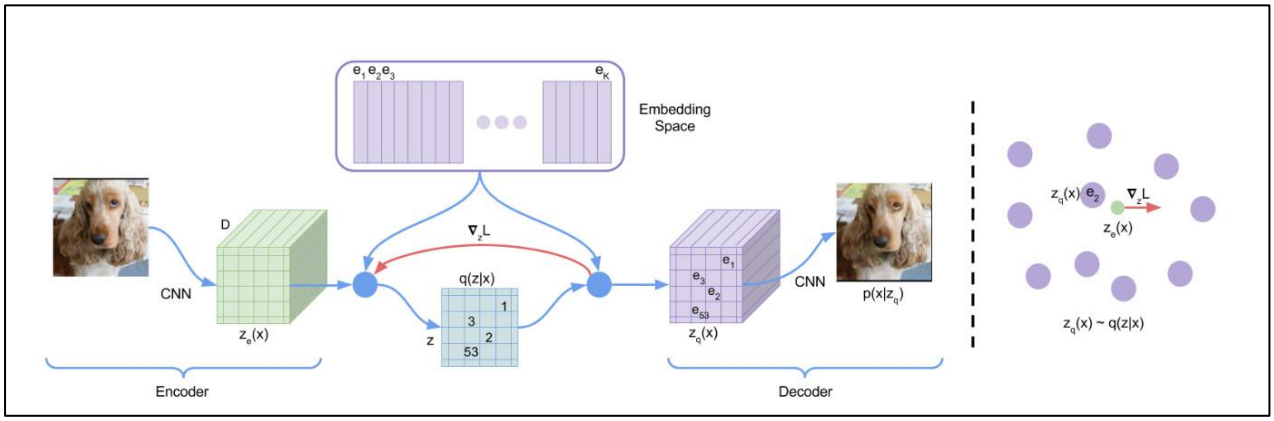
To implement an image in BigSleep, you need to press two buttons in the Colab and provide a text prompt.

In the absence of the GAN tooling that OpenAI has not released for DALL-E, Murdock has connected CLIP's predictive capabilities to the generative capabilities of BigGAN.

BigSleep has two specific eccentricities: firstly, BigGAN specializes in 1000 particular ImageNet categories[29] (a great many of which are animal species). A prompt that includes any of those categories is likely to fare better than one outside that scope.

Secondly, BigSleep's 'seed' image is always a random picture of a dog breed, and sometimes the iterations have difficulty letting go of that canine element. Subsequently, people in BigSleep images can end up wearing an unfashionable number of fur coats (see the 'Humphrey Bogart' images later in the article).

A more recent project from Murdock is Aleph2Image[30], which exploits DALL-E's Vector Quantised-Variational AutoEncoder (VQ-VAE).



Conceptual map of the VQ-VAE, with the embedding space visualized on the right. The encoder and decoder share the same dimensional space, an aid to iteratively lowering reconstruction loss.
(https://arxiv.org/pdf/1711.00937.pdf)

The text-prompted results are more impressionistic even than BigSleep, but tend to have more visual consistency:

(https://old.reddit.com/r/MediaSynthesis/)

Aleph2Image's Colab is currently less user-friendly than BigSleep's, but the author promises a better user experience in future versions.

The Reddit Media Synthesis group maintains a list of other image synthesis Colabs and GitHub repositories that leverage CLIP for image generation, though some have been broken by framework updates that the older notebooks don't support.

Murdock believes that the use of CLIP+GAN deployments like this in production environments, even for general design exploration, is 'pretty rare' at the moment, and concedes that there are some hard roadblocks to overcome in developing these systems further.

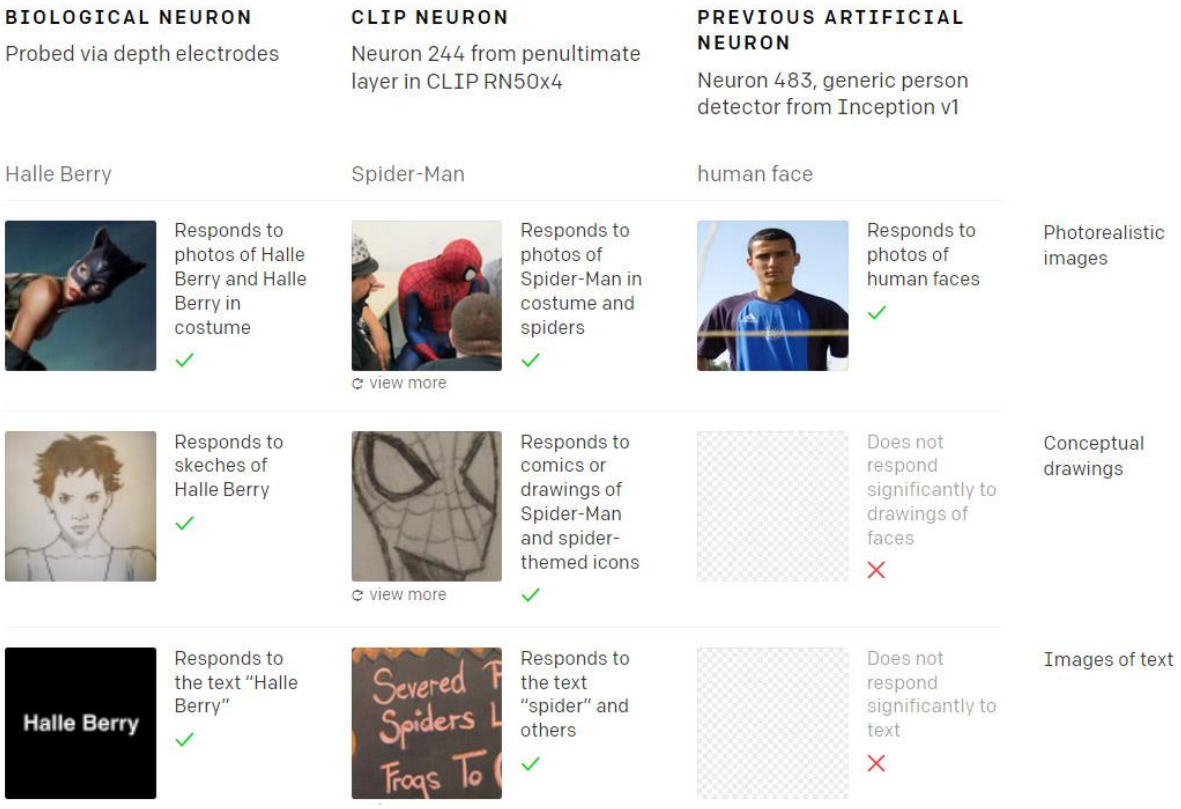
'I think hardware is a huge bottleneck,' he told me. 'I doubt it's possible for anyone with any software to emulate something like CLIP without at least a number of GPUs and plenty of time.

'But data is the most serious bottleneck and the unsung hero of most advances. Software may be a bottleneck, but it already has the brunt of people's attention.'

Navigating the Strange Obsessions of CLIP

Even where text is not directly involved in image synthesis, as it is with the CLIP Colabs, it's deeply bound up in the semantic segmentation that defines objects and entities, and which underpins many of the generative systems described here.

CLIP's neural network is a perfect construct trained by data from an imperfect world. Since the world is very interested in Spider-Man, CLIP's multimodal neurons will fire for even the vaguest reference to your friendly neighborhood web-slinger:



In 2005 researchers discovered a human neuron in test subjects that responded across an extraordinary range of possible triggers regarding the actress Halle Berry. Likewise, CLIP has, among other eccentricities, a multimodal 'Spider-Man' neuron that is easy to activate, even through very oblique references.
(https://openai.com/blog/multimodal-neurons/)

Leaving aside whether CLIP should be more interested in Spider-Man than Shakespeare or Gandhi, it should be considered that image synthesis systems which rely on it will also have to suffer CLIP's eccentricities.

Neural networks that have been trained by the internet are prone to demonstrate obsessions and biases in accordance with the volume and breadth of material that they might find on any particular topic. The more pervasive an entity is in an image synthesis system's contributing dataset, the more likely it is to appear unbidden, or to be triggered by apparently unrelated words or images:

In 2020 the face of actor Ryan Gosling appeared[31] *randomly in the output of Gigapixel's AI-based image upscaling program, whose machine learning models had been partially trained on the widely-used Celeb A dataset*[32].

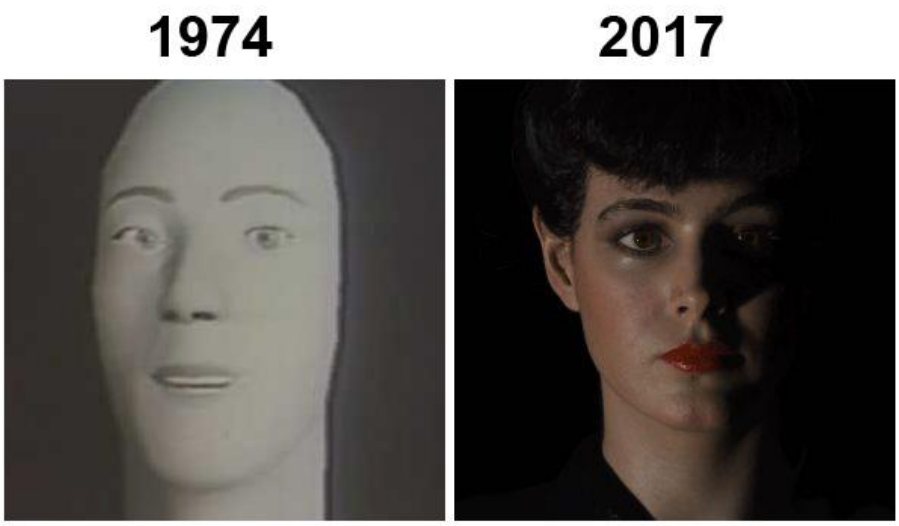
If you experiment long enough with CLIP-based text-to-image tools, you'll quickly build up a picture of its 'interests', associations and obsessions. It's not a neutral system – it has a lot of opinions, and it didn't necessarily form them from the best sources, as OpenAI recently conceded in a blog post.

Therefore, in terms of the development of professional VFX image synthesis systems, CLIP implementations should perhaps be seen as a proof-of-concept playground for future and more rigorous data architectures.

Pure Image Synthesis in Professional VFX?

However, it doesn't seem that VFX professionals are likely to either feel threatened by the psychedelic ramblings of CLIP-based GAN image generators such as BigSleep, or even necessarily feel able to satisfy their curiosity and enthusiasm about image synthesis, so long as the instrumentality is so constrained and the ability to control the output is so limited.

Perhaps Amara's Law[33] is applicable here; the late American futurist and scientist stated *'We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run.'*



How CGI played out in terms of Amara's law – left, an experimental parametric facial animation from the University of Utah in 1974; forty-three years later, MPC's recreation of the youth of actress Sean Young, for Blade Runner 2049 *(2017).*

Continuing the analogy with CGI, we can see that the evolution of disruptive VFX technologies is sporadic and inconsistent; and that new technologies often languish as industry curiosities and niche pursuits until an unpredictable breakthrough renders a cherished VFX role extinct, and makes it necessary to catch up and adapt, as stop-motion master Phil Tippett did during the making of *Jurassic Park* (1993).

In the second regard – let's take a look at new approaches that could bring a greater measure of control and instrumentality to pure image synthesis, facilitating the development of new tools, and perhaps even new professions within the sector.

The Power and Pitfalls of Paired Datasets

"Once you get to the stage where you're no longer even necessarily designing a tool, but you're asking the AI to do something for you, and it does that…that opens up a lot of possibilities."

Valentine Kozin is the lead technical artist at UK games studio Rare[34]. His passion for the possibilities that machine learning can open up for VFX pipelines led him to host a popular presentation[35] on practical deep learning for the industry in December 2020.

"With GPT-3," he told me. "you can even use it for game systems. Instead of codifying the rules of what happens when a flaming torch hits a body of water, you don't have to code in the fact that the fire gets extinguished – you can just ask the AI 'What happens if I combine fire and water?', and the AI will tell you 'The fire gets extinguished', because it has some of that sort of interaction knowledge."

One of Kozin's key research interests at the moment is the potential of paired datasets to help generative adversarial networks transform sketches (or other forms of rudimentary, indicative imagery) into fully-realized, AI-generated images.

Paired dataset training input for the video game Sea Pirates. In the first two columns, the contributing 'ground truth' – renders and generated sketches. In the third column, the synthesized images produced by the neural network. (Practical Deep Learning for VFX and Technical Artists – https://www.youtube.com/watch?v=miLIwQ7yPkA)

To accomplish this, two datasets are needed for training – one where the character faces are traditionally-rendered CGI, and one where the same faces appear in the form of crude sketches. Generating the latter can be done through open source GitHub projects such as PhotoSketch[36].

*PhotoSketch uses AI-driven boundary detection to infer contours for a sketch-like equivalent to an input photo. Output from the project's trained models and default algorithms can help train new machine learning systems to recognize and categorize domains, objects and segmented sections of objects (such as faces), and subsequently associate those labels with images. (*https://github.com/mtli/PhotoSketch*)*

But the current gold standard for sketch>image inference is Pix2Pix[37], a project out of the Berkeley AI Research (BAIR) Laboratory, which uses conditional adversarial networks[38] as a Rosetta stone for image translation.

From Berkeley's 2018 paper. Pix2Pix learns and assimilates the mapping from input to output images, but also provides a loss function to train the mapping. The result is a generic transformative capability between very different qualities and categories of image types, though usually within a target domain ('houses', 'cats'. 'people', etc.). (https://www.tensorflow.org/tutorials/generative/pix2pix)

Contributing Datasets That Know Too Little or Too Much

Though the potential for previz and asset generation using these techniques is obvious, the path to true instrumentality and fine-grained control for high-resolution, photorealistic output is almost as constrained by considerations of data design and architecture as by current hardware limitations for neural networks.

For example, casting Humphrey Bogart in *Blade Runner* is a predictably impressionistic experience in the new range of text-to-image Colab playgrounds such as BigSleep and Aleph2Image.

Various BigSleep prompts regarding 'Humphrey Bogart in Blade Runner'. Note that he is wearing a fur coat in the second image, and that fur recurs in other images- residue from the fact that each BigSleep render starts off with an image derived from a dataset of dog breeds!

Clearly the 1.3 million images in the ImageNet dataset powering BigGAN have enough material to understand Bogart's basic facial lineaments and iconography, and enough production sketches, videos and stills from the 1982 movie for the Colabs to synthesize the style of *Blade Runner* quite well. Why then, is it so hard to obtain coherent imagery?

Partly the problem is semantic, as many image synthesis enthusiasts report of the popular new T2I Colabs: if you prompt the AI with 'A man in a dark suit', you could, despite the 'obvious' intent of the phrase, end up with a) a courtroom scene b) a single-occupant apartment made out of gabardine or c) Darth Vader.

Apart from the aforementioned issues with the frequency of 'popular' high level concepts in the contributing database/s, this is a problem linked to inference of intent in natural language processing (NLP); and even, arguably, to the field of sentiment analysis. Improvements in this respect are likely to come out of upstream NLP research that has loftier aims (or at least more immediately commercial ones).

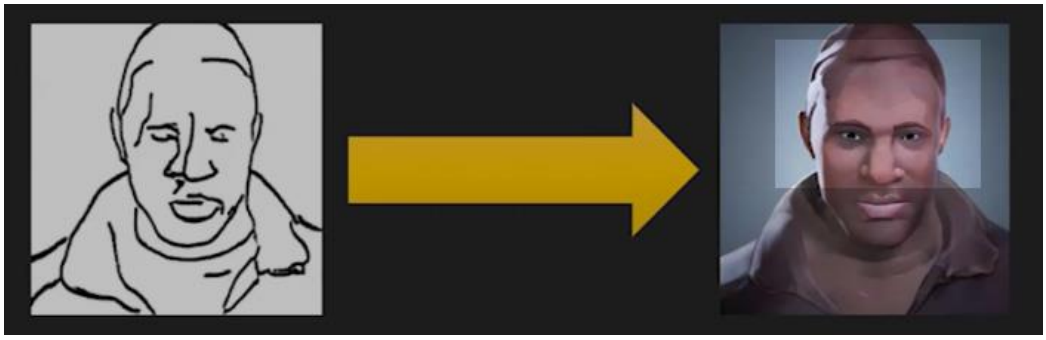
But in terms of developing VFX-centered machine learning workflows, the more pressing issue is one of the scope and breadth of the contributing image datasets.

If your generator uses vast and generalized data repositories such as ImageNet or the Common Crawl[40], you'll be summoning up just a small subsection of available imagery related to your intent – and these representative datasets do not have enough space or remit to exhaustively catalog every Humphrey Bogart or *Blade Runner* image on the internet.

Therefore why not curate your own Bogart and *Blade Runner* dataset, and train the model exclusively on these?

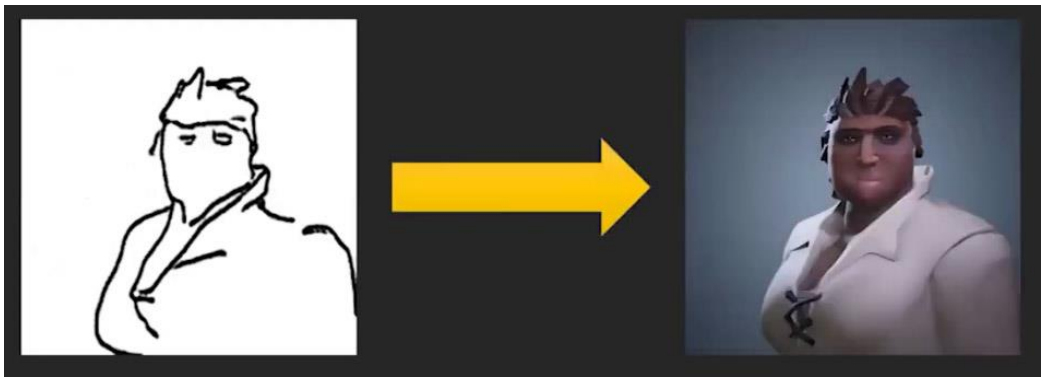
Brittle Image Synthesis Output From Overfitted Datasets

As Valentine Kozin notes, the excessive specificity of such datasets, whilst improving image quality, makes for brittle and inflexible output[39], without the 'random seed' that characterizes the more vivacious output of BigSleep and its fellow T2I Colabs:



Here Kozin has drawn a face with a high forehead, but the paired dataset model he has trained (with characters from the video-game) features faces that have exclusively low foreheads, and we see the ghostly artifact of where the AI thought the forehead 'should' have ended.

In the end Kozin was forced to compromise on a more generalized and diversely trained model, resulting in an algorithm that's more robust and interpretive; for instance, it will produce faces with mouths even where the user did not draw a mouth:



But Kozin observes that now the model is incapable of drawing a face *without* a mouth, if the user should desire that, and that, in a sense, this restored 'flexibility' comes with its own set of frustrating constraints.

Potential Data Architectures for Image Synthesis

"The biggest unsolved part of what OpenAI are doing with things like GPT-3," Kozin says. "is that it's very much a static knowledge bank. It's got a thousand, two thousand character attention window…it's not an agent that can dynamically learn more about the task and then adapt, so that there's a 'snapshot' that we can apply to different things.

"The solution to that would be AI architectures that are able to have this large data set, but then are also able to learn more about the *specifics* of the problem they're trying to solve, and able to apply both that generalized and specific knowledge. And then, that's kind of getting close to human level creativity, I suppose.

"I think that's a model that we don't really have yet. Once we do have it, that's probably going to be very powerful."

Ryan Murdock believes that for the moment, using far-ranging, high-volume image sets remains necessary for generating coherent and novel structures, and that training a separate GAN for every element of a potential image (movement, pose structure, lighting, etc.) would be labor-intensive.

'The benefit of the huge data,' he says, 'is allowing the system to fill in gaps through its large associative knowledge. Hopefully it'll be possible someday to get a very specific [image], but for now I think the closest thing is to guide CLIP by giving it a sort of structure.

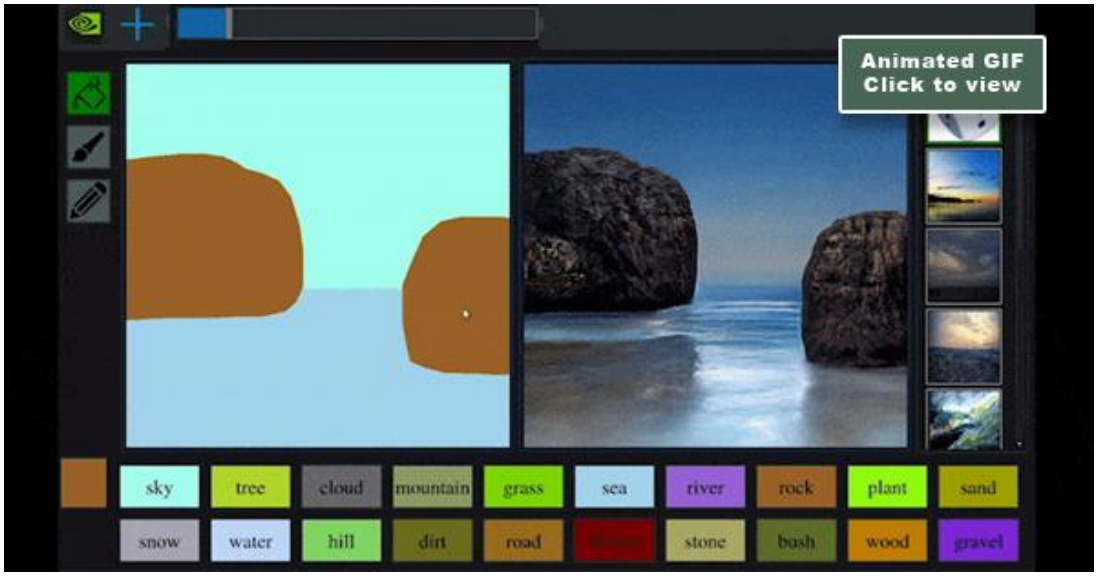
'You could basically just give any image a structure and then have CLIP fill in the rest, either altering the image or giving it a specific area to fill, allowing it to have some potentially useful constraint.

'I haven't done this yet, because I'm super busy, but it's the type of thing that is surprisingly low-hanging fruit for putting humans closer into the loop.'

A number of research projects are currently addressing the challenge of training generative adversarial networks with more limited data, without producing brittle results or overfitting. These include initiatives from NVIDIA[40], MIT[41], and the MetaGAN few-shot learning study[42].

Landscape Synthesis

In 2019 NVIDIA's research arm released details of GauGAN, an AI art application that allows the user to develop photorealistic landscapes from crude sketches.



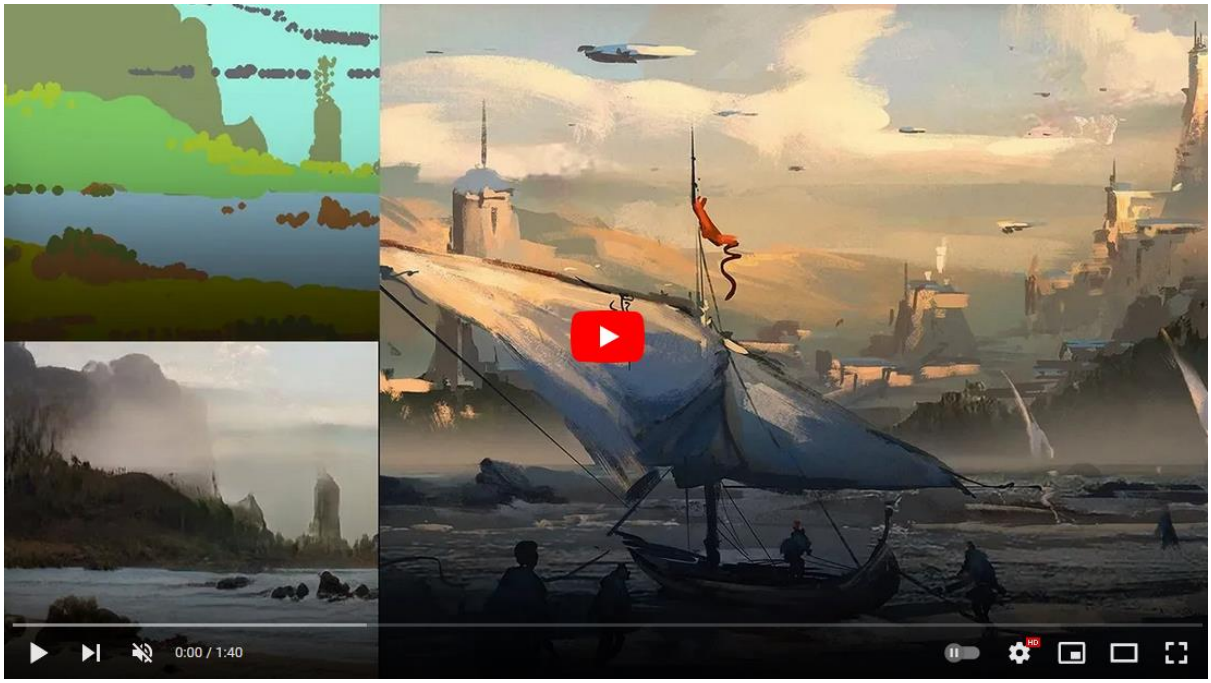
(https://blogs.nvidia.com/blog/2019/03/18/gaugan-photorealistic-landscapes-nvidia-research/)

GauGAN is arguably a perfect example of the kind of topically-constrained media synthesis system that can be developed by training paired datasets on a small number of specific domains – in this case, landscape-related imagery, pair-trained against sketch imagery to develop an interpretive interchange between crude daubs and photo-real images.

Though it's not clear how one can create anything other than landscape imagery with GauGAN, NVIDIA has promoted[43] VFX concept artist and modeler Colie Wertz's[44] use of the tool as the basis of a more complex spaceship sketch.

(https://blogs.nvidia.com/blog/2019/07/30/gaugan-ai-painting/)

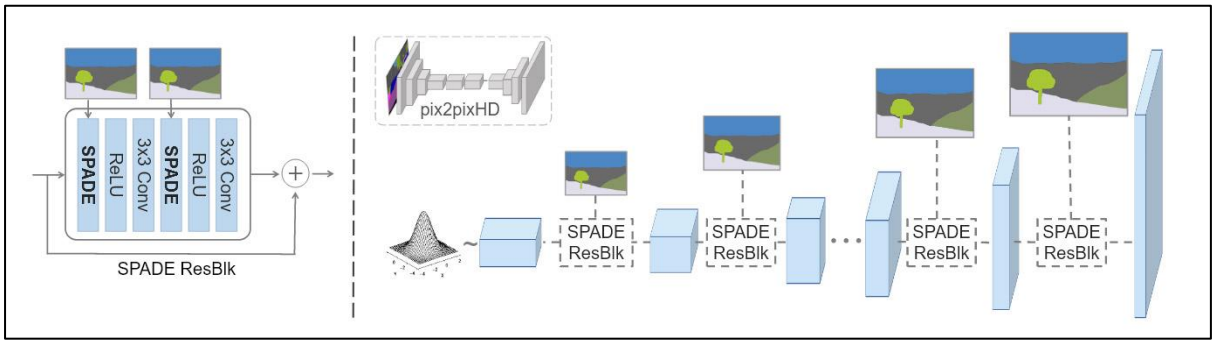
A later promotional video features output created with GauGAN by VFX practitioners from Lucasfilm, Ubisoft and Marvel Studios:



In the live interactive demo, GauGAN allows the user to define a segmentation map, where colors are anchored to different types of output, such as sea, rocks, trees, and other facets of landscape imagery.

The map can either be drawn directly into the interface, or generated elsewhere by the user and uploaded as a direct input into the synthesis system.

The neural network in the SPADE generator (see 'Google's Infinite Flyovers' below) controls the layer activations via the segmentation mask, bypassing the downsampling layers used by the Pix2PixHD model (pictured upper center in the image below), achieving better results and improved performance with fewer parameters.



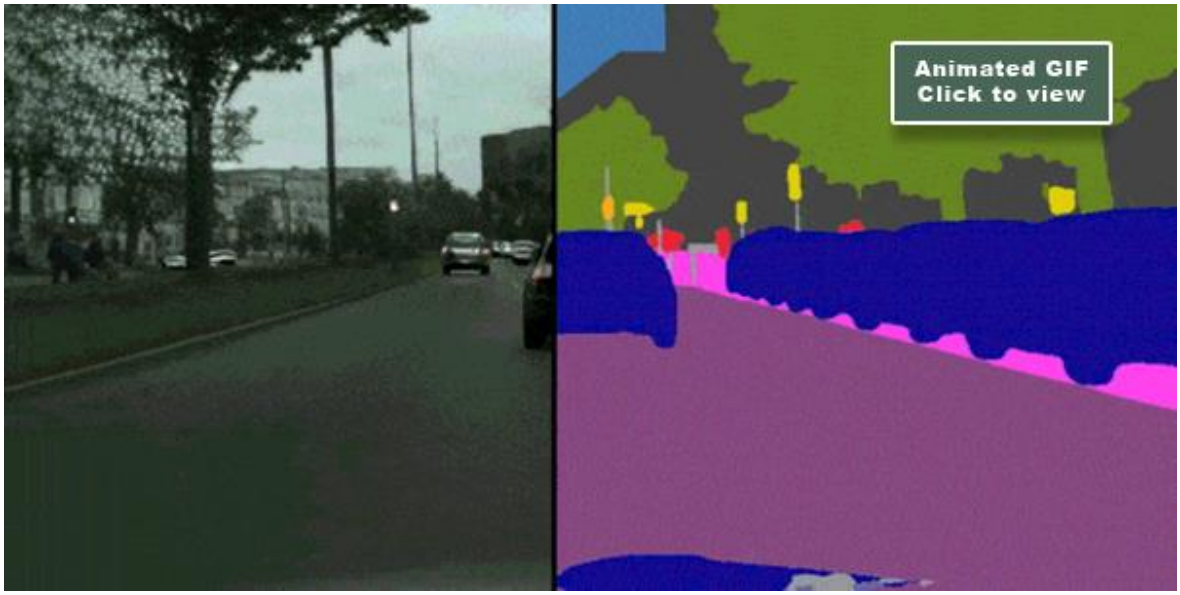
(https://arxiv.org/pdf/1903.07291.pdf)

Extended Interpretation in Image and Video Synthesis

Semantic Segmentation: The Key to AI World Building?

The year before GauGAN was released, NVIDIA showed the potential of semantic segmentation to generate randomized urban street scenes, rather than idyllic pastures:

This system, called Vid2Vid[45], was trained[46] on hundreds of hours of car-level street footage, and then against a separate segmentation neural network, in order to derive object classes for the contents of 'fictional' street videos – cars, roads, markings, street lamps, and other environmental details.



Deriving semantic segmentation from real world footage. (https://github.com/NVIDIA/vid2vid)

Once segmentation is established, it's possible to attach many different aspects to generated footage, including surface changes…

Multi-modal video synthesis with NVIDIA's Vid2Vid. Once segmentation is established, style transfer makes it possible to change the surface and other aspects of object appearance. (https://arxiv.org/pdf/1808.06601.pdf)

…time of day…

(https://www.youtube.com/watch?v=tSOCHisHIAw)

…and time of year.

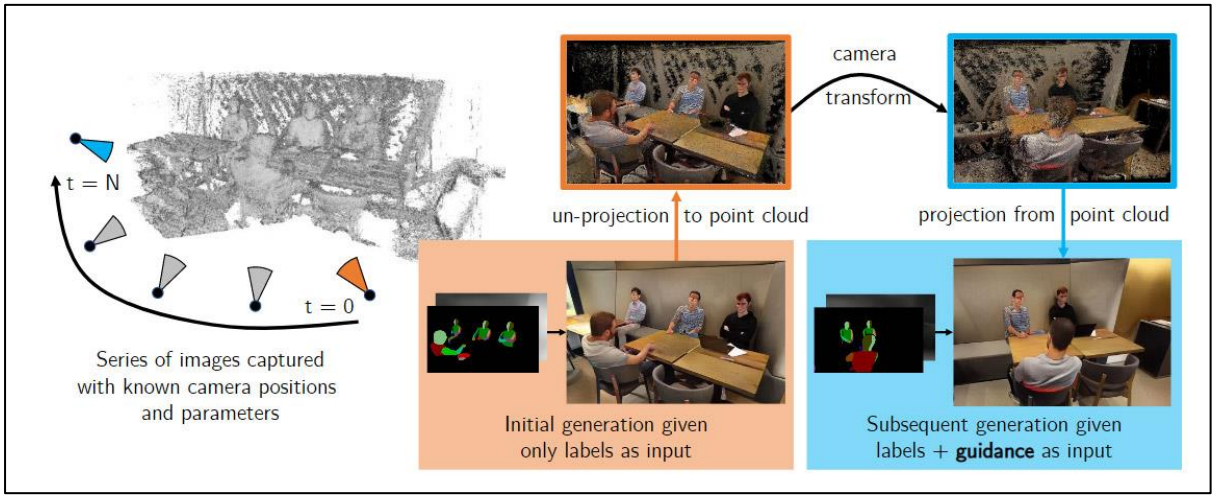
(https://www.youtube.com/watch?v=e3aXvYIWIq0)

Achieving Temporal Consistency in Video Synthesis

With NVIDIA's Vid2Vid, as with so many of the emerging image synthesis tools, VFX professionals have been prone to express concern about the lack of control, and the ability to apply consistent logic or continuity to the output.

In the summer of 2020 new research from NVIDIA, titled *World-Consistent Video-to-Video Synthesis*, addressed this issue with a new neural network architecture capable of achieving long-term temporal consistency.

The new architecture generates 'guidance images' – a kind of 'key-frame' that's evaluated from the motion data in the scene flow, and which collates all the previous viewpoints instead of merely calculating the next image from the previous frame, as with the original Vid2Vid, or with Google's 'Infinite Nature' system (which we will look at shortly).



Colors and structures tend to change in original Vid2Vid output, because the renderer has only the 'memory' of the previous frame with which to build the next frame. Here we see the generation of a guidance image that can carry forward stabilized information about objects in the scene. (https://arxiv.org/pdf/2007.08509.pdf)

Attention and focus are critical areas in general AI research, and while it's not clear how much cohesive and consistent output NVIDIA's new implementation of Vid2Vid can achieve for longer

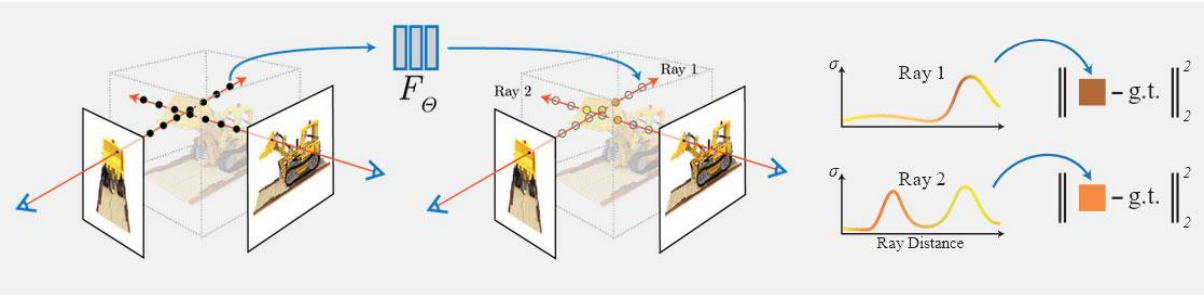
video synthesis footage, this is exactly the kind of effective corralling of semantic logic that AI in VFX needs in order to move forward into production environments.

NeRF

In August of 2020 researchers from UC Berkeley, UC San Diego and Google released details[47] of a new machine learning method capable called Neural Radiance Fields (NeRF) – an innovative take on transforming 2D imagery into a navigable, AI-driven 3D space:

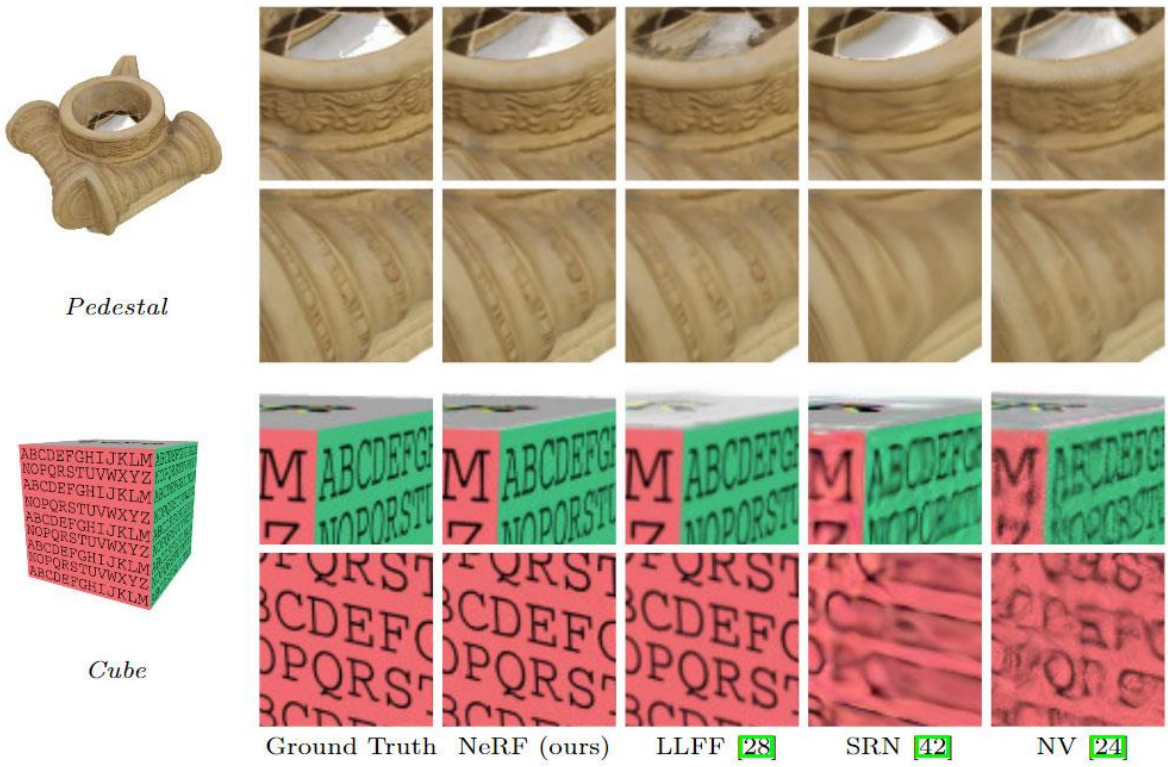
(*NeRF: Neural Radiance Fields*)

Led by Berkeley researcher Matthew Tancik[48], the approach uses a simple neural network (without perceptrons[49] or convolutional layers[50]) to translate a limited number of photos of a scene or object into what is effectively a volumetric 3D space, where each RGB pixel has an assigned X/Y/Z coordinate.



The NeRF space synthesizes a solid cloud of points by querying 5D coordinates on the path of camera rays, and projecting the output densities and colors with the use of established volume rendering techniques. In this sense, it is in some ways analogous to traditional CGI ray-tracing. (https://www.matthewtancik.com/nerf)

NeRF notably improves upon four predecessors in this field:



Ground truth source images (left column), followed by NeRF's interpretive capabilities, and equivalent results from three prior algorithms. (https://arxiv.org/pdf/2003.08934.pdf)

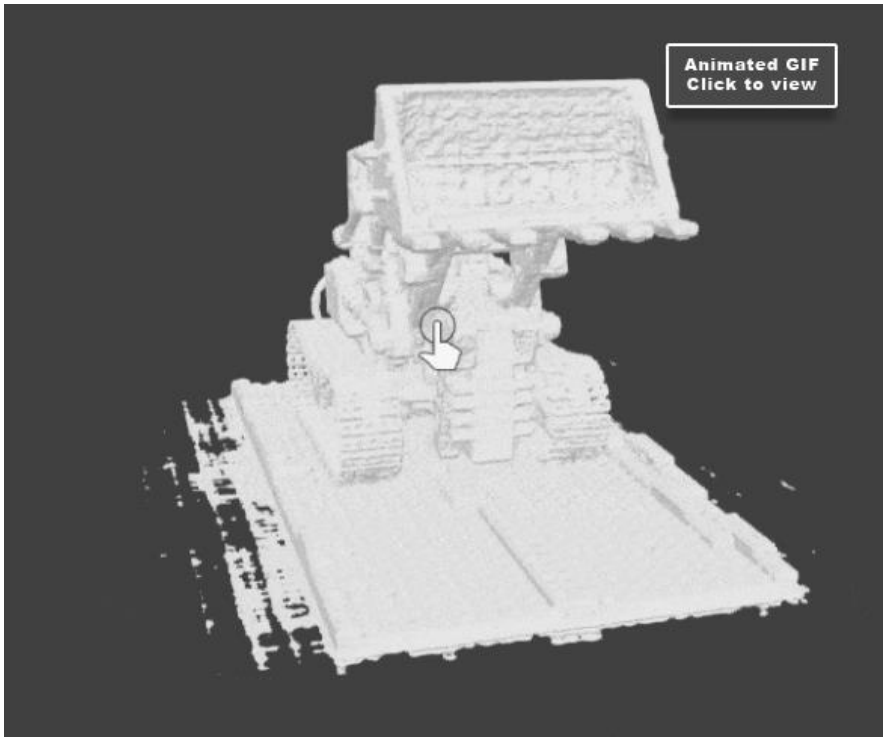
This is not mere 'in-betweening' or image interpolation, but equivalent to the creation of a points cloud that's so consistent it can even be used to generate a 3D mesh (see image below), and allows of panning, tilting and craning, with highly accurate lighting transitions and occlusion depth maps.

'I think the NeRF paper might be the first domino in a new breed of machine learning papers that will bear more fruit for visual effects.' says VFX supervisor[51] and machine learning enthusiast[52] **Charlie Winter**. 'The spinoffs so far demonstrate the ability to do things that would simply be impossible with a traditional GAN.'

'The way that a convolutional neural network generates an image.' Winter told me, 'is by progressively up-scaling and running computations on a 2D piece of information. Given a similar task, the limitation to this is that if you don't have many samples around a particular angle, the results get much worse and even unstable, as it can only use tricks learned in 2D to roughly approximate what might be there.

'With NeRF, what's being learned is an approximation of 3D shape and light reflectance, and an image is only generated by marching rays through the scene and sampling the network at each point.

'It's similar to the way a CG scene is rendered. Using this technique, there can still be angles with less quality. However, because this technique doesn't "cut the corner" of learning in 3D, it's much more likely that coherent information comes out.'



The NeRF space can generate a marching cube triangle mesh. (https://www.matthewtancik.com/nerf)

In contrast to approaches that use semantic segmentation or inferred/metadata depth maps, NeRF provides geometry-based occlusion for compositing purposes.

NeRF's complete 3D space makes occlusion mapping relatively trivial.

The pi-GAN[53] NeRF implementation from Stanford's Computational Imaging Lab covers similar ground, in terms of mesh extraction and continuity, while NeRF has been adapted for several other purposes, including the creation of alternate viewpoints from monocular video[54]; the generation of

avatars from monocular[55] and multi-view input[56]; and the extraction of game-ready meshes and textures from monocular video[57].

Winter notes that NeRF is very demanding in terms of processing time, wherein one complicated scene might take a month to render. This can be alleviated through the use of multiple TPUs/TPUs, but this only shortens render time – it doesn't cut the processing costs.

DyNeRF: Facebook's Neural 3D Video Synthesis

In March 2021, a subsequent collaboration[58] between Facebook and USC, entitled *Neural 3D Synthesis*, evolved the principles of NeRF into a new iteration called Dynamic Neural Radiance Field (DyNeRF). The system is capable of producing navigable 3D environments from multi-camera video inputs.

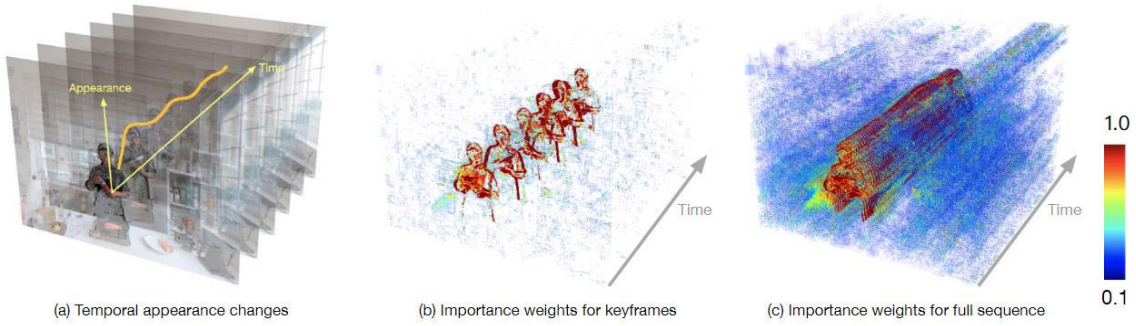
The capture system behind this DyNeRF clip uses 21 GoPro Black Hero 7 cameras (see image below) recording at 2.7k resolution at 30fps. (https://neural-3d-video.github.io/)

18 camera viewpoints are used for training the neural networks to create a radiance environment for each video clip.

By default, a 10-second 1014×1352 multiple view video sequence using 18 cameras generates a staggering amount of computational data – 7.4 billion ray samples for each epoch in the training cycle.

With 8 GPUs fielding 24,576 samples for each iteration and 300,000 total iterations required, training would normally require 3-4 days – notably more for successive iterations to improve the quality of the video.

To improve these calculation times, DyNeRF evaluates the priority of each pixel, and assigns it a weight at the expense of other pixels that will contribute less to the rendered output. In this way, convergence times (how long it takes the training session to achieve an acceptable minimum loss result) are kept within workable bounds.



Hierarchical training is performed sequentially, first on the individual key-frames, and then on the continuous sequence, with temporal appearance changes transformed into weights for the final algorithm. (https://neural-3d-video.github.io/)

Nonetheless, the researchers admit that the results are 'near' to photorealism; that the computational power needed to generate this level of interpretability is formidable; and that extrapolating objects beyond the bounds of the camera views will be an additional challenge.

DyNeRF is able to cope well with traditionally problematic visual elements such as fire, steam and water, and with anisotropic elements such as flat material reflections, and other light-based environmental considerations.

Once the synthesis is achieved, it is remarkably versatile. The frame rate can be upscaled from the native 30fps to 150fps or more, and reversing that mechanism allows extreme slow-motion and 'bullet-time' effects to be achieved.

Facebook SynSin – View Synthesis From a Single Image

A 2020 research paper from Facebook AI Research (FAIR) offers a less exhausting approach to dynamic view synthesis, albeit a less dazzling one.

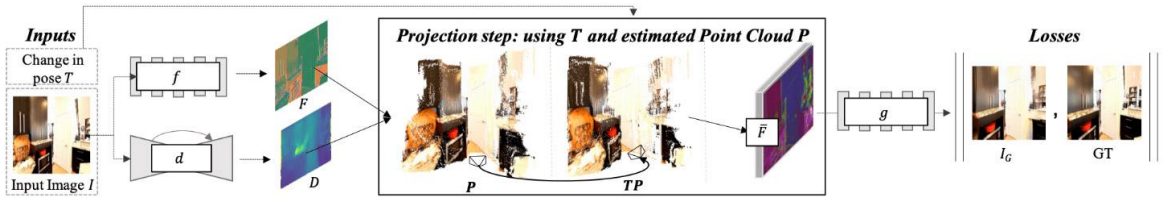
SynSin derives moving images from one static, single image through inference and context.

(https://www.robots.ox.ac.uk/~ow/synsin.htm)

Here the model must incorporate an understanding of 3D structure and semantics, since it must deduce and in-paint the parts of the scene that are revealed as the motion tracks beyond the bounds of the original photo.

Similar work in recent years relies on multi-camera views (such as DyNeRF), or on images that are accompanied by depth map data, synthesized externally (to the training model) or else captured at the time the image was taken.

Instead, SynSin estimates a more limited point cloud than DyNeRF, and uses a depth regressor to cover the amount of movement necessary. The features learned by the neural network are projected into the calculated 3D space, and those rendered features are then passed to a refinement network for final output.



(*SynSin: End-to-end View Synthesis from a Single Image*)

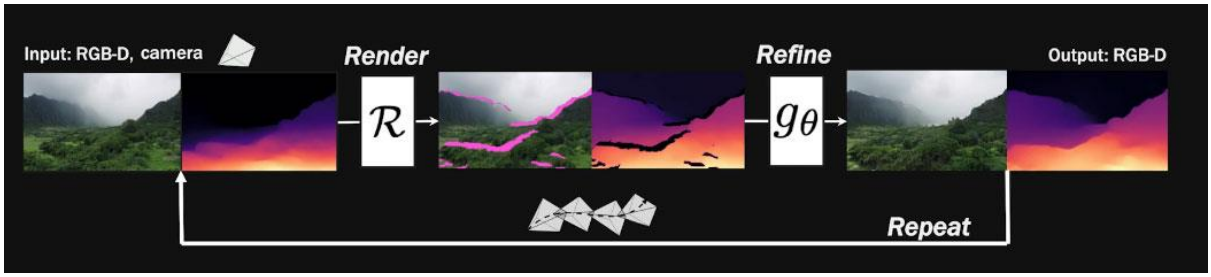
Google's Infinite Flyovers

While this kind of 'moving image' extends a little beyond Apple's iOS Live Photos feature[62], (and Adobe also moved into this territory in 2018 with its Project Moving Stills initiative[63]) it's not exactly the 'infinite zoom' we were promised in *Blade Runner*.

Google Research offers something nearer to this in the form of Infinite Nature[59] a machine learning system capable of generating 'perpetual views' – endless flyovers which are seeded by a single image, and which use domain knowledge of aerial views to visualize a flight right through the initial picture and beyond, into and over imaginary landscapes:

Infinite Nature is trained on 700 videos of drone footage from coastlines and nature scenes, a database of more than two million frames. The researchers ran a structure-from-motion (SfM[65]) pipeline to establish the 3D camera trajectories. The resultant database, called the Aerial Coastline Imagery Dataset (ACID), has been made publicly available[60].

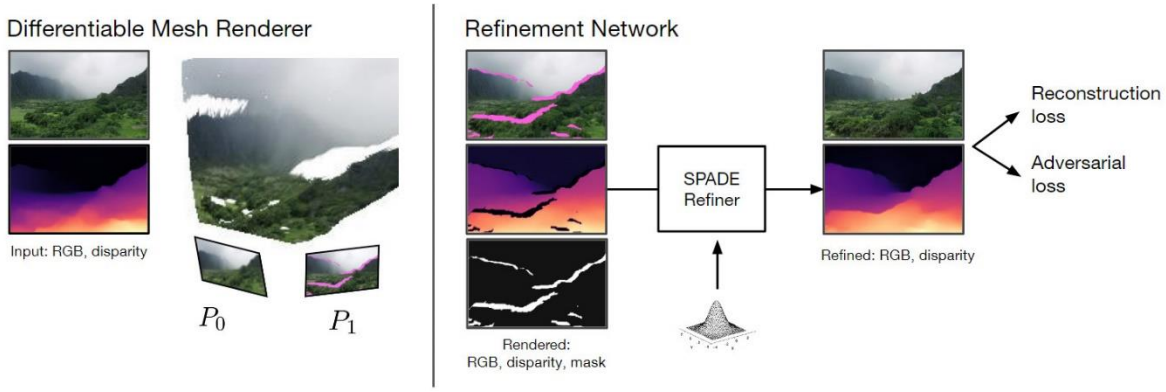
Though the system takes some inspiration from SynSin's complete computation of a motion path, it instead performs an infinite loop, dubbed by the researchers as *Render – Refine – Repeat*.



(https://www.youtube.com/watch?v=oXUf6anNAtc)

From the source frame to each 'fictitious' frame, each image is evaluated from a calculated next-step in a trajectory. The newly-revealed geography is then inpainted based on the model's landscape feature data, and the entire process repeated in an auto-regressive manner.

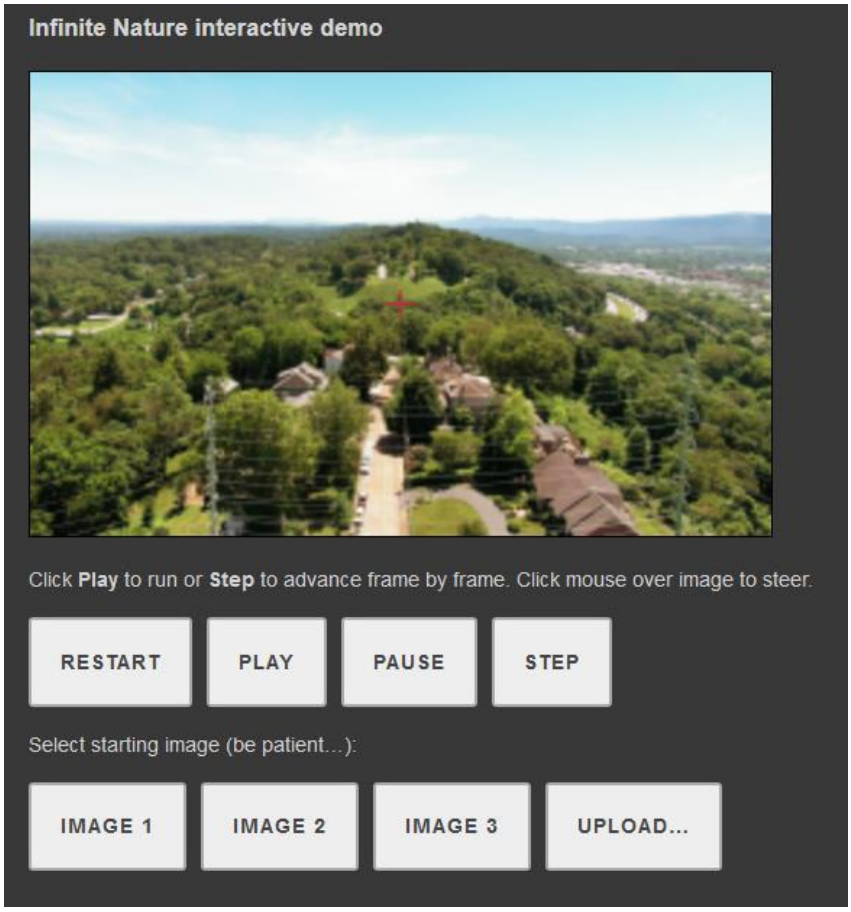
The system's differentiable renderer compares the existing frame to a projected next frame and generates a 'disparity map', which reveals where the holes are in the landscape.



(https://arxiv.org/pdf/2012.09855.pdf)

The refinement network that fills in these gaps uses NVIDIA's Semantic Image Synthesis with Spatially-Adaptive Normalization (SPADE) framework[61], which also powers the GauGAN system mentioned earlier.

Google provides an interactive Colab notebook, wherein you can take a (geographically inaccurate!) journey from your own uploaded starting image.



However, the sequence will transform very quickly into aerial landscape imagery, should you choose to provide any other kind of image.

There is nothing to stop future developers from applying this logic to other types of 'infinite journey' in different domains, such as Vid2Vid-style urban street imagery, or to use artificial datasets to produce randomly-generated voyages through other worlds or alien cultures.

If you can gather together enough gritty photos of early 1980s downbeat New York tenements, you could even finish Deckard's Esper[62] analysis of his retro Polaroid, and finally find out just how far that 'enhance' function goes.

Conclusion

The development costs for GPT-3 and DALL-E range in the low millions, in contrast to big budget VFX outings such as *Avengers: Infinity War*, which is reported to have spent $220 million USD on visual effects alone. Therefore it seems likely that the major studios will funnel some research money into comparable image synthesis frameworks, if only to ascertain the current limitations of this kind of architecture – but hopefully, eventually, to radically advance the state of the art.

As an analogy to how long CGI took to gain a true foothold in the VFX industry, image synthesis is currently nearer to *Westworld* (1973) than *Wrath Of Khan* (1982); but there is the sound of distant thunder.

---

[1] *https://old.reddit.com/r/MediaSynthesis/*

[2] *Since this off-beat stream of imagery was overwhelming the more general Media Synthesis forum, the posts were eventually requested to move elsewhere: https://old.reddit.com/r/bigsleep/ and https://old.reddit.com/r/deepdream/*

[3] *BigSleep: Bloody river painting – Metacognition, Reddit, 10th Feb 2021. https://old.reddit.com/r/MediaSynthesis/comments/lgw3n5/bigsleep_bloody_river_painting/*

[4] *King Kong in the style of Picasso, BigSleep – HerbChii, Reddit, 26th January 2021. https://old.reddit.com/r/MediaSynthesis/comments/l5ay28/king_kong_in_the_style_of_picasso_bigsleep/*

[5] *A Pikachu chasing Mark Zuckerberg (BigSleep) – Ubizwa, Reddit, 27th January 2021. https://old.reddit.com/r/MediaSynthesis/comments/l6c6z9/a_pikachu_chasing_mark_zuckerberg_bigsleep/*

[6] *BigSleep: Witches around a cauldron in the style of Synthwave – motionphi2, Reddit, 22nd February 2021. https://old.reddit.com/r/MediaSynthesis/comments/lpzxl8/bigsleep_witches_around_a_cauldron_in_the_style/*

[7] *BigSleep: Psychedelic rainbow kitties getting intoxicated in Wonderland – motionphi2, Reddit, 28th February 2021. https://old.reddit.com/r/MediaSynthesis/comments/luqcj3/bigsleep_psychedelic_rainbow_kitties_getting/*

[8] *Aleph2Image: Interdimensional portal in the middle of the street – jdude, Reddit, 1st March 2021. https://old.reddit.com/r/MediaSynthesis/comments/lvoj8i/aleph2image_interdimensional_portal_in_the_middle/*

[9] *DALL-E x CLIP – "The Industrial Revolution and its consequences." – LaserbeamSharks, Reddit, 27th February 2021. https://old.reddit.com/r/MediaSynthesis/comments/ltxpmx/dalle_x_clip_the_industrial_revolution_and_its/*

[10] *When asking Clip-GLaSS AI for "The son of Obama and Trump" I've got a perfect Draco Malfoy casting – navalguijo, Reddit, 4th February 2021. https://old.reddit.com/r/MediaSynthesis/comments/lcbsjq/when_asking_clipglass_ai_for_the_son_of_obama_and/*

[11] *The Big Sleep: BigGANxCLIP. https://colab.research.google.com/drive/1NCceX2mbiKOSlAd_o7IU7nA9UskKN5WR?usp=sharing#scrollTo=NRUouuHptoRp*

[12] 'DeepFaceDrawing' AI can turn simple sketches into detailed photo portraits*, Rachel England, Engadget, June 17th 2020. https://www.engadget.com/ai-can-produce-detailed-photos-of-faces-from-simple-sketches-122924655.html*

[13] *OpenAI's latest AI text generator GPT-3 amazes early adopters – Mike Wheatley, Silicon Angle, 19th July 2020. https://siliconangle.com/2020/07/19/openais-latest-ai-text-generator-gpt-3-amazes-early-adopters/*

[14] *Multimodal Machine Learning: A Survey and Taxonomy – Tadas Baltrusaitis, Chaitanya Ahuja, Louis-Philippe Morency, Arxiv, 1st August 2017. https://arxiv.org/pdf/1705.09406.pdf*

[15] *CLIP – OpenAI, GitHub. https://github.com/OpenAI/CLIP 'Learning Transferable Visual Models From Natural Language Supervision' Alec Radford et al., Arxiv, 26th February 2021. https://arxiv.org/pdf/2103.00020.pdf*
[16] *Zero-Shot Learning – A Comprehensive Evaluation of the Good, the Bad and the Ugly – Yongqin Xian et al., Arxiv, 23rd September 2020. https://arxiv.org/pdf/1707.00600.pdf*
[17] Deep Residual Learning for Image Recognition – *Kaiming He et al., Open Access, undated. https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf*
[18] *According to the creators, the project is named both for the surrealist painter Salvador Dali and the fictional Pixar robot WALL-E – reflecting the bizarre nature of the imagery that it generates. DALL·E: Creating Images from Text – openai.com, January 5 2021. https://openai.com/blog/dall-e/#fn1*
[19] *Unsurprisingly the web-crawled database fueling CLIP-GlaSS is likely to have a higher number of contributing images of prominent people from the post-analogue age (such as Donald Trump) than archival material of older or more obscure figures.*
[20] Image GPT – *openai.com, 17th June 2021. https://openai.com/blog/image-gpt/*
Generative Pretraining from Pixels – *Mark Chen et al., Arxiv, 2020. https://cdn.openai.com/papers/Generative_Pretraining_from_Pixels_V1_ICML.pdf*
[21] DALL-E *-openai, GitHub. https://github.com/openai/DALL-E*
[22] DALLE-PyTorch – *openai, GitHub. https://github.com/lucidrains/DALLE-pytorch*
[23] DALL-E – colab. *https://colab.research.google.com/drive/1dWvA54k4fH8zAmiix3VXbg95uEIMfqQM?usp=sharing*
[24] A Gentle Introduction to BigGAN the Big Generative Adversarial Network – *Jason Brownlee, Machine Learning Mastery, 23rd August 2019. https://machinelearningmastery.com/a-gentle-introduction-to-the-biggan/*
[25] stylegan2 – *NVIDIA Labs, GitHub. https://github.com/NVlabs/stylegan2*
[26] gpt-2 *open.ai, GitHub. https://github.com/openai/gpt-2*
[27] *Ryan Murdock – GitHub. https://github.com/rynmurdock*
[28] Ryan Murdock – *Google Scholar. https://scholar.google.com/citations?user=L2Fmp1AAAAAJ&hl=en*
[29] imagenet1000_clsidx_to_labels.txt – *yrevar, GitHub. https://gist.github.com/yrevar/942d3a0ac09ec9e5eb3a*
[30] Thoughts on DeepDaze, BigSleep, and Aleph2Image – *Ryan Murdock, rynmurdock.github.io, 26th February 2021. https://rynmurdock.github.io/2021/02/26/Aleph2Image.html*
[P] A Colab notebook from Ryan Murdock that creates an image from a given text description using SIREN and OpenAI's CLIP – *Wiskkey, Reddit, 16th January 2020. https://old.reddit.com/r/MachineLearning/comments/ky8fq8/p_a_colab_notebook_from_ryan_murdock_that_creates/*
[31] Gigapixel AI Accidentally Added Ryan Gosling's Face to This Photo – *DL Cade, PetaPixel, 17th August 2020. https://petapixel.com/2020/08/17/gigapixel-ai-accidentally-added-ryan-goslings-face-to-this-photo/*
[32] Large-scale CelebFaces Attributes (CelebA) Dataset – *http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html*
[33] Amara's law – *PCMag. https://www.pcmag.com/encyclopedia/term/amaras-law*
[34] Rare – *https://www.rare.co.uk/*
[35] Practical Deep Learning for VFX and Technical Artists – *Valentine Kozin, YouTube, Dec 23 2020. https://www.youtube.com/watch?v=miLIwQ7yPkA*
36 PhotoSketch – mtli, GitHub. https://github.com/mtli/PhotoSketch
37 Image-to-Image Translation with Conditional Adversarial Networks – Phillip Isola et al., Arxiv, 26th November 2018. https://arxiv.org/pdf/1611.07004.pdf / pix2pix – Phillip Isola, GitHub. https://github.com/phillipi/pix2pix
38 An Introduction To Conditional GANs (CGANs) – Manish Nayak, DataDrivenInvestor, 9th May 2019. https://medium.datadriveninvestor.com/an-introduction-to-conditional-gans-cgans-727d1f5bb011
[39] Investigating Under and Overfitting in Wasserstein Generative Adversarial Networks – *Ben Adlam, Charles Weill, Amol Kapoor, Arxiv, 30th October 2019. https://arxiv.org/pdf/1910.14137.pdf*
[40] Training Generative Adversarial Networks with Limited Data – *Tero Karras et al., Arxiv, 2020. https://papers.nips.cc/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf*
Training GANs With Limited Data – *Mayank Agarwal, Medium, 28th October 2020. https://medium.com/swlh/training-gans-with-limited-data-22a7c8ffce78*
[41] MIT researchers claim augmentation technique can train GANs with less data – *Kyle Wiggers, VentureBeat, 24th June 2020. https://venturebeat.com/2020/06/24/mits-technique-trains-state-of-the-art-gans-with-less-data/*
Data-efficient-gans – *MIT HAN Lab, GitHub. https://github.com/mit-han-lab/data-efficient-gans*
[42] MetaGAN: An Adversarial Approach to Few-Shot Learning – *Ruixiang Zhang et al., Arxiv, 2018. http://www.cse.ust.hk/~yqsong/papers/2018-NIPS-MetaGAN-long.pdf*
[43] A Pigment of Your Imagination: GauGAN AI Art Tool Receives 'Best of Show,' 'Audience Choice' Awards at SIGGRAPH – *Isha Salian, NVIDIA blog, 30th July 2019. https://blogs.nvidia.com/blog/2019/07/30/gaugan-ai-painting/*
[44] Colie Wertz – *IMDB. https://www.imdb.com/name/nm0921643/*
[45] vid2vid – *NVIDIA. https://github.com/NVIDIA/vid2vid*
[46] NVIDIA Invents AI Interactive Graphics – *NVIDIA developer news center, December 3, 2018. https://news.developer.nvidia.com/nvidia-invents-ai-interactive-graphics/*
Video-to-Video Synthesis – *Ting-Chun Wang et al., Arxiv, 3rd December 2018. https://arxiv.org/pdf/1808.06601.pdf*
[47] NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis – *Ben Mildenhall et al., Arxiv, 3rd August 2020. https://arxiv.org/pdf/2003.08934.pdf / https://www.youtube.com/watch?v=JuH79E8rdKc*
NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis – *Pratul P. Srinivasan et al., GitHub, 2020. https://pratulsrinivasan.github.io/nerv/*
[48] Matthew Tancik – *Google Scholar. https://scholar.google.com/citations?user=l0Bj7U8AAAAJ&hl=en*

[49] Perceptron – *deepai.org. https://deepai.org/machine-learning-glossary-and-terms/perceptron*
[50] How Do Convolutional Layers Work in Deep Learning Neural Networks? – *Jason Brownlee , Machine Learning Mastery, April 17 2019 . https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/*
[51] Charlie Winter – *IMDB. https://www.imdb.com/name/nm1138639/*
[52] Neural VFX – A scratchbook for machine learning and AI in visual effects – *http://neuralvfx.com/*
[53] pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-AwareImage Synthesis – *Eric R. Chan et al., Arxiv, 2nd December 2020. https://arxiv.org/pdf/2012.00926.pdf*
[54] Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video – *Edgar Tretschk, Arxiv, 26th February 2021. https://arxiv.org/pdf/2012.12247.pdf*
Also see a similar, non-NeRF collaboration between NVIDIA and the University Of Minnesota:
Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera – *Jae Shin Yoon et al.. University of Minnesota, 2020. https://www-users.cs.umn.edu/~jsyoon/dynamic_synth/*
[55] Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction – *Guy Gafni et al., GitHub, 2020. https://gafniguy.github.io/4D-Facial-Avatars/*
[56] Learning Compositional Radiance Fields of Dynamic Human Heads – *Ziyan Wang et al., GitHub, 2021. https://ziyanw1.github.io/hybrid_nerf/*
[57] NeRD: Neural Reflectance Decomposition from Image Collections – *Mark Boss, YouTube, December 7th 2020. https://www.youtube.com/watch?v=JL-qMTXw9VU*
[58] Neural 3D Video Synthesis – *Tianye Li et al., Arxiv, 3rd March 2021. https://arxiv.org/pdf/2103.02597.pdf*
[59] Infinite_nature – *Google Research, GitHub. https://github.com/google-research/google-research/tree/master/infinite_nature*
[60] *ACID database download. http://storage.googleapis.com/gresearch/aerial-coastline-imagery-dataset/acid_v1_release.tar.gz*
[61] Semantic Image Synthesis with Spatially-Adaptive Normalization – *Taesung Park et al., Arxiv, 5 Nov 2019. https://arxiv.org/pdf/1903.07291.pdf*
[62] Esper – *The Blade Runner Wiki. https://bladerunner.fandom.com/wiki/Esper*