How AI Is Transforming the Visual Effects Industry (2020)

By Martin Anderson

First published ***May 13th, 2020*** *at:*
https://rossdawson.com/futurist/implications-of-ai/comprehensive-guide-ai-artificial-intelligence-visual-effects-vfx/ | *Web-archived version*

It's over twenty five years since the ground-breaking CGI effects of *Jurassic Park* usurped 100 years of visual effects tradition. When Steven Spielberg showed the first rushes of computer-generated dinosaurs to acclaimed traditional stop-motion animator Phil Tippett (who had been hired to create the dinosaurs in the same way they had been done since the 1920s) he announced "I think I'm extinct." It's a line so significant that it made it into the movie itself, in reference to a paleontologist envisaging a world where no-one would need him to theorize about dinosaurs any longer.

Though a quantum leap, the visual effects of *Jurassic Park* did not represent an overnight upheaval. They had been presaged sporadically throughout the 1970s, and at greater length in the 1980s, in cinematic curios such as *Tron*, *The Last Starfighter* and *Flight of The Navigator*. In the few years directly prior, James Cameron had brought renewed interest to the possibilities of CGI with the 'liquid' effects of *The Abyss* and *Terminator 2: Judgement Day*.

But *Jurassic Park* was different: computers had achieved the ability to generate solid, *photo-real* objects, promising to relegate the uncomfortable burdens of the physical, photochemical VFX world. It set the trend for the decades ahead, and reinvented the visual effects industry — not without many casualties among the old guard.

Many influential movie makers and VFX studios were unable or unwilling to read the signs of the times in the years leading up to *Jurassic Park*. It now seems that the water is rippling again for the current state of the art in visual effects, as new machine learning techniques slowly encroach on the now-established workflows of CGI – and that a new 'disruptive event' may be coming to shake up the industry.

1: Deepfakes

Born in Porn

In late 2017, not for the first time, porn proved a prime mover for a relatively obscure new technology. In that period a new sub-Reddit appeared, dedicated to publishing short pornographic video clips which had been convincingly altered to feature the faces of celebrities.

This apparent alchemy, now packaged by a pseudonymous user into a public code repository called Deepfakes, had been achieved with the use of a Convolutional Neural Network (CNN) and autoencoders (but not, as widely reported in mainstream articles and in Wikipedia, using a Generative Adversarial Network [GAN] – a machine learning technique first proposed by lead Google researcher Ian Goodfellow over three years earlier which was then gaining traction in other image-generation projects [1]).

Something seismic had begun to occur in machine learning research in this period. Recent advances in GPU-based machine learning had begun to facilitate the processing of large amounts of data in increasingly efficient time-frames. Almost overnight (in terms of the often-hindered history of AI), a great deal of global governmental and industry-led research into computer vision and object recognition, research centered around well-funded sectors such as robotics, logistics and security footage analysis, had become actionable and accessible to less 'serious' purposes.

GANs, CNNs and autoencoders began to crop up in headline-grabbing experiments around style-transfer and the inference and generation, using data from the public domain, of 'unreal', yet photorealistic images.

Research into Deep Photo Style Transfer in 2017, a collaboration between Cornell University and Adobe. Reference images and stylized secondary images are combined via a neural network to output photorealistic combinations.

But Deepfakes, which allowed casual users to assault our longstanding faith in the authenticity of casual video footage, became the villainous totem which revealed the extent and nature of the coming revolution.

The port of the code was rough, but solid. It soon led to the availability of various, slightly more user-friendly DeepFake applications. However, those wishing to create videos were (and are) required to become (ironically) meticulous data scientists, gathering and curating large face-sets of celebrities to feed into the neural network, and then waiting up to a week for the neural net to cook the data into a model capable of making the transformation from one face to another.

Face-swapping software in action, as a journalist trains a machine learning model designed to swap faces between Jeff Bezos and Elon Musk. The preview window shows how near the model is getting to becoming capable of a photorealistic swap between the two subjects. Full training can take anywhere between six hours and a week, depending on the configuration. Once trained, the model can spit out face-swapped images in seconds, and videos in minutes.

That notwithstanding, results which were now stunning and shocking the world *could* be obtained via some diligence, publicly available photos and a mid-level PC with a NVIDIA graphics card.

As would soon become clear, the technology seemed capable of rivalling or exceeding any comparable work out of Hollywood, within the limits of its own ambition: the convincing digital manipulation and transposition of faces in video footage.

Permission to Fake

Though the internet ruminates on it almost daily, the implications of post-truth video are a subject for another day. However, it's worth noting, in a legal climate that has yet to catch up with DeepFake technology, that the New York bill attempting to criminalize Deepfakes has been repudiated by the MPAA. The organization believes such a sweeping law would limit Hollywood's ability to replicate historical personages, even with pre-DeepFake technology. A more general bill is currently working its way through Congress, though this relates to exclusively criminal usage of DeepFake tech.

But even in the event that the New York bill passes, it seems reasonable to assume that U.S.-based film-makers will be able to obtain permission to replicate actors using machine learning. For sure, the possibilities that AI-

based technologies offer to the visual effects industry far exceed the aims of the implementations that made them famous.

Automation of VFX Roles

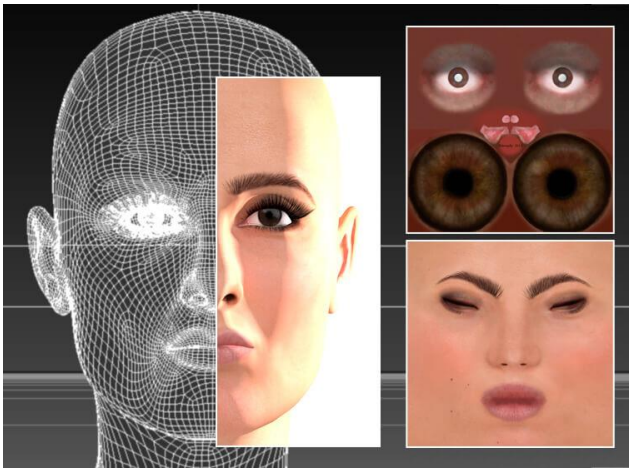
Much as the march of AI is threatening radiologists more than doctors, the use of machine learning imperils certain trades within the VFX industry more than others — at least in the early years of tentative adoption. Its eventual potential scope extends to nearly every facet of VFX production currently handled under a traditional CGI pipeline.

As with most trades which AI is encroaching upon, it's the layer of 'interpretation' which is most subject to automation: the artisanal process of collating and creatively manipulating data into the desired results.

In terms of CGI vs. AI, it's useful to understand which parts of the process are susceptible to a machine learning approach.

The Difference Between a CGI Mesh and a Deep Learning 'Model'

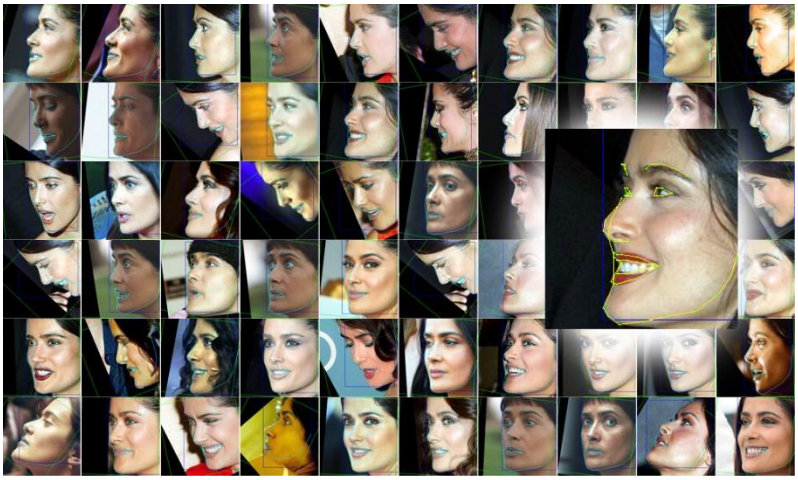
A traditional CGI approach to generating a human face involves creating or generating a 3D 'mesh' of the person, and mapping appropriate texture images onto the model. If the face needs to move, such as blinking or smiling, these variations will have to be painstakingly sculpted into the model as parameters. Muscle and skin simulations may need to be devised, in addition to hair or fur systems, to simulate eyebrows and facial hair.



A traditional CGI head comprising vector-based mesh, point information and texture detailing, among other facets

A machine learning-based model is much more abstract in nature. It's created by the process of analyzing and assimilating thousands of real-world source images of the two subjects being processed (the 'target' person who will feature in the final work, and the 'source' person that they will be transformed from).

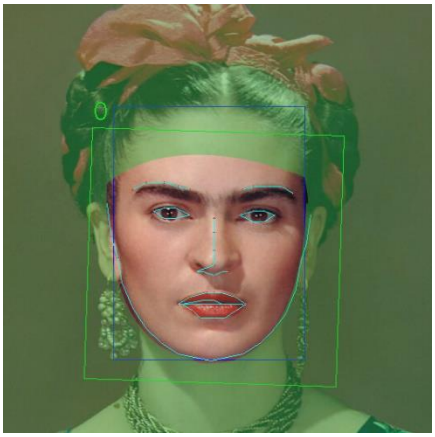
During extraction of the data from source images, the software applies facial pose estimation to gain an understanding of the angle and expression of the face in each image (see colored lines in the image-set below). These 'facial landmarks are used to make effective conversions, and, optionally, to train the model more efficiently.



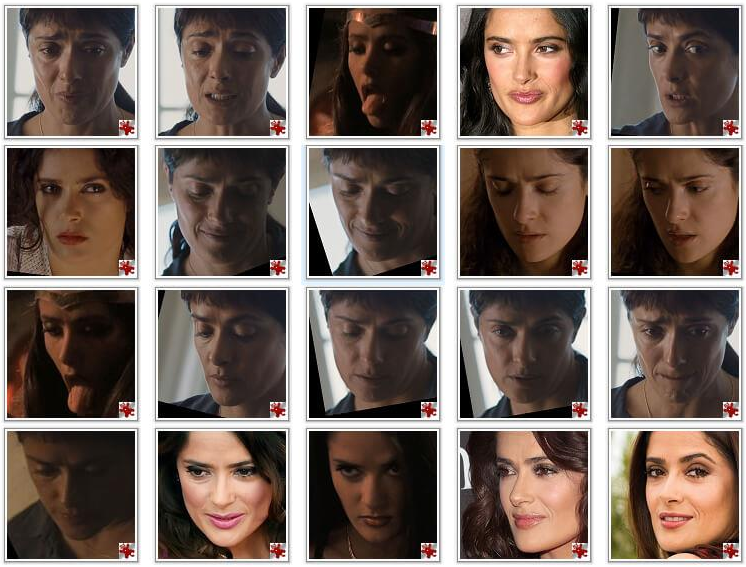
A machine learning model applies pose recognition to a large database of source photos, identifying which way the person is facing, and some characteristics of expression.

The resulting model learns implicitly how to transfer facial characteristics between the two subjects in the training data, without referring to any specific facial angles that it learned during training.

DeepFake tools (such as the FaceSwap repository) can optionally use a 'mask' to delineate the area that the system should concentrate on. This is a rudimentary segmentation map, about which we'll hear more later. The spline-like alignment lines (in pale blue, below) delineate the curve and location of the features of the face. The compiled knowledge of a very large image database will be corralled into a well-defined area of pixels:



The 'destination' training set (photo of the person who you want to insert into an image or video) represents the generalized 'ground truth':



The specific lighting or characteristics of the actual photo you are trying to alter represents an additional, case-specific ground truth:

The system's ability to apply domain transfer will ensure that the imposed face matches the face it is replacing, at least roughly, in terms of shadow, texture, color and general lighting.

Domain transfer in this case extends beyond the facial features; the 'swap' takes account of the target lighting and environmental integration in a way that is very hard (and much more expensive) for older CGI techniques to match.

One of the developers behind the FaceSwap project explains how the model arrives at this equivalency:

> *'There is no explicit input of 20 degrees left, 30 degrees down and go find a similar face in the training corpus. [There] is just a string of 1024 numbers that we tell the algorithm to adjust in the most intelligent way to describe the entire face…then reconstruct the entire face from those 1024 numbers. [One] of those numbers may be a vector encoding hair color, one may be nose length, one may be happy vs. sad, one may be shadowy vs. brightly lit, one may be the multi-component combination of several correlated underliers that we call attractiveness, etc.…one may very well be face angle/pose. [But] we don't bother to know. [And] each number in that 1024 will describe a different concept for every new training model.'*

Learning to Fish

In the CGI workflow, you gather data and end up with, for instance, one CGI head. In the machine learning model, the gathered data is effectively compiled into software which can render out that likeness endlessly, accounting for diverse lighting and other environmental conditions.

The CGI method not only involves a significant cast of contributors (texture artists, mesh designers and riggers, among many others), but also requires expensive and time-consuming resources to render the model once it is complete, often to little appreciation.

With the machine learning technique, nearly all the resources are spent in compiling the data, one time, into a functioning model. The process of rendering is a trivial concern by comparison. Almost the only people remaining from the CGI workflow are the photographers who gathered the reference photos.

Essentially it's the difference between buying a fish or a fishing rod.

Doug Roble, the senior director of software at award-winning VFX studio Digital Domain, recently discussed new fluid dynamics simulation software the company is developing as part of a deepening interest in machine learning, and clarified the significance of the new AI paradigm in visual effects:

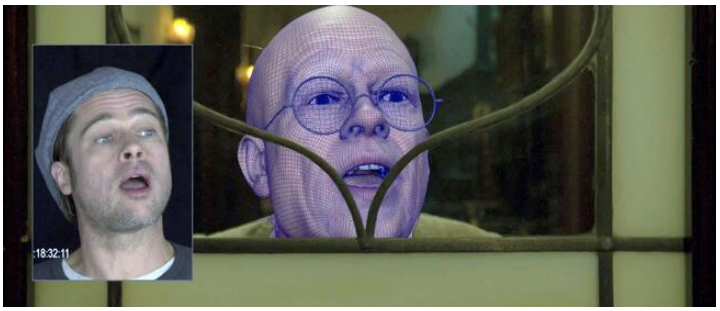
> *"If you're just going to do a one-off shot, machine learning is probably not the way to go. But if you're going to do fifty shots or a hundred shots, or a whole series of something where these shots are going to be coming up over and over again, then you take the time to simulate a whole bunch of generic shots — enough that you can train a machine learning system, which is not cheap either. It takes a long time to actually build one of these machine learning things.*
>
> *"But at the end of it you get something where you can do fluid simulations in real time. Once the computer understands how water moves, it can run staggering effects."*

Role Reversal?

Besides its obvious significance for the traditional CGI sector, machine learning may also have ramifications for the make-up and prosthetics trades in film production.

One future possible use for face-swapping at a professional VFX level could be the superimposition of the faces of real historical figures over that of the actor playing them. Besides a long history of award-laden

performances from an actor buried under unrecognizable prosthetics (including Gary Oldman, Charlize Theron, John Hurt, Christian Bale and Marlon Brando), several actors, such as Andy Serkis and Doug Jones, even specialize in having their performances translated into CGI-driven creations.



The old way: actor Brad Pitt's facial expressions being mimicked into a real-time flexible mesh by VFX lead El Ulbrich and team, for David Fincher's The Curious Case of Benjamin Button *(2009)*

Though it would depend on the film-makers' and lead actors' commitment to a project, a historical biopic such as *Frida* (2002) could actually feature the face of the movie's subject, subject to any necessary permissions from the estate. This technique could also potentially make available such roles to actors that would otherwise be considered too poor a facial match to play the part.

It could even make possible the casting of male lead roles to actors of another gender (with or without the aid of new developments in voice cloning); or to make more flexible the racial aspect of casting, where that's essential to the plot, and where only flexible casting can otherwise solve the problem.

The political implications of such a capability are probably more difficult to resolve.

Cautious Optimism From VFX Studios

Complete facial substitution at this level is likely to need more R&D from the major VFX labs, since openly available face swapping software does not currently provide adequate, or adequately consistent high resolution surface detail to allow close-ups (however, in this respect, it shares a problem with CGI itself).

Rob Bredow, the recently appointed new chief at Industrial Light and Magic, commented last year that Deep Fake technology can't currently provide that necessary resolution, but added that "it won't be too long before we'll be able to use something like that at much higher quality [in motion picture production]. I don't think we're 10 years away from that. We may only be one or two."

Not everyone at ILM is so optimistic. Speaking to me, CG supervisor Jordi Cardus described the state of the art of Generative Adversarial Networks as 'more speculative than actually being used in production in any meaningful way'. Washington-based VFX Shader Artist William Mauritzen is also wary of the current hype around GANs, saying "I don't see how adversarial systems are going to change things as much as just clever engineering of other network solutions."

ILM's Yannick Lorvo, who has worked as CG Generalist/Lead on productions such as this year's *Aladdin*, as well as *Avengers: Infinity War*, *Ready Player One* and *Star Wars: Episode VIII – The Last Jedi*, indicated to me that the company is actively using machine learning in noise reduction and texture generation from photos,

with upcoming ML-based developments around rotoscoping (of which, more later), animation and concept creation.

Hollywood Upstaged by Open Source Software and Low-End Computers

Face swapping in mainstream movies seems likely to start small, much as CGI itself did in *WestWorld* (1973) and *FutureWorld* (1976). However, this time the public has an unprecedented lead on a new visual effects technology.

In 2017 emergency reshoots for the DC superhero outing *Justice League* required additional late-stage footage of Superman actor Henry Cavill, who was by then contractually obliged to retain a moustache he had since grown for the shooting of another movie. Ultimately the offending whiskers were removed via CGI, but to very little appreciation from viewers.

Several months later a fan published a YouTube video showing moustache removal results for Cavill which were certainly no worse, and arguably better, than what a major effects house had produced for the $300 million movie, emphasizing that this was achieved with open source software and a $500 home computer.

The incident was only one example of the new ML-based face swapping technique upstaging a Hollywood VFX publicity machine which was very proud of its recent advances in face-based CGI. The 2016 *Star Wars* spin-off entry *Rogue One* went to great efforts to recreate original series actors Peter Cushing and Carrie Fisher, attaching CGI facial recreations of their 1977 appearance onto stand-in actors. The work got mixed reviews.

Many were surprised when the apparently easier job of recreating a young and flawless Fisher proved the less effective of the two efforts, prompting another DeepFake user to give it a rather effective re-working with home hardware and free software in 2018.

The same faker made further inroads into the *Star Wars* VFX sphere by inserting a young Harrison Ford over the face of the actor portraying his most iconic role in 2018's *Solo: A Star Wars Story*.

Though not an actual trumping of the film-makers' original intentions this time, it still seemed that the audience, for once, was significantly ahead of the studios in terms of understanding and acting upon what machine learning could do for movie visual effects.

Commercial Deepfake Usage

In 2018 singer Charli XCX and Troye Sivan used the open source Deepfakes code to simulate multiple versions of the singer in the form of 1990s pop stars The Spice Girls.

California-based feature-film director and visual effects supervisor David Gidali also used the Deepfakes code to project the likenesses of film stars George Clooney and Rachel McAdams in a short film about a couple using DeepFake technology to spice up their sex life.

2: Environment and Object Generation

Having been CG supervisor on nine movies to date, Charlie Winter is a veteran of the CGI pipeline. In his spare time, he also develops experimental neural network-based solutions to some of the problems he has to face when working in visual effects.

Winter believes that the field of environment creation is a prime candidate for machine learning. "In light of current developments, I think that environments will get hit very quickly, because they always have a hundred different types of rocks, trees…things that are all very similar. And if you can find an algorithm that can make one nice, it can make a million nice."

Though interesting experiments and research around the potential of neural networks to create environmental AI imagery have been popping up over the last few years, a far more portentous demo of its capability came to light a couple of months after we talked.

The graphics chip manufacturer NVIDIA has become central to the entire AI revolution since researchers began utilizing the GPU to iterate faster through complex data sets. Its CUDA machine learning hardware (together with the enabling cuDNN libraries) is the academic and industry standard, hard-wired into the company's latest high-end graphics boards. The appropriation of the graphics card from the gaming set by the AI research community, combined with a surge of interest in using the GPU to accelerate crypto-mining,

drove NVIDIA's unit prices through the roof in recent years (however, as we'll see later, the eventual bursting of that crypto-mining bubble may have some implications for AI VFX research).

In mid-March NVIDIA unveiled a machine learning system called GauGAN, capable of assimilating vast amounts of visual data from world environments into a tool that lets the user 'paint' rough swathes of color, which the AI will then replace with breathtakingly realistic imagery:

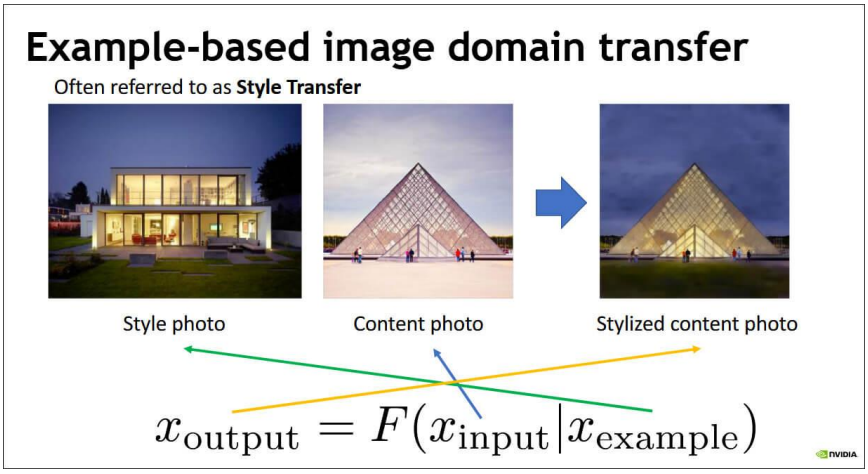
The work on GauGAN builds on (and arguably supersedes) NVIDIA's own previous achievements with Pix2PixHD.

Initial commentary on GauGAN has been divided, but apocalyptic in tone: for everyone noting that the experiment is still just an artistic tool with a Photoshop-like GUI, others argue that it lowers entry barriers (from 'professional' to 'dabbler') so comprehensively as to represent, in concept, a disruptive ground-shift for artists and photographers.

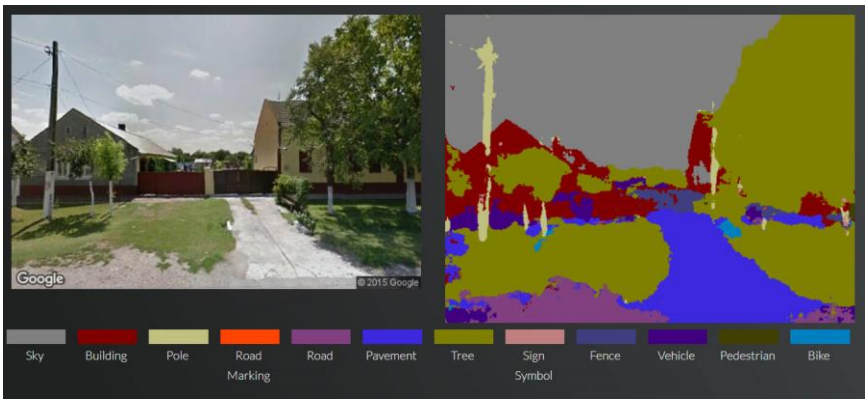
The 'Magic Wand' of Segmentation, Ground Truth and Domain Transfer

More importantly, GauGAN quite literally illustrates the power of simple, *user-defined* segmentation in orchestrating vast compiled arrays of images into a cohesive, synthesized visual representation. It posits a general system of 'indicative illustration' that could represent a template for the future machine learning-driven systems in the VFX industry: automation for image creation, in an unprecedented capacity.

Jan Kautz, VP of Learning and Perception Research at NVIDIA, explained the image domain transfer principle underpinning software such as GauGAN best in visual terms, in this presentation.



Insofar as these workflows depend on segmentation mapping, this new concept in visual processing has developed almost directly from research into more mainstream pursuits such as autonomous vehicles, security analysis and medical imaging. Much of the open source code (TensorFlow, Keras, Scikit-Image, OpenCV et al) in GAN-based image-creation software (commercial, proprietary and open source) remains intact from those non-entertainment projects.



A random example of segmentation from Cambridge University's SegNet project.

With the advent of tools such as GauGAN, segmentation has become creative and bi-directional, instead of merely analytical.

A Product on the Streets

As usual, key machine learning player NVIDIA is conducting pioneering work in this aspect of GAN-driven image creation. At the end of 2018 the company revealed a system that uses high-level semantic segmentation (i.e. green = 'lamp-post', blue = 'sidewalk'), in combination with a deep network that had analyzed many hours of driving footage, to create completely fictional road footage:

The paradigm is becoming clear. Though not currently trivial to build or train, these 'magic wands' based around domain transfer offer the chance to translate sketchy objectives into extremely specific and photo-real imagery, with minimal human intervention.

It's little wonder that Doug Roble, while cautious about machine learning's current limitations, is so enthused about this basic potential VFX production model:

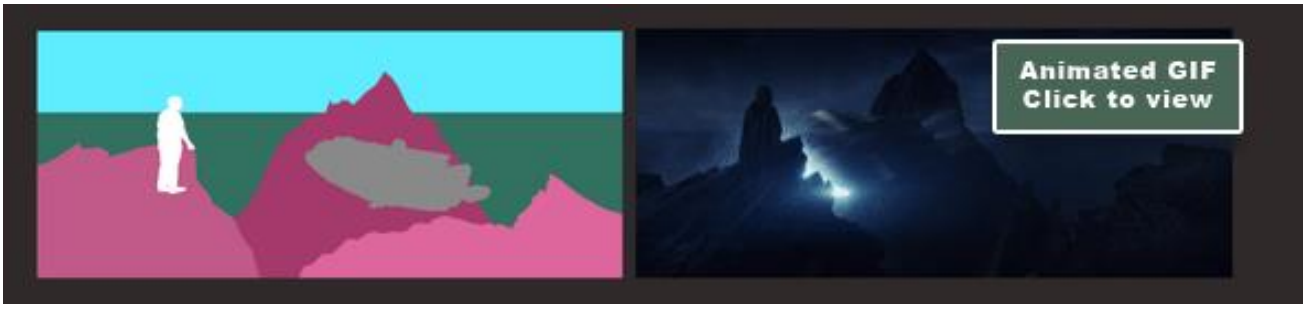
> *"The amazing thing about this is that it's a totally different way of looking at problems. Over the last twenty years, whenever we wanted to do a problem like simulating muscle groups, we would get into the physics of it. We'd do the mathematics and the physics and start thinking about how the muscles deform and then we'd have to model it and use finite element techniques and all sorts of mathematical tricks in order to get this thing to work.*
>
> *"Now, if we want to simulate muscles, a totally different way of looking at the problem is to go and say, well, maybe if I just get a couple of people in, and I watch how their muscles move and I actually measure how their muscles move. And I do this enough, I can generate enough data so that I can train a system that will automatically figure out how muscles move."*

From Animatics Directly to Production-Quality VFX?

Though the idea is currently nothing more than speculation about the way these facets of image-based machine learning might coalesce in future projects, ML's ability to recognize and transpose objects into existing (or entirely new) imagery could even offer a previously undreamt-of workflow: from a director's vision to a finished shot.

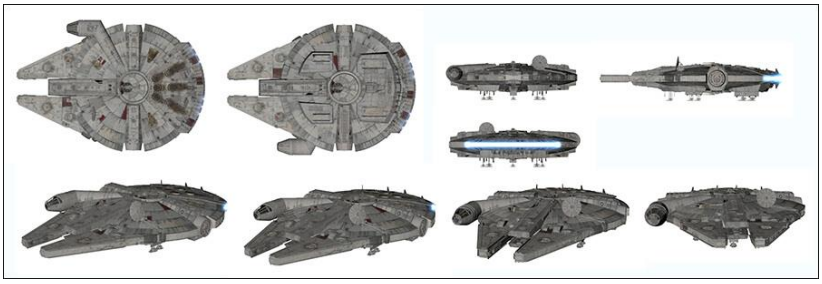
In effect, for instance, artificial semantic grouping and object recognition could be made to combine with pre-trained machine learning models to transform simple animatics directly into high-resolution, fully-finished visual effects shots. It's a similar concept to the AI-generated NVIDIA street footage (see above).

Much as Phil Tippet's contribution to *Jurassic Park* was demoted from head of animation to 'dinosaur supervisor', it's possible that the future use of actual models (CGI or physical), will be limited to indicating what's wanted from the shot to machine learning systems which have been fed with all the necessary data to integrate the required objects into the footage.

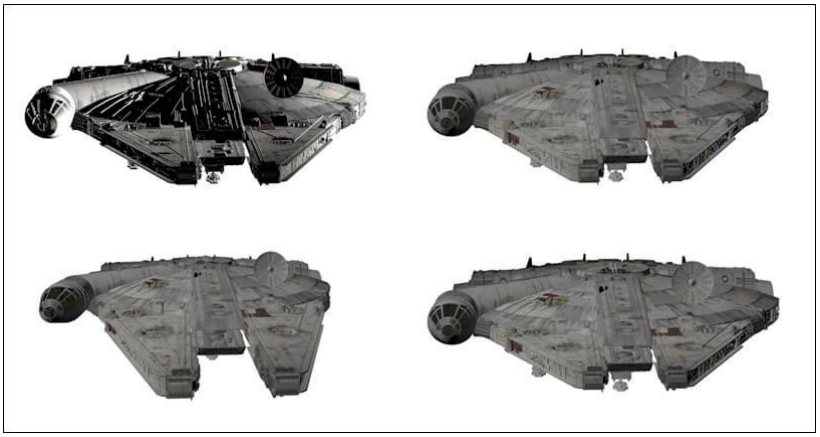


In the above re-imagining of a (supposedly machine learning-driven) VFX shot from *The Last Jedi*, deliberate semantic segmentation-mapping defines several elements in the rough animatic. A deep network can then populate these segments from large databases of visual resources: white ('person', else matted in afterwards from real-world footage); pink (rocks, with diminishing brightness indicating distance from camera); green (sea); and blue (sky).

In this imagining, the neural net has already assimilated multiple viewpoints and photos of the Falcon, either as a CGI or a photographed physical model.



Our putative Falcon database, likely to comprise around 5-10,000 images, features shots with every possible usable length of lens, from wide-angle to telephoto, with and without motion blur, and in a sufficient number of lighting styles that the model will probably never have to be made again, unless to depict damage or design changes.



The segmentation in the sketchy animatic for the spaceship is tagged 'FALCON' at a semantic level, with its associated, defining light-grey color. The machine learning system recognizes the shape and angle of the rough CGI model from the animatic footage and knows to substitute an interpreted view of the ship from that angle (although, as we've seen, it will accomplish this by a learned, implicit understanding of the object, based on analysis of thousands of images, and not just by 'pasting' an image whose alignment matches the animatic).

Incorporating custom objects, such as a spaceship, into a semantic recognition framework in this way would require genuine innovation from VFX AI software developers. As we'll see later, changing or adding to fundamental aspects of open source machine learning models promises a burden of commitment, along with rewards.

Much as GauGAN does, this fictitious system would also draw in textural and structural imagery around rocks, rough seas and bad weather, informed by the same global and proprietary databases that currently fuel other image-based GAN projects.

Further ground truth for the shot, such as weather and general lighting, is provided by existing or specially-shot reference videos – cues for style transfer.

3: Rotoscoping

Cutting Out the Middleman With AI

The visual effects industry has always benefitted directly or indirectly from mainstream consumer and government/military R&D. However, the intensity of current global research around CCTV footage analysis now offers truly ground-breaking machine learning approaches to some old problems. The fruit of these solutions are already feeding into commercial film production and VFX software, and promise to affect several sub-trades and sectors in the long term.

One of those traditional challenges is rotoscoping – the practice of extracting individual elements, such as people and objects, from their background environments. In the traditional photochemical world, this was pioneered as Blue Screen Process for RKO's *The Thief of Baghdad* in the late 1930s.

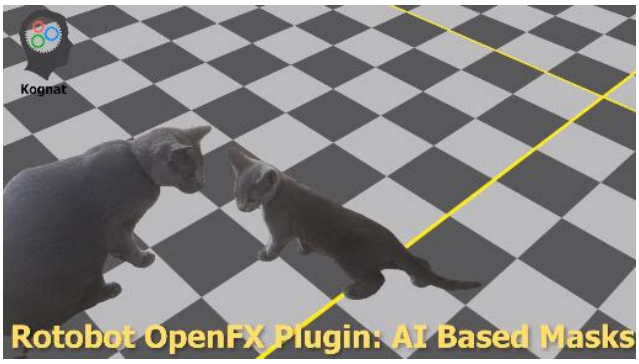
Though Walt Disney innovated a superior 'yellow screen' matte extraction process using sodium vapor (later renting it to unlikely clients such as Alfred Hitchcock), and though the preferred knock-out color evolved from blue to green by the late 1990s, the physical inconvenience of creating, placing and lighting large monotone backdrops was unaffected.

Even though VFX legend Richard Edlund has often commented that the digital revolution of the early 1990s (prior to CGI) released the industry from the 'sumo wrestling' of photo-chemical mattes, the prospect of abandoning actual physical backdrops is very new.

An Industrial Background

It's now possible for neural networks to extract distinct multiple entities from a background through image segmentation, wherein the network seeks to identify and consistently track individual facets in a video stream.

The OpenFX plugin Rotobot is capable of extracting multiple elements from footage via machine learning, and piggy-backs on an avalanche of government and private sector security research of the last 5-10 years — research chiefly concerned with the analysis of security and civic camera footage for the purposes of gait and facial recognition.



That the Kognat code has been developed from wider-ranging global open source research can be seen from the limitations of category that it can provide for segmentation (where an entire class of object can be defined, and multiple instances of it isolated in one frame); namely 'aeroplane', 'bicycle', 'bird', 'boat', 'bottle', 'bus', 'car', 'cat', 'chair', 'cow', 'dining table', 'dog', 'horse', 'motorbike', 'person', 'potted plant', 'sheep', 'sofa', 'train', and 'TV' — categories which date back to the Visual Object Classes Challenge of ten years ago, and which recur regularly on GitHub and in other computer vision research streams.

Bad Hair Days

Palo Alto-based VFX startup Arraiy recently raised $10 million in series A funding based on an emerging portfolio of AI-driven VFX services, including an automated rotoscoping functionality, which the company promises to release to market in Q4.

Arraiy has trained its rotoscoping machine learning model by inputting vast amounts of specially-shot video footage, recreating in the process some of the thorniest problems in background extraction — such as getting a good matte around strands of hair.

In the image below we see some of the studio-based data setups which are being used to train the model:

Arraiy Founder Dr. Gary Bradski said:

> *"One of the things we want to do is those hard rotoscoping jobs, and this is how we're gathering the data for that. We're putting a lot of [challenging] objects in front of our data capture, and running systems of equations that allow us to extract the exact ground truth. Then we're compositing or blending or putting physical environments behind to get a…ground truth."*

Arraiy data gathering: establishing a 'ground truth' not via color-keying, but by machine learning that can apply the same principles to 'live' examples without the need for flat-color backgrounds later.

Ironically, Arraiy uses green screens extensively in the generation of its datasets, with the tacit objective of making the methodology obsolete.

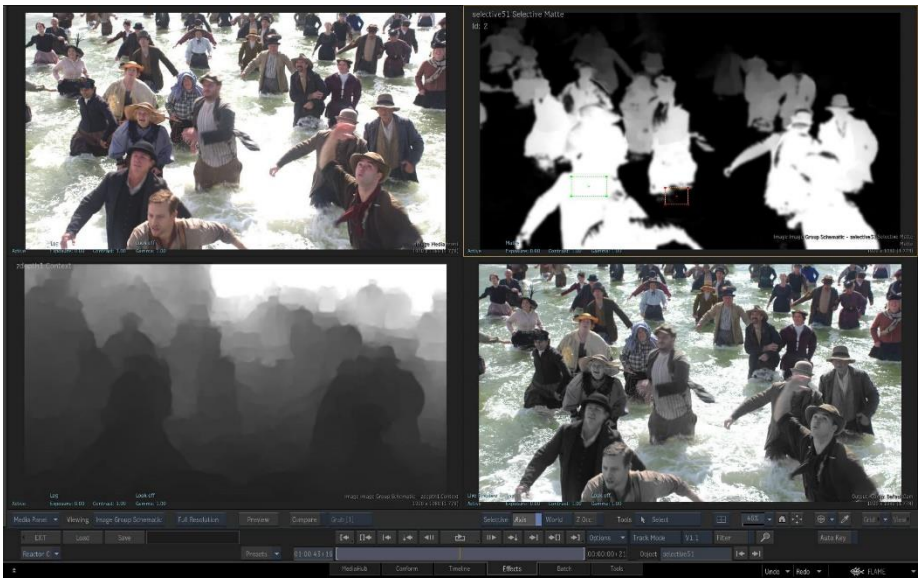
One of many sessions at Arraiy which produces foreground data for machine learning models to analyze and eventually extract, without the need for green screens.

In an earlier demonstration, Arraiy shows a human being extracted from a relatively complicated background in real time, without color keying. Bradski observes that the company has had to avoid hiring rotoscope artists that are used to a traditional pipeline, claiming that they're unwilling to adapt to the new freedoms offered by an ML-driven rotoscoping system.

Though Arraiy is conducting extensive original research and data generation, its projects rely also on OpenCV and Open3D repositories, among others.

Commercial and Academic Applications

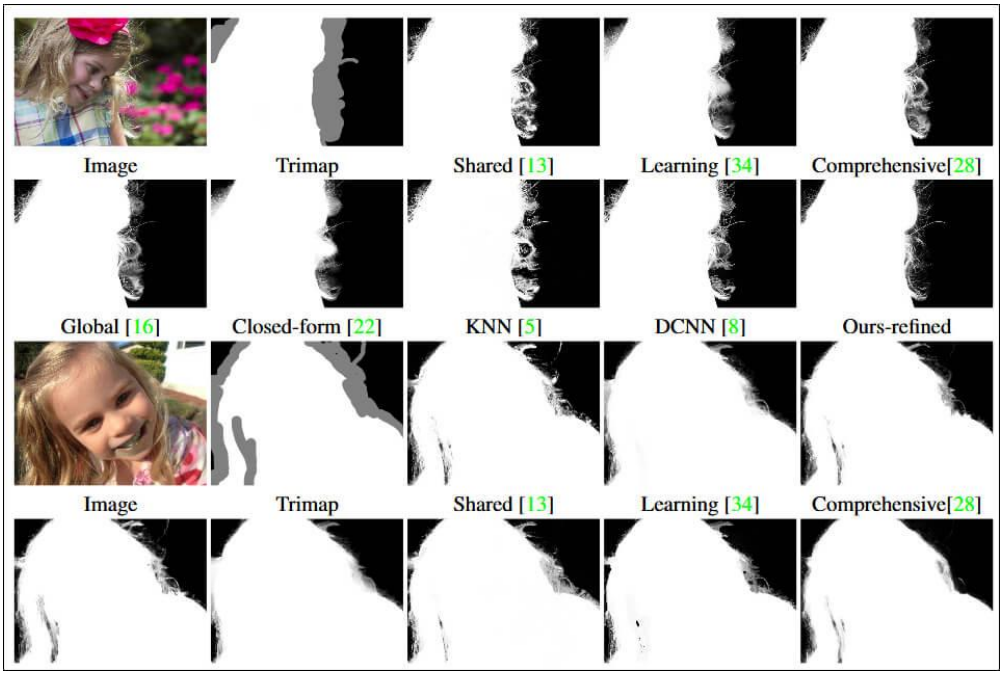
The 2020 edition of Autodesk's Flame video suite features a host of machine learning-enhanced features, including AI-driven segmentation and extraction. Additionally, the software's ability to individuate facets from a crowd in an image combines with scene recognition to allow a user to generate a depth map of scenes with multiple people, permitting the convincing addition of elements such as fog:



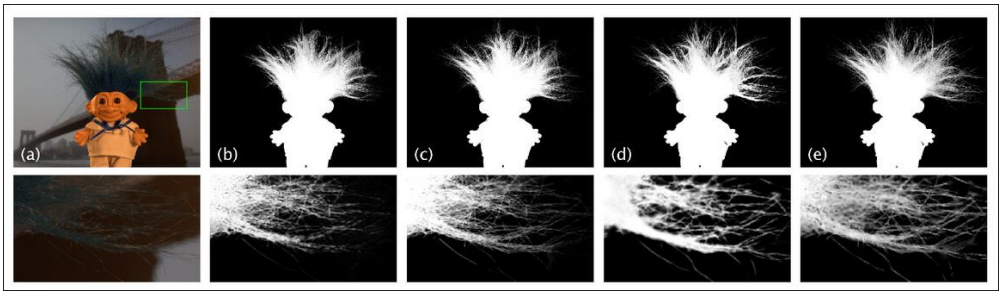
Autodesk Flame can generate a depth map from a real world image without additional on-set equipment. Here the neural net has identified those subjects further away, which have less luminosity in the processed z-channel.

Testing a beta version of the software, VFX supervisor Craig Russo said: "I recently worked on a virtual set comp where they forgot to add depth-of-field on the backgrounds. It took me two hours per shot to roto and add motion blur; I ran the same shots through the Z Depth Map Generator and got results in two seconds."

As one might expect, Adobe has been developing AI-driven rotoscoping techniques for several years. In 2018 it published a paper around Deep Image Matting, building on work that had been announced the previous year, though no product integration has been announced for the company's image and video editing software suites.



Trinity College Dublin also has an ongoing research project to develop a Generative Adversarial Network capable of extracting very fine detail without the need for knock-out matte background colors. The source code is publicly available.



Though numerous tools have been developed over the last thirty years to speed up or otherwise facilitate the process of rotoscoping, it remains, as a branch of pure animation, one of the most laborious and painstaking tasks in movie production. A rotoscope artist averages just 15 frames a day, with clean-up and tweaking on complex and expensive shots employing large numbers of rotoscope artists and facilities.

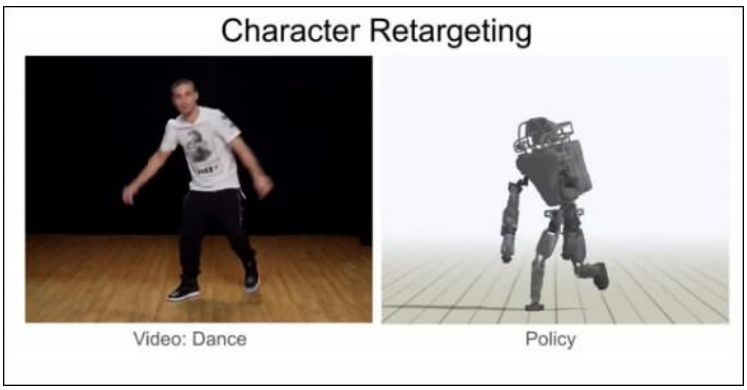
4: Machine Learning and Motion Capture

Asked which people within VFX are most likely be affected by GANs and machine learning workflows, Charlie Winter replies "A lot of people. I think probably face and body tracking is going out the window [as a sub-profession of VFX] pretty soon. As long as you have your track already done, if there's humans in there, in a couple of years you're going to get an automatic body and face-track."

Current solutions and ongoing research have benefitted greatly from general interest in sentiment analysis and general facial recognition technologies capable of mapping and interpreting what's going on in a face in a non-controlled environment.

Pose Estimation as a Solution to Sensor-Based Motion Capture

Research out of the University of California applies reinforcement learning to generate motion capture data even from so limited a source as videos on YouTube. The use of machine learning in this case is to clean up analogue-style noise based on a pre-learned understanding of the way the human body moves.



Carnegie Mellon University and DeepMotion Inc. have collaborated on a machine learning system designed to infer the physics of movements in basketball game footage to transfer these sporting 'skills' to virtual avatars. The project anticipates future applications in animation, gaming, motion analysis, and robotics.

Clean-up and Resolution Enhancement on Existing MoCap Data

This up-rezzing of distorted data is also being applied to the shortcomings of traditional, sensor-based motion capture data. The 'de-noising' of obscured MoCap data is becoming such a standard practice in machine learning that you can even find generic tutorials on the subject.

Data compression and de-noising via an autoencoder (from DeepLearner3d)

The highest-profile case of this new practice to date has been Digital Domain's treatment of Josh Brolin's character work as Thanos in the 2018 *Avengers: Infinity War* release. Using the Medusa Performance Capture system from Disney Research, Digital Domain built up a high-resolution database of the actor's face undergoing a range of expressions and distortions.

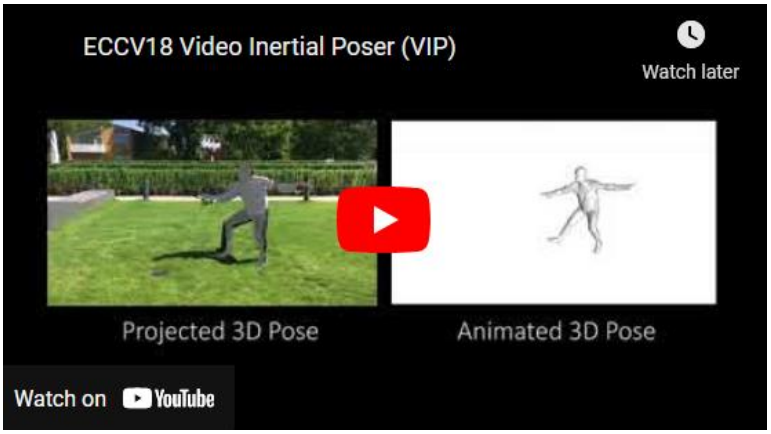
The Medusa data was then used to convert the mere 150 points of marker information from Brolin's final performance into 40,000 points of hi-res facial motion capture data, via a proprietary machine learning system called Masquerade.

Similar work is being undertaken by various academic departments around the world, including KTH Royal Institute of Technology at Sweden, which has developed a neural net capable of reconstructing missing motion capture data without the use of forward-frame analysis, and even in cases where an essential element of the capture has dropped out for a non-trivial amount of time.

Besides its endeavors in rotoscoping, Arraiy has also launched a commercial machine learning facial tracking system called DeepTrack, aimed at sports-casting and news and studio-based environments.

Academic Studies in Machine Learning and Motion Capture

As a contribution to general research in this area, a collaboration between the Leibniz University of Hannover and the Max Planck Institute for Intelligent Systems has made available a dataset of 'in the wild' 3D poses, using footage from mobile phone cameras.



Ubisoft also conducts ongoing work in this particular application of ML-driven motion capture research.

Research from the Federal University of Santa Catarina uses machine learning to develop a new paradigm for inference of body mechanics from existing video-driven data by 'decoupling' the expressiveness of captured movement (which effectively constitutes 'noise') from the actual derived movement.

However, algorithms derived from neural networks are also bridging the gap between current methodologies and the marker-free, ML-driven future that machine learning promise. Research at University of Basel uses deep learning to change the expression on mesh-based facial avatars.

An Early Victory for Machine Learning?

Doug Roble, whose likeness — capable of real-time response to his own movements via machine learning — is now the totem for work in AI-driven facial capture at Digital Domain, envisages early market capture in the field.

> *"Digital humans are going to start to be in the vernacular. This is going to be a thing. There's lots of uses for this, but obviously for visual effects previz and for virtual production, having a character that an actor can control…super-easily. There's no dots, there's no nothing — just a single camera looking at my face. All of a sudden I can create and control this character."*

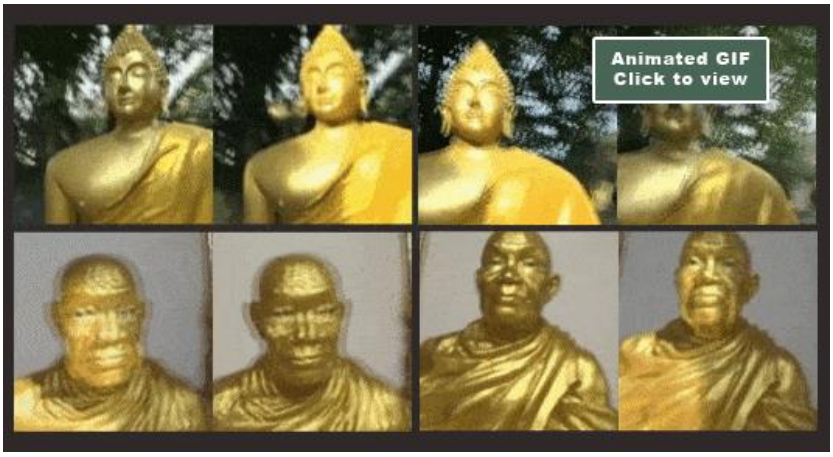
5: Other Applications

Lengthy as it is, it's beyond the scope of this overview to cover in depth all the fields to which machine learning is currently being developed and applied to challenges in visual effects. But it's worth considering, at least in passing, some of those other possibilities and projects.

Color Grading

Given the differences in weather that can plague an exterior shoot even in the course of one day (never mind attempting to represent a single day with footage shot over weeks), color grading has always been an essential post-production tool to cohere the various lighting conditions between set-ups.

Charlie Winter made his own experiments in 'changing the weather' using UC Berkeley's CycleGAN framework.



Winter says: "That's like light-swapping…I thought this was a fun experiment – probably not too useful. But then I showed it to my VFX supervisor friend and he's like, 'Dude, you've done auto-grading — that was exactly what I need!' I could just re-grade a hundred shots with this type of thing."

Though the prospect of ML-enhanced color grading must face issues of resolution and consistency (which we'll discuss shortly), the theory of this kind of transformation is already well within the purview of domain transfer.

Work from Microsoft and the Hong Kong University of Science and Technology has also examined the prospect not only of changing weather conditions in footage, but even using a neural network to make the traditionally ineffective day-for-night shooting techniques into something more convincing.

In 2017 research from NVIDIA used unsupervised image-to-image translation via a Convolutional Neural Network to render superficially convincing day-for-night footage.

In-Betweening for Animation and Match-Move

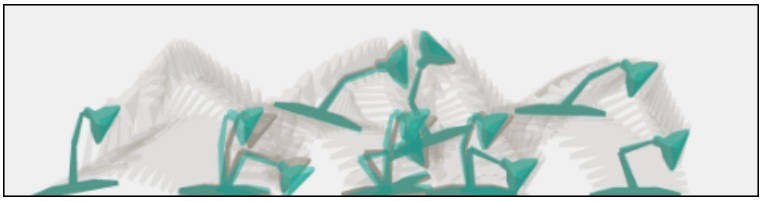
Frame interpolation is one of the oldest jobs in animation, wherein keyframes in a sequence are established, and automated or manual processes used to generate the 'in-between' frames that create an impression of consistent movement. The general principle is used in video codecs to reduce file size by throwing away 'repeated' or redundant data, and in the generation of vector movement paths in 3D and 2D video and CGI-based software.

In 2017 Japanese telecommunications and media giant Dwango postulated a framework for AI to create interpolative frames for animation in 2017, also demonstrating the product on YouTube.

Legendary Studio Ghibli animator Hayao Miyazaki reacted to the new system negatively.

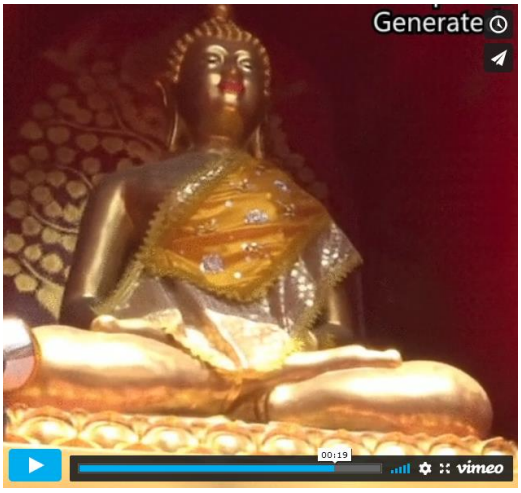
Returning briefly to the subject of motion capture, researchers from China's Xidian University have also suggested that Convolutional Neural Networks could provide a solution to establishing keyframes in MoCap data — another effort to cut through the 'noise' that such data-streams tend to generate.

At the other end of the problem, a team from the University of Columbia published a paper outlining an ML-based system potentially capable of deriving full-length MoCap sequences only from keyframes.



Charlie Winter has experimented with full-frame video interpolation, with the aim of addressing the problem of match moving — altering in-the-can footage after the fact.

Since a Generative Adversarial Network is inclined to distraction (a problem we'll cover shortly) Winter used a discriminator similar to the Pix2PixHD solution to regulate the interpolated frames. This ensures the temporal coherence of the output, allowing the creation of new and configurable camera motion from what was originally jerky, handheld footage



Normal and Depth Map Generation

Besides the obvious possibilities that style transfer offers for the purposes of concept art, machine learning is already making inroads into several 2D areas associated with the old style of CGI workflow: 'normal' and depth map generation.

A depth map is a discrete layer in an image (or sequence of images) which conveys directional and relief information about the objects depicted.

Since depth estimation from single-camera sources is a furiously active field of computer vision research outside of the entertainment industry, there is an unusual amount of prior achievement to kick-start VFX projects dealing with this challenge.

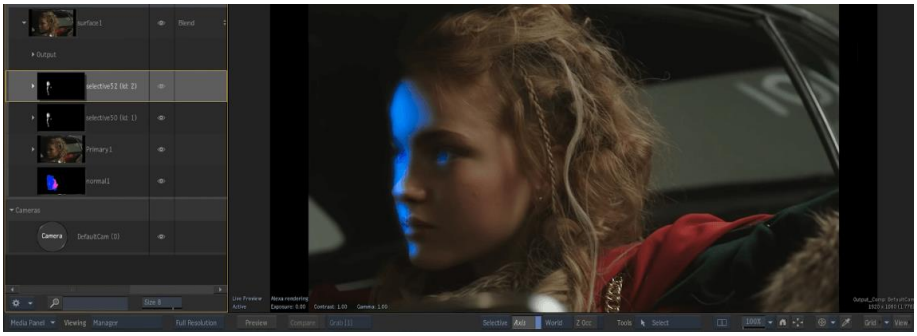
As mentioned earlier, Flame 2020 uses object recognition and machine learning algorithms to create a z-order (depth order) of crowded scenes. However, this is a different type of 'depth map'. If z-order is more akin to mapping the receding layers of 'flat' actors in a Victorian shadow theatre, depth/normal mapping is more like understanding the form of a shallow, yet still three-dimensional shape by turning an acute light on it.

Research out of Trinity College Dublin in 2018 presented an automatic pipeline, facilitated by Convolutional Neural Networks, that's capable of generating depth-map information from hand-drawn 2D characters in 'traditional' animation.

The resulting layer is a 'normal' map – effectively a limited description of the 3D substance of the subject matter depicted, and something that is usually impossible to obtain automatically from real world footage or from 'flat' animation.

The ability to enhance 'live' footage with normal maps without manually creating them offers post-production the chance to apply lighting and grading effects without passing the footage through any other departments.

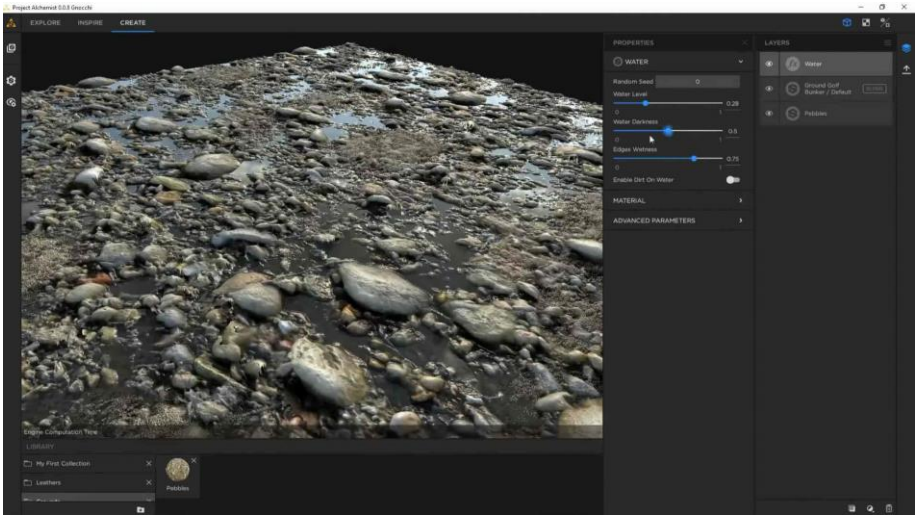
Machine learning's ability to recognize faces suits it well to this task. Returning to Flame 2020, the new release is also capable of generating normal maps directly from video footage of human faces via the suite's new AI-enhanced functionality.



Flame 2020's new ability to generate normal maps on faces removes the need to send 2D footage to interstitial post-production departments for evaluation and conversion into 3D-centric data.

Emerging commercial applications in the field of normal map generation have a good body of academic work to spring from. Columbia University has undertaken significant work on this; researchers from Illinois and China have assembled a large body of academic references around the subject of image synthesis and normal generation; and lone researcher and AI/AR enthusiast George Duan has created an auto-encoding normal generator, complete with an online demo.

Allegorithmic's Project Alchemist, debuted at SIGGRAPH 2018, provides an eye-opening suite of machine learning-enabled tools for texture artists, including the ability to generate depth maps from photos and then remove any bitmapped shadows left over from the source (delighting). Also on the roadmap for Project Alchemist is the integration of NVIDIA's AI-powered UpRes re-scaling process.



Editing Movie Trailers

A great deal of the AI research currently benefitting VFX innovation comes from more unilateral marketing-led objectives — using data to understand how people react emotionally to stimuli, in order to craft more effective campaigns.

In 2016 this ambit crossed over into cinema in the form of a trailer edited by a machine learning system underpinned by IBM's Watson framework. The system took 24 hours to evaluate the movie, before distilling six minutes of it which, based on its study of hundreds of other movie trailers, it judged were most likely to engage and interest a potential audience.

6: The Road Ahead for AI in VFX

"I've talked to people at some of the best studios," says Charlie Winter. "and they are actually hiring PhD. research scientists to come up with ways to improve the pipeline. The majority of departments are incorporating deep learning in one way or another…Ultimately it'll probably change everything. "

Though Weta Digital has made the fewest public declarations about its research into AI, one industry insider told me that Peter Jackson's New Zealand VFX house is "heavily invested" in developing machine learning solutions.

Digital Domain's Pipeline Technical Director Dmytro Korolov envisages a lot of initial inertia, given the potential scale of the re-tooling needed to convert to machine learning-driven workflows.

"The VFX pipeline is very conservative." He told me. "It will take years until neural networks will be a native part of it. There are a lot of compatibility and technical problems. AI tools should be controllable and predictable. No VFX artist likes a tool that creates complex things with 'one magic button. That's not really controllable, and hard to tweak.

"Artists need to have at least some idea about how AI tools work. And they need to trust those tools."

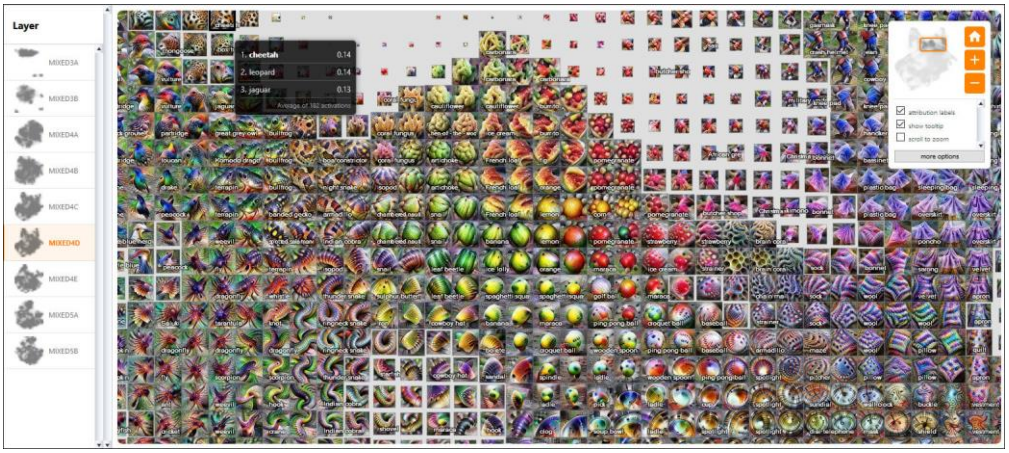
Indeed, the enthusiasm for the potential of machine learning in visual effects seems routinely tempered by skepticism around several specific issues, which are worth considering here.

Lack of Control

At the current state of the art, Generative Adversarial Networks, Convolutional Neural Networks, autoencoders and most of the other AI-based solutions of interest to visual effects houses obtain their amazing results in ways that are not always entirely understood. When problems occur in the workflow, it can be difficult to say if the error was in the data, the model configuration, a fault in the open source code, a fault in a proprietary revision to that code…or even a fundamental misconception about how the core source code approaches a particular problem one might give to it. For sure, misinterpreted data in neural nets can pose an embarrassing problem.

The issue is very general to AI. The potential for machine learning systems to reflect bias or prejudice from source data inspired the EU to incorporate a 'right to explanation' about algorithmic decisions into the GDPR's wide-ranging regulations. It's a request that many have argued presents semantic and practical difficulties, because understanding what happens inside a neural net is in itself a work in progress.

Google has just launched a collaboration with OpenAI entitled 'Exploring Neural Networks with Activation Atlases' — an effort that reveals what individual neurons 'see' when they approach a dataset, and one that could lead to greater understanding of how networks arrive at their decisions and transformations.



Activation Atlases use Feature Visualization to put AI in the psychiatrist's chair.

Analysis of the 'psychology' of neural networks is becoming a distinct field of study in AI research. Naftali Tishby, an AI researcher and neuroscientist from the Hebrew University of Jerusalem, caused a ground stir in 2017 by postulating a theory of an 'information bottleneck' that may occur in neural nets as they seek to sift essential information from the mountains of data thrown at them.

The work has led to new experimental practices in limiting training time for machine learning models, so that the generalized nature of the model's understanding does not begin to erode into over-specificity around the data — a type of 'psychotic break', where the overall scope of the task is degraded into strange obsessions about facets or aspects of the training input.

Aberrations in GAN-based image synthesis programs can be mitigated by controlling the purity of the source data, defining clearly the guiding ground truths, limiting the scope of the objective, and designing discriminators that help to classify input to the net, to discount irrelevant data or approaches — to maintain focus on the overall task, while still dealing with the fine-grained detail of its transformations.

Charlie Winter envisages greater uptake, development and enthusiasm from the studios when machine learning visual effects solutions begin to become more accessible in terms of offering clear parameters and controls.

"You probably want to have a more structured pipeline," he says. "or wait until algorithms get to where there's actually adjustment sliders built into them. That's totally doable, with conditional GANs. There's multiple types of GANs, but the ones that are actually truly conditional give the ability to provide kind of a switch, like 'Okay, right now I want a rabbit and now I want a turtle, so give me either one.

"An interface [like that] would give a visual effects supervisor the ability to sit down and say 'These are the things that I need to be able to dial in order for it to actually be useful. Train a GAN and build that into it'."

The Future of Open Source Code in VFX Machine Learning-Based Software

Some of the problems around predictability that confront machine learning-based VFX solutions center on the way that these solutions have come into being — and the fact that the code being exploited was not originally developed for the purposes that VFX houses want.

Much of the innovation discussed here derives directly from the world's wider-ranging research groups: massive collective efforts that are seeking to make self-driving cars which don't crash, advertising insights that work, robots that can replace Amazon warehouse workers, and security systems that can track individuals by face, gait and other characteristics. Serious work, in trillion-dollar sectors.

In this initial period of revelation, VFX houses are effectively kicking the tires of autoencoders, GANs, CNNs and Recurrent Neural Networks. But deeper involvement with machine learning seems likely to necessitate the same kind of open source, intra-studio cooperation around machine learning which has led to partly or fully open source collaborative projects such as Pixar Animation Studio's Universal Scene Description (USD) format; continuing development on the OpenVDB C++ library initiated by DreamWorks; the OpenColorIO (OCIO) color management solution; the interchangeable computer graphics file format Alembic, developed by Sony Pictures Imageworks and Industrial Light & Magic; Sony Imageworks' Field3D voxel data storage system; and the OpenEXR projected initiated by ILM – among many others.

By way of example: forking a popular public implementation of a GAN framework so that it better fits the needs of a VFX house is fraught with long-term obligation, and the likelihood that the branch will periodically become incompatible with new developments in the powerful open source libraries that made the project viable in the first place. Every time TensorFlow takes a version leap, or Keras gets a new commit, in-house proprietary VFX software developed from such sources must either adapt, if it can, or accept the need to re-tool its own workflow from the revised code, and fix everything that just broke — again.

Therefore, given the provenance of the mainstream research that's feeding new innovation in AI-driven VFX, and the relatively unilateral way that it flows to Hollywood effects houses from the economic and state-sponsored powerhouses of the world, version control, still a nascent field in machine learning, seems likely to become an essential aspect of a 'satellite culture' of collaborative open source AI efforts in the industry.

NVIDIA and the Future of the GPU in Large-Scale Neural Net Pipelines

NVIDIA is now so bound to the continuing evolution of machine learning that its future prospects and intentions are a variable to be considered for anyone committing to a GPU-based machine learning pipeline. When the company released its cuDNN machine learning library in 2014, it proved such a powerful aid in machine learning processing that it was quickly adopted by all the most important schemes and frameworks in global AI research, including Caffe, Theano, Keras, Torch, TensorFlow, and CNTK.

In considering a future driven by machine learning, effects houses must currently regard the prospect of a long term lock-in to a single provider that is splitting its resources and roadmaps among a vast range of sectors, and which seems likely to be challenged soon by new innovations in ASIC chip designs specifically aimed at machine learning.

The company's recent overstocking, due to misreading of the crypto-mining market, represents a coincidental short-term abundance rather than any kind of provisioning strategy for the future of GPU in general AI, or to meet the potential needs of the VFX industry. NVIDIA might well pick its battles and its markets quite selectively from now on, making the availability of potential new GPU-based network architectures either limited or very expensive in the medium term – likely both. The volatility of the market, where demand can plummet from famine to glut inside a year, seems likely to lead to a circumspect attitude on the part of NVIDIA.

If you're wondering how the platform of a single company came to a position of such influence in one of the most important revolutions in the history of technology, you're not alone. Though cuDNN can be exploited in

a similar manner to a genuinely open source platform, it remains linked to NVIDIA's hardware units, prompting calls from various critics to disassociate the evolution of AI from the destiny of NVIDIA.

While there are open source alternatives (and even a CUDA emulator or two) available, cuDNN currently links the fate of neural networks to the whims, wishes and general fortunes of NVIDIA. With no clear open source and hardware-agnostic challenger, and with AMD's machine learning hardware and frameworks a generation behind CUDA, only players at the level of Google seem to have a chance of diversifying the market via cloud-based services (though, in the case of Google, that seems unlikely to be the intent).

Conclusion

Considering the timeframe of the potential transformation from CGI to ML-based visual effects, Charlie Winter says "It won't take too long, I think. The studios who choose to embrace AI will ultimately end up being able to bid lower on projects, leaving the other studios struggling to stay afloat.

"What's happened this year is just a small wave of ML tools, mostly doing primitive kind of things kind which roughly relate to non-domain specific research papers. Looking at the amount of money that's been invested into domain-specific VFX-AI R&D over the past year, I'd say there is a tsunami on the horizon of the next two or three years.

"My advice to any artist or studio who hasn't tried to get a handle on the current wave of change is to try to get a grip before the big wave shows up."

It seems likely that those who have made a career in the low-hanging fruit for AI-driven VFX processes should think quickly about re-skilling. Talking about the potential of AI to transform new ML-based image extraction systems, Digital Domain's Doug Roble says:

> *"[Once] they're built, it's just going to change the industry — and it already is starting to change the industry."*