# Deepfake Detectors Pursue New Ground: Latent Diffusion Models and GANs

Originally published at https://www.unite.ai/deepfake-detectors-pursue-new-ground-latent-diffusion-models-and-gans/

***Opinion***

Of late, the deepfake detection research community, which has since late 2017 been occupied almost exclusively with the autoencoder-based framework that premiered at that time to such public awe (and dismay), has begun to take a forensic interest in less stagnant architectures, including latent diffusion models such as DALL-E 2 and Stable Diffusion, as well as the output of Generative Adversarial Networks (GANs). For instance, in June, UC Berkeley published the results of its research into the development of a detector for the output of the then-dominant DALL-E 2.

What seems to be driving this growing interest is the sudden evolutionary jump in the capability and availability of latent diffusion models in 2022, with the closed-source and limited-access release of DALL-E 2 in spring, followed in late summer by the sensational open sourcing of Stable Diffusion by stability.ai.

GANs have also been long-studied in this context, though less intensively, since it is very difficult to use them for convincing and elaborate video-based recreations of people; at least, compared to the by-now venerable autoencoder packages such as FaceSwap and DeepFaceLab – and the latter's live-streaming cousin, DeepFaceLive.

## Moving Pictures

In either case, the galvanizing factor appears to be the prospect of a subsequent developmental sprint for *video* synthesis. The start of October – and 2022's major conference season – was characterized by an avalanche of sudden and unexpected solutions to various longstanding video synthesis bugbears: no sooner had Facebook released samples of its own text-to-video platform, than Google Research quickly drowned out that initial acclaim by announcing its new Imagen-to-Video T2V architecture, capable of outputting high resolution footage (albeit only via a 7-layer network of upscalers).

If you believe that this kind of thing comes in threes, consider also stability.ai's enigmatic promise that 'video is coming' to Stable Diffusion, apparently later this year, while Stable Diffusion co-developer Runway have made a similar promise, though it is unclear whether they are referring to the same system. The Discord message from Stability's CEO Emad Mostaque also promised *'audio, video [and] 3d'*.

What with an out-of-the-blue offering of several new audio generation frameworks (some based on latent diffusion), and a new diffusion model that can generate authentic character motion, the idea that 'static' frameworks such as GANs and diffusers will finally take their place as supporting *adjuncts* to external animation frameworks is starting to gain real traction.

In short, its seems likely that the hamstrung world of autoencoder-based video deepfakes, which can only effectively substitute the central portion of a face, could by this time next year be eclipsed by a new generation of diffusion-based deepfake-capable technologies – popular, open source approaches with the potential to photorealistically fake not just entire bodies, but entire scenes.

For this reason, perhaps, the anti-deepfake research community is beginning to take image synthesis seriously, and to realize that it might serve more ends than just generating fake LinkedIn profile photos; and that if all their intractable latent spaces can accomplish in terms of temporal motion is to act as a really great texture renderer, that might actually be more than enough.

## Blade Runner

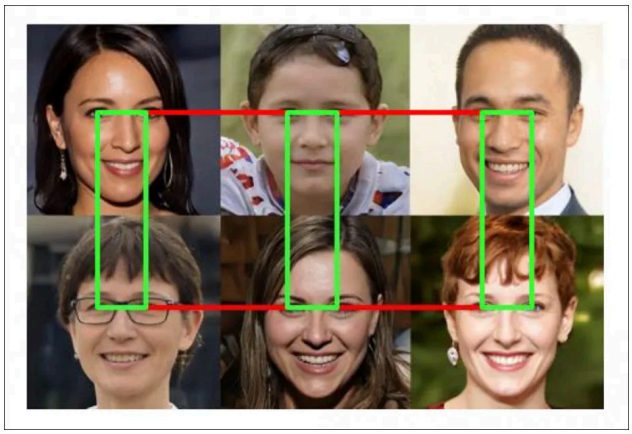
The latest two papers to address, respectively, latent diffusion and GAN-based deepfake detection, are, respectively, *DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models*, a collaboration between the CISPA Helmholtz Center for Information Security and Salesforce

; and *BLADERUNNER: Rapid Countermeasure for Synthetic (AI-Generated) StyleGAN Faces*, from Adam Dorian Wong at MIT's Lincoln

Laboratory.

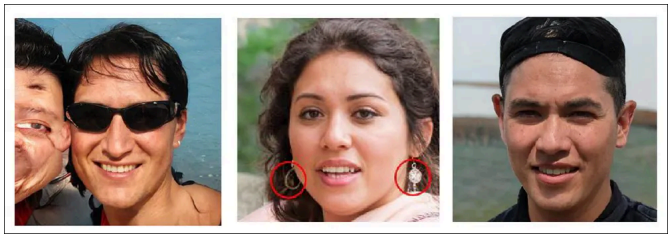
Before explaining its new method, the latter paper takes some time to examine previous approaches to determining whether or not an image was generated by a GAN (the paper deals specifically with NVIDIA's StyleGAN family).

The 'Brady Bunch' method – perhaps a meaningless reference for anyone who was not watching TV in the 1970s, or who missed the 1990s movie adaptations – identifies GAN-faked content based on the fixed positions that particular parts of a GAN face are certain to occupy, due to the rote and templated nature of the 'production process'.

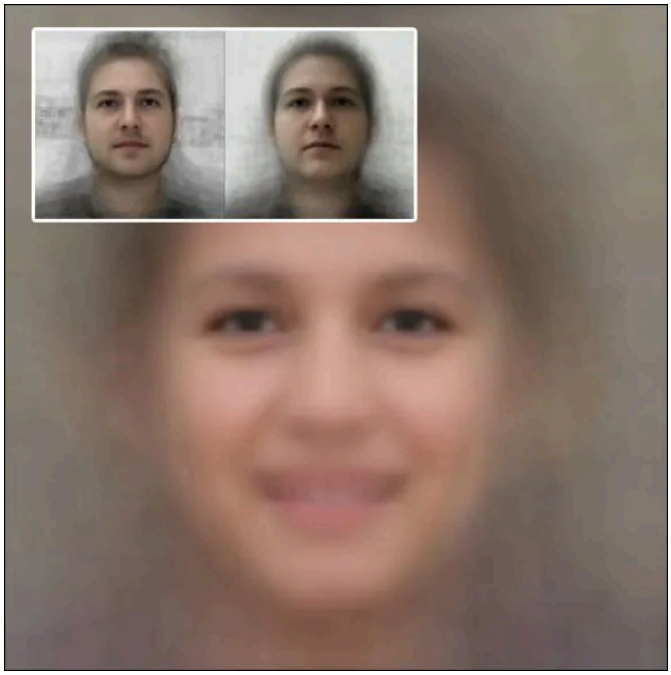


*The 'Brady Bunch' method propounded by a webcast from the SANS institute in 2022: a GAN-based face generator will perform improbably uniform placement of certain facial features, belying the origin of the photo, in certain cases.* Source: https://arxiv.org/ftp/arxiv/papers/2210/2210.06587.pdf

Another useful known indication is StyleGAN's frequent inability to render multiple faces (first image below), if necessary, as well as its lack of talent in accessory coordination (middle image below), and a tendency to use a hairline as the start of an impromptu hat (third image below).

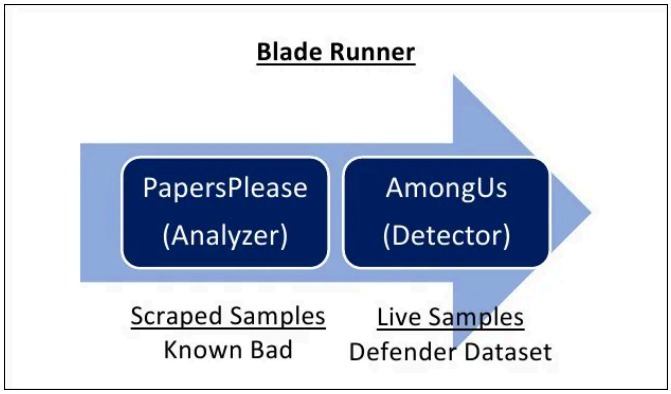


The third method that the researcher draws attention to is *photo overlay* (an example of which can be seen in our August article on AI-aided diagnosis of mental health disorders), which uses compositional 'image blending' software such as the CombineZ series to concatenate multiple images into a single image, often revealing underlying commonalities in structure – a potential indication of synthesis.



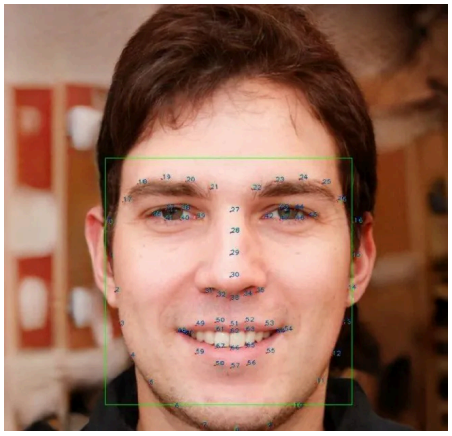
The architecture proposed in the new paper is titled (possibly against all SEO advice) *Blade Runner*, referencing the Voight-Kampff test that determines whether antagonists in the sci-fi franchise are 'fake' or not.

The pipeline is composed of two phases, the first of which is the PapersPlease analyzer, which can evaluate data scraped from known GAN-face websites such as thispersondoesnotexist.com, or generated.photos.




Though a cut-down version of the code can be inspected at GitHub (see below) few details are provided about this module, except that OpenCV and DLIB are used to outline and detect faces in the gathered material.

The second module is the *AmongUs* detector. The system is designed to search for coordinated eye placement in photos, a persistent feature of StyleGAN's face output, typified in the 'Brady Bunch' scenario detailed above. AmongUs is powered by a standard 68-landmark detector.



*Facial point annotations via the Intelligent Behaviour Understanding Group (IBUG), whose facial landmark plotting code is used in the Blade Runner package.*

AmongUs depends on pre-trained landmarks based on the known 'Brady bunch' coordinates from PapersPlease, and is intended for use against live, web-facing samples of StyleGAN-based face images.

Blade Runner, the author suggests, is a plug-and-play solution intended for companies or organizations that lack resources to develop in-house solutions for the kind of deepfake detection dealt with here, and a 'stop-gap measure to buy time for more permanent countermeasures'.

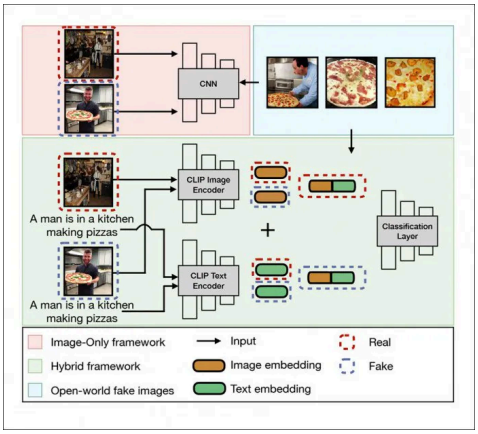
In fact, in a security sector this volatile and fast-growing, there are not many bespoke *or* off-the-rack cloud vendor solutions to which an under-resourced company can currently turn to with confidence.

Though Blade Runner performs poorly against *bespectacled* StyleGAN-faked people, this is a relatively common problem across similar systems, which are expecting to be able to evaluate eye delineations as core points of reference, obscured in such cases.

A reduced version of Blade Runner has been released to open source on GitHub. A more feature-rich proprietary version exists, which can process multiple photos, rather than the single photo per operation of the open source repository. The author intends, he says, to upgrade the GitHub version to the same standard eventually, as time allows. He also concedes that StyleGAN is likely to evolve beyond its known or current weaknesses, and the software will likewise need to develop in tandem.

## DE-FAKE

The DE-FAKE architecture aims not only to achieve 'universal detection' for images produced by text-to-image diffusion models, but to provide a method to discern *which* latent diffusion (LD) model produced the image.




*The universal detection framework in DE-FAKE addresses local images, a hybrid framework (green), and open world images (blue).* Source: http://export.arxiv.org/pdf/2210.06998

To be honest, at the moment, this is a fairly facile task, since all of the popular LD models – closed or open source – have notable distinguishing characteristics.

Additionally, most share some common weaknesses, such as a predisposition to cut off heads, because of the arbitrary way that non-square web-scraped images are ingested into the massive datasets that power systems such as DALL-E 2, Stable Diffusion and MidJourney:


*Latent diffusion models, in common with all computer vision models, require square-format input; but the aggregate web-scraping that fuels the LAION5B dataset ‘luxury extras’ such as the ability to recognize and focus on faces (or anything else), and truncates images quite brutally instead of padding them out (which would entire source image, but at a lower resolution). Once trained in, these ‘crops’ become normalized, and very frequently occur in the output of latent diffusion system Stable Diffusion.* Sources: https://blog.novelai.net/novelai-improvements-on-stable-diffusion-e10d38db82ac and Stable Diffusion.

DE-FAKE is intended to be algorithm-agnostic, a long-cherished goal of autoencoder anti-deepfake researchers, and, right now, quite an achievable one in regard to LD systems.

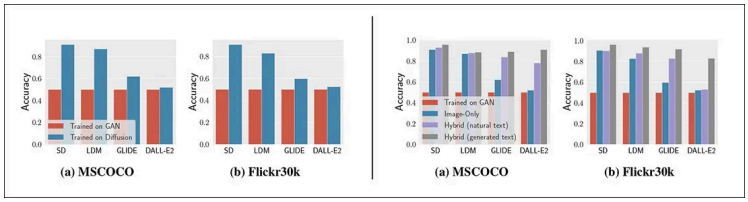
The architecture uses OpenAI’s Contrastive Language-Image Pretraining (CLIP) multimodal library – an essential element in Stable Diffusion, and fast becoming the heart of the new wave of image/video synthesis systems – as a way to extract embeddings from ‘forged’ LD images and train a classifier on the observed patterns and classes.

In a more ‘black box’ scenario, where the PNG chunks that hold information about the generation process have long been stripped away by uploading processes and for other reasons, the researchers use the Salesforce BLIP framework (also a component in at least one distribution of Stable Diffusion) to ‘blindly’ poll the images for the likely semantic structure of the prompts that created them.

| Model | Dataset | Images | Image size |
|---|---|---|---|
| Stable Diffusion | MSCOCO<br>Flickr30k | 59247<br>13231 | 512*512<br>512*512 |
| Latent Diffusion | MSCOCO<br>Flickr30k | 29276<br>17969 | 256*256<br>256*256 |
| GLIDE | MSCOCO<br>Flickr30k | 41685<br>27210 | 256*256<br>256*256 |
| DALL-E2 | MSCOCO<br>Flickr30k | 1028<br>300 | 256*256<br>256*256 |

*The researchers used Stable Diffusion, Latent Diffusion (itself a discrete product), GLIDE and DALL-E 2 to populate a training and testing dataset leveraging MSCOCO and Flickr30k.*

Normally we would take quite an extensive look at the results of the researchers’ experiments for a new framework; but in truth, DE-FAKE’s findings seem likely to be more useful as a future benchmark for later iterations and similar projects, rather than as a meaningful metric of project success, considering the volatile environment that it is operating in, and that the system it is competing against in the paper’s trials is nearly three years old – from back when the image synthesis scene was truly nascent.




*Left-most two images: the 'challenged' prior framework, originated in 2019, predictably faring less well against DE-FAKE (rightmost two images) across the four LD*

The team's results are overwhelmingly positive for two reasons: there is scant prior work against which to compare it (and none at all that offers a fair comparison, i.e., that covers the mere twelve weeks since Stable Diffusion was released to open source).

Secondly, as mentioned above, though the LD image synthesis field is developing at exponential speed, the output content of current offerings effectively watermarks itself by dint its own structural (and very predictable) shortcomings and eccentricities – many of which are likely to be remediated, in the case of Stable Diffusion at least, by the release of the better-performing 1.5 checkpoint (i.e. the 4GB trained model powering the system).

At the same time, Stability has already indicated that it has a clear roadmap for V2 and V3 of the system. Given the headline-grabbing events of the last three months, any corporate torpor on the part of OpenAI and other competing players in the image synthesis space is likely to have been evaporated, meaning that we can expect a similarly brisk pace of progress also in the closed-source image synthesis space.

*First published 14th October 2022.*

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**