GAN as a Face Renderer for ‘Traditional’ CGI

Originally published at https://www.unite.ai/gan-as-a-face-renderer-for-traditional-cgi/

Opinion When Generative Adversarial Networks (GANs) first demonstrated their capability to reproduce stunningly realistic 3D faces, the advent triggered a gold rush for the unmined potential of GANs to create temporally consistent video featuring human faces.

Somewhere in the GAN’s latent space, it seemed that there *must* be hidden order and rationality – a schema of nascent semantic logic, buried in the latent codes, that would allow a GAN to generate consistent multiple views and multiple interpretations (such as expression changes) of the *same* face – and subsequently offer a temporally-convincing deepfake video method that would blow autoencoders out of the water.

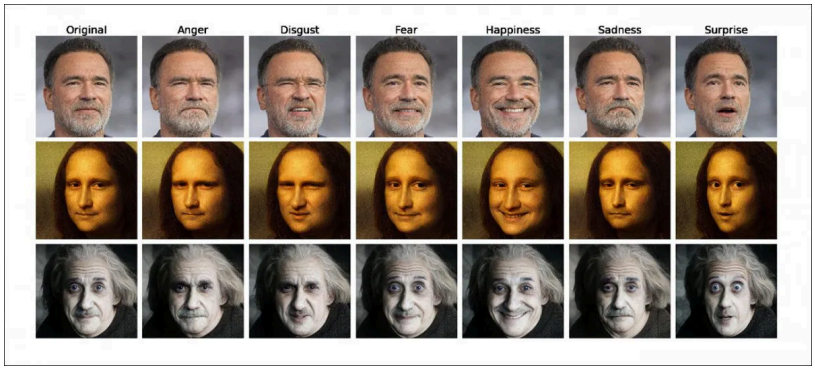
High-resolution output would be trivial, compared to the slum-like low-res environments in which GPU constraints force DeepFaceLab and FaceSwap to operate, while the ‘swap zone’ of a face (in autoencoder workflows) would become the ‘creation zone’ of a GAN, informed by a handful of input images, or even just a single image.

There would be no more mismatch between the ‘swap’ and ‘host’ faces, because the *entirety* of the image would be generated from scratch, including hair, jawlines, and the outermost extremities of the facial lineaments, which frequently prove a challenge for ‘traditional’ autoencoder deepfakes.

The GAN Facial Video Winter

As it transpired, it was not going to be nearly that easy. Ultimately, disentanglement proved the central issue, and remains the primary challenge. How can you keep a distinct facial identity, and change its pose or expression without gathering together a corpus of thousands of reference images that teach a neural network what happens when these changes are enacted, the way that autoencoder systems so laboriously do?

Rather, subsequent thinking in GAN facial enactment and synthesis research was that an input identity could perhaps be made subject to teleological, generic, *templated* transformations that are not identity-specific. An example of this would be to apply an expression to a GAN face that was not present in any of the images of that person that the GAN knows about.

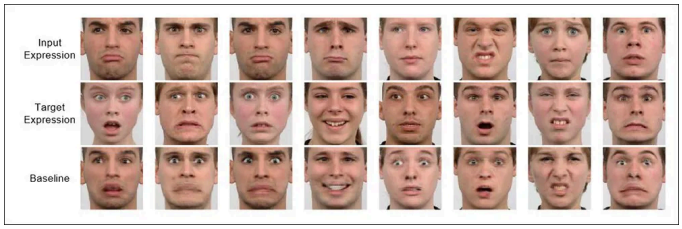


From the 2022 paper Tensor-based Emotion Editing in the StyleGAN Latent Space, templated expressions are applied to an input face from the FFHQ dataset. Sc
https://arxiv.org/pdf/2205.06102.pdf

It is obvious that a ‘one size fits all’ approach can’t cover the diversity of facial expressions unique to an individual. We have to wonder if a smile as unique as that of Jack Nicholson or Willem Dafoe could ever receive a faithful interpretation under the influence of such ‘mean average expression’ latent codes.

Who is this charming Latin stranger? Though the GAN method produces a more 'realistic' and higher-resolution face, the transformation is not informed by multiple real-world images of the actor, as is the case with DeepFaceLab, which trains extensively on a database of thousands of such images, and consequently the resemblance is compromised. Here (background) a DeepFaceLab model is imported into DeepFaceLive, a streaming implementation of the popular and controversial software. Examples are from https://www.youtube.com/watch?v=9tr35y-yQRY (2022) and https://arxiv.org/pdf/2205.06102.pdf.

A number of GAN facial expression editors have been put forward over the last few years, most of them dealing with unknown identities, where the fidelity of the transformations is impossible for the casual reader to know, since these are not familiar faces.



Obscure identities transformed in the 2020 offering Cascade-EF-GAN. Source: https://arxiv.org/pdf/2003.05905.pdf

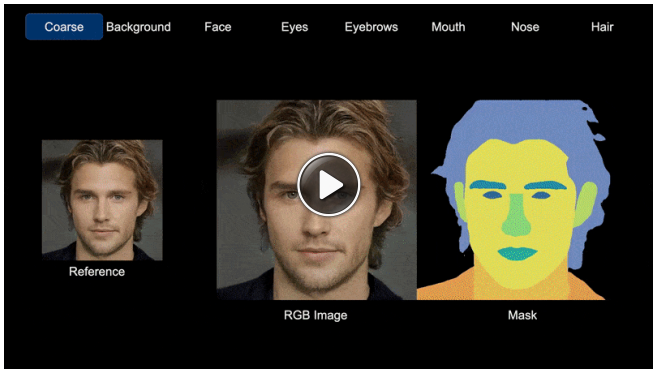
Perhaps the GAN face editor that has received the most interest (and citations) in the last three years is InterFaceGAN, which can perform latent space traversals in latent codes relating to pose (angle of the camera/face), expression, age, race, gender, and other essential qualities.

InterFaceGAN Demo (CVPR 2020)

The 1980s-style 'morphing' capabilities of InterFaceGAN and similar frameworks are mainly a way to illustrate the path towards transformation as an image is reprojected back through an apposite latent code (such as 'age'). In terms of producing video footage with temporal continuity, such schemes to date have qualified as 'impressive disasters'.

If you add to that the difficulty of creating temporally-consistent hair, and the fact that the technique of latent code exploration/manipulation has no innate temporal guidelines to work with (and it is difficult to know how to inject such guidelines into a framework designed to accommodate and generate still images, and which has no native provision for video output), it might be logical to conclude that GAN is not All You Need™ for facial video synthesis.

Therefore, subsequent efforts have yielded incremental improvements in disentanglement, while others have bolted on other conventions in computer vision as a 'guidance layer', such as the use of semantic segmentation as a control mechanism in the late 2021 paper *SemanticStyleGAN: Learning Compositional Generative Priors for Controllable Image Synthesis and Editing*.



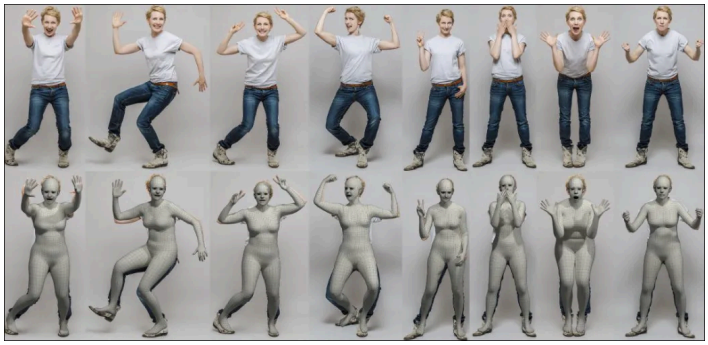
CLICK TO VIEW ANIMATED GIF

Semantic segmentation as a method of latent space instrumentality in SemanticStyleGAN. Source: https://semanticstylegan.github.io/

Parametric Guidance

The GAN facial synthesis research community is steering increasingly towards the use of 'traditional' parametric CGI faces as a method to guide and bring order to the impressive but unruly latent codes in a GAN's latent space.

Though parametric facial primitives have been a staple of computer vision research for over twenty years, interest in this approach has grown lately, with the increased use of Skinned Multi-Person Linear Model (SMPL) CGI primitives, an approach pioneered by the Max Planck Institute and ILM, and since improved upon with the Sparse Trained Articulated Human Body Regressor (STAR) framework.

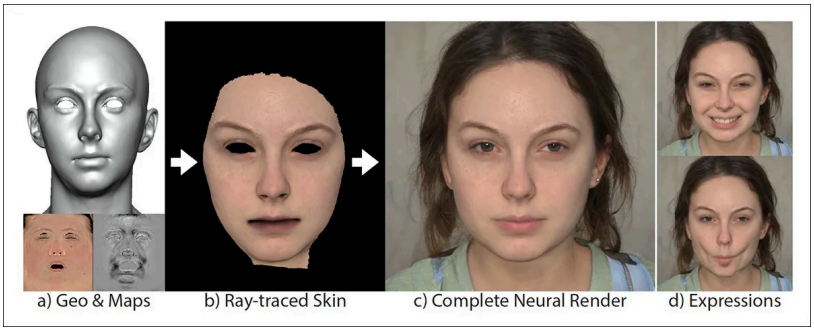


SMPL (in this case a variant called SMPL-X) can impose a CGI parametric mesh that accords with the estimated pose (including expressions, as necessary) of the entirety of the human body featured in an image, allowing new operations to be performed on the image using the parametric mesh as a volumetric or perceptual guideline. Source: https://arxiv.org/pdf/1904.05866.pdf

The most acclaimed development in this line has been Disney's 2019 *Rendering with Style* initiative, which melded the use of traditional texture-maps with GAN-generated imagery, in an attempt to create improved, 'deepfake-style' animated output.

Old meets new, in Disney's hybrid approach to GAN-generated deepfakes. Source: https://www.youtube.com/watch?v=TwpLqTmvqVk

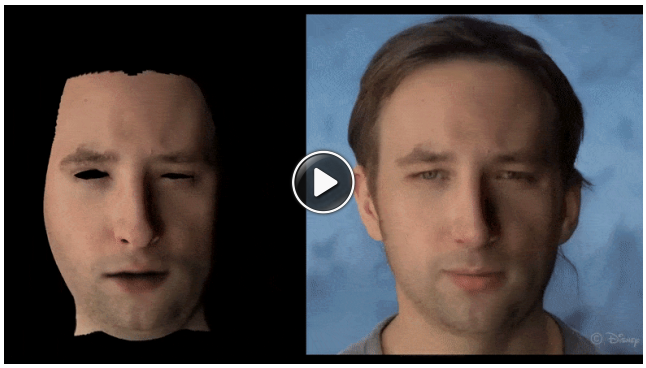
The Disney approach imposes traditionally rendered CGI facets into a StyleGAN2 network to 'inpaint' human facial subjects in 'problem areas', where temporal consistency is an issue for video generation – areas such as skin texture.



The Rendering with Style workflow.

Since the parametric CGI head that guides this process can be tweaked and changed to suit the user, the GAN-generated face is able to reflect those changes, including changes of head pose and expression.

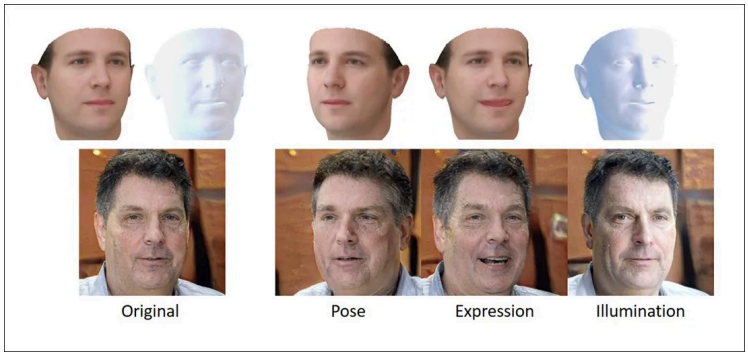
Though designed to marry the instrumentality of CGI with the natural realism of GAN faces, in the end, the results demonstrate the worst of both worlds, and still fail to keep hair texture and even basic feature positioning consistent:



CLICK TO VIEW ANIMATED GIF

A new kind of uncanny valley emerges from Rendering with Style, though the principle still holds some potential.

The 2020 paper *StyleRig: Rigging StyleGAN for 3D Control over Portrait Images* takes an increasingly popular approach, with the use of three-dimensional morphable face models (3DMMs) as proxies for altering characteristics in a StyleGAN environment, in this case through a novel rigging network called RigNet:



3DMMs stand in as proxies for latent space interpretations in StyleRig. Source: https://arxiv.org/pdf/2004.00121.pdf

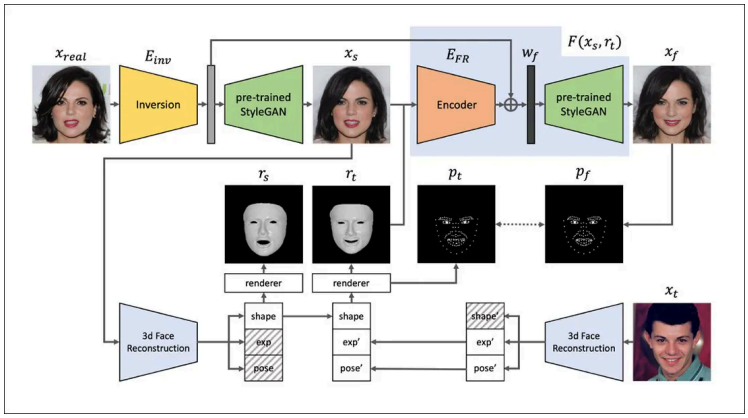
However, as usual with these initiatives, the results to date seem limited to minimal pose manipulations, and 'uninformed' expression/affect changes.

CLICK TO VIEW ANIMATED GIF

StyleRig improves on the level of control, though temporally consistent hair remains an unsolved challenge.
Source: https://www.youtube.com/watch?v=eaW_P85wQ9k

Similar output can be found from Mitsubishi Research's *MOST-GAN*, a 2021 paper that uses nonlinear 3DMMs as a disentanglement architecture, but which also struggles to achieve dynamic and consistent motion.

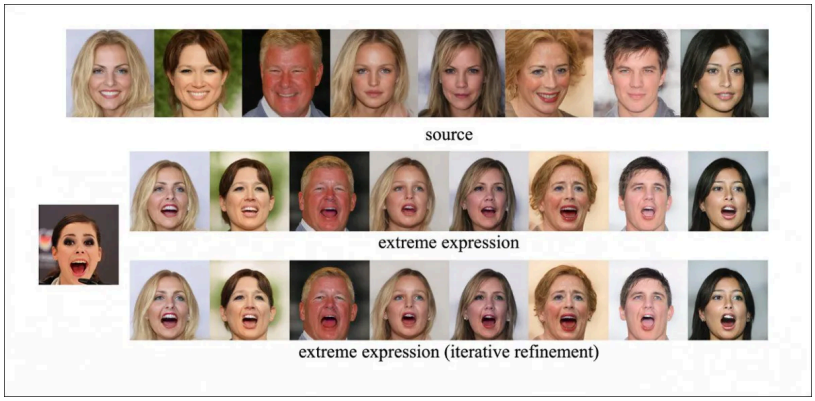
The latest research to attempt instrumentality and disentanglement is *One-Shot Face Reenactment on Megapixels*, which again uses 3DMM parametric heads as a friendly interface for StyleGAN.



In the MegaFR workflow of One-Shot Face Reenactment, the network performs facial synthesis by combining an inverted real-world image with parameters taken 3DMM model. Source: https://arxiv.org/pdf/2205.13368.pdf

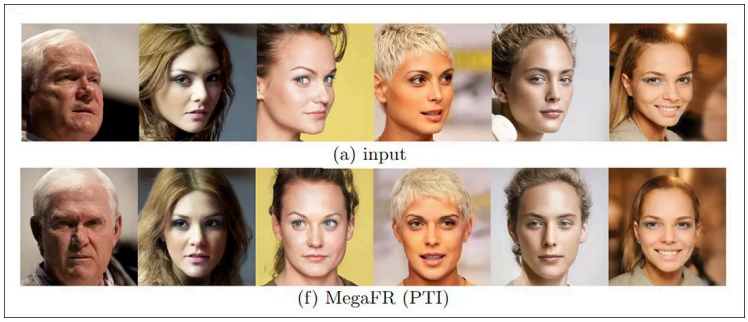
OSFR belongs to a growing class of GAN face editors that seek to develop Photoshop/After Effects-style linear editing workflows where the user can input a desired image on which transformations can be applied, rather than hunting through the latent space for latent codes relating to an identity.

Again, parametric expressions represent an overarching and non-personalized method of injecting expression, leading to manipulations that seem 'uncanny' in their own, not always positive way.



Injected expressions in OSFR.

Like prior work, OSFR can infer near-original poses from a single image, and also perform 'frontalization', where an off-center posed image is translated into a mugshot:



Original (above) and inferred mugshot images from one of the implementations of OSFR detailed in the new paper.

In practice, this kind of inference is similar to some of the photogrammetry principles that underpin Neural Radiance Fields (NeRF), except that the geometry here must be defined by a single photo, rather than the 3-4 viewpoints that allow NeRF to interpret the missing interstitial poses and create explorable neural 3D scenes featuring humans.

(However, NeRF is not All You Need™ either, as it bears an almost entirely different set of roadblocks to GANs in terms of producing facial video synthesis)

Does GAN Have a Place in Facial Video Synthesis?

Achieving dynamic expressions and out-of-distribution poses from a single source image seems to be an alchemy-like obsession in GAN facial synthesis research at the moment, chiefly because GANs are the only method currently capable of outputting quite high resolution and relatively high-fidelity neural faces: though autoencoder deepfake frameworks can train on a multitude of real-world poses and expressions, they must operate at VRAM-restricted input/output resolutions, and require a 'host'; while NeRF is similarly constrained, and – unlike the other two approaches – currently has no established methodologies for changing facial expressions, and suffers from limited editability in general.

It seems that the only way forward for an accurate CGI/GAN face synthesis system is for a new initiative to find some way of assembling a multi-photo identity entity inside the latent space, where a latent code for a person's identity does not have to travel all the way across the latent space to exploit unrelated pose parameters, but can refer to its own related (real world) images as references for transformations.

Even in such a case, or even if an entire StyleGAN network were trained on a single-identity face-set (similar to the training sets that autoencoders use), the lacking semantic logic would still likely need to be provided by adjunct technologies such as semantic segmentation or parametric 3DMM faces, which, in such a scenario, would at least have more material to work with.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More