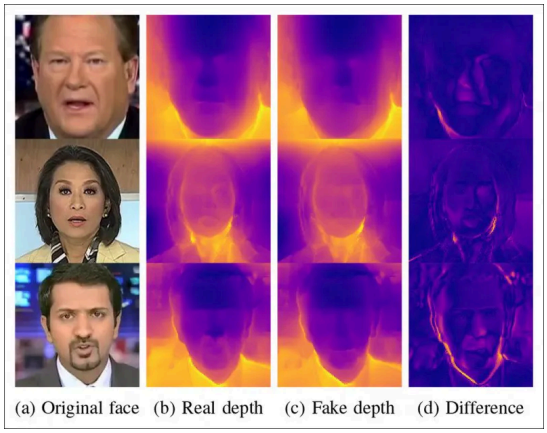
Depth Information Can Reveal Deepfakes in Real-Time

Originally published at https://www.unite.ai/depth-information-can-reveal-deepfakes-in-real-time/

New research from Italy has found that depth information obtained from images can be a useful tool to detect deepfakes – even in real-time.

Whereas the majority of research into deepfake detection over the past five years has concentrated on artifact identification (which can be mitigated by improved techniques, or mistaken for poor video codec compression), ambient lighting, biometric traits, temporal disruption, and even human instinct, the new study is the first to suggest that depth information could be a valuable cipher for deepfake content.



Examples of derived depth-maps, and the difference in perceptual depth information between real and fake images. Source: https://arxiv.org/pdf/2208.11074.pdf

Critically, detection frameworks developed for the new study operate very well on a lightweight network such as Xception, and acceptably well on MobileNet, and the new paper acknowledges that the low latency of inference offered through such networks can enable real-time deepfake detection against the new trend towards live deepfake fraud, exemplified by the recent attack on Binance.

Greater economy in inference time can be achieved because the system does not need full-color images in order to determine the difference between fake and real depth maps, but can operate surprisingly efficiently solely on grayscale images of the depth information.

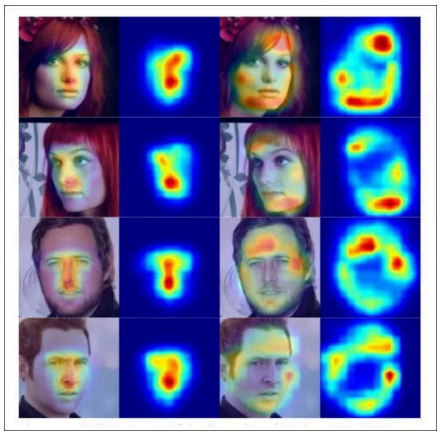
The authors state: *'This result suggests that depth in this case adds a more relevant contribution to classification than color artifacts.'*

The findings represent part of a new wave of deepfake detection research directed against real-time facial synthesis systems such as DeepFaceLive – a locus of effort that has accelerated notably in the last 3-4 months, in the wake of the FBI's warning in March about the risk of real-time video and audio deepfakes.

The paper is titled *DepthFake: a depth-based strategy for detecting Deepfake videos*, and comes from five researchers at the Sapienza University of Rome.

Edge Cases

During training, autoencoder-based deepfake models prioritize the inner regions of the face, such as eyes, nose and mouth. In most cases, across open source distributions such as DeepFaceLab and FaceSwap (both forked from the original 2017 Reddit code prior to its deletion), the outer lineaments of the face do not become well-defined until a very late stage in training, and are unlikely to match the quality of synthesis in the inner face area.



From a previous study, we see a visualization of 'saliency maps' of the face. Source: https://arxiv.org/pdf/2203.01318.pdf

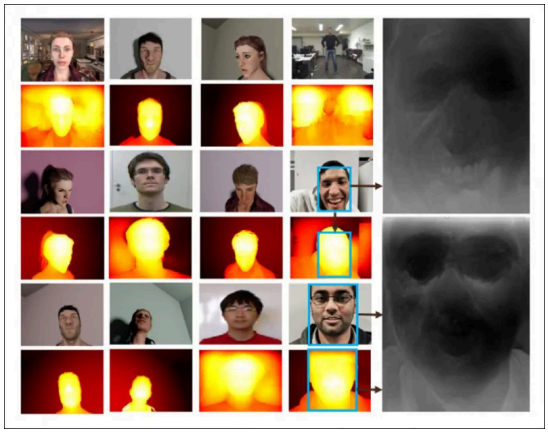
Normally, this is not important, since our tendency to focus first on eyes and prioritize, 'outwards' at diminishing levels of attention means that we are unlikely to be perturbed by these drops in peripheral quality – most especially if we are talking live to the person who is faking another identity, which triggers social conventions and processing limitations not present when we evaluate 'rendered' deepfake footage.

However, the lack of detail or accuracy in the affected margin regions of a deepfaked face can be detected algorithmically. In March, a system that keys on the peripheral face area was announced. However, since it requires an above-average amount of training data, it's only intended for celebrities who are likely to feature in popular facial datasets (such as ImageNet) that have provenance in current computer vision and deepfake detection techniques.

Instead, the new system, titled *DepthFake*, can operate generically even on obscure or unknown identities, by distinguishing the quality of estimated depth map information in real and fake video content.

Going Deep

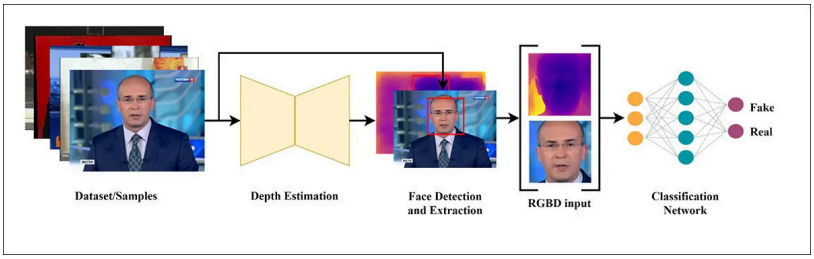
Depth map information is increasingly being baked into smartphones, including AI-assisted stereo implementations that are particularly useful for computer vision studies. In the new study, the authors have used the National University of Ireland's FaceDepth model, a convolutional encoder/decoder network which can efficiently estimate depth maps from single-source images.



The FaceDepth model in action. Source: https://tinyurl.com/3ctcazma

Next, the pipeline for the Italian researchers' new framework extracts a 224×224 pixel patch of the subject's face from both the original RGB image and the derived depth map. Critically, this allows the process to copy over core content without resizing it; this is important, as size standard resizing algorithms will adversely affect the quality of the targeted areas.

Using this information, from both real and deepfaked sources, the researchers then trained a convolutional neural network (CNN) capable of distinguishing real from faked instances, based on the differences between the perceptual quality of the respective depth maps.



Conceptual pipeline for DepthFake.

The FaceDepth model is trained on realistic and synthetic data using a hybrid function that offers greater detail at the outer margins of the face, making it well-suited for the DepthFake. It uses a MobileNet instance as a feature extractor, and was trained with 480×640 input images outputting 240×320 depth maps. Each depth map represents a quarter of the four input channels used in the new project's discriminator.

The depth map is automatically embedded into the original RGB image to provide the kind of RGBD image, replete with depth information, that modern smartphone cameras can output.

Training

The model was trained on an Xception network already pretrained on ImageNet, though the architecture needed some adaptation in order to accommodate the additional depth information while maintaining the correct initialization of weights.

Additionally, a mismatch in value ranges between the depth information and what the network is expecting necessitated that the researchers normalized the values to 0-255.

During training, only flipping and rotation was applied. In many cases various other visual perturbations would be presented to the model in order to develop robust inference, but the necessity to preserve the limited and very fragile edge depth map information in the source photos forced the researchers to adopt a pare-down regime.

The system was additionally trained on simple 2-channel grayscale, in order to determine how complex the source images needed to be in order to obtain a workable algorithm.

Training took place via the TensorFlow API on a NVIDIA GTX 1080 with 8GB of VRAM, using the ADAMAX optimizer, for 25 epochs, at a batch size of 32. Input resolution was fixed at 224×224 during cropping, and face detection and extraction was accomplished with the dlib C++ library.

Results

Accuracy of results was tested against Deepfake, Face2Face, FaceSwap, Neural Texture, and the full dataset with RGB and RGBD inputs, using the FaceForensic++ framework.

	ResNet50		MobileNet–V1		XceptioNet	
	RGB	RGBD	RGB	RGBD	RGB	RGBD
DF	93.91%	**94.71%** (+0.8)	95.14%	**95.86%** (+0.72)	97.65%	**97.76%** (+0.11)
F2F	96.42%	**96.58%** (+0.16)	97.07%	**98.44%** (+1.37)	95.82%	**97.41%** (+1.59)
FS	97.14%	**96.95%** (−0.19)	97.12%	**97.87%** (+0.75)	97.84%	**98.80%** (+0.96)
NT	75.81%	**77.42%** (+1.61)	70.47%	**82.26%** (+11.79)	76.87%	**85.09%** (+8.22)
FULL	85.68%	**90.48%** (+4.80)	90.60%	**91.12%** (+0.52)	86.80%	**91.93%** (+5.13)

Results on accuracy over four deepfake methods, and against the entire unsplit dataset. The results are split between analysis of source RGB images, and the same images with an embedded inferred depth-map. Best results are in bold, with percentage figures underneath demonstrating the extent to which the depth map information improves the outcome.

In all cases, the depth channel improves the model's performance across all configurations. Xception obtains the best results, with the nimble MobileNet close behind. On this, the authors comment:

> *'[It] is interesting to note that the MobileNet is slightly inferior to the Xception and outperforms the deeper ResNet50. This is a notable result when considering the goal of reducing inference times for real-time applications. While this is not the main contribution of this work, we still consider it an encouraging result for future developments.'*

The researchers also note a consistent advantage of RGBD and 2-channel grayscale input over RGB and straight grayscale input, observing that the grayscale conversions of depth inferences, which are computationally very cheap, allow the model to obtain improved results with very limited local resources, facilitating the future development of real-time deepfake detection based on depth information.

First published 24th August 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