

A New Social Image-Sharing System Deepfakes All People by Default

Originally published at https://arquivo.pt/wayback/20240203194458oe_/https://blog.metaphysic.ai/a-new-social-image-sharing-system-deepfakes-all-people-by-default/

November 14, 2022



A new collaboration between Binghamton University and Intel offers a novel take on the problem of the unauthorized use of social media photos for facial recognition purposes, as well as AI and [deepfake](#) training – by using deepfake techniques to subtly alter the appearance of people in posted photos, so that only their friends and permitted contacts are able to see their original face images.



Depending on how deeply entrenched you are in a friends' network that's represented in a photo, you may see all or none of the 'real' faces from the original capture, while the subtly altered faces proposed by the new system are not jarringly dissimilar – but not accurate enough to form usable training data for deepfakes or latent diffusion models, and other image synthesis systems. Source: <http://arxiv.org/pdf/2211.01361>

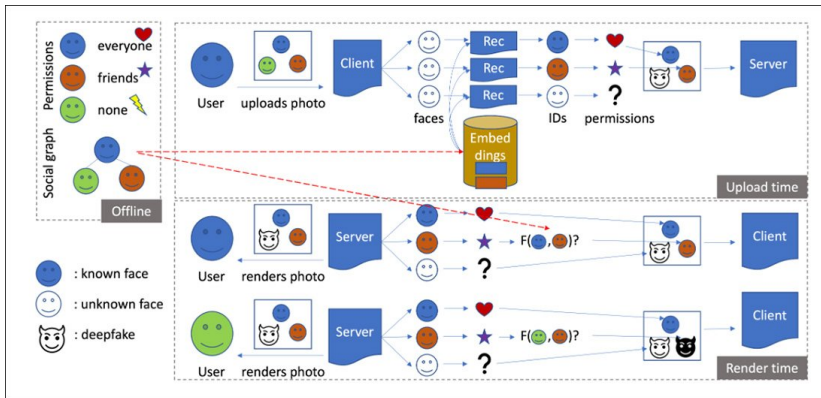
The [new paper](#), titled *My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization*, proposes various 'face-access' models capable of replacing 'unapproved' faces with quantitatively dissimilar faces – though the faces created retain the gender, age, pose and basic disposition of the original people depicted.

The system is designed to be adoptable by existing social networks, rather than to constitute the basis of a novel social network.

My Face My Choice (MFMC) alters the image on upload so that by default, 'outsiders' see all the faces in it as broadly representative deepfakes. The transformations are effected by [ArcFace](#), a 2022 project led by Imperial College London. ArcFace extracts face embeddings with 512 features, but also optimizes feature embeddings so that the derived face has approximate visual parity with the 'overwritten' face, without allowing extractable (real) features to be copied over to the amended photo.

In the new framework, gender and age classifications are performed by [InsightFace](#). However, the classification that InsightFace makes is not exposed as public data, but rather is used as a direction in the latent space of the image synthesis system that powers MFMC.

Under MFMC, only the submitter can decide to reveal their real face in the image. Any friends or other people in the photo who wish their real face to be revealed have to ask the submitter to 'unlock' that face.



The workflow for a three-user network under MPMC. Here, the green person does not want to be seen by anybody, and will always be deepfaked; the blue person wants to be seen by everyone, and will never be deepfaked; and the red person wants to be seen only by friends, and so will be deepfaked for everyone except their own friends.

The system also allows identities which are tagged by the user to be revealed, with the tag acting as a default 'unlock' of the face. When the tag is removed, the face is replaced by a deepfaked face.

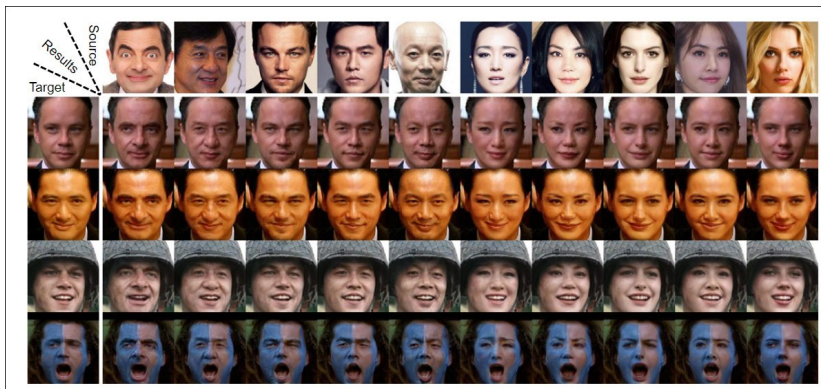
	Original	Furthest	Furthest/[age, gender]	Closest/[age, gender]	Random/[age, gender]
Social photo uploaded by					
Person 1					
Person 2					
Person 3					
Person 4					
	Source	Target	Result	Target	Result

Source images and the derived deepfakes created from random target queries, restricted to age and gender, under MPMC. The paper concedes that the flexible and reversible nature of this approach increases both the processing requirements and storage needs, compared to current systems.

Multiple Deepfake Approaches

The objective of MPMC is to avoid storing real identities, which is in opposition to traditional autoencoder-based deepfakes such as DeepFaceLab and FaceSwap, which trains on two datasets, each of which contains images of a real person.

Therefore the researchers have incorporated four separate approaches into MPMC: a [technique](#) developed in 2017, for a project led by the Open University of Israel, which uses only a standard fully convolutional network (FCN) to effect face transfer between subjects, without storing identity information, and aided by [Poisson Blending](#); the better-known 'few-shot face translation' model [FTGAN](#), which uses unsupervised image-to-image translation powered by NVIDIA's [SPADE](#) framework; [FSGAN](#), an Israeli academic collaboration from 2019, a two stage subject-agnostic architecture that uses facial inpainting to insert a substituted face; and [SimSwap](#), a 2021 initiative from China, which adapts traditional [autoencoder](#) deepfakes to accommodate arbitrary faces.



Examples of SimSwap in action, one of the four deepfake modules used in MPMC. <https://arxiv.org/pdf/2106.06340.pdf>

Data and Approach

Though the researchers ultimately selected InsightFace as the face detection module, rejecting Carnegie Mellon's [OpenFace](#) and Google's [FaceNet](#), the detector can be easily swapped out to fine-tune the balance between speed and accuracy.

Both training and inference processes are run on a NVIDIA GTX 1060 GPU, while a generalization suitable to typical user uploads was found in the 'party' subset of the [Facial Descriptors Dataset](#).



Facial Descriptors: extracted faces from interactions, used as training data for MFMC. Source: <https://arxiv.org/pdf/1509.05366.pdf>

The Facial Descriptors Dataset consists of various configurations of 1282 people across 193 images, at diverse resolutions, lighting conditions, and in variegated poses.

To generate a corpus of deepfake imagery on which to help train and calibrate the system, the researchers used two datasets: one containing 10,000 fake faces generated by [StyleGAN](#), and another containing a further 10,000 faces originated by [generated.photos](#).

The targeted use of similarity metrics is crucial to ensuring that the basic group composition of an uploaded photo is retained, while also making certain that the individual faces are not so similar to the source faces that they could potentially be used as trainable material to reproduce the real identities depicted.

Different methods of evaluating similarity place emphasis on different characteristics. Below we see the varying results when we're looking for a face that's similar to the source face (top row), according to different methods, including Fréchet inception distance ([FID](#)), Root Mean Square Error ([RMSE](#)), Structural Similarity Index Measure ([SSIM](#)), Peak signal-to-noise ratio ([PSNR](#)), and native [embedding distance](#).

	Person 1	Person 2	Person 3	Person 4	Person 5
Source Images					
Furthest by Embedding Distance					
Furthest by FID					
Furthest by RMSE					
Furthest by SSIM					
Furthest by PSNR					

As we can see, the different metrics measure dissimilarity in different ways, with PSNR favoring color, SSIM and RMSE favoring head-pose, and FID creating some of the most extreme counterpoints to the original image, notably in the change of age-range.

Tests

Testing MMFC involved ensuring that the transformed identities could not be picked up by facial recognition systems that would ordinarily be capable of doing so. Achieving this goal also signifies that the original faces cannot be reconstructed by training them into image synthesis systems such as autoencoder deepfakes or [latent diffusion frameworks](#).

The output from MFMC was tested against seven state-of-the-art facial recognition systems: [DeepID](#), OpenFace, DeepFace, FaceNet, [DLib](#), and ArcFace. The test data was 1282 faces in 193 images, this time generated by MFMC.

Face Detector	Source vs. Target		Source vs. Result	
	SSIM	RMSE	SSIM	RMSE
FaceNet512	0.001	0.0	0.12	0.16
OpenFace	0.001	0.003	0.17	0.24
FaceNet	0.03	0.02	0.27	0.34
DLib	0.02	0.05	0.30	0.44
ArcFace	0.05	0.03	0.29	0.45
DeepID	0.005	0.01	0.44	0.52
DeepFace	0.03	0.06	0.47	0.53
Average	0.02	0.01	0.29	0.23

Very low face recognition scores for seven SOTA recognition systems, indicating that the systems were unable to recognize the subtly-altered deepfake faces.

Scores this low are scarcely better than chance, even though, as can be seen by the examples of MFMC earlier in this article, the faces appear to have high resemblance to the original source photos. This illustrates the extent to which facial recognition (and reproducible identity characteristics) relies on low-level structural features in faces, and that moving or altering these features can produce a striking similarity that, to the human eye, is still ‘not quite right’.

The paper concludes*:

‘FMC also demonstrates a responsible use for deepfakes by design, especially for protecting faces of minors and vulnerable populations, in contrast [to] [dystopian scenarios](#).’

And continues:

‘We believe that current social media platforms would free the users if similar approaches to MFMC are implemented for face privacy and contextual integrity.’

The Face of the Future?

MFMC may seem an extreme solution to the problem of publicly available images of people being exploited for deepfake or image synthesis systems; however, it’s only a more granular application of user-access rights systems currently in place for a social media site such as Facebook, where users can hide channel content selectively from people that they are not connected to, as they prefer.

In terms of the change to user experience, one can consider that all the participants in a group photo of themselves will all see their original faces from the photo, with no deepfaked faces present; and that a user who is connected to one friend in the photo will see their friend’s original face, and an approximation of the other faces (and presumably, only the friend’s face is of interest to that viewer, else they would already be connected to the six ‘deepfaked’ people standing around their friend).

At the same time, any itinerant web-scraper that managed to cajole its way past a platform login, or which is making use of ‘publicly available’ images from the platform, will either get the real faces that the people in the photo have *willingly* surrendered, or a group of faces that don’t exist (which is abstractly useful for latent diffusion and other ‘imaginative’ image synthesis training, but completely obstructs *ad hoc* facial recognition web-scrapers).

** My conversion of the authors’ inline citations to italics.*