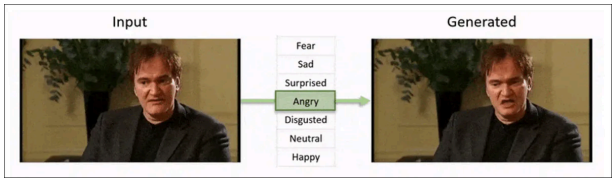
Exposing Small but Significant AI Edits in Real Video

Originally published at https://www.unite.ai/exposing-small-but-significant-ai-edits-in-real-video/

In 2019, US House of Representatives Speaker Nancy Pelosi was the subject of a targeted and pretty low-tech deepfake-style attack, when real video of her was edited to make her appear drunk – an unreal incident that was shared several million times before the truth about it came out (and, potentially, after some stubborn damage to her political capital was effected by those who did not stay in touch with the story).

Though this misrepresentation required only some simple audio-visual editing, rather than any AI, it remains a key example of how subtle changes in real audio-visual output can have a devastating effect.

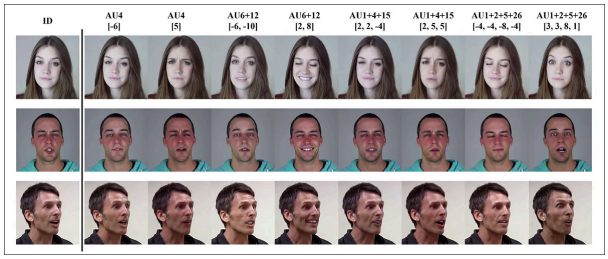
At the time, the deepfake scene was dominated by the autoencoder-based face-replacement systems which had debuted in late 2017, and which had not significantly improved in quality since then. Such early systems would have been hard-pressed to create this kind of small but significant alterations, or to realistically pursue modern research strands such as expression editing:



The 2022 'Neural Emotion Director' framework changes the mood of a famous face. Source: https://www.youtube.com/watch?v=Li6W8pRDMJQ

Things are now quite different. The movie and TV industry is seriously interested in post-production alteration of real performances using machine learning approaches, and AI's facilitation of *post facto* perfectionism has even come under recent criticism.

Anticipating (or arguably creating) this demand, the image and video synthesis research scene has thrown forward a wide range of projects that offer 'local edits' of facial captures, rather than outright replacements: projects of this kind include Diffusion Video Autoencoders; Stitch it in Time; ChatFace; MagicFace; and DISCO, among others.

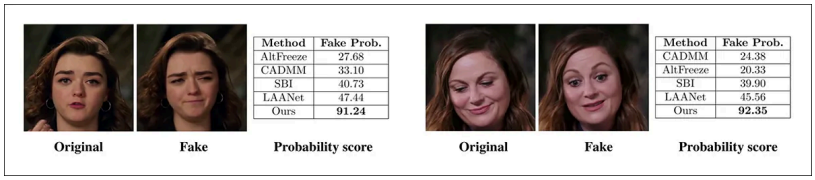


Expression-editing with the January 2025 project MagicFace. Source: https://arxiv.org/pdf/2501.02260

New Faces, New Wrinkles

However, the enabling technologies are developing far more rapidly than methods of detecting them. Nearly all the deepfake detection methods that surface in the literature are chasing yesterday's deepfake methods with yesterday's datasets. Until this week, none of them had addressed the creeping potential of AI systems to create small and topical local alterations in video.

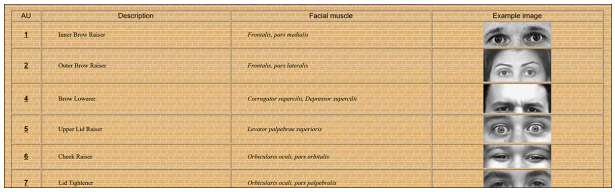
Now, a new paper from India has redressed this, with a system that seeks to identify faces that have been *edited* (rather than replaced) through AI-based techniques:



Detection of Subtle Local Edits in Deepfakes: A real video is altered to produce fakes with nuanced changes such as raised eyebrows, modified gender traits, and disgust (illustrated here with a single frame). Source: https://arxiv.org/pdf/2503.22121

The authors' system is aimed at identifying deepfakes that involve subtle, localized facial manipulations – an otherwise neglected class of forgery. Rather than focusing on global inconsistencies or identity mismatches, the approach targets fine-grained changes such as slight expression shifts or small edits to specific facial features.

The method makes use of the Action Units (AUs) delimiter in the Facial Action Coding System (FACS), which defines 64 possible individual mutable areas in the face, which which together form expressions.



Some of the constituent 64 expression parts in FACS. Source: https://www.cs.cmu.edu/~face/facs.htm

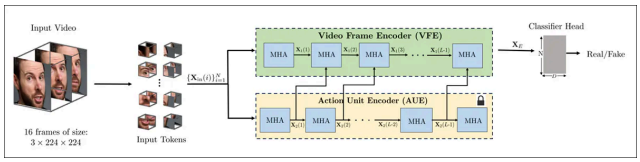
The authors evaluated their approach against a variety of recent editing methods and report consistent performance gains, both with older datasets and with much more recent attack vectors:

> *'By using AU-based features to guide video representations learned through Masked Autoencoders [(MAE)], our method effectively captures localized changes crucial for detecting subtle facial edits.*
>
> *'This approach enables us to construct a unified latent representation that encodes both localized edits and broader alterations in face-centered videos, providing a comprehensive and adaptable solution for deepfake detection.'*

The new paper is titled *Detecting Localized Deepfake Manipulations Using Action Unit-Guided Video Representations*, and comes from three authors at the Indian Institute of Technology at Madras.

Method

In line with the approach taken by VideoMAE, the new method begins by applying face detection to a video and sampling evenly spaced frames centered on the detected faces. These frames are then divided into small 3D divisions (i.e., temporally-enabled patches), each capturing local spatial and temporal detail.



Schema for the new method. The input video is processed with face detection to extract evenly spaced, face-centered frames, which are then divided into 'tubular' patches and passed through an encoder that fuses latent representations from two pretrained pretext tasks. The resulting vector is then used by a classifier to determine whether the video is real or fake.

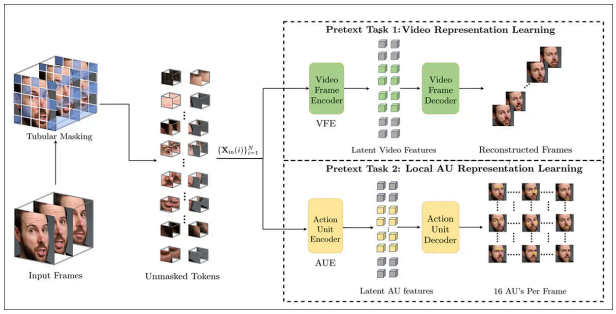
Each 3D patch contains a fixed-size window of pixels (i.e., 16×16) from a small number of successive frames (i.e., 2). This lets the model learn short-term motion and expression changes – not just what the face looks like, but *how it moves*.

The patches are embedded and positionally encoded before being passed into an encoder designed to extract features that can distinguish real from fake.

The authors acknowledge that this is particularly difficult when dealing with subtle manipulations, and address this issue by constructing an encoder that combines two separate types of learned representations, using a cross-attention mechanism to fuse them. This is intended to produce a more sensitive and generalizable feature space for detecting localized edits.

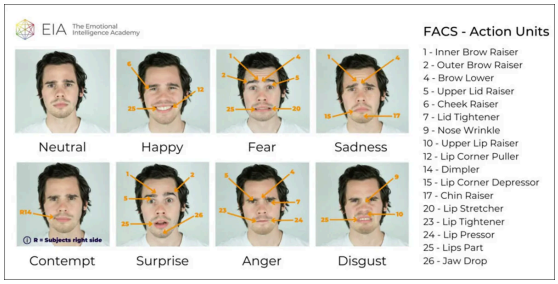
Pretext Tasks

The first of these representations is an encoder trained with a masked autoencoding task. With the video split into 3D patches (most of which are hidden), the encoder then learns to reconstruct the missing parts, forcing it to capture important spatiotemporal patterns, such as facial motion or consistency over time.



Pretext task training involves masking parts of the video input and using an encoder-decoder setup to reconstruct either the original frames or per-frame action unit maps, depending on the task.

However, the paper observes, this alone does not provide enough sensitivity to detect fine-grained edits, and the authors therefore introduce a second encoder trained to detect facial action units (AUs). For this task, the model learns to reconstruct dense AU maps for each frame, again from partially masked inputs. This encourages it to focus on localized muscle activity, which is where many subtle deepfake edits occur.



Further examples of Facial Action Units (FAUs, or AUs). Source: https://www.eiagroup.com/the-facial-action-coding-system/

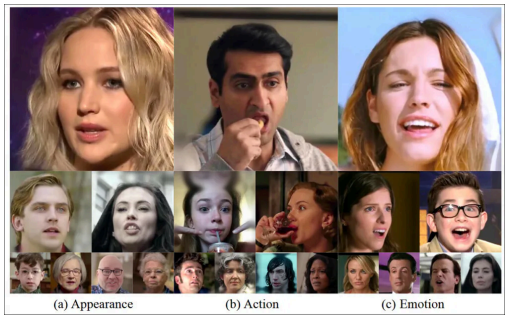
Once both encoders are pretrained, their outputs are combined using cross-attention. Instead of simply merging the two sets of features, the model uses the AU-based features as *queries* that guide attention over the spatial-temporal features learned from masked autoencoding. In effect, the action unit encoder tells the model where to look.

The result is a fused latent representation that is meant to capture both the broader motion context and the localized expression-level detail. This combined feature space is then used for the final classification task: predicting whether a video is real or manipulated.

Data and Tests

Implementation

The authors implemented the system by preprocessing input videos with the FaceXZoo PyTorch-based face detection framework, obtaining 16 face-centered frames from each clip. The pretext tasks outlined above were then trained on the CelebV-HQ dataset, comprising 35,000 high-quality facial videos.



From the source paper, examples from the CelebV-HQ dataset used in the new project. Source: https://arxiv.org/pdf/2207.12393

Half of the data examples were masked, forcing the system to learn general principles instead of overfitting to the source data.

For the masked frame reconstruction task, the model was trained to predict missing regions of video frames using an L1 loss, minimizing the difference between the original and reconstructed content.

For the second task, the model was trained to generate maps for 16 facial action units, each representing subtle muscle movements in areas such including eyebrows, eyelids, nose, and lips, again supervised by L1 loss.

After pretraining, the two encoders were fused and fine-tuned for deepfake detection using the FaceForensics++ dataset, which contains both real and manipulated videos.

The FaceForensics++ dataset has been the cornerstone of deepfake detection since 2017, though it is now considerably out of date, in regards to the latest facial synthesis techniques. Source: https://www.youtube.com/watch?v=x2g48Q2l2ZQ

To account for class imbalance, the authors used Focal Loss (a variant of cross-entropy loss), which emphasizes more challenging examples during training.

All training was conducted on a single RTX 4090 GPU with 24Gb of VRAM, with a batch size of 8 for 600 epochs (complete reviews of the data), using pre-trained checkpoints from VideoMAE to initialize the weights for each of the pretext tasks.

Tests

Quantitative and qualitative evaluations were carried out against a variety of deepfake detection methods: FTCN; RealForensics; Lip Forensics; EfficientNet+ViT; Face X-Ray; Alt-Freezing; CADMM; LAANet; and BlendFace's SBI. In all cases, source code was available for these frameworks.

The tests centered on locally-edited deepfakes, where only part of a source clip was altered. Architectures used were Diffusion Video Autoencoders (DVA); Stitch It In Time (STIT); Disentangled Face Editing (DFE); Tokenflow; VideoP2P; Text2Live; and FateZero. These methods employ a diversity of approaches (diffusion for DVA and StyleGAN2 for STIT and DFE, for instance)

The authors state:

> *'To ensure comprehensive coverage of different facial manipulations, we incorporated a wide variety of facial features and attribute edits. For facial feature editing, we modified eye size, eye-eyebrow distance, nose ratio, nose-mouth distance, lip ratio, and cheek ratio. For facial attribute editing, we varied expressions such as smile, anger, disgust, and sadness.*
>
> *'This diversity is essential for validating the robustness of our model over a wide range of localized edits. In total, we generated 50 videos for each of the above-mentioned editing methods and validated our method's strong generalization for deepfake detection.'*

Older deepfake datasets were also included in the rounds, namely Celeb-DFv2 (CDF2); DeepFake Detection (DFD); DeepFake Detection Challenge (DFDC); and WildDeepfake (DFW).

Evaluation metrics were Area Under Curve (AUC); Average Precision; and Mean F1 Score.

Detection Methods	DVA CVPR'23		STIT SIGGRAPH'22		DFE ICCV'21		Tokenflow ICLR'23		VideoP2P CVPR'24		TextLive ECCV'22		Fatezero ICCV'23	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
FTCN	27.1	30.1	34.8	37.0	33.5	35.8	29.5	32.1	31.0	33.7	30.2	32.9	28.7	31.3
RealForensics	37.5	40.3	46.9	49.8	45.6	48.6	41.2	44.2	42.8	45.8	42.0	45.1	40.5	43.4
Lip Forensics	33.8	37.1	42.0	45.8	40.9	44.6	36.5	40.3	38.0	41.9	37.3	41.2	35.8	39.5
EfficientNet+ViT	36.2	38.4	44.7	47.1	43.5	45.9	39.0	41.5	40.8	43.2	40.1	42.5	38.6	40.8
Face X-Ray	33.4	36.1	41.1	43.6	40.3	42.9	36.0	39.1	37.5	40.7	36.8	40.0	35.2	38.2
LAA Net	61.5	58.0	72.5	69.3	71.2	68.1	66.4	62.9	68.0	64.5	67.1	63.7	65.7	61.8
SBI	65.2	62.8	75.5	73.2	73.3	71.5	69.0	66.4	70.8	68.1	70.2	67.5	68.6	65.9
AltFreezing	41.1	44.0	51.8	53.6	51.2	52.8	45.6	47.1	47.5	48.9	46.9	48.3	45.0	46.5
CADMM	44.5	47.2	55.6	57.4	54.3	56.2	49.0	50.9	50.5	52.3	49.9	51.7	48.2	49.9
Ours	**87.2**	**85.8**	**92.5**	**90.7**	**93.1**	**91.6**	**91.7**	**89.4**	**90.3**	**90.2**	**89.1**	**87.9**	**88.5**	**86.0**

From the paper: comparison on recent localized deepfakes shows that the proposed method outperformed all others, with a 15 to 20 percent gain in both AUC and average precision over the next-best approach.

The authors additionally provide a visual detection comparison for locally manipulated views (reproduced only in part below, due to lack of space):

A real video was altered using three different localized manipulations to produce fakes that remained visually similar to the original. Shown here are representative frames along with the average fake detection scores for each method. While existing detectors struggled with these subtle edits, the proposed model consistently assigned high fake probabilities, indicating greater sensitivity to localized changes.

The researchers comment:

'[The] existing SOTA detection methods, [LAANet], [SBI], [AltFreezing] and [CADMM], experience a significant drop in performance on the latest deepfake generation methods. The current SOTA methods exhibit AUCs as low as 48-71%, demonstrating their poor generalization capabilities to the recent deepfakes.

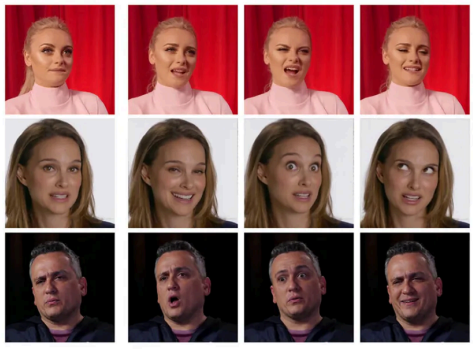
'On the other hand, our method demonstrates robust generalization, achieving an AUC in the range 87-93%. A similar trend is noticeable in the case of average precision as well. As shown [below], our method also consistently achieves high performance on standard datasets, exceeding 90% AUC and are competitive with recent deepfake detection models.'

Method	CDF2		DFD		DFW		DFDC	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP
FTCN [72]	86.9	86.0	94.4	90.33	64.73	65.5	86.0	87.48
RealForensics [24]	85.6	85.2	82.24	84.62	66.72	66.5	89.7	88.46
Lip Forensics [23]	82.4	82.66	90.2	89.37	62.3	60.75	92.53	93.41
EfficientNet+ViT [11]	79.0	75.61	87.0	88.09	72.0	68.74	91.0	85.12
Face X-Ray [39]	79.5	80.41	95.4	94.7	61.5	60.94	85.27	85.0
LAANet[50]	**95.4**	**97.64**	**99.5**	**99.8**	87.6	85.08	86.94	**97.7**
SBI [61]	93.18	85.16	97.56	92.79	84.83	**88.37**	86.16	93.24
AltFreezing [69]	89.5	88.46	98.50	93.17	72.6	72.0	**94.0**	88.11
CADMM [16]	93.0	91.12	99.03	99.59	75.0	79.42	88.3	89.71
Ours	93.84	95.27	97.15	95.28	**91.0**	88.25	93.0	91.93

Performance on traditional deepfake datasets shows that the proposed method remained competitive with leading approaches, indicating strong generalization across a range of manipulation types.

The authors observe that these last tests involve models that could reasonably be seen as outmoded, and which were introduced prior to 2020.

By way of a more extensive visual depiction of the performance of the new model, the authors provide an extensive table at the end, only part of which we have space to reproduce here:



Sad, Anger, Disgust

Method	Fake Prob.
AltFreeze	34.47
CADMM	40.30
SBI	68.10
LAANet	59.17
Ours	**90.36**

Smile, EyeRaise, EyeGaze

Method	Fake Prob.
AltFreeze	32.13
CADMM	36.0
SBI	51.19
LAANet	53.74
Ours	**87.89**

Shock, EyeRaise, Smile

Method	Fake Prob.
AltFreeze	33.0
CADMM	33.69
SBI	66.25
LAANet	59.07
Ours	**89.13**

In these examples, a real video was modified using three localized edits to produce fakes that were visually similar to the original. The average confidence scores across these manipulations show, the authors state, that the proposed method detected the forgeries more reliably than other leading approaches. Please refer to the final page of the source PDF for the complete results.

The authors contend that their method achieves confidence scores above 90 percent for the detection of localized edits, while existing detection methods remained below 50 percent on the same task. They interpret this gap as evidence of both the sensitivity and generalizability of their approach, and as an indication of the challenges faced by current techniques in dealing with these kinds of subtle facial manipulations.

To assess the model's reliability under real-world conditions, and according to the method established by CADMM, the authors tested its performance on videos modified with common distortions, including adjustments to saturation and contrast, Gaussian blur, pixelation, and block-based compression artifacts, as well as additive noise.

The results showed that detection accuracy remained largely stable across these perturbations. The only notable decline occurred with the addition of Gaussian noise, which caused a modest drop in performance. Other alterations had minimal effect.

Perturbation	DVA	STIT	DFE	TF	T2L	FZ	V2P
Gaussian Noise	82.3	83.1	83.8	84.5	85.0	84.2	83.7
Saturation Change	87.2	88.1	88.5	88.9	89.3	89.0	88.6
Blockwise Distortion	86.7	87.5	87.8	88.3	88.6	88.4	88.0
Contrast Change	86.8	87.6	88.0	88.4	88.8	88.5	88.1
Pixelation	86.0	86.8	87.3	87.8	88.1	87.9	87.5
Gaussian Blur	86.3	87.0	87.5	87.9	88.2	88.0	87.6
No Perturbation	**88.5**	**89.3**	**89.7**	**90.1**	**90.6**	**90.3**	**89.8**

An illustration of how detection accuracy changes under different video distortions. The new method remained resilient in most cases, with only a small decline in AUC. The most significant drop occurred when Gaussian noise was introduced.

These findings, the authors propose, suggest that the method's ability to detect localized manipulations is not easily disrupted by typical degradations in video quality, supporting its potential robustness in practical settings.

Conclusion

AI manipulation exists in the public consciousness chiefly in the traditional notion of deepfakes, where a person's identity is imposed onto the body of another person, who may be performing actions antithetical to the identity-owner's principles. This conception is slowly becoming updated to acknowledge the more insidious capabilities of generative video systems (in the new breed of video deepfakes), and to the capabilities of latent diffusion models (LDMs) in general.

Thus it is reasonable to expect that the kind of local editing that the new paper is concerned with may not rise to the public's attention until a Pelosi-style pivotal event occurs, since people are distracted from this possibility by easier headline-grabbing topics such as video deepfake fraud.

Nonetheless much as the actor Nic Cage has expressed consistent concern about the possibility of post-production processes 'revising' an actor's performance, we too should perhaps encourage greater awareness of this kind of 'subtle' video adjustment – not least because we are by nature incredibly sensitive to very small variations of facial expression, and because context can significantly change the impact of small facial movements (consider the disruptive effect of even smirking at a funeral, for instance).

First published Wednesday, April 2, 2025

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