Detecting Real-Time Deep-Fake Videos Using Active Illumination

Candice R. Gerstner
National Security Agency
Fort George G. Meade MD, USA
crgerst@uwe.nsa.gov

Hany Farid
University of California, Berkeley
Berkeley CA, USA
hfarid@berkeley.edu

Abstract

While many have grown suspicious of viral images and videos found online, there is a general sense that we can and should trust that the person on the other end of our video-conferencing call is who it purports to be. The real-time creation of sophisticated deep fakes, however, is making it more difficult to trust even live video calls. Detecting deep fakes in real time introduces new challenges as compared to off-line forensic analyses. We describe a technique for detecting, in real-time, deep-fake videos transmitted over a live video-conferencing application. This technique leverages the fact that a video call typically places a user in front of a light source (the computer display) which can be manipulated to induce a controlled change in the appearance of the user's face. Deviations of the expected change in appearance over time can be measured in real time and used to verify the authenticity of a video-call participant.

1. Introduction

In early 2020, a United Arab Emirates' bank was swindled out of $35 million (USD). The bank teller was convinced to transfer the funds after receiving a phone call from the purported director of a company whom the bank manager knew and with whom he had previously done business. The voice on the other end of the phone instructed the bank manager to transfer the funds as part of a corporate acquisition. Because the request was consistent with previously received emails describing the acquisition, and because the purported director's voice was familiar to him, the bank manager transferred the funds. It was later revealed that the voice was that of an AI-synthesized voice made to sound like the director.

This was not the first time AI-synthesized content was used to steal large sums of money. In 2019, a United Kingdom based company suffered a similar fate when an imposter used an AI-synthesized voice to steal $243,000 (USD) in a similar type of scam.

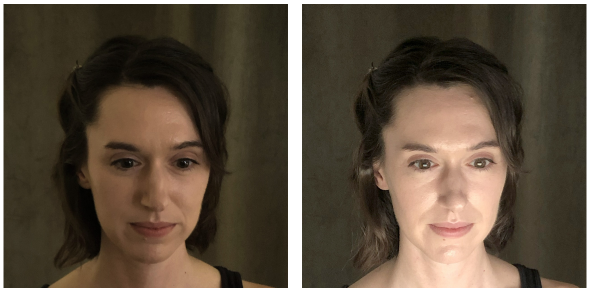

Figure 1. A computer display acts as an area light source that can be controlled in real time to induce an authenticating pattern on the face of a video-call participant. This simple example shows the impact of switching between viewing a dark (left) and bright (right) browser window.

These two incidents are almost certainly the canaries in the coal mine. As AI-synthesized audio and video continue to improve in quality and accessibility, it is reasonable to predict that these technologies will continue to be used to commit a range of small- to large-scale frauds, among other potentially nefarious uses.

All forms of deep fakes pose potential threats from non-consensual sexual imagery to fraud, and disinformation campaigns. The creation of real-time deep fakes, however, poses unique threats because of the general sense of trust surrounding a live video or phone call, and the challenge of detecting deep fakes in real time, as a call is unfolding.

Over the past two pandemic years, we have grown accustomed to video calls replacing previously in-person meetings and phone calls. Although not yet perfected, deep fakes can be synthesized in real time and piped through a virtual camera (e.g., github.com/alievk/avatarify-python and github.com/iperov/DeepFacelive), meaning that it will become increasingly more difficult to distinguish a real person from an AI-synthesized person at the other end of a video call.

One approach to detecting deep-fake video calls is to employ any of a plethora of passive deep-fake forensic techniques (see Section 2). Most of these approaches, however, struggle to run in real time, and most struggle to achieve the levels of accuracy that would be needed to be incorporated into a video-conferencing application.

In contrast to passive forensic techniques, active forensic techniques (a.k.a., control-capture [2, 25]) focus on authenticating content at the point of recording. By extracting a compact, digital signature at the point of recording and packaging this signature alongside the recording, audio, images, and videos can be efficiently and accurately authenticated.

Motivated by the reliability of active forensic approaches and the unique constrained environment afforded by a video-conferencing call, we describe an active approach for detecting real-time, deep-fake video calls. In particular, instead of explicitly trying to distinguish an authentic video from a deep-fake video, we authenticate videos by projecting a distinct illumination pattern onto the face of each call participant. This pattern can be induced by a call participant displaying the temporally varying pattern on a shared screen, or directly integrated into the video-call client. In either case, no specialized imaging or lighting hardware is required.

Through large-scale simulations, we evaluate the reliability of this approach under a range of imaging scenarios, and validate this approach in a variety of real-world settings. We begin by framing our technique within previous work.

2. Related Work

We provide an overview on the state of the art in creating and detecting deep-fake videos, with an emphasis on real-time deep fakes.

2.1. Creation

Since they splashed onto the scene in 2017 with full force (an earlier incarnation dates to a decade earlier [9]), AI-synthesized content – so-called deep fakes – have continued their rapid trajectory of increased sophistication, realism [26], and accessibility. This includes images of fully fabricated people [14, 15], audio recordings mimicking another voice [29], and videos of people saying anything the creator wants them to say [35].

Within this broad range of different types of AI-synthesized content, so-called puppet-master deep fakes (e.g., [13, 32]) are particularly intriguing for their power to create a deep fake in real time from a single source image. In particular, starting with a single image of a person (the puppet), and a recorded or live video of another person (the puppet master), a video of the puppet is synthesized to mimic the expressions, mouth movement, and head movements of the puppet master. The resulting synthesized video can then be piped into a live video call through a virtual camera. Although not yet perfected, this type of puppet-master deep fake holds the potential to deceive someone into believing they are talking with anyone that a fraudster wants to impersonate [7]. In contrast, this same technology holds the potential to significantly reduce the bandwidth necessary for a video call [1].

As compared to a puppet-master deep fake, a face-swap deep fake (github.com/deepfakes/faceswap, github.com/shaoanlu/faceswap-GAN) replaces the face – from eyebrows to chin and cheek to cheek – of the impersonator with that of another. The viral deep-fake Tom Cruise videos (www.tiktok.com/@deeptomcruise) are a particularly compelling example of this type of deep fake. While the creation of deep Tom Cruise is the result of a talented impersonator and a highly-skilled, special-effects artist, open-source software for creating real-time, face-swap deep fakes are emerging on the scene. DeepFaceLive (github.com/iperov/DeepFacelive), for example, allows the creator to swap their face with a celebrity in real time and, according to their documentation, also incorporates a color transfer that maps the creator's environmental lighting onto the deep fake.

2.2. Detection

Forensic techniques for detecting deep fakes can be broadly categorized into low- and high-level approaches. Low-level techniques detect pixel-level, synthesis artifacts, ranging from general artifacts [23, 36, 38–40], to warping artifacts [21], and blending artifacts [19]. High-level techniques focus on semantically meaningful features, including inconsistencies in eye blinks [20], head-pose [37], physiological signals [10], mouth shape and movement [5], and distinct mannerisms [3, 6].

While some of these techniques might be applicable to detecting real-time deep-fake videos, in our view, a class of particularly promising approaches takes advantage of the unique physical constraints of a live video call: call participants are in front of a light source (the computer display) that can be actively adjusted in real time to induce specific lighting patterns on a user's face. The consistency of a call participant's appearance under this induced lighting can then be used to verify their liveness and physical presence in front of the camera. We posit that this approach will be effective because either the deep-fake video simply fails to transfer the active illumination, or there is a temporal delay in transferring the active illumination. We will show that both of these scenarios are easily detected.

Exploiting the computer display as an active light source has previously been leveraged as an inexpensive light stage in which, after recording a user's appearance under a time-varying illumination pattern, her face can be synthetically re-lit under an arbitrary lighting environment [30]. This type of active approach has also previously been explored particularly for the purpose of thwarting playback or re-broadcast attacks [4]. FaceRevelio [11], for example, uses

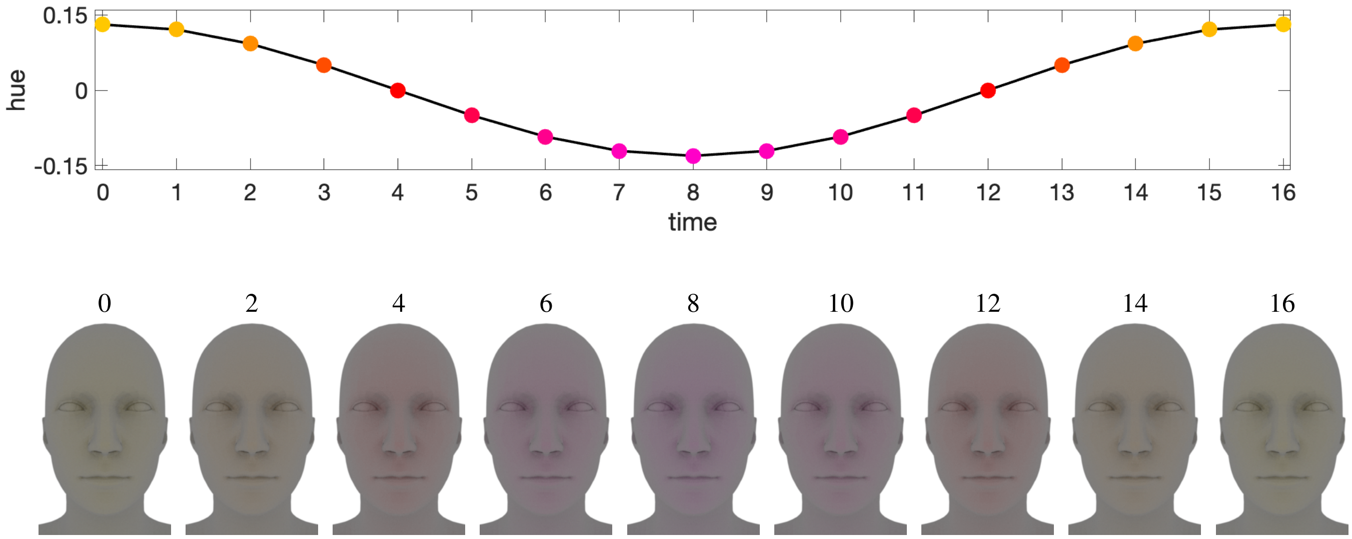

Figure 2. Shown in the top panel is a visualization of the dynamic change in the hue of a uniform-colored area light source (simulating a computer screen). Shown below are nine renderings of a 3-D model illuminated with a different light-source hue at nine distinct moments in time. In this simulation, the face has a monochromatic reflectance and there is a 1:1 ratio of area- to ambient-light intensity (for the purposes of visualization, the image saturation was boosted by 50%).

a smartphone screen to illuminate a user's face from multiple directions from which a 3D facial model is constructed. Here the active illumination is a means to an end to construct a 3D model with the goal of distinguishing a real person from a 2D rebroadcast attack version.

LiveScreen [22], uses a device's display to induce an inconspicuous lighting pattern on a user's face. The liveness detection system then tracks the weak facial appearance changes with the goal of thwarting rebroadcast attacks. Although this system, running on a laptop, achieves reasonable detection accuracy (94.8%) with relatively low false detection (1.6%), their focus on an inconspicuous lighting pattern makes detection of the active illumination challenging, particularly in an otherwise well-illuminated environment.

Whereas these earlier and related works focus on rebroadcast attacks, the work of Shang and Wu [31] focuses specifically on detecting deep-fake videos. In this work, the authors place a user in front of a 27-inch display which flashes between white and black at 0.2 Hz. The detection system then measures the correlation between this active illumination pattern and the brightness of the facial appearance.

Our system follows a similar structure to [31], but with some important differences that make authentication of the active illumination pattern more robust. Because all modern webcams perform auto exposure, the type of high intensity active illumination of [31] is likely to trigger the camera's auto exposure which in turn will confound the recorded facial appearance. To avoid this, we employ an active illumination consisting of an isoluminant change in hue. While this avoids the camera's auto exposure, it could trigger the camera's white balancing which would again confound the recorded facial appearance. To avoid this, we operate in a hue range that we empirically determined does not trigger white balancing. We show that this choice of active illumination affords a particularly simple mechanism for separating the impact of the active illumination from the surrounding environmental lighting. We show the efficacy of our approach in large-scale simulations under a range of imaging configurations, and in a range of different real-world configurations. We also evaluate the robustness of our system to expected adversarial attacks.

3. Methods

We describe the underlying methodology for generating the active illumination, localizing, and measuring the pattern of illumination on a face, and determining the consistency between this measured and the expected illumination.

3.1. Active Illumination

An active illumination source is achieved by displaying a fixed-size image on the same screen as the video call. As shown in Figure 2, the hue, $H(t)$, of a uniform-color image is shifted over time, t – and can be synchronized with the display frame-rate – as follows:

$$H(t) = 0.1307 \times \cos(t/8), \qquad (1)$$

where, $t \in [0, 16]$, yielding a hue value in the range of 0.1307 (yellow-ish) to -0.1307 (magenta-ish). Because hue is circular, a negative hue of $-h$ is the same as $1 - h$. The light-source hue is modulated sinusoidally to avoid abrupt and distracting changes. And, the hue is constrained to the range of yellow to magenta because we found that these hues, unlike the blues and greens, do not induce an automatic white balancing found in some video conferencing.

With this change in hue over time, the value, V, and saturation, S, of the light source are fixed at unit value, resulting in an isoluminant change in color in which the brightness does not change over time. For purposes of rendering, the specified HSV is converted to RGB using a standard conversion (Python's `colorsys.hsv_to_rgb`).

3.2. Face Detection

A face is automatically localized in each video frame using Dlib [17], yielding both a bounding box and 68 keypoints delineating the facial features and facial outline. A bounding ellipse – parameterized by a center, two scale factors along the major and minor axes, and an orientation – is fitted to 4 facial keypoints on the bridge of the nose, the base of the chin, and each cheek bone.

3.3. Source Separation

A person sitting in front of a computer display is illuminated by the surrounding environmental lighting and our active illumination. Because we are only interested in the impact of the active illumination, we next describe how to separate the contributions of these two illumination sources.

A Lambertian surface with surface normal $\vec{N}$ and surface reflectance α_s, illuminated with a single distant point light source with orientation $\vec{L}$ and color α_l, will be imaged as $I_k = \alpha_s \alpha_l (\vec{N}_k^T \cdot \vec{L})$, where the color I_k at pixel k and α_s and α_l are each specified as a triple of RGB values. Generally speaking, separating the contribution of the surface reflectance and lighting is a difficult problem [18]. If, however, either the reflectance or lighting terms are known, the other quantity can be trivially estimated by dividing the measured image by the known quantity.

In our case, we make the simplifying assumption that the face – sans active illumination – is illuminated with a non-directional white light. The addition of the active illumination, modeled as an area light source with constant color α_a, yields the appearance model:

$$I_k = \hat{\alpha_s} \alpha_a \sum_{\omega \in \Omega} \vec{N}_k^T \cdot \vec{L}_\omega, \tag{2}$$

where $\hat{\alpha_s}$ is a scaled – by the surrounding illumination – version of the underlying facial reflectance α_s, and where the summation is performed over the area Ω of the active light source. The summation term consists of a monochromatic multiplicative factor, and can therefore be ignored because we will only measure the active illumination hue which itself is invariant to an overall scale factor. To separate the contribution of the scaled facial reflectance $\hat{\alpha}_s$ and the active illumination α_a, an estimate of $\hat{\alpha}_s$ is acquired by measuring the average color of the face before the active illumination sequence begins. Once illuminated by the active illumination, the measured color I_k at facial pixel k is divided by the measured quantity $\hat{\alpha}_s$ to yield the desired hue of the active illumination α_a.

3.4. Measurement

After extracting the face from a video frame, the RGB value of each facial pixel is divided by the average facial RGB pixel value measured with no active illumination. This, as described in the previous section, extracts the facial reflectance (assumed to be constant across the face). Each of these adjusted facial RGB pixel values is then converted to HSV (using Python's `colorsys.rgb_to_hsv`), and the facial hues H are averaged to yield an estimate of the hue of the active illumination.

For simplicity, we assume that the person of interest does not have bangs covering their forehead, facial hair, or eye glasses, each of which violate our assumption of a constant facial reflectance function. A more sophisticated facial segmentation could eventually be deployed that isolates the facial pixels with a constant reflectance.

Because hue, specified in the range $[0, 1]$, is circular (i.e., a hue value of 0 corresponds to the same color as a hue value of 1), we must account for this circularity when computing the mean hue across the face. The circular mean hue, $\tilde{H}(t)$, at time t, from n hue values, $h_i(t)$, corresponding to pixel i in the detected face, is:

$$\tilde{H}(t) = \tfrac{1}{2\pi} \operatorname{atan2} \left(\sum_{i=1}^{n} \sin(2\pi h_i(t)), \sum_{i=1}^{n} \cos(2\pi h_i(t)) \right). \tag{3}$$

3.5. Comparison

The difference between the expected facial hue $H(t)$, Equation (1), and the measured facial hue $\tilde{H}(t)$, Equation (3), can be quantified with the Pearson correlation coefficient, where a maximum correlation of 1 corresponds to perfect correlation, and a value 0 corresponds to a lack of correlation. As we saw above, however, the measured hue is circular, so we perform this correlation on the unit circle [8].

Because the facial detection, hue computation, and circular correlation are each computationally efficient, at a standard video frame rate of 30 Hz, a short 30-frame illumination pattern can be validated with as little as a one-second delay in the live video stream.

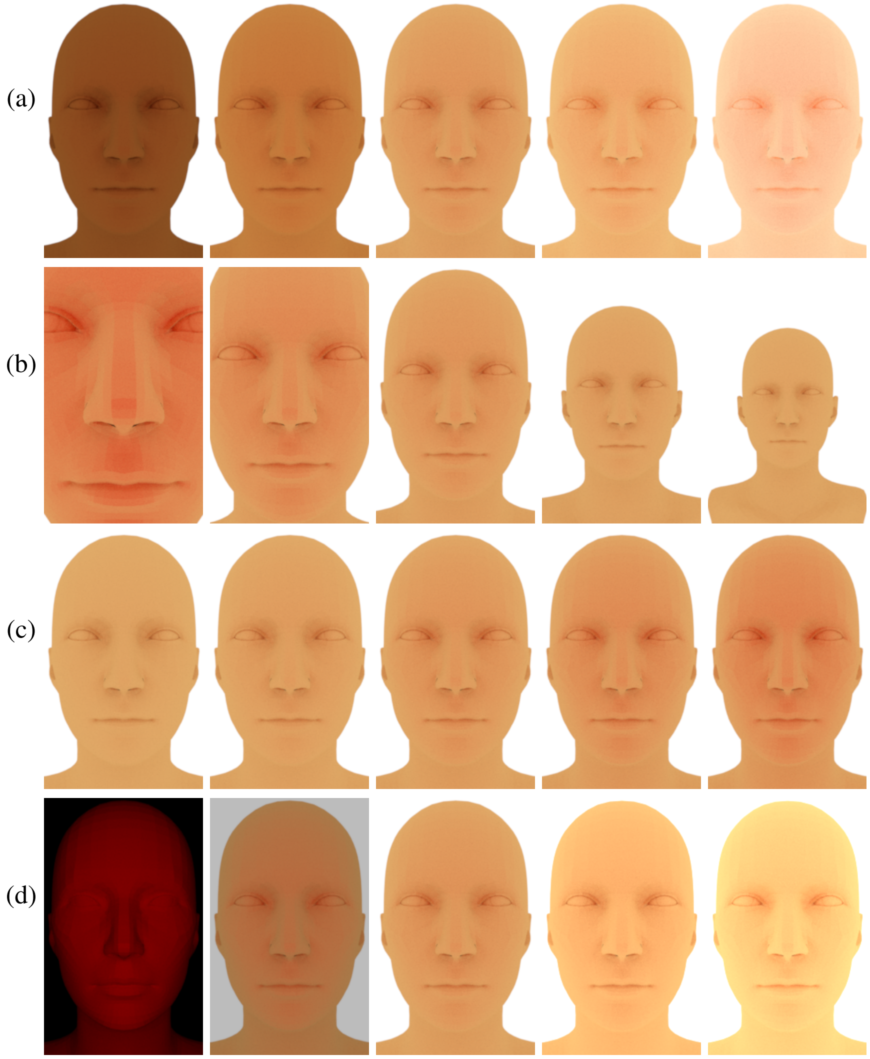

Figure 3. A representative sample of our simulated data set with varying: (a) skin tone; (b) head to camera distance (increasing from left to right); (c) size of active light source (increasing from left to right); and (d) intensity of ambient light (increasing from left to right).

3.6. Counter Measures

The effectiveness and robustness of our approach hinges on the real-time (30 Hz) generation of an active illumination pattern, and the assumption that the synthesis engine will either not transfer the illumination onto the deep-fake face, or will have a temporal delay in transferring the illumination.

If, as in our case, the active illumination is deterministic, then an adversary could easily predict the illumination pattern and add it to the generated deep-fake video without a temporal delay. A simple way to avoid this adversarial attack is to randomly interject blank frames in the temporal illumination sequence. An added benefit of this counter measure is that the baseline reflectance can be reestimated during these moments.

4. Results

We evaluate the efficacy of our technique on two data sets. The first simulated data set allows us to evaluate our technique across a broad range of assumptions and environmental conditions, while the second real-world data set validates our technique in realistic and variable environments.

4.1. Simulation

This data set is created using the physically-based renderer Mitsuba [27]. The basic scene geometry consists of a camera with a 90° field of view, a 3-D head with a Lambertian reflectance and neutral skin-tone and with an ear-to-ear distance of 6 in, placed 2 ft directly in front of the camera. This scene is illuminated with a unit-value ambient light source and an area light of size 9×9 in with unit intensity, placed alongside the camera. The area light simulates the eventual implementation of displaying an image on the computer display.

To evaluate our ability to measure the active illumination under a range of environmental conditions, we rendered this basic scene geometry varying, one at a time, the skin tone, Figure 3(a), a head to camera distance of $16, 20, 24, 30, 36$ in, Figure 3(b), an illumination light size of 5×5, 7×7, 9×9, 11×11, 13×13 in, Figure 3(c), and an ambient intensity of 0.0, 0.5, 1.0, 1.5, 2.0, Figure 3(d).

Across all of these imaging parameters, the average correlation between the measured hue and the induced light hue is 0.99 and a minimum correlation of 0.988. In simulation, with our various assumptions satisfied, our proposed technique is highly robust to a broad range of imaging configurations. We next evaluate the robustness in different real-world settings.

4.2. Real world

The real-world data set was recorded from 15 users with a range of skin tones and in a range of different environments. Users were placed approximately 24 inches away from the display and camera, with the active illumination ranging in size from 13×13 to 3×3 in (in steps of 2 in). The active illumination consisted of two cycles of the hue pattern shown in Figure 2 (i.e. with a 30 Hz display refresh rate synchronized with the camera, the entire active illumination pattern would only be visible for one second).

Shown in the top panel of Figure 4 is the correlation of the measured facial hue and the active-illumination hue. At the largest illumination size of 13 in, the average correlation is 0.93. As the illumination reduces in size from 11 to 3 in, the average correlation decreases from 0.92, to 0.85, 0.83, 0.68, and 0.33.

By way of comparison, we also measured the correlation between the facial appearance in the absence of an illumination pattern – as might occur if the deep-fake synthesis does not transfer the lighting environment of the imposter. Across all 15 users, the average correlation in this baseline conditions is 0.09 with a variance of 0.01, and a maximum correlation of 0.34. By comparison, for a light source of size greater than 5×5 in, the vast majority of correlations are greater than 0.5.

4.3. Adversarial Attack

We saw in the previous section that if a deep-fake creator fails to transfer the environmental lighting onto a deep-fake video, then the resulting synthesized video will be easily detected. If, on the other hand, the creator measures the environmental lighting in real time and transfers this into the deep-fake video, then detection may be more difficult. Assuming that there will be some temporal delay from the moment when the environmental lighting is measured and the frame-by-frame video synthesis, the effectiveness of our defense depends on the impact of a temporal phase shift on the hue correlation in the face. Shown in the bottom panel of Figure 4 is the correlation of the measured facial hue and the

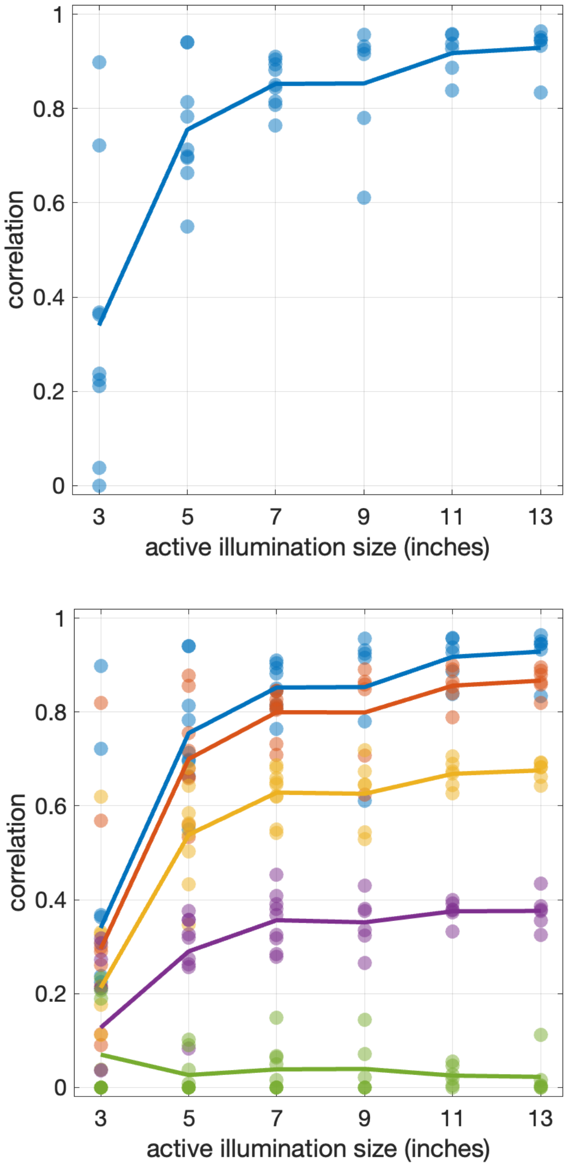

Figure 4. Shown in the upper panel is the correlation of the measured facial hue and the active-illumination hue from 15 different users and six different light sizes. As expected, the correlation is stronger for the larger light sources. Shown in the lower panel is the same correlation if the measured hue was shifted by zero (blue), one (red), two (yelllow), three (purple), or four (green) frames. In each panel, the solid line corresponds to the average correlation across all users.

active-illumination hue from our 15 users, where now there is a temporal shift of 1 to 4 frames between the induced active illumination and the measured facial hue.

For the four largest illumination sizes (13, 11, 9, and 7 in), the average correlation is 0.89 when the facial and illumination hue are synchronized. By comparison, the average correlation is 0.83 for a one-frame delay, 0.65 for a two-frame delay, 0.37 for a three-frame delay, and 0.03 for a four-frame delay. This rapid loss of correlation means that, at an assumed frame rate of 30 frame/sec, a deep fake will be detectable if there is a delay of more than $2/30$-th of a second in the synthesis, at which point the correlation slips well below 0.5.

We verified that the deep fakes created by Avatarify (github.com/alievk/avatarify-python) do not incorporate the environmental lighting and are therefore easily identifiable because in the presence of our active illumination, their temporal facial hue is flatlined with a nearly zero correlation.

5. Discussion

Although the creation of artifact-free, real-time deep fakes are not yet upon us, it is reasonable to predict that they soon will be. While standard passive forensic techniques are at their best when they assume as little as possible about the imaging hardware and environment, detection of a deep fake, video-conference participant poses a unique opportunity to exploit the typical imaging configuration in which call participants are sitting in front of a computer display (i.e., controllable light source).

By displaying a simple, dynamic, colored square on the display and then measuring the temporal impact on the call participant's face, we have exploited just one aspect of this unique imaging configuration. In particular, we have only considered the impact of the lighting on the 2-D facial appearance. A more sophisticated 3-D estimation of lighting [16] would likely provide a richer appearance model which would be even more difficult for a forger to circumvent. While we focused only on the face, the computer display also illuminates the neck, upper body, and surrounding background, from which similar measurements could be made. These additional measurements would force the forger to consider the entire 3-D scene, not just the face. Similarly, because of the proximity of the call participant to the display, and the high-resolution of most webcams, it might be possible to make even more fine-grained measurements of the color and shape of the display reflected in the participant's eyes [28]. This again would make circumvention even more difficult.

Our proposed intervention could either be realized by a call participant who simply shares her screen and displays the temporally varying pattern, or, ideally, it could be directly integrated into the video-call client. Any real-time system would need to ensure the camera frame rate and display refresh rate are synchronized or any delay be calibrated and adjusted for.

We have assumed that the call participant's face is not obscured by, for example, bangs, facial hair, or glasses. In order to be more broadly applicable, our approach would benefit from automatically segmenting the face into regions of uniform reflectance, from which the required hue measurements can be made.

Beyond the visual, if the deep-fake synthesis includes a synthetic voice, then an audio correlate to our active illumination pattern may be used to determine if the voice is being directly recorded. However, while it may not be overly distracting to show a glowing square on the display for a short period of time, an – even occasional – audible sound may be prohibitively distracting. An active auditory signal can, however, be played in the ultrasonic range, outside of the range of the human auditory system [12].

Because of the reasonable trust we place on live video calls, and the growing ubiquity of video calls in our personal and professional lives, we propose that techniques for authenticating video (and audio) calls will only grow in importance. This more narrow forensic application is, therefore, worthy of increased attention from the media-forensics community.

References

[1] NVIDIA MAXINE. https://developer.nvidia.com/maxine. 2

[2] The Coalition for Content Provenance and Authenticity (C2PA). https://c2pa.org. 2

[3] Shruti Agarwal, Tarek El-Gaaly, Hany Farid, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In *IEEE Workshop on Image Forensics and Security*, 2020. 2

[4] Shruti Agarwal, Wei Fan, and Hany Farid. A diverse large-scale dataset for evaluating rebroadcast attacks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1997–2001. IEEE, 2018. 2

[5] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 660–661, 2020. 2

[6] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–45, 2019. 2

[7] Ali Aliev. Avatarify python. https://github.com/alievk/avatarify-python, 2021. 2

[8] Philipp Berens. CircStat: a MATLAB toolbox for circular statistics. *Journal of Statistical Software*, 31:1–21, 2009. 4

[9] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *24th*

Annual Conference on Computer Graphics and Interactive Techniques, pages 353–360, 1997. 2

[10] Umur Aybars Ciftci and Ilke Demir. FakeCatcher: Detection of synthetic portrait videos using biological signals. arXiv: 1901.02212, 2019. 2

[11] Habiba Farrukh, Reham Mohamed Aburas, Siyuan Cao, and He Wang. FaceRevelio: a face liveness detection system for smartphones with a single front camera. In *Annual International Conference on Mobile Computing and Networking*, pages 1–13, 2020. 2

[12] Michael Hanspach and Michael Goetz. On covert acoustical mesh networks in air. arXiv: 1406.1213, 2014. 7

[13] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics*, 36(6):1–14, 2017. 2

[14] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 2

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. arXiv:1912.04958, 2019. 2

[16] Eric Kee and Hany Farid. Exposing digital forgeries from 3-D lighting environments. In *IEEE International Workshop on Information Forensics and Security*, pages 1–6, 2010. 7

[17] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 2009. 4

[18] Edwin H Land and John J McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61(1):1–11, 1971. 4

[19] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-ray for more general face forgery detection. arXiv: 1912.13458, 2019. 2

[20] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security*, pages 1–7, 2018. 2

[21] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. arXiv: 1811.00656, 2018. 2

[22] Hongbo Liu, Zhihua Li, Yucheng Xie, Ruizhe Jiang, Yan Wang, Xiaonan Guo, and Yingying Chen. LiveScreen: Video chat liveness detection leveraging skin reflection. In *IEEE Conference on Computer Communications*, pages 1083–1092, 2020. 3

[23] Vineet Mehta, Parul Gupta, Ramanathan Subramanian, and Abhinav Dhall. FakeBuster: A deepfakes detection tool for video conferencing scenarios. arXiv: 2101.03321, 2021. 2

[24] Vineet Mehta, Parul Gupta, Ramanathan Subramanian, and Abhinav Dhall. Fakebuster: a deepfakes detection tool for video conferencing scenarios. In *26th International Conference on Intelligent User Interfaces-Companion*, pages 61–63, 2021. 3

[25] Assa Naveh and Eran Tromer. PhotoProof: Cryptographic image authentication for any set of permissible transformations. In *Symposium on Security and Privacy*, pages 255–271, 2016. 2

[26] Sophie J Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8), 2022. 2

[27] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. Mitsuba 2: A retargetable forward and inverse renderer. *Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 38(6), Dec. 2019. 5

[28] Ko Nishino and Shree K Nayar. The world in an eye. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I, 2004. 7

[29] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. arXiv: 1609.03499, 2016. 2

[30] Soumyadip Sengupta, Brian Curless, Ira Kemelmacher-Shlizerman, and Steven M Seitz. A light stage on every desk. In *IEEE/CVF International Conference on Computer Vision*, pages 2420–2429, 2021. 2

[31] Jiacheng Shang and Jie Wu. Protecting real-time video chat against fake facial videos generated by face reenactment. In *International Conference on Distributed Computing Systems*, pages 689–699, 2020. 3

[32] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[33] Jeremy Speth, Nathan Vance, Patrick Flynn, Kevin W. Bowyer, and dam Czajka. Digital and physical-world attacks on remote pulse detection. arXiv: 2110.11525, 2021. 5

[34] Di Tang, Zhe Zhou, Yinqian Zhang, and Kehuan Zhang. Face flashing: a secure liveness detection protocol based on light reflections. arXiv:1801.01949, 2018. 3

[35] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. DeepFakes and beyond: A survey of face manipulation and fake detection. arXiv: 2001.00179, 2020. 2

[36] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot...for now. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[37] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8261–8265, 2019. 2

[38] Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *IEEE International Conference on Computer Vision*, 2018. 2

[39] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. arXiv: 1907.06515, 2019. 2

[40] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Two-stream neural networks for tampered face detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 2