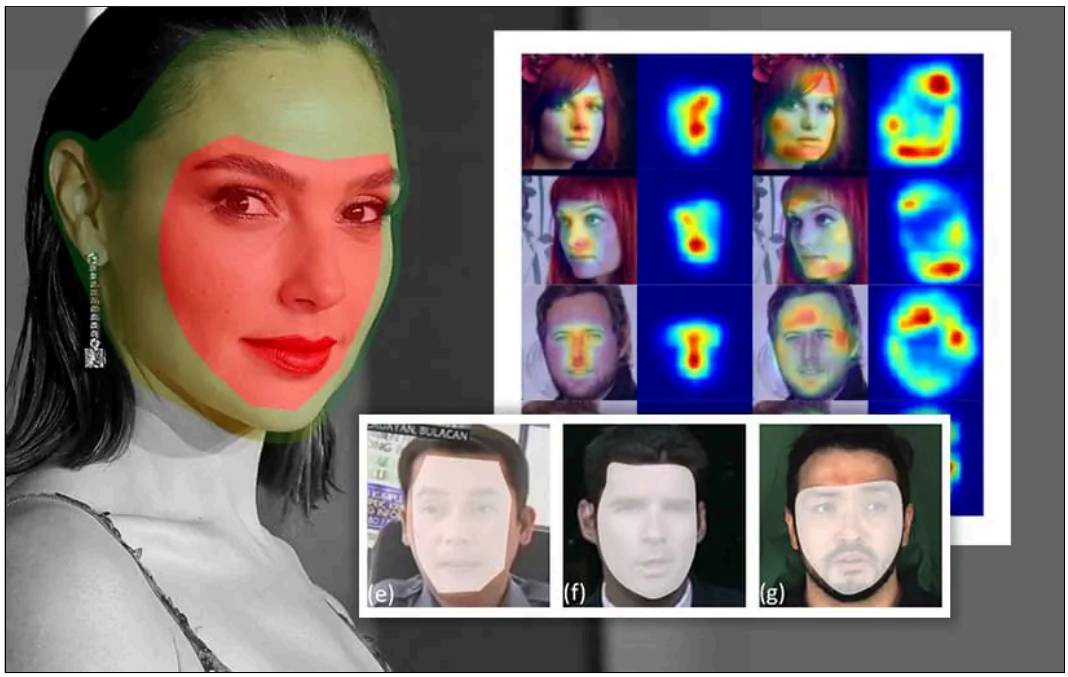
Identifying Celebrity Deepfakes From Outer Face Regions

Originally published at https://www.unite.ai/identifying-celebrity-deepfakes-from-outer-face-regions/



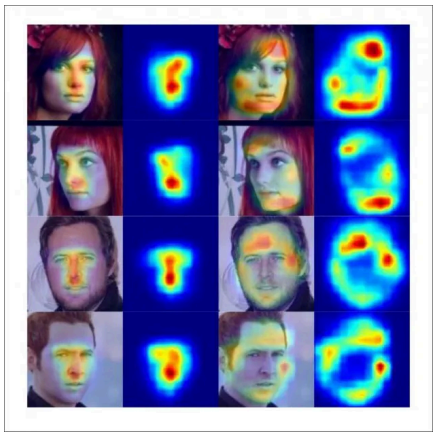
A new collaboration between Microsoft and a Chinese university has proposed a novel way of identifying celebrity deepfakes, by leveraging the shortcomings of current deepfake techniques to recognize identities that have been 'projected' onto other people.

The approach is called *Identity Consistency Transformer* (ICT), and works by comparing the outermost parts of the face (jaw, cheekbones, hairline, and other outer marginal lineaments) to the interior of the face. The system exploits commonly available public image data of famous people, which limits its effectiveness to popular celebrities, whose images are available in high numbers in widely available computer vision datasets, and on the internet.

The forgery coverage of faked faces across seven techniques: DeepFake in FF+; DeepFake in Google DeepFake Detection; DeepFaceLab; Face2Face; FSGAN; Popular packages such as DeepFaceLab and FaceSwap provide similarly constrained coverage. Source: https://arxiv.org/pdf/2203.01318.pdf

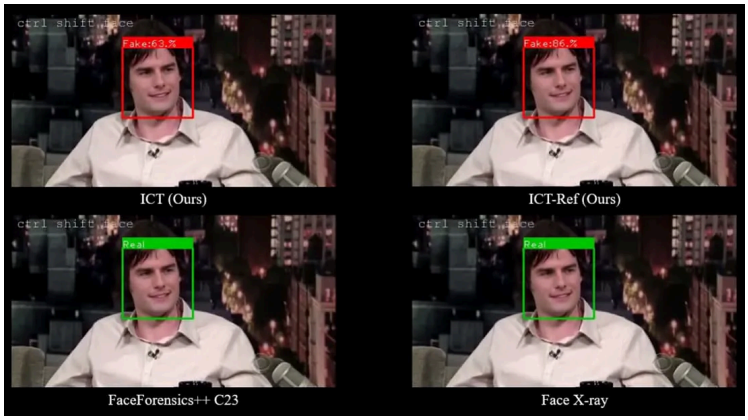
As the image above illustrates, currently popular methods for deepfaking are quite resource-constrained, and rely on apposite host-faces (the image or video of a person who will have their identity replaced by the deepfake) to minimize evidence of face substitution.

Though varying methods may encompass the full forehead and a large part of the chin and cheekbone areas, all are more or less constrained inside the frame of the host face.



A saliency map that emphasizes the 'inner' and 'outer' identities calculated by ICT. Where an inner facial match is established but an outer identity does not correspond, ICT evaluates the image as false.

In tests, ICT proved able to detect deepfake content in fake-friendly confines such as low resolution video, where the content of the entire video is degraded by compression artifacts, helping to hide residual evidence of the deepfake process – a circumstance that confounds many competing deepfake detection methods.



ICT outperforms contenders in recognizing deepfake content. See video embedded at end of article for more examples and better resolution. See embedded sour of article for further examples. Source: https://www.youtube.com/watch?v=zgF50dcymj8

The paper is titled *Protecting Celebrities with Identity Consistency Transformer*, and comes from nine researchers variously affiliated to the University of Science and Technology of China, Microsoft Research Asia, and Microsoft Cloud + AI.

The Credibility Gap

There are at least a couple of reasons why popular face-swapping algorithms such as DeepFaceLab and FaceSwap neglect the outermost area of the swapped facial identities.

Firstly, training deepfake models is time-consuming and resource-critical, and the adoption of 'compatible' host faces/bodies frees up GPU cycles and epochs to concentrate on the relatively immutable inner areas of the face which we use to distinguish identity (since variables such as weight fluctuation and ageing are least likely to change these core facial traits in the short term).

Secondly, most deepfake approaches (and this is certainly the case with DeepFaceLab, the software used by the most popular or notorious practitioners) have limited ability to replicate 'end of face' margins such as cheek and jaw areas, and are constrained by the fact that their upstream (2017) code did not extensively address this issue.

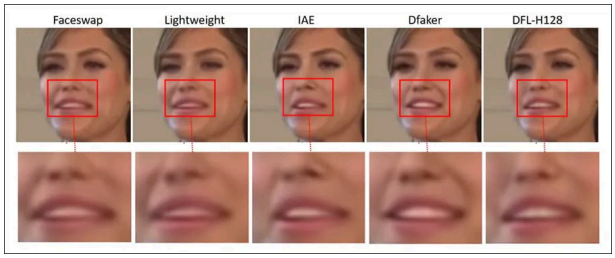
In cases where the identities don't match well, the deepfake algorithm must 'inpaint' background areas around the face, which it does clumsily at best, even in the hands of the best deepfakers, such as Ctrl Shift Face, whose output was used in the paper's studies.

The best of the best: stills from a deepfake video from acclaimed deepfaker Ctrl-Shift-Face, swapping Jim Carrey over Gary Oldman. This work arguably represen best output currently available via DeepFaceLab and post-processing techniques. Nonetheless, the swaps remain limited to the relatively scant attention that DFL outer face, requiring a Herculean effort of data curation and training to address the outermost lineaments. Source: https://www.youtube.com/watch?v=x8igrh1eyLk

This 'sleight of hand', or deflection of attention largely escapes public attention in the current concern over the growing realism of deepfakes, because our critical faculties around deepfakes are still developing past the 'shock and awe' stage.

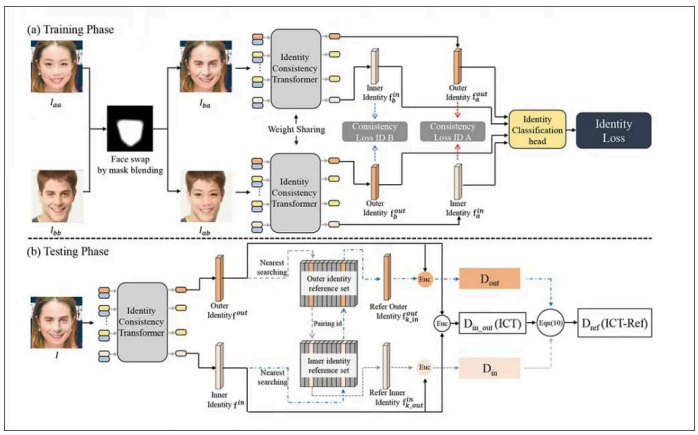
Split Identities

The new paper notes that most prior methods of deepfake detection rely on artifacts that betray the swap process, such as inconsistent head poses and blinking, among numerous other techniques. Only this week, another new deepfake detection paper has proposed using the 'signature' of the varying model types in the FaceSwap framework to help identify forged video created with it (see image below).



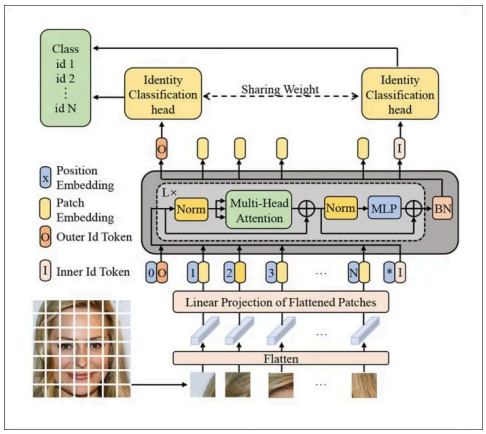
Identifying deepfakes by characterizing the signatures of different model types in the FaceSwap framework. Source: https://arxiv.org/pdf/2202.12951.pdf

By contrast, ICT's architecture creates two separate nested identities for a person, each of which must be verified before the entire identity is concluded to be 'true' footage or imagery.



Architecture for the training and testing phases of ICT.

The split of identities is facilitated by a vision Transformer, which performs facial identification before splitting the surveyed regions into tokens belonging to the inner or outer identities.



Distributing patches among the two parallel identity signifiers.

The paper states:

> *'Unfortunately existing face verification [methods] tend to characterize the most discriminative region, i.e., the inner face for verification and fail to capture the identity information in the outer face. With Identity Consistency Transformer, we train a model to learn a pair of identity vectors, one for the inner face and the other for the outer face, by designing a Transformer such that the inner and the outer identities can be learned simultaneously in a seamlessly unified model.'*

Since there is no existing model for this identification protocol, the authors have devised a new kind of consistency loss that can act as a metric for authenticity. The 'inner token' and 'outer token' that result from the identity extraction model are added to the more conventional patch embeddings produced by facial identification frameworks.

Data and Training

The ICT network was trained on Microsoft Research's MS-Celeb-1M dataset, which contains 10 million celebrity face images covering one million identities, including actors, politicians, and many other types of prominent figures. According to the procedure of prior method Face X-ray (another Microsoft Research initiative), ICT's own fake-generation routine swaps inner and outer regions of faces drawn from this dataset in order to create material on which to test the algorithm.

To perform these internal swaps, ICT identifies two images in the dataset that exhibit similar head poses and facial landmarks, generates a mask region of the central features (into which a swap can be performed), and performs a deepfake swap with RGB color correction.

The reason that ICT is limited to celebrity identification is that it relies (in its most effective variation) on a novel reference set that incorporates derived facial vectors from a central corpus (in this case MS-Celeb-1M, though the referencing could be extended to network-available imagery, which would only likely exist in sufficient quality and quantity for well-known public figures).

These derived vector-set couplets act as authenticity tokens to verify the inner and outer face regions in tandem.

The authors note that the tokens obtained from these methods represent 'high-level' features, resulting in a deepfake detection process that is more likely to survive challenging environments such as low-resolution or otherwise degraded video.

Crucially, ICT is *not* looking for artifact-based evidence, but rather is focused on identity verification methods more in accord with facial recognition techniques – an approach which is difficult with low volume data, as is the case with the investigation of incidents of deepfake revenge porn against non-famous targets.

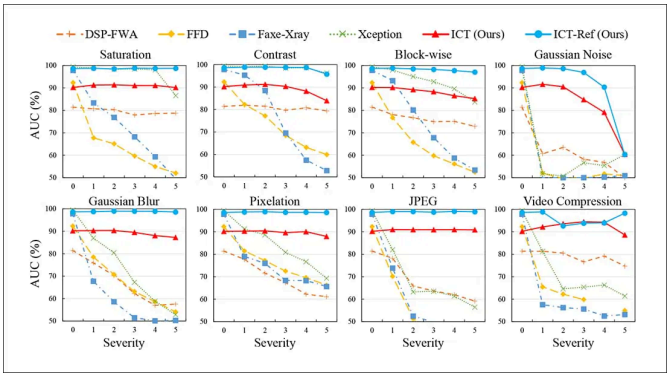
Tests

Trained on MS-Celeb-1M, ICT was then divided into reference-assisted and 'blind' versions of the algorithm, and tested against a range of competing datasets and methods. These included FaceForensics++ (FF++), a dataset of 1000 authentic and deepfake videos created across four methods including Face2Face and FaceSwap; Google's Deepfake Detection (DFD), also comprised of thousands of Google-generated deepfake videos; Celeb-DeepFake v1 (CD1), which features 408 real and 795 synthesized, low-artifact videos; Celeb-DeepFake v2, an extension of V1 that contains 590 real and 5,639 fake videos; and China's 2020 Deeper-Forensics (Deeper).

Those are the datasets; the detection methods in the test challenges were Multi-task, MesoInc4, Capsule, Xception-c0, c2 (a method employed in FF++), FWA/DSP-FW from the University at Albany, Two-Branch, PCL+I2G, and Yuval Nirkin's context-discrepancy method.

The aforementioned detection methods are aimed at detecting particular types of facial manipulation. In addition to these, the new paper's authors tested more general deepfake detection offerings Face X-ray, Michigan State University's FFD, CNNDetection, and Patch-Forensics from MIT CSAIL.

The most evident results from the test are that the competing methods drastically drop in effectiveness as video resolution and quality lowers. Since some of the most severe potential for deepfake penetrating of our discriminative powers lies (not least at the current time) in non-HD or otherwise quality-compromised video, this would seem to be a significant result.



In the results graph above, the blue and red lines indicate the resilience of ICT methods to image degradation in all areas except the roadblock of Gaussian noise (not a likelihood in Zoom and webcam-style footage), while the competing methods' reliability plummets.

In the table of results below, we see the effectiveness of the varied deepfake detection methods on the unseen datasets. Grey and asterisked results indicate comparison from originally published results in closed-source projects, which cannot be externally verified. Across nearly all comparable frameworks, ICT outperforms the rival deepfake detection approaches (shown in bold) over the trialed datasets.

Method	DFD	FF++	Deeper	CD1	CD2	Avg
Multi-task [43]	65.21	72.23	65.32	72.28	61.06	65.96
MesoInc4 [7]	59.06	63.41	51.41	42.26	53.60	53.95
Capsule [44]	69.70	96.50	68.44	69.98	63.65	67.94
Xcep-c0 [51]	89.05	99.26	57.76	48.08	50.37	61.32
Xcep-c23 [51]	**95.60**	98.54	69.85	74.97	77.82	79.56
FWA [34]	80.59	74.82	45.46	72.88	64.87	67.72
DSP-FWA [34]	90.99	81.90	60.00	78.51	81.41	78.56
CNNDetect [60]	60.12	71.08	57.16	56.12	57.17	60.33
Patch-Foren [13]	49.91	73.75	55.35	59.66	57.16	55.52
FFD [20]	76.61	**92.32**	46.64	74.15	77.80	73.50
Face X-ray [32]	94.14	98.44	72.35	74.76	75.39	79.16
Two-Branch [38]*	-	-	-	-	73.41	-
PCL+I2G [63]*	-	-	-	-	81.80	-
Nirkin.*et al.* [47]*	-	99.70	-	-	66.00	-
ICT (Ours)	84.13	90.22	**93.57**	**81.43**	**85.71**	**87.01**
ICT-Ref (Ours)	93.17	98.56	99.25	96.41	94.43	96.34

As an additional test, the authors ran content from the YouTube channel of acclaimed deepfaker Ctrl Shift Face, and found competing methods achieved notably inferior identification scores:

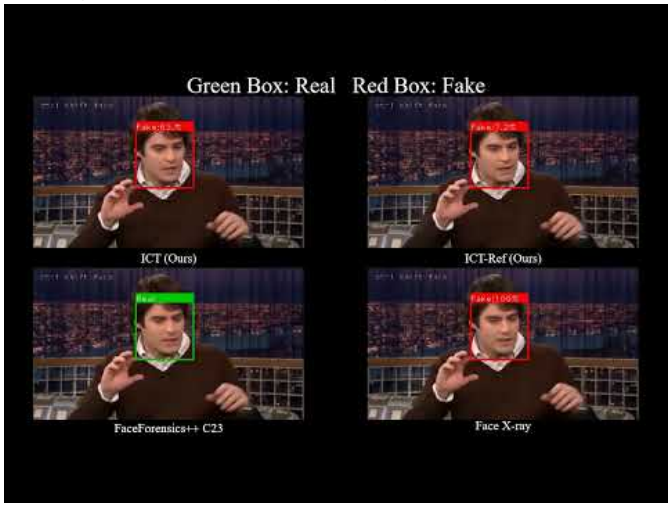
Method	Video1	Video2	Video3	Avg
Multi-task [43]	62.50	26.45	78.50	55.81
DSP-FWA [34]	23.08	33.14	51.81	36.01
Xception-c23 [51]	44.56	83.85	99.31	75.90
FFD [20]	56.04	71.16	79.07	68.75
Face X-ray [32]	66.49	94.03	87.10	82.54
ICT (Ours)	82.27	97.78	98.05	94.36
ICT-Ref (Ours)	100.0	100.0	100.0	100.0

Notable here is that FF++ methods (Xception-c23) and FFD, which achieve a few of the highest scores across some of the testing data in the new paper's general tests, here achieve a far lower score than ICT in a 'real world' context of high-effort deepfake content.

The authors conclude the paper with the hope that its results steer the deepfake detection community towards similar initiatives that concentrate on more easily generalizable high-level features, and away from the 'cold war' of artifact detection, wherein the latest methods are routinely obviated by developments in deepfake frameworks, or by other factors that make such methods less resilient.

Check out the accompanying supplementary video below for more examples of ICT identifying deepfake content that often outfoxes alternative methods.

Protecting Celebrities with Identity Consistency Transformer-CVPR2022



First published 4th March 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