To Uncover a Deepfake Video Call, Ask the Caller to Turn Sideways

By Martin Anderson

First published ***August 8th 2022*** *at:*
https://metaphysic.ai/to-uncover-a-deepfake-video-call-ask-the-caller-to-turn-sideways/
Web-archived version

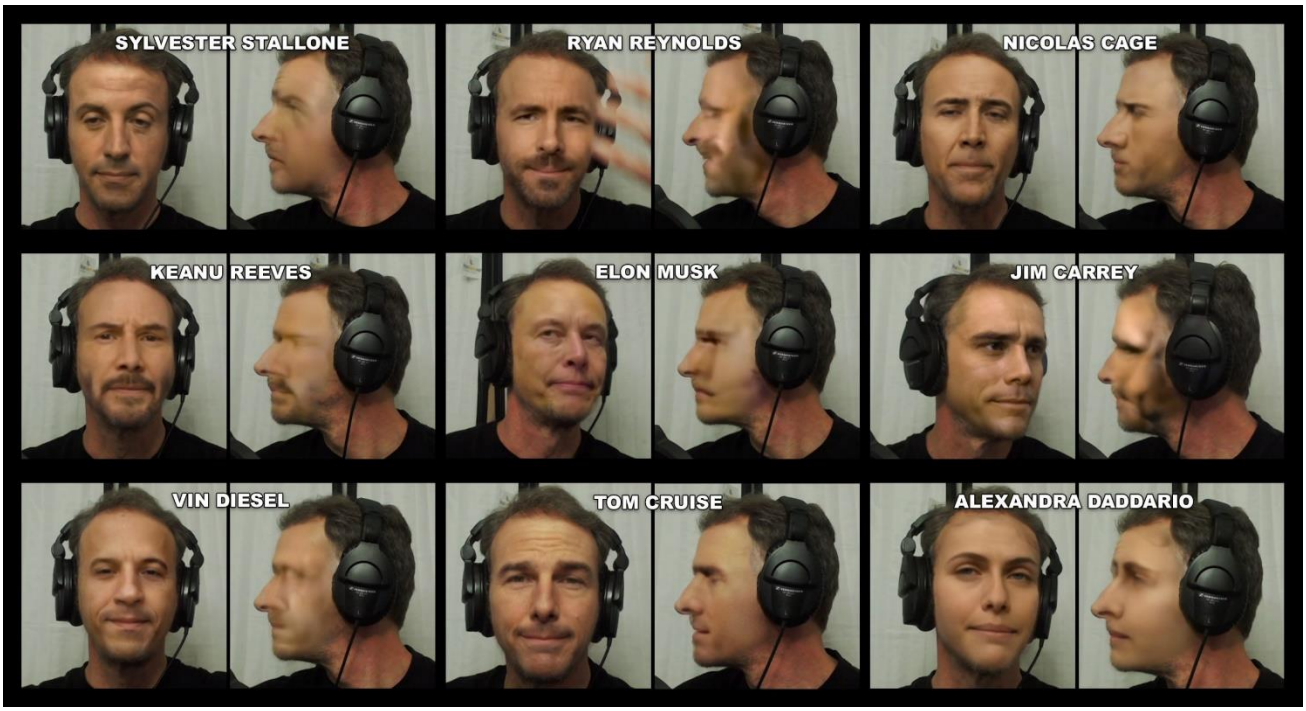
There is an interesting vulnerability in video deepfakes that, to date, has been generally overlooked by the security research community, perhaps because 'live', real-time deepfakes in video calls have not been a major cause for concern until very recently.

For a number of reasons, which we'll examine in this article, deepfakes are not usually very good at recreating profile views:

The above examples are taken* from a session with tech exponent and commenter Bob Doyle, who agreed to run some tests with us, using DeepFaceLive to change his appearance to that of a series of popular celebrities.

DeepFaceLive is a live-streaming version of the popular DeepFaceLab software, and is capable of creating alternate video identities in real-time.

From more or less face-on viewpoints, most of the celebrity recreations are quite effective, and some are very convincing even at fairly acute angles – until the facial angle hits a full 90°.



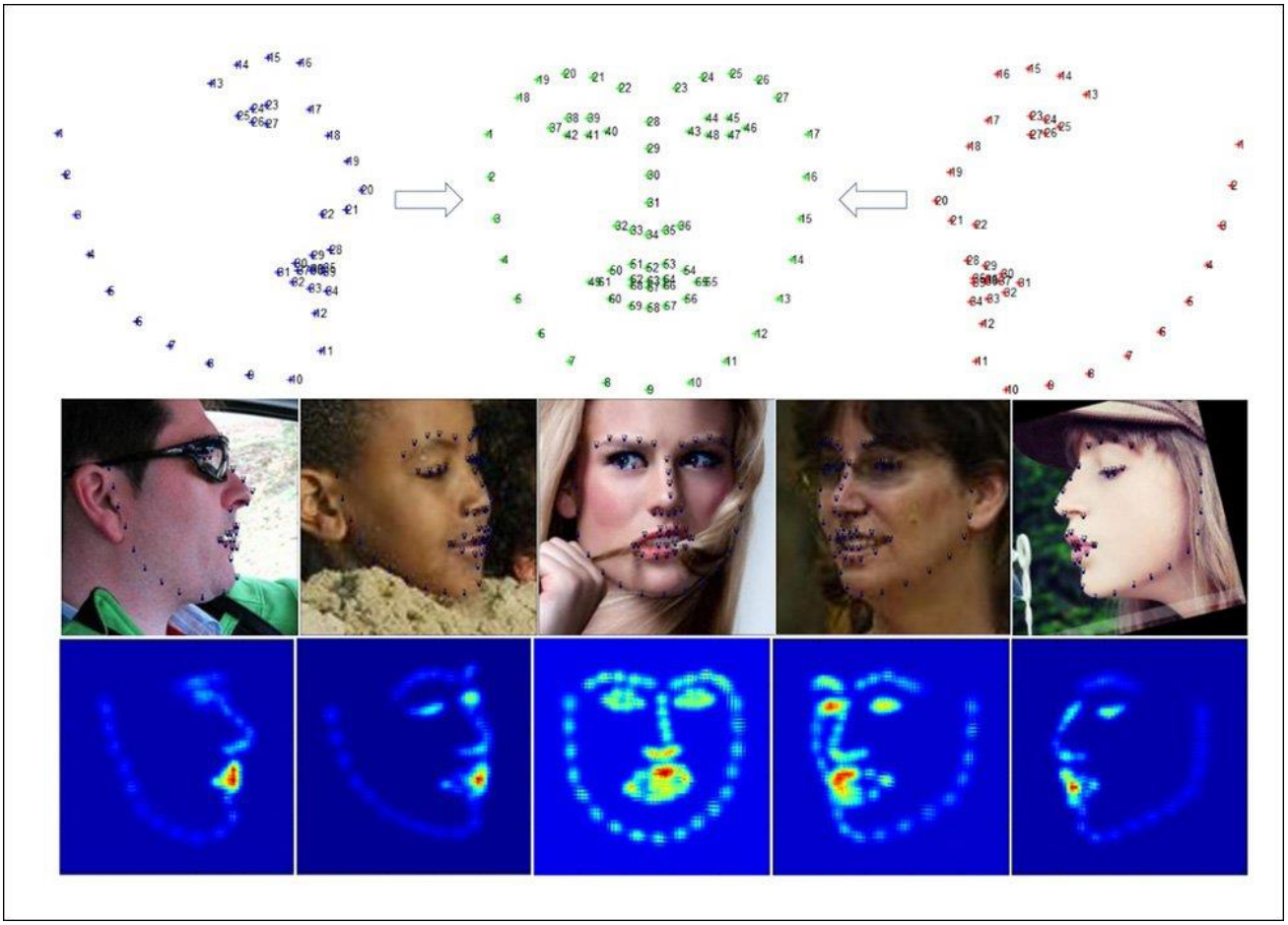
It's evident also that Bob's real profile lineaments entirely survive the deepfake process for all these models, none of which have been trained with enough good-quality profile data to be capable of either transforming the distinct boundaries of the face, or performing the inpainting necessary to simulate 'revealed background' (i.e. when the 'host's' profile extends further than that of the 'guest' identity, and more of the background needs to be 'invented' – see image below).

Any shortfall between the facial profile of the target identity and that of the host will have to be 'invented' by the deepfake model, through inpainting. This kind of capability is only likely to develop in the model if it has been trained on abundant profile views, to a very high number of iterations (usually over a million, which may take 6-10 days, depending on resolution and settings).

Why does recreation fail so steeply when a deepfaked subject goes into sharp 90° profile? Is this a possible way to detect whether or not the person you're talking to in a videoconference is real or deepfaked? And what, if anything, can deepfakers themselves do to 'fix' it?

Lateral Limitations

The standard software (Facial Alignment Network) that estimates facial poses in images, in deepfakes packages, does not work reliably at acute angles. In fact, most 2D-based facial alignments algorithms assign only 50-60% of the number of landmarks from a front-on face view to a profile view.



From the 2015 paper 'Joint Multi-view Face Alignment in the Wild', which showcases the Multi-view Hourglass facial alignment model. Frontal alignments contain 68 landmarks, while profile alignments have only 39. Source: https://arxiv.org/pdf/1708.06023.pdf

Typical 2D alignment packages consider a profile view to be 50% hidden, which hinders recognition, as well as accurate training and subsequent face synthesis.

Frequently the generated profile landmarks will 'leap out' to any possible group of pixels that may represent a 'missing eye' or other facial detail that's obscured in a profile view:

In the above example, we see some typical profile extraction issues in a clip from *All The President's Men* (1976), in spite of the fact that there are no distracting background patterns (such as wallpaper or patterned curtains) to 'fool' FAN Align into believing that they might constitute part of a face (which is a common problem).

For this reason, in spite of manual intervention and various clean-up processes that can be used, sheer profile shots are likely to be less temporally consistent than most frames in extracted videos, as well as in the training of deepfake models based on those extracted images, and in the final reconstruction, where a new identity is finally superimposed on a 'target'.

The 'Profile Data' Desert Outside Hollywood

Cognizant of this weak spot in deepfakes, most viral deepfakers tend to avoid using clips that require well-trained, accurate and temporally consistent profile views.

Some, however, have made a special effort. YouTube deepfaker DesiFakes told us that he achieved an extraordinary profile view of Jerry Seinfeld inserted into a tense scene from *Pulp Fiction* (1994) through extensive post-processing, and that the above-average side-view of Seinfeld is aided also by a close resemblance between the comedian and original actor Alexis Arquette:

For such heroic (and untypical) profile efforts, we need to consider the high availability of data for notable Hollywood TV and movie actors. By itself, the TV show *Seinfeld* represents 66 hours of available footage, the majority featuring Jerry Seinfeld, with abundant profile footage on display due to the frequent multi-person conversations.

Matt Damon's current movie output alone, likewise, has a rough combined runtime of 144 hours, most of it available in high-definition.

By contrast, how many profile shots do you have of yourself?

Unless you've been arrested at some point, it's likely that you don't have even *one* such image, either on social media or in an offline collection.

Jane Fonda in profile, aged 32, courtesy of the Cleveland police department, who arrested her in November of 1970 on (contested) charges of drug smuggling. Note that even this right-hand image is not a pure profile view, but at best an 80-85° stance (Fonda herself has long since embraced the iconic status of this image). Source: https://clevelandmemory.contentdm.oclc.org/digital/collection/general/id/8076/rec/2

Besides clinicians, VFX artists, forensics experts and police authorities, nobody wants profile shots. Photographers will fight a crowd to escape them; picture editors can't sell them (or can't sell them as well as a 'real' photo); and they are in general charmless and prosaic representations of us that many of us literally would barely even recognize as ourselves.

Ioana Grecu, Content Manager for the stock image domain Dreamstime, comments to us on the slim demand for the side-on view:

> *'Most customers of stock sites will look for images that connect with their end subject. My guess is that a profile shot will do this job in only very few and specific cases and mostly only in instances where a concept image is needed.'*

And continues:

> *'[The] profile shot is one of the least flattering ones and the one where not much can be expressed either by the model or the camera angle. While a photographer can use light to express emotion, such an image*

> *would likely not be fit for your expressed purpose…The main reason photographers don't take these shots is because there are so many more angles that have more to offer in a traditional photo shoot.'*

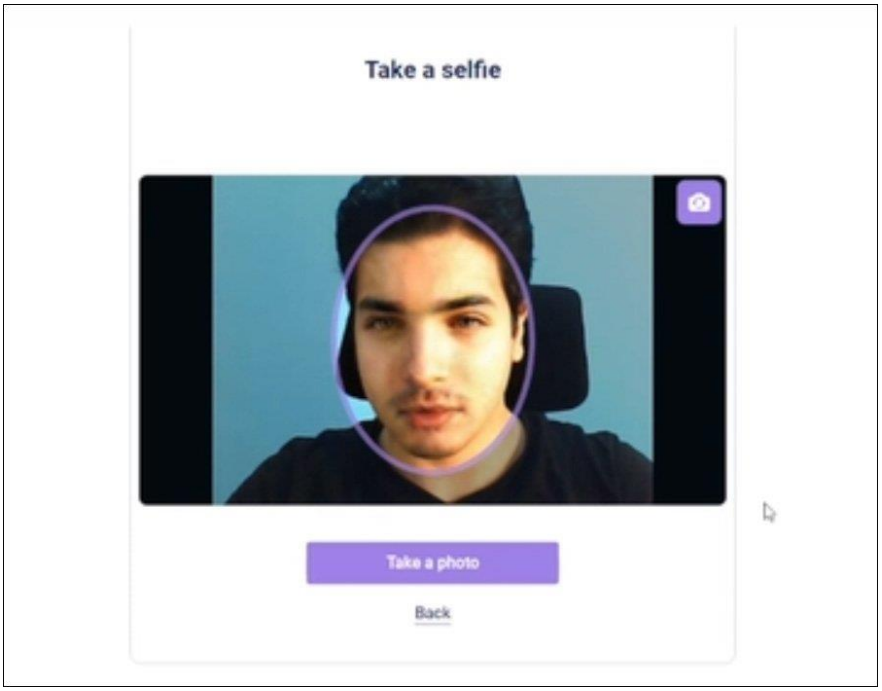
Using Profile Requests for Deepfake Detection in Video Calls

That paucity of available data makes it difficult to obtain a range of profile images on *non*-celebrities that's diverse and extensive enough to train a deepfake model to reproduce profile views convincingly.

Consequently, this weakness in deepfakes offers a potential way of uncovering 'simulated' correspondents in live video calls, recently classified as an emergent risk by the FBI: if you suspect that the person you're talking to might be a 'deepfake clone', you could ask them to turn sideways for more than a second or two, and see if you're still convinced by their appearance.

Arguably, this approach could be extended to automated systems that ask the user to adopt various poses in order to authenticate their entry into banking and other security-critical systems.

In May of this year, AI-based security company Sensity released a report, and a supporting video, demonstrating a DeepFaceLive-style system that apparently succeeds in fooling a liveness detector by superimposing a deepfaked identity onto a potential attacker in real-time.



A snapshot of a part of a Sensity video demonstrating a deepfake-driven attack on a liveness detector. Source: https://www.theverge.com/2022/5/18/23092964/deepfake-attack-facial-recognition-liveness-test-banks-sensity-report

However, none of the accompanying videos show the subject in acute profile. We asked Sensity's CEO and Chief Scientist Giorgio Patrini if the experiments and tests included the subject making 90° turns from camera as part of the deception technique, and he confirmed† that they did not.

We also asked him if he considers that there is any possible merit in soliciting profile views as an anti-deepfake measure during videoconferencing calls. Patrini responded:

> *'Lateral views of people faces, when used as a form of identity verification, may indeed provide some additional protection against deepfakes. As pointed out, the lack of widely available profile view data make the training of deepfake detector very challenging.*
>
> *'Additionally, I'd argue that most state of the art deepfake software simply fails if applied for faceswapping or re-enacting faces fully rotated on their side. This is because deepfake software needs to accurately detect faces and their landmarks on the target video; profile views make this more difficult as the detector has to work with only half of the facial key points.*
>
> *'Indeed, one tip for performing deepfake detection "by eye" today is to check whether one can spot face artefacts or flickering while a person is turning completely to their side — where it's more likely that a face landmarks detector would have failed.'*

Further, we asked deepfake expert Dr. Siwei Lyu, Professor of Computer Science and Engineering at the University at Buffalo School of Engineering and Applied Sciences, if this 'lateral' approach potentially has any value in deepfake detection in a live video scenario. He agreed that it does:

> *'The profile is a big problem for current deepfake technologies. The FAN network works extremely well for frontal faces, but not very well for side-on faces.*

> *'The lack of available data is certainly an obstacle. Another aspect is that these algorithms do have a fundamental limitation: the alignment mechanism works well if you cover only part of your face, and is quite robust in those circumstances – but when you turn around, more than half the landmarks are missing.*
>
> *'Probably the best an algorithm can do is to roughly estimate the profile, particularly if the person is enacting various expressions, or taking requests from the other correspondent in a video call, or from an automated liveness detection system.*
>
> *'In a case like that, profile estimation is going to be kind of a 'guess', even if you were to have some depth information from some of the more recent sensors in smartphones.'*

However, Dr. Lyu believes that the emerging new generation of 3D landmark location systems could improve on FAN's performance (though FAN itself can also enact 3D landmarks, and is used by DeepFaceLab in this way for 'full head' pose capture), but notes that this would not solve the problem of the lack of profile data for 'non-famous' people who deepfake attackers might want to train into models for deceptive purposes in a videoconference scenario.

In the absence of high-quality profile images as source training input, Dr. Lyu does not feel that novel-view synthesis systems such as NeRF, Generative Adversarial Networks (GANs) and Signed Distance Fields (SDF) are likely to be able to provide the necessary level of inference and detail in order to accurately imitate a person's profile views – at least, to a level that could stand up to the reasonably high-resolution capabilities of modern smartphone cameras and laptop webcams, which represent the likeliest environments for a deepfake attack.

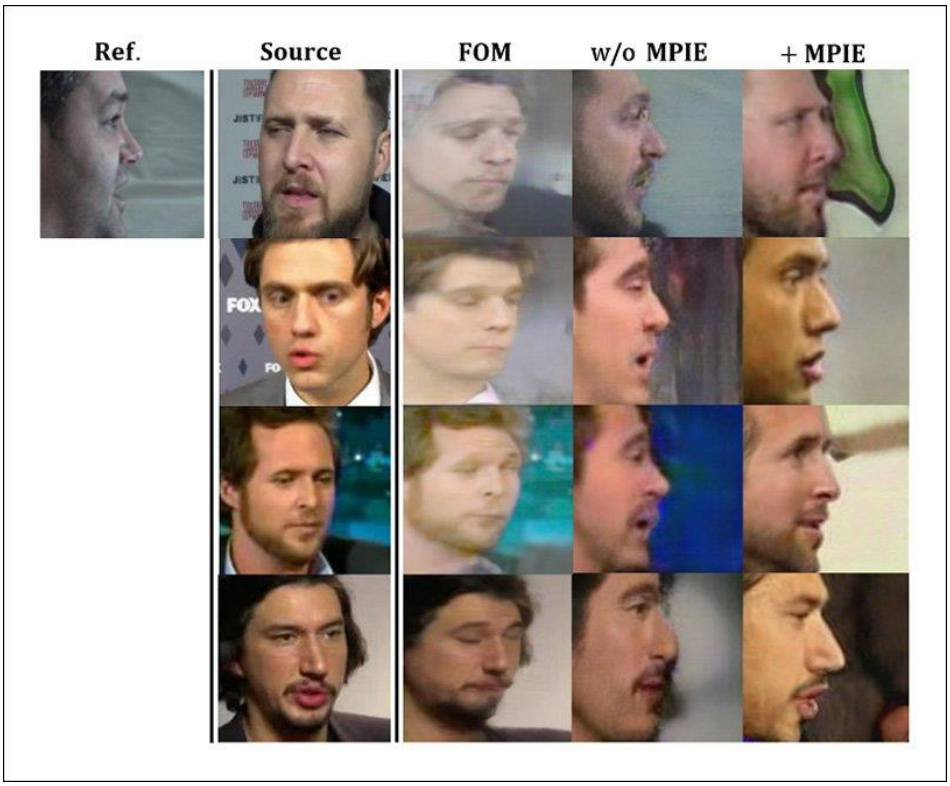
> *'The problem is that you would have to make up information on the basis of inadequate, estimated data. Perhaps you could use a GAN model and get something similar, but the data would not be likely to stand up to cross-checking on a user who has already been legitimately enrolled into a liveness detection system, for instance.'*

A new generation of facial alignment models are using parametric CGI templates to help estimate facial poses more accurately, including profile poses. However, there is no evidence that these more sophisticated models will be integrated into popular deepfakes packages, nor could such systems solve the remaining issues around profile quality by themselves. This example is from Version 2 of the 3DDFA package, a collaboration between various Chinese universities. Source: https://arxiv.org/pdf/2009.09960.pdf

Could Profile Data Be Synthesized?

Facial profile synthesis is an unusual pursuit in computer vision, not least because demand is slim. Nonetheless, there is an ongoing strand of research that concentrates on profiles.

This year the University of Taipei released a research paper called *Dual-Generator Face Reenactment*, the accompanying material of which provides a few rare samples of entirely 90° deviations generated from less oblique material, notably in the below example, which also includes actor Adam Driver:



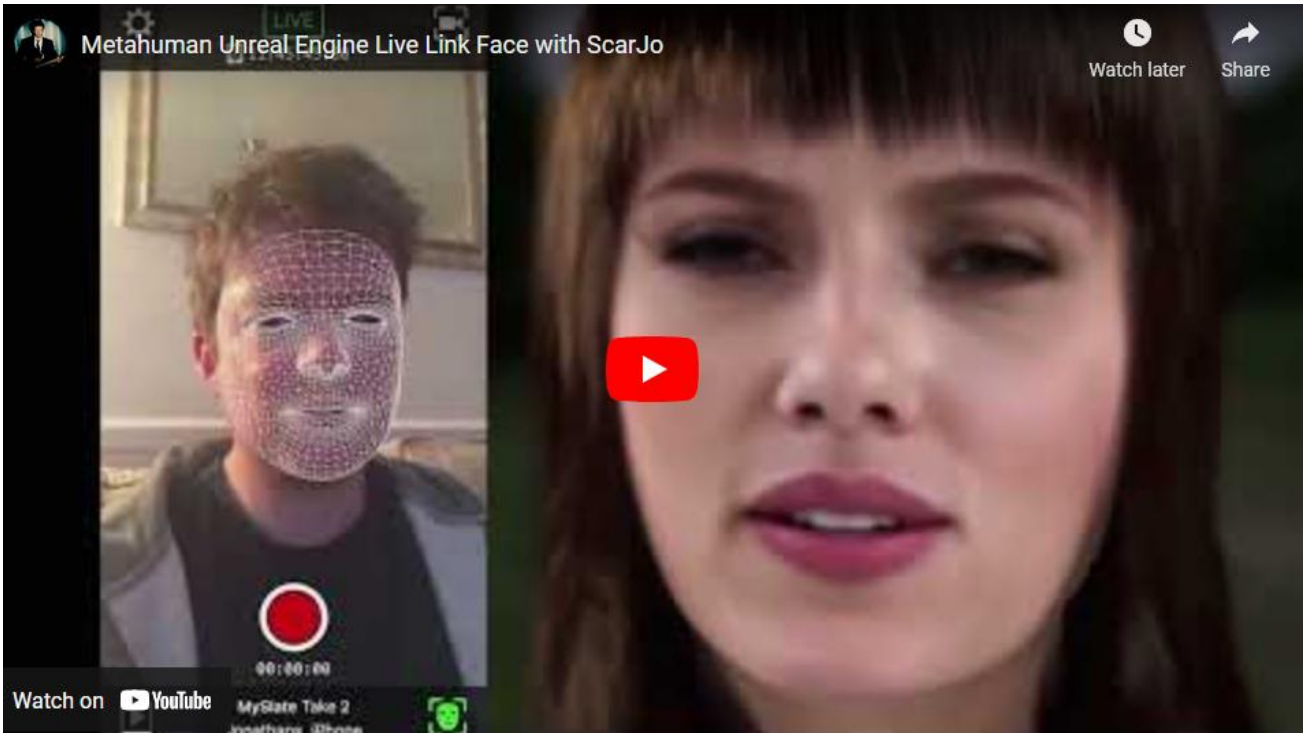
Profile faces (far right) synthesized from more frontal source input material. Source: https://openaccess.thecvf.com/content/CVPR2022/papers/Hsu_Dual-Generator_Face_Reenactment_CVPR_2022_paper.pdf

The majority of the paper's accompanying examples stop short of this extreme angle, at around the 80° mark also found in the Sensity material from May.

A little short of 90 degrees, pose variations created by Dual Generator Face Reenactment. Source: https://github.com/AvLab-CV/Dual_Generator_Face_Reenactment/blob/main/result.gif

It's just a few degrees, but it seems to make all the difference, and getting there reliably and with authenticity would be no minor milestone for a live deepfake streaming system, or the models that power it:

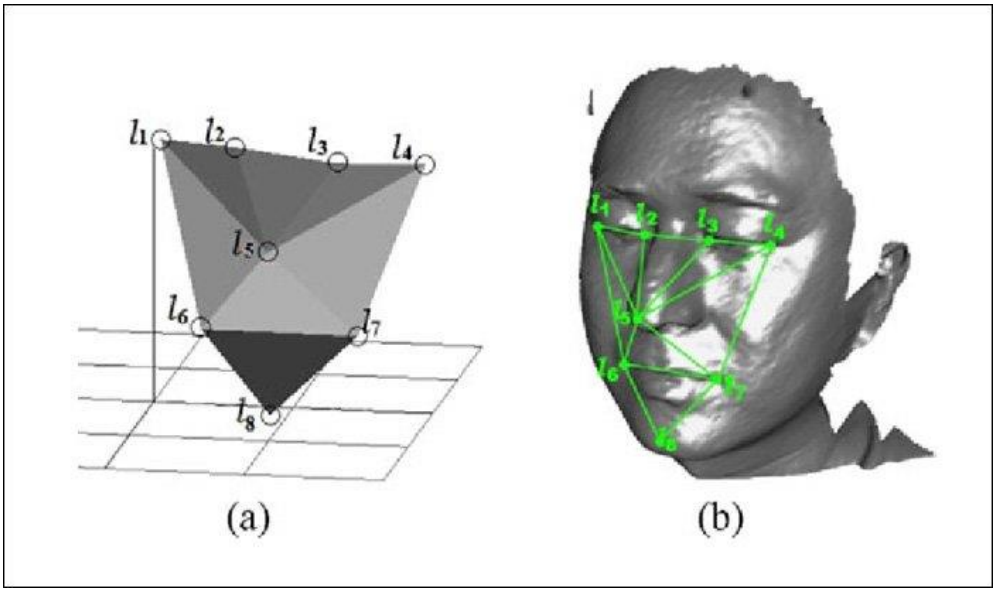
Deepfakers have been experimenting with the use of CGI-human simulations as training data practically since the technology emerged in 2017. Currently that interest has switched to applying deepfakes to Unreal Engine MetaHuman simulations – a 'CGI replacement' scenario where a real person controls a completely simulated person who is being deepfaked in real time:



The converse case is where a hyper-realistic CGI head is created for the purpose of providing 'missing' angles (such as profile views) in a training dataset for a deepfake model.

However, deepfakes have been capable, almost from the start, of producing more realistic and convincing images than CGI can, because they're based on real data rather than artistic interpretation. Therefore, not only is this a backward approach to the challenge, but it would require a level of industry-standard artistry even to 'fail well'.

The 2005 *Handbook of Face Recognition* includes some interesting material regarding profile recognition and synthesis, and features a use-case of a security system that only has profile access to a target, who is driving a car, and can only be seen in profile. To this end, the system seeks to extrapolate profile views from other views at a prior stage.



From the Handbook of Face Recognition, the basic principles of estimating profile geometry based on a very different point of view. Source: http://what-when-how.com/face-recognition/face-recognition-using-3d-images-face-recognition-techniques-part-1/

The image above illustrates the central idea of novel viewpoint synthesis (including the generation of profile views) – that alternate-viewpoint estimation can potentially infer a model accurately enough to resolve well from any angle.

Short of volunteering to be 3D-bodyscanned by a visual effects house (in which case, one could expect that the data would be highly controlled), this can be achieved by using depth-map information on frontal views, LIDAR information, or simply, as in the case of NeRF (see below), visual geometry estimation.

There are a number of frameworks offering such conversion processes, such as the Heges iOS app, which supports LIDAR as of iOS 14.

Profile views obtained from the Heges app. Source: https://hege.sh/

Facial inference of this type, though not always powered by LIDAR, is now trivial technology: in 2017, Apple hit the headlines when it was revealed that scans from the iPhone X's front sensors, which can obtain 30,000 points from an iPhone user's face, was apparently being made available to app developers.

Neural Radiance Fields (NeRF) can, in theory, extrapolate any number of facial angles from just a handful of pictures, which could be used to train an autoencoder model (presuming that NeRF-related technologies do not advance beyond autoencoders in the next few years).

NVIDIA's InstantNeRF extrapolates an impressive range of facial views from just four images, but resolution, expression accuracy and mobility remain major challenges to high-resolution inference. Source: https://www.youtube.com/watch?v=DJ2hcC1orc4

However, as we noted in our June feature on NeRF, at the moment, issues around resolution, facial mobility and temporal stability hinder NeRF from producing the rich data needed to train an autoencoder model that can handle profile images well. Additionally, really accurate NeRF profile inferences are likely to need original, genuine profile data that includes a range of expressions; if that's available, you won't need NeRF anyway.

In fact, the hallmarks of inferred profiles using *any* existing technology are a) rigid faces, b) inadequate resolution and c) poor fidelity of expression and facial pose. To effectively simulate a wide range of profile data at a suitably high resolution, inferred data is not adequate – at least not at the current state of the art.

Nor is it possible, as it so often is in computer vision research, to simply bolt on an existing Python library or open source dataset and benefit from 'upstream efforts'. *No-one* has that profile-centric data, and no-one has *ever* had that data; and, unless you're quite famous, no-one has access to your particular side-views, because your profile view doesn't interest anyone (usually not even you). At least, not yet.

Edge Cases

There are possible exceptions to this 'security-by-obscurity' safeguard, even for potential deepfake victims that are not famous. For instance, we mentioned earlier that the social set-up of *Seinfeld* (i.e., frequent group conversations) tends to put its star into profile more often than leading actors in a typical headlining role.

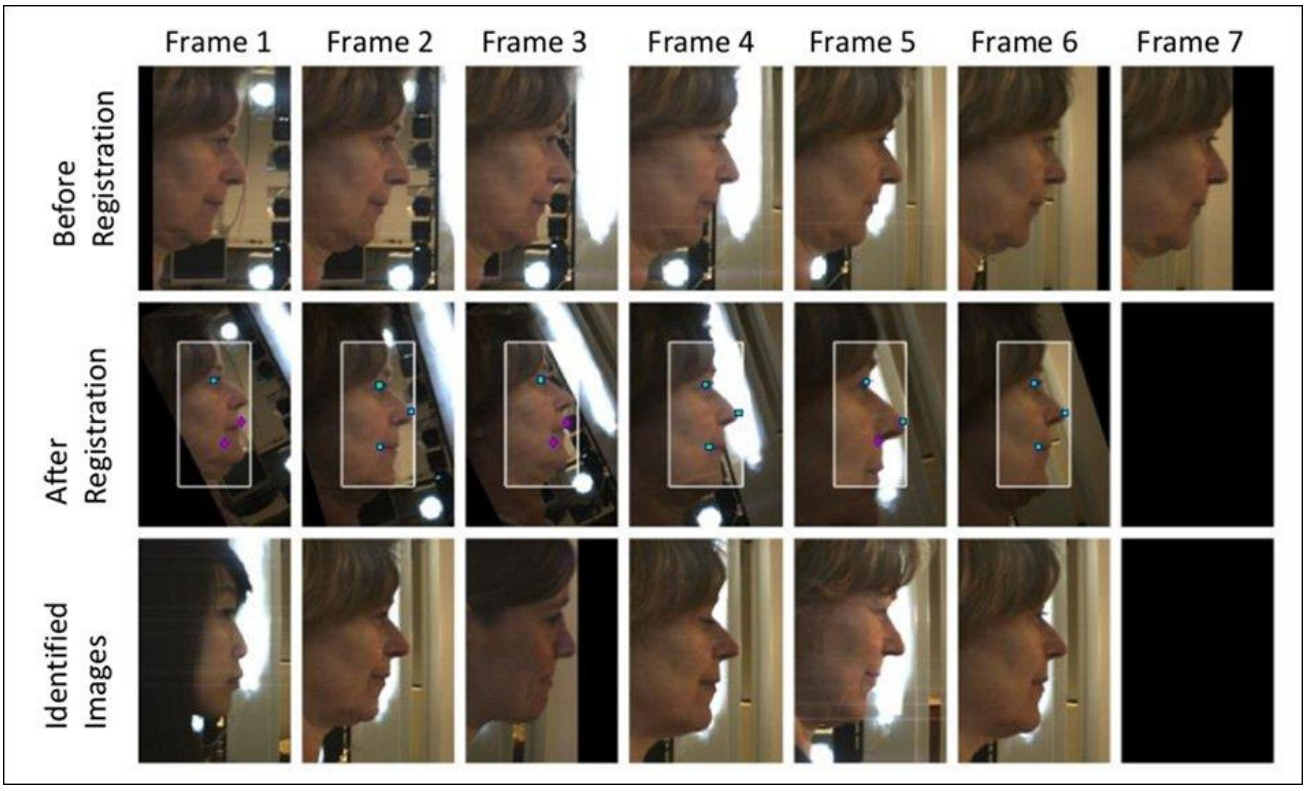
Likewise, one social media video of this type (perhaps an interview scenario), at a reasonably high resolution and featuring a sustained profile view, could provide a deepfaker with enough material to generate the data needed to exhaustively train lateral viewpoints to a high standard.

Even so, a single video is unlikely to provide every possible facial expression necessary. It would be trivial for a liveness detection system to request the person at the other end of a potential deepfake authentication session to enact certain expressions and poses that it has on record, such as mouth aperture, tightly-shut eyes, and particularly smiling (which affects the topography of the face in unpredictable ways that are difficult to estimate from a limited viewpoint), and expect that the subject could reproduce expressions that the deepfake model cannot.

Including 'rare' facial poses and dispositions in facial enrolment is already part of facial ID culture. Adding the 'obscurity' of these facial changes to already rare facial profile data could further help liveness systems to distinguish real from fake face-based login attempts.

Profile-Based Recognition Systems

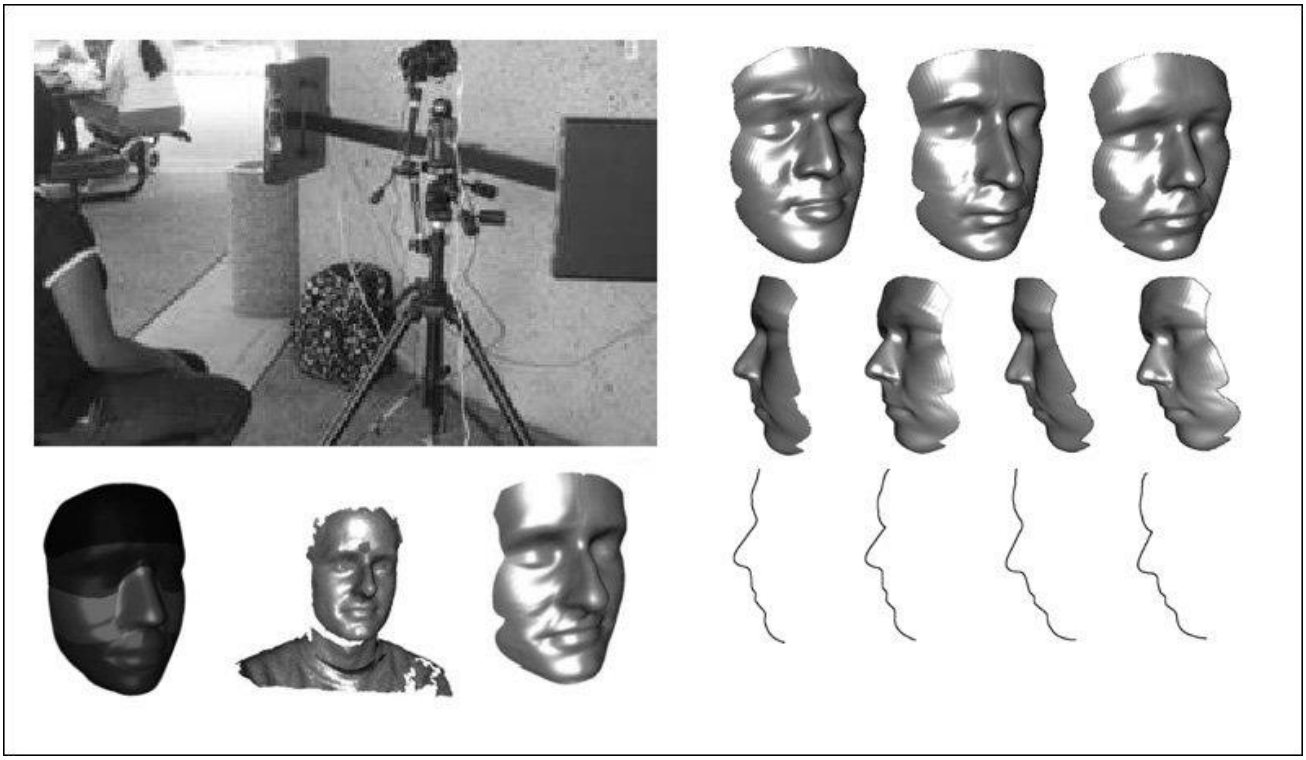
In 2017, researchers from the Netherlands developed a face-based security system that keys on profiles.



The Dutch system keys on profile recognition, and has to account for obstruction of ears, and the various other challenges that make profile data acquisition normally so difficult. Source: https://ris.utwente.nl/ws/files/71635819/automaticfacerecognition.pdf

Though the work does not aid potential deepfake attackers, it does offer a rare example of a curated database that concentrates on pure profile views, and addresses the challenge of the limited number of landmarks that are available in a side-view, and represents an unusual foray into the facial profile as a potential enrolment tool.

In the late 2010s, the University of Houston also offered some seminal work on a profile-based facial recognition system, which acquires an estimated 3D model of a user at enrolment time, and uses this as a reference for estimated profile comparisons of subjects that are acquired from the side whilst driving.

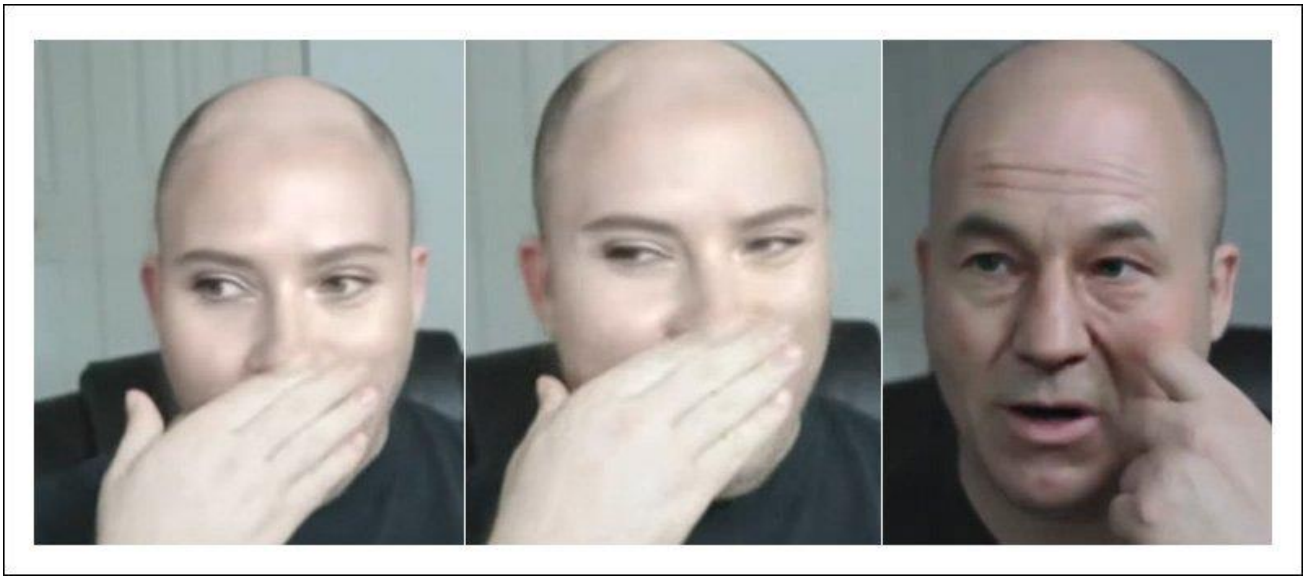


Illustrations from the Houston paper. Top left, the data capture equipment; top right, select fitted 3D face models which will eventually yield profile view data; bottom left, the base 3D template, the raw data, and the fitted facial capture; bottom right, different poses that fit within the latitude of a 'profile' definition. The method retains an active patent. Source: https://cgi.di.uoa.gr/~graphics/Downloads/papers/conferences/s38.pdf

Mattes Matter

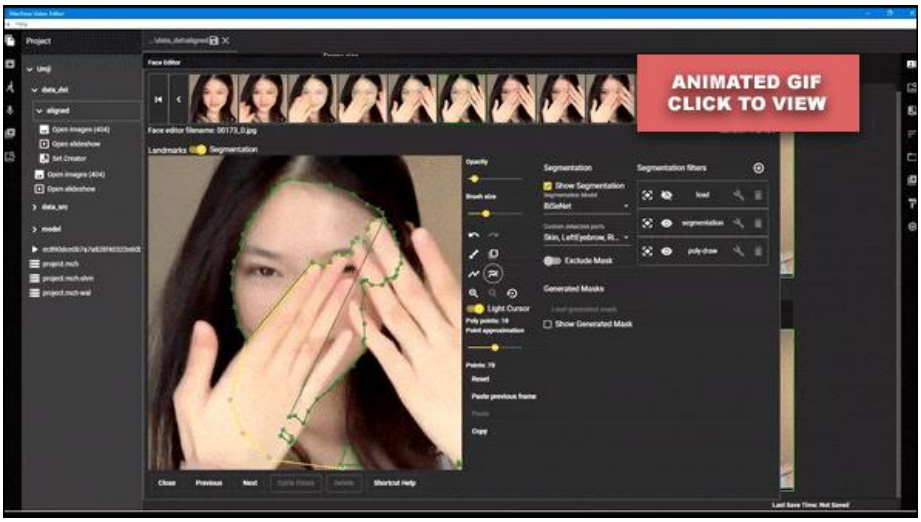
There are other possible requests one can make to a potential deepfake video caller that can be helpful in determining their authenticity, though none of them are quite as data-starved or (for the moment) impervious to improvement as profile examination.

Live deepfake video shares a common problem with augmented reality – the need to superimpose real hand images and other kinds of real obstruction onto a non-real image. In the visual effects world, these superimpositions are known as 'mattes', *rotoscoping*, or 'background removal'.



YouTube deepfaker Druuzil Tech and Games stumbles across some digital disruption while impersonating Scarlett Johansson and Patrick Stewart. Sources: https://youtu.be/C4kU4emoowk?t=501 and https://youtu.be/kjEWX67jS6U?t=102

In rendered-out deepfake videos (i.e. viral videos on YouTube and TikTok, that cannot be altered), the deepfaker is able to elaborately rotoscope away any facial obstructions – such as hands or fingers – that proved too complex for the automated matting process of XSeg (DeepFaceLab/Machine Video Editor) or trained BiseNet weights (FaceSwap).



Masking out obstructions in Machine Video Editor (MVE), an extended platform that primarily supports the DeepFaceLab workflow. Source: https://www.youtube.com/watch?v=EpWmdGfvXR0

On the other hand, a 'live' deepfake model needs the capacity to perform matting at will, and on request, to an acceptably convincing level. This may involve including many images that deliberately feature facial obstructions (including artificially-generated obstructions) in the training dataset, so that the model is at least a little better prepared to handle sudden facial intrusions and obfuscations.

However, no matter how well it's trained, asking a video caller to wave their hands in front of their face creates a critical situation for the model, which is likely to demonstrate poor latency and quality of superimposition over the deepfaked face:

Such matte glitches are currently a useful indicator of a potential deepfake video caller, but they're potentially a more solvable problem than the absence of authentic profile data, and not likely to endure as a reliable 'tell' in the long term.

Conclusion

In recent weeks, the FBI's warning about potential live deepfake fraud has been reiterated by the Better Business Bureau, and it seems that what was once an improbable attack vector is set to grow, and to continue to capture headlines.

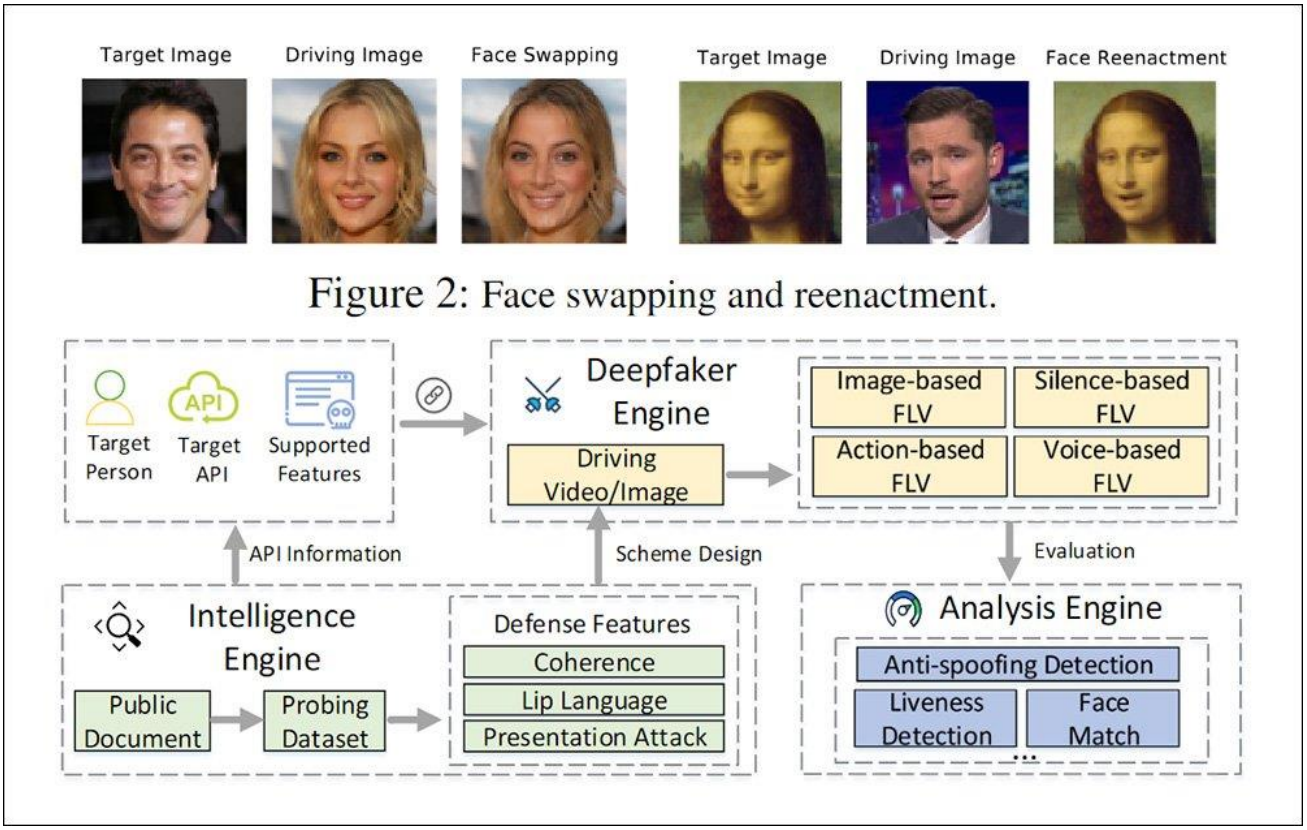
The research community has been deeply engaged in the development of deepfake detection technologies since the inception of the phenomenon, but are, to an extent, hindered by the fact that they cannot interact with the material that they're investigating.

Nonetheless, there are a growing number of solutions emerging that could be applied as a security layer in video calls, including measuring monitor illumination; evaluating inconsistency in facial regions; using the recognized parameters of known deepfake models as a security signature; cataloguing and detecting consistent artefacts in deepfake video; embedding facets of a known and trusted video into a detection system; and comparing potential deepfaked video content against known biometric traits, among many other approaches.

In live video conversations, there are additional possible ways to ask video correspondents to help authenticate themselves, such as by turning to hard profile, and checking if facial obstructions are poorly matted. Algorithmic security systems can add to this the possibility of ear recognition, and the plethora of new approaches to deepfake detection, many of which are likely to be as effective on 'live' content as they are on rendered video.

From the university of Berkeley, one recent suggestion for an additional deepfake recognition vector is based on ear recognition. Source: https://farid.berkeley.edu/downloads/publications/cvpr21a.pdf

The real threat of such attacks could be that we are not expecting them, and are likely to actually *aid* deepfake attackers by dismissing artefacts and glitches that might have raised our state of alert if we were more aware of how fallible video and audio content is becoming.



Schematic of the workflow of a recent system designed to defeat Facial Liveness Verification (FLV) through the injection of deepfakes into a live video scenario, created as a security research tool by researchers in the US and China. Source: https://arxiv.org/pdf/2202.10673.pdf

Perhaps in the future, video content of any kind will be relegated to the same status as the 'on-the-spot' engravings that adorned newspapers and books prior to the advent of reproducible photography, and viewed as potentially *representative* of the truth, rather than as an authentic token of truth in itself.

** My great thanks to Bob Doyle for running these tests with me. The video showing profile failures is a local capture taken at Bob's end by OBS Studio, and is therefore not affected by encoding errors that might occur with Zoom and other videoconferencing software. Bob's DeepFaceLive implementation was running on a NVIDIA 2070S graphics card with 8GB of VRAM.*

† *In an email dated July 3rd 2022.*