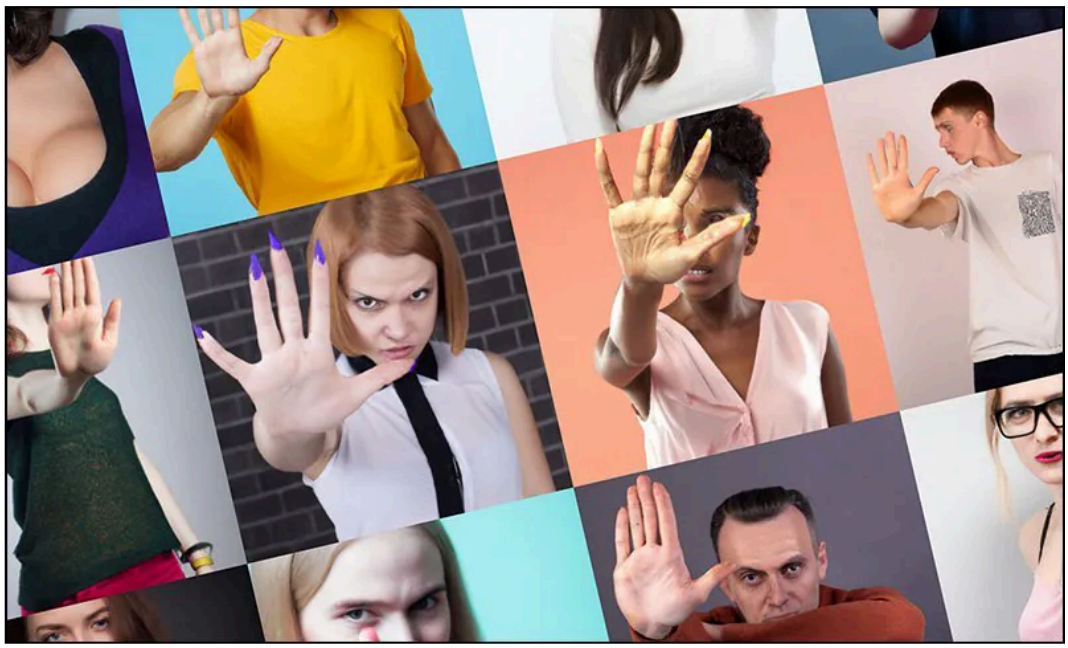
Encoding Images Against Use in Deepfake and Image Synthesis Systems

Originally published at https://www.unite.ai/encoding-images-against-use-in-deepfake-and-image-synthesis-systems/



The most well-known line of inquiry in the growing anti-deepfake research sector involves systems that can recognize artifacts or other supposedly distinguishing characteristics of deepfaked, synthesized, or otherwise falsified or 'edited' faces in video and image content.

Such approaches use a variety of tactics, including depth detection, video regularity disruption, variations in monitor illumination (in potentially deepfaked live video calls), biometric traits, outer face regions, and even the hidden powers of the human subconscious system.

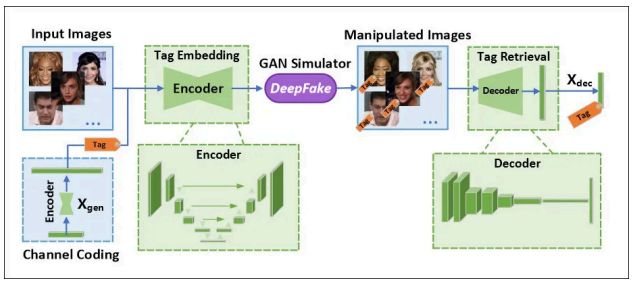
What these, and similar methods have in common is that by the time they are deployed, the central mechanisms they're fighting have already been successfully trained on thousands, or hundreds of thousands of images scraped from the web – images from which autoencoder systems can easily derive key features, and create models that can accurately impose a false identity into video footage or synthesized images – even in real time.

In short, by the time such systems are active, the horse has already bolted.

Images That Are Hostile to Deepfake/Synthesis Architectures

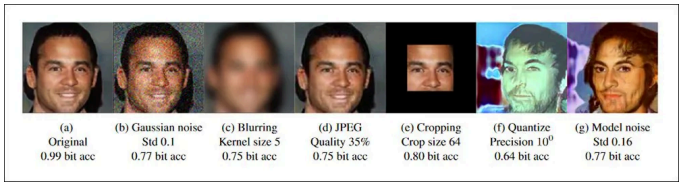
By way of a more *preventative* attitude to the threat of deepfakes and image synthesis, a less well-known strand of research in this sector involves the possibilities inherent in making all those source photos *unfriendly* towards AI image synthesis systems, usually in imperceptible, or barely perceptible ways.

Examples include FakeTagger, a 2021 proposal from various institutions in the US and Asia, which encodes messages into images; these encodings are resistant to the process of generalization, and can subsequently be recovered even after the images have been scraped from the web and trained into a Generative Adversarial Network (GAN) of the type most famously embodied by thispersondoesnotexist.com, and its numerous derivatives.



FakeTagger encodes information that can survive the process of generalization when training a GAN, making it possible to know if a particular image contributed to the system's generative capabilities. Source: https://arxiv.org/pdf/2009.09869.pdf

For ICCV 2021, another international effort likewise instituted artificial fingerprints for generative models, (see image below) which again produces recoverable 'fingerprints' from the output of an image synthesis GAN such as StyleGAN2.



Even under a variety of extreme manipulations, cropping, and face-swapping, the fingerprints passed through ProGAN remain recoverable. Source: https://arxiv.org/pdf/2007.08457.pdf

Other iterations of this concept include a 2018 project from IBM and a digital watermarking scheme in the same year, from Japan.

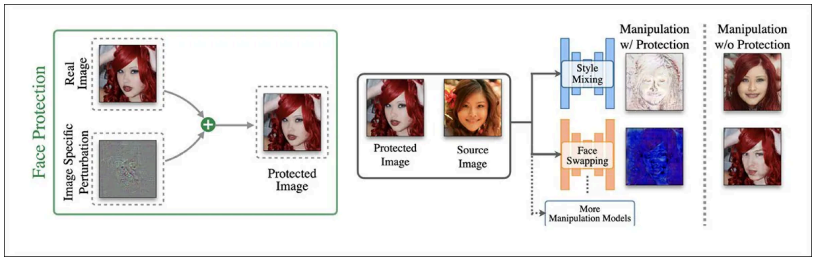
More innovatively, a 2021 initiative from the Nanjing University of Aeronautics and Astronautics sought to 'encrypt' training images in such a way that they would train effectively only on authorized systems, but would fail catastrophically if used as source data in a generic image synthesis training pipeline.

Effectively all these methods fall under the category of steganography, but in all cases the unique identifying information in the images needs to be encoded as such an essential 'feature' of an image that there is no chance that an autoencoder or GAN architecture would discard such fingerprints as 'noise' or outlier and inessential data, but rather will encode it along with other facial features.

At the same time, the process cannot be allowed to distort or otherwise visually affect the image so much that it is perceived by casual viewers to have defects or to be of low quality.

TAFIM

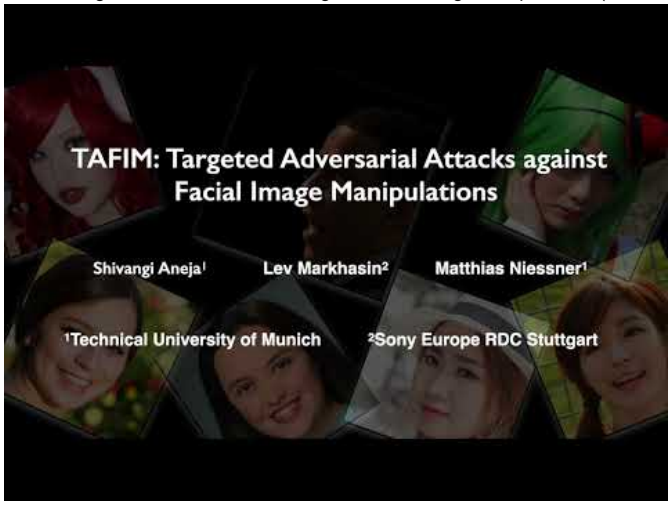
Now, a new German research effort (from the Technical University of Munich and Sony Europe RDC Stuttgart) has proposed an image-encoding technique whereby deepfake models or StyleGAN-type frameworks that are trained on processed images will produce unusable blue or white output, respectively.



TAFIM's low-level image perturbations address several possible types of face distortion/substitution, forcing models trained on the images to produce distorted out authors to be applicable even in real-time scenarios such as DeepFaceLive's real-time deepfake streaming. Source: https://arxiv.org/pdf/2112.09151.pdf

The paper, titled *TAFIM: Targeted Adversarial Attacks against Facial Image Manipulations*, uses a neural network to encode barely-perceptible perturbations into images. After the images are trained and generalized into a synthesis architecture, the resulting model will produce discolored output for the input identity if used in either style mixing or straightforward face-swapping.

TAFIM: Targeted Adversarial Attacks against Face Image Manipulations (ECCV'22)



Re-Encoding the Web..?

However, in this case, we're not here to examine the minutiae and architecture of the latest version of this popular concept, but rather to consider the practicality of the whole idea – particularly in light of the growing controversy about the use of publicly-scraped images to power image synthesis frameworks such as Stable Diffusion, and the subsequent downstream legal implications of deriving commercial software from content that may (at least in some jurisdictions) eventually prove to have legal protection against ingestion into AI synthesis architectures.

Proactive, encoding-based approaches of the kind described above come at no small cost. At the very least, they would involve instituting new and extended compression routines into standard web-based processing libraries such as ImageMagick, which power a large number of upload processes, including many social media upload interfaces, tasked with converting over-sized original user images into optimized versions that are more suitable for lightweight sharing and network distribution, and also for effecting transformations such as crops, and other augmentations.

The primary question that this raises is: would such a scheme be implemented 'going forward', or would some wider and retroactive deployment be intended, that addresses historical media that may have been available, 'uncorrupted', for decades?

Platforms such as Netflix are not averse to the expense of re-encoding a back catalogue with new codecs that may be more efficient, or could otherwise provide user or provider benefits; likewise, YouTube's conversion of its historic content to the H.264 codec, apparently to accommodate Apple TV, a logistically monumental task, was not considered prohibitively difficult, despite the scale.

Ironically, even if large portions of media content on the internet were to become subject to re-encoding into a format that resists training, the limited cadre of influential computer vision datasets would remain unaffected. However, presumably, systems that use them as upstream data would begin to diminish in quality of output, as watermarked content would interfere with the architectures' transformative processes.

Political Conflict

In political terms, there is an apparent tension between the determination of governments not to fall behind in AI development, and to make concessions to public concern about the ad hoc use of openly available audio, video and image content on the internet as an abundant resource for transformative AI systems.

Officially, western governments are inclined to leniency in regards to the ability of the computer vision research sector to make use of publicly available media, not least because some of the more autocratic Asian countries have far greater leeway to shape their development workflows in a way that benefits their own research efforts – just one of the factors that suggests China is becoming the global leader in AI.

In April of 2022, the US Appeals Court affirmed that public-facing web data is fair game for research purposes, despite the ongoing protests of LinkedIn, which wishes its user profiles to be protected from such processes.

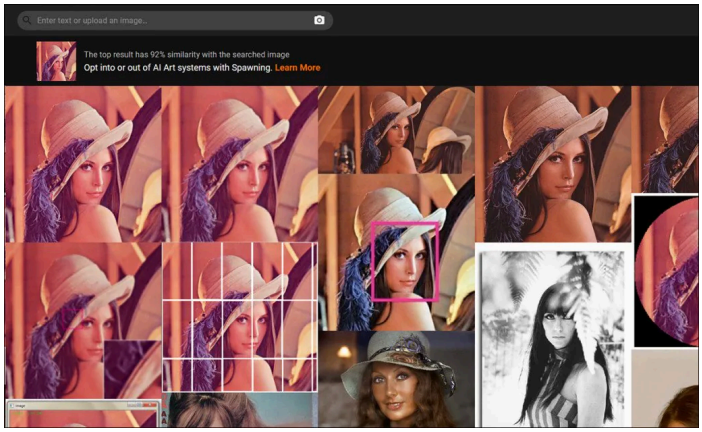
If AI-resistant imagery is therefore not to become a system-wide standard, there is nothing to prevent some of the major sources of training data from implementing such systems, so that their own output becomes unproductive in the latent space.

The essential factor in such company-specific deployments is that images should be *innately resistant* to training. Blockchain-based provenance techniques, and movements such as the Content Authenticity Initiative, are more concerned with proving that image have been faked or 'styleGANned', rather than preventing the mechanisms that make such transformations possible.

Casual Inspection

While proposals have been put forward to use blockchain methods to authenticate the true provenance and appearance of a source image that may have been later ingested into a training dataset, this does not in itself prevent the training of images, or provide any way to prove, from the output of such systems, that the images were included in the training dataset.

In a watermarking approach to excluding images from training, it would be important not to rely on the source images of an influential dataset being publicly available for inspection. In response to artists' outcries about Stable Diffusion's liberal ingestion of their work, the website haveibeentrained.com allows users to upload images and check if they are likely to have been included in the LAION5B dataset that powers Stable Diffusion:



'Lenna', literally the poster girl for computer vision research until recently, is certainly a contributor to Stable Diffusion. Source: https://haveibeentrained.com/

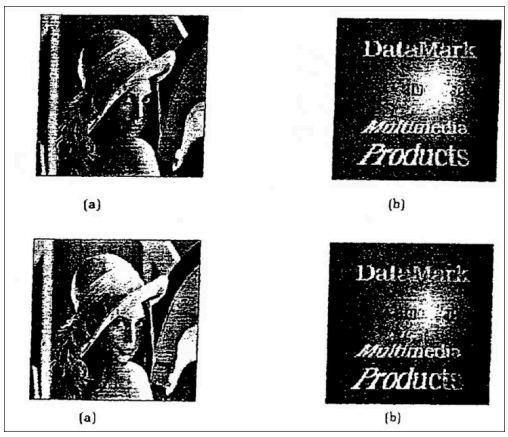
However, nearly all traditional deepfake datasets, for instance, are casually drawn from extracted video and images on the internet, into non-public databases where only some kind of neurally-resistant watermarking could possibly expose the use of specific images to create the derived images and video.

Further, Stable Diffusion users are beginning to add content – either through fine-tuning (continuing the training of the official model checkpoint with additional image/text pairs) or Textual Inversion, which adds one specific element or person – that will not appear in any search through LAION's billions of images.

Embedding Watermarks at Source

An even more extreme potential application of source image watermarking is to include obscured and non-obvious information into the raw capture output, video or images, of commercial cameras. Though the concept was experimented with and even implemented with some vigor in the early 2000s, as a response to the emerging 'threat' of multimedia piracy, the principle is technically applicable also for the purpose of making media content resistant or repellant to machine learning training systems.

One implementation, mooted in a patent application from the late 1990s, proposed using Discrete Cosine Transforms to embed steganographic 'sub images' into video and still images, suggesting that the routine could be 'incorporated as a built-in feature for digital recording devices, such as still and video cameras'.



In a patent application from the late 1990s, Lenna is imbued with occult watermarks that can be recovered as necessary.
Source: https://www.freepatentsonline.com/6983057.pdf

A less sophisticated approach is to impose clearly visible watermarks onto images at device-level – a feature that's unappealing to most users, and redundant in the case of artists and professional media practitioners, who are able to protect the source data and add such branding or prohibitions as they deem fit (not least, stock image companies).

Though at least one camera currently allows for optional logo-based watermark imposition that could signal unauthorized use in a derived AI model, logo removal via AI is becoming quite trivial, and even casually commercialized.

First published 25th September 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