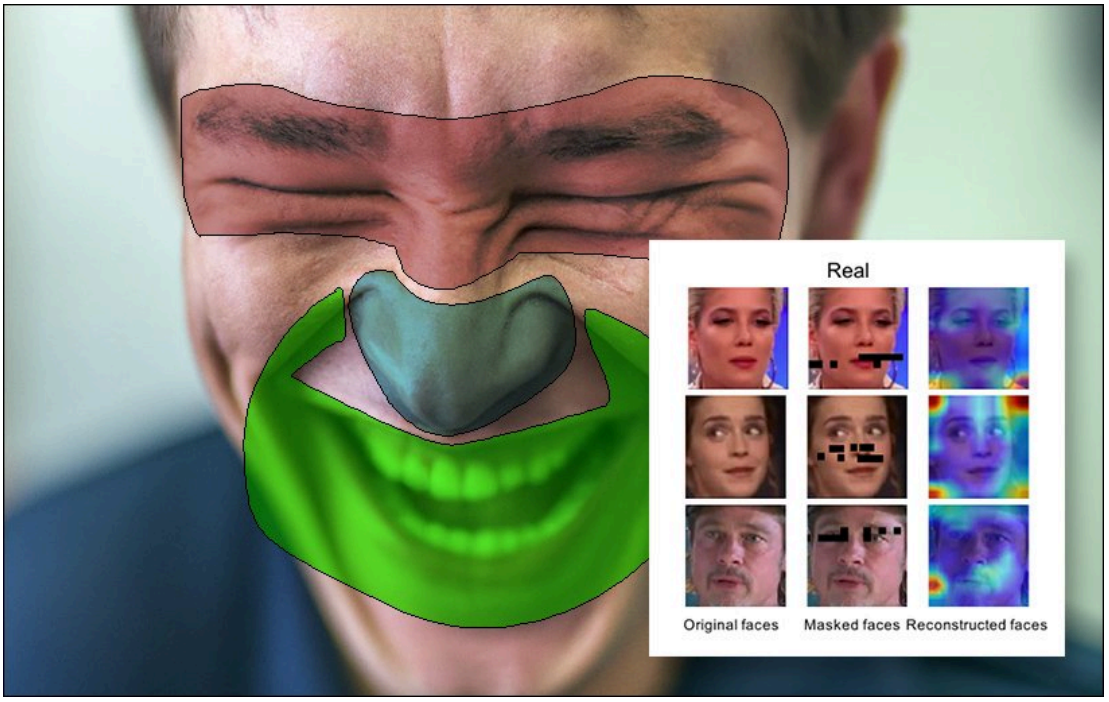
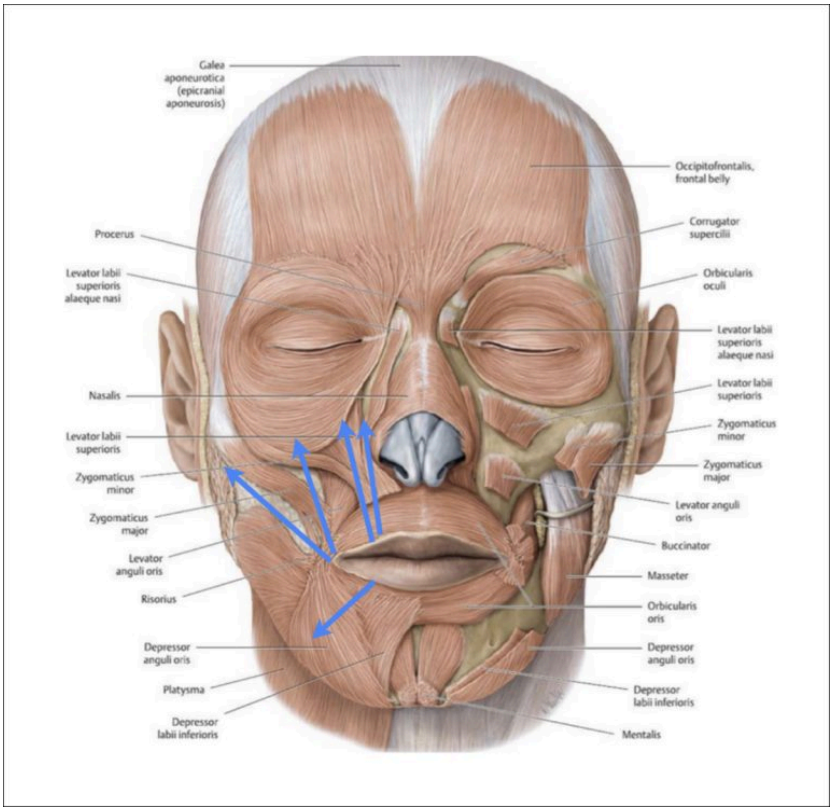
Detecting Deepfakes by Studying ‘Disconnected’ Facial Expressions

Originally published at https://blog.metaphysic.ai/detecting-deepfakes-by-studying-disconnected-facial-expressions/

March 7, 2023



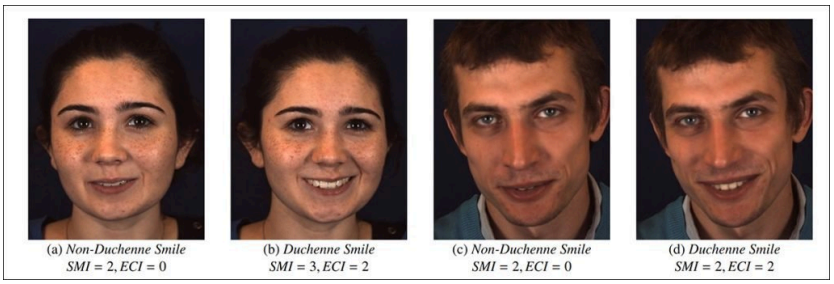
A new academic collaboration between China and Singapore takes a novel and higher-level approach to detecting deepfaked faces than the general impetus of research in the last few years.
Rather than seeking out artefacts or badly-rendered parts of a deepfake face, the system studies the way that different parts of the face interact and correspond with each other during the normal production of facial expressions, or speaking.



In the medical sphere, the involuntary 'agreement' of other parts of the face with whatever is happening elsewhere in the face is known as facial synkinesis – one example being the tendency of the eyes to close a little (or entirely) during an intense smile or laughter. Source: https://www.facialparalysisinstitute.com/blog/the-anatomy-of-a-smile-and-facial-synkinesis/

For instance, if an AI system is used to turn a smile into a frown, there are many interconnected muscles away from the mouth that are affected by the movements in the mouth muscles. However, deepfake systems tend to partition off areas of the face, rather than considering the synergistic effect of a smile or any other facial expression.

This means that an artificially-altered image of a person may have 'smiling lips', but not 'smiling eyes'. This is the difference (for instance) between an authentic Duchenne smile and the kind of 'mouth only' smile that strikes us as 'insincere' or 'contrived', rather than natural.

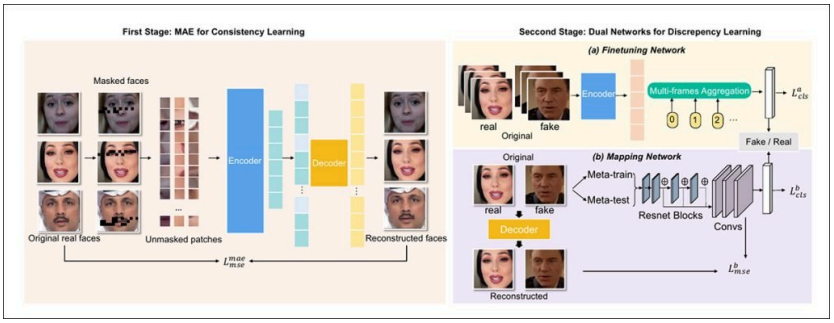


Insincere and genuine smiles, in a study of the Duchenne smile. Source: https://link.springer.com/article/10.1007/s42761-020-00030-w

If one can quantify these relationships, it's possible to create an analytical framework that can detect when these subtle collateral effects are not present, and therefore to produce a system that is based on the way faces work, and not on the way that the latest deepfake algorithm functions. Such a system would be highly resistant to the typical quality of improvements that occur in deepfaking technologies, due to the 'Balkanized' way that face-swapping systems (particularly autoencoder deepfake systems such as DeepFaceLab) tend to mix-and-match facial characteristics. Though this kind of disentanglement is generally a great benefit to the quality of image synthesis, it can be a notable 'tell' that a face may not be authentic.
This is the central idea between *DeepfakeMAE*, a new initiative from researchers at China's Hunan University, the National University of Singapore, and the Nanyang Technological University at Singapore.

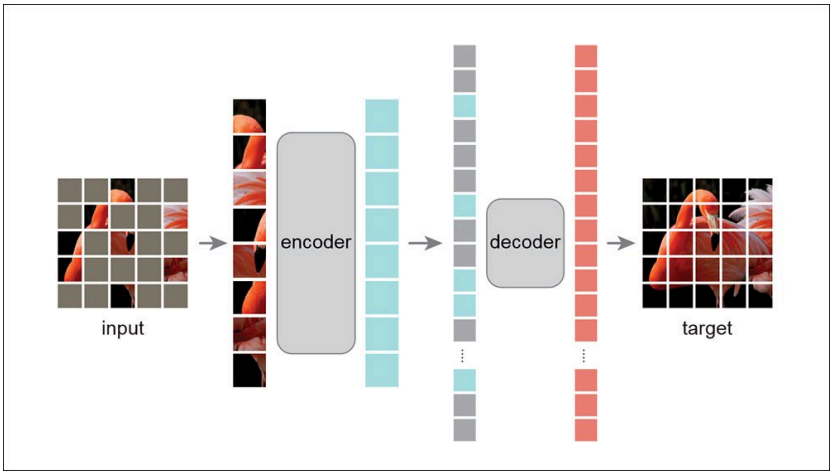
Approach

The workflow for DeepfakeMAE (with MAE standing for 'Masked Autoencoder') is divided into two stages. First, a dedicated network trains exclusively on 'real' videos, learning how the different parts of the face relate to each other, by masking out sections of the face and building up a complete picture of those intra-facial relationships; and then, a second fine-tuning network applies the results of the first stage to a second round of training, this time adding known fake videos into the mix.



DeepfakeMAE's pipeline initially learns general facial consistencies exclusively from genuine videos (left, in image above). The second phase adds fake videos into a fine-tuning stage, and simultaneously creates a mapping network that can identify discrepancies between real and fake video sources. Source: https://arxiv.org/pdf/2303.01740.pdf

The masked autoencoder used in the system is based on a Facebook AI Research (FAIR) 2021 project, which learns reconstruction tasks by obscuring sections of the source image during training – an adversarial approach, not entirely dissimilar to the discriminator system in Generative Adversarial Networks (GANs).

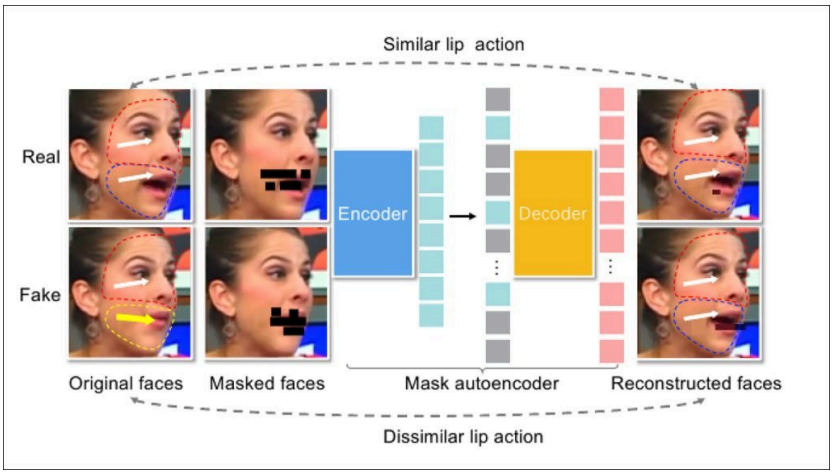


The FAIR masked autoencoder initiative iteratively teaches image reconstruction by challenging the system to recreate masked-out areas of a source image – a task at which it progressively improves through training. Source: https://arxiv.org/pdf/2111.06377.pdf

The encoder is ViT-based, but applied only to *un*masked patches, which is not the standard Vision Transformer method.
The reconstruction is evaluated with the Mean Squared Error (MSE) loss function. Discussing the initial real>real reconstruction face, the authors state:

'Our facial part masking strategy makes each part selected randomly, which enforces the model to learn the representation unspecific to any facial part. Furthermore, because this stage only uses real videos and does not use any Deepfake videos, it can prevent the model from over-fitting to any specific tampering pattern.'

During the *real>real* stage, a standard 68 facial landmarks are identified via Dlib, which are then used to divide the face up into three patches: eyes, nose & cheek, and lips. For each iteration during training, one of these segments is randomly selected for comparison between the various real videos.
The second stage consists of two parts: a fine-tuning network and a mapping network. Cross-entropy loss is used to identify the difference between real and fake videos. After this, fully-patched sets of fake and real faces are put into the encoder from the first stage, using the average output of five frames in the last layer of the decoder.



Lip movement notably affects the rest of the face, and DeepfakeMAE can use this as one of a number of keys and characteristics from which to extract evidence of deepfaked video.

Data, Training, and Tests

DeepfakeMAE was evaluated using four publicly-available deepfake datasets: Celeb-DF; WildDeepfake; The DeepFake Detection Challenge (DFDC) dataset; and FaceForensics++. The latter consists of 4000 videos using four different facial synthesis algorithms, including DeepFakes, Face2Face, FaceSwap and Neural Textures.
The Celeb-DF dataset features 5639 deepfaked videos; WildDeepfake is a web-scraped collection of 707 deepfaked videos; and DFDC is a high-scale (124k+) collection frequently used in evaluation of new deepfake detector systems.
For training the first of the two stages in the architecture, the AdamW optimizer was set to an initial learning rate of $1.5 – 10^{-4}$, a weight decay of 0.05, and a momentum of 0.9. For stage 2, the fine-tuning network used AdamW at an initial learning rate of 10^{-3} for video detection, with an SGD optimizer used to optimize the mapping network at a learning rate of 0.1, a momentum of 0.9 and a weight decay of 5.10^{-4}.
The system was tested first for cross-domain generalization, against an extensive range of prior approaches: Face Warping Artifacts (FWA); ADDnet (the accompaniment to WildDeepfake); Face X-ray; CNN-aug; Multi-attentional Deepfake Detection (MultiAtt); Spatial-Phase Shallow Learning (SPSL); Learning-to-weight (LTW); LipForensics; Frame inference-based deepfake detection (FInffer); Hierarchical Contrastive Inconsistency Learning for Deepfake Video Detection (HCIL); Reconstruction-classification learning (RECCE); and Self-supervised Learning of Adversarial Example (noted in results as '*Chin et al.*').

Method	Testing datasets			
	Celeb-DF	WildDeepfake	DFDC	Avg
FWA [27]	69.5	68.2	67.3	68.3
ADDNet [52]	60.9	54.2	65.9	60.3
Face X-ray [24]	79.5	77.3	65.5	74.1
CNN-aug [45]	75.6	*	72.1	*
MultiAtt [50]	67.4	65.7	71.0	68.0
SPSL [29]	76.9	70.9	66.2	71.3
LTW [40]	64.1	*	69.0	*
LipForensics [15]	82.4	78.1	73.5	78.0
FInfer [18]	70.6	69.5	70.4	70.2
HCIL [14]	79.0	*	69.2	*
Chen et al. [4]	79.7	*	*	*
RECCE [2]	68.7	64.3	69.1	67.4
DeepfakeMAE	**86.3**	**81.8**	**75.2**	**81.1**

Initial results against prior frameworks.

Of these initial test, the authors state:

'DeepfakeMAE achieves a satisfactory generalization to unseen datasets surpassing previous methods. It also outperforms the previous state-of-the-art method, [LipForensics], by 3.9% AUC on Celeb-DF, by 3.7% AUC on WildDeepfake, by 1.7% AUC on DFDC.
'We note that [LipForensics] focuses on the features of the lips area, but the proposed DeepfakeMAE enforces the model to automatically learn the consistency features of all parts'

The second round of tests were for cross-manipulation generalization, where systems were asked to detect deepfake content for unknown manipulation methods (i.e., methods that may crop up in the future, but which either do not currently exist, or currently have different characteristics).
Here, the forgery technologies available in the samples from FaceForensics++ were split into a training and test set.

Method	Train	F2F	FS	NT	Avg
Freq-SCL [23]	DF	58.9	66.9	63.6	63.1
MultiAtt [50]		66.4	67.3	66.0	66.6
RECCE [2]		70.7	74.3	67.3	70.8
DeepfakeMAE		**86.6**	**77.2**	**69.4**	**77.7**
Method	Train	DF	FS	NT	Avg
Freq-SCL [23]	F2F	67.6	55.4	66.7	63.2
MultiAtt [50]		73.0	65.1	71.9	70.0
RECCE [2]		76.0	64.5	**72.3**	71.0
DeepfakeMAE		**88.9**	**71.4**	55.1	**71.8**
Method	Train	DF	F2F	NT	Avg
Freq-SCL [23]	FS	75.9	54.6	49.7	60.1
MultiAtt [50]		82.3	61.7	54.8	66.3
RECCE [2]		**82.4**	64.4	**56.7**	67.8
DeepfakeMAE		80.0	**80.9**	47.3	**69.4**
Method	Train	DF	F2F	FS	Avg
Freq-SCL [23]	NT	79.1	74.2	54.0	69.1
MultiAtt [50]		74.6	80.6	60.9	72.0
RECCE [2]		78.8	**80.9**	**63.7**	**74.5**
DeepfakeMAE		**85.7**	63.4	62.4	70.5

Results for cross-domain generalization, where the systems must detect generic possible deepfake approaches, without prior knowledge of the signatures of such methods. F2F = Face2Face; FS = FaceSwap; DF = Deepfakes; and NT = Neural Textures.

DeepfakeMAE's performance in this this round is explained by the authors:

'[Our] method has a little bit more difficulty in detecting the Deepfake videos using neuraltextures than the other methods. NeuralTextures learns specific neural texture [features] to train a model to generate faces leaving specific artifacts.
'On the contrary, the proposed DeepfakeMAE focuses on unspecific features, which might be the reason why our method does not generalize well on NeuralTextures.'

Finally, DeepfakeMAE was tested for intra-dataset detection performance, using four subsets of FaceForensics++. Frameworks tested included Xception (the adjunct architecture for FaceForensics++), ADDnet, Sharp Multiple Instance Learning for DeepFake Video Detection (S-MIL), Spatiotemporal Inconsistency Learning for DeepFake Video Detection (STIL), and HCIL.

Method	DF	FS	F2F	NT	Avg
Xception [37]	98.9	99.6	98.9	95.0	98.1
ADDNet [52]	92.1	92.5	83.9	78.2	86.7
S-MIL [25]	98.6	99.3	99.3	95.7	98.2
STIL [12]	99.6	**100**	99.3	95.4	98.6
HCIL [14]	**100**	**100**	99.3	**96.8**	**99.0**
DeepfakeMAE	99.6	99.4	**99.4**	96.3	98.7

Results from the intra-dataset detection performance test.

While noting that DeepfakeMAE achieves the best overall performance in this round, the authors observe:

'[We] highlight that the cross-dataset performance of DeepfakeMAE…significantly surpasses that of [HCIL] by 7.3% on Celeb-DF and by 6% on DFDC datasets. Overall, DeepfakeMAE owns satisfactory intra-dataset detection performance in addition to state-of-the-art cross-dataset detection performance, benefiting from the mechanism that learns robust features from masking and reconstructing facial parts.'

Conclusion

DeepfakeMAE is one of a new breed of deepfake detection research projects that are attempting to key not on the behavior of deepfake algorithms, but on the behavior of the resulting deepfaked faces. This is an important strand of research, since a slew of once-effective deepfake detection methods have since been surpassed by improvements in facial synthesis technologies, which amend errant behavior such as lack of eye-blinking or various other kinds of bizarre artefacts.

A truly exhaustive deepfake dataset of images could defeat a system of this type, but only if it contained an *extraordinary* and improbable variety of facial expressions in an exhaustive array of possible facial poses.

That's the kind of dataset you are only likely to obtain in feature film production, with the cooperation of actors who are willing and able to run through a wide variety of emotions in a clinical, light stage-style scenario, and where the resulting dataset will not ever reach the public.

Additionally, the recognition system that extracts the faces would need to make some progress on an aspect that is long overdue in professional, VFX-centric deepfaking: emotion recognition. Unsupervised, pixel-based latent reconstruction of faces via the popular deepfake frameworks makes no account of this, and has no such facility; therefore, these publicly-available systems are currently unable to even flag facial poses such as the Duchenne smile, so that the end-user can be alerted to the need to reproduce 'knock-on effects' in other part of the face.

One possible weakness of DeepfakeMAE's system is that it seems likely to 'flag' instances of insincere smiles, which are not a meager resource in the real world; and that its generalized approach may fail to account for the vagaries and eccentricities of a person's own facial affective behavior, which may deviate from the trained norms.