ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094720

DETECTION OF REAL-TIME DEEPFAKES IN VIDEO CONFERENCING WITH ACTIVE PROBING AND CORNEAL REFLECTION

Hui Guo, Xin Wang, Siwei Lyu

Department of Computer Science and Engineering
University at Buffalo, State University of New York, USA.
{hguo8,xwang264,siweilyu}@buffalo.edu

ABSTRACT

The COVID pandemic has led to the wide adoption of online video calls in recent years. However, the increasing reliance on video calls provides opportunities for new impersonation attacks by fraudsters using the advanced real-time DeepFakes. Real-time DeepFakes pose new challenges to detection methods, which have to run in real-time as a video call is ongoing. In this paper, we describe a new active forensic method to detect real-time DeepFakes. Specifically, we authenticate video calls by displaying a distinct pattern on the screen and using the corneal reflection extracted from the images of the call participant's face. This pattern can be induced by a call participant displaying 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 in a variety of real-world imaging scenarios.

Index Terms— Real-time DeepFake, Corneal Reflection

1. INTRODUCTION

Video calls have been increasingly replacing in-person meetings and phone calls in recent years, mainly due to the high demand for remote working during the COVID pandemic. For instance, at the end of 2019, the Zoom video conferencing platform had only about 10 million users. By late April of 2021, that figure had surged to over 200 million, a 20-fold increase. However, the wide adoption of video calls as a means of meeting and inter-person communication has also given rise to new forms of deception. In particular, the lack of physical presence opens the gate for digital impersonation in video calls using DeepFakes (*i.e.*, AI-synthesized human face videos). The most recent tools (*e.g.*, Avatarify [1] and DeepFaceLive [2]) have enabled the synthesis of DeepFakes in real-time and piped through a virtual camera. The DeepFakes are either in the form of face-swap or face puppetry [3]. Although there are still artifacts in the real-time DeepFakes [4], the continuing improvement of the synthesis technology means that it will become increasingly difficult to distinguish a real person from an AI-synthesized person at the other end of a video call. Indeed, recent years have seen such frauds emerge at an alarming speed and start causing real damage [5].

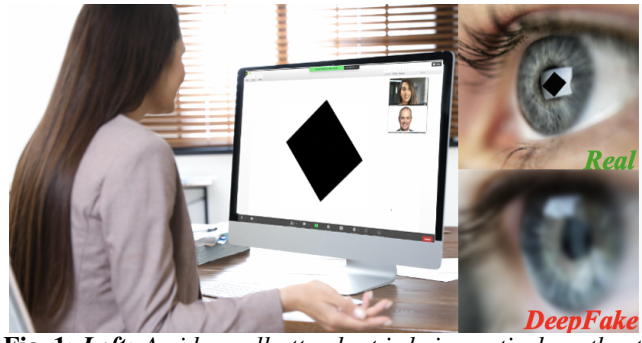

Fig. 1: ***Left:*** *A video call attendant is being actively authenticated with the live patterns shown on the screen.* ***Right:*** *A real person's cornea will produce an image of the pattern shown on the screen, while a real-time DeepFake cannot. The figures are for demonstration. For actual results, see Fig. 5.*

The real-time DeepFakes pose new challenges to existing detection methods, which are mostly *passive*, in that they classify an input video into the category of authentic or DeepFake. Most of these methods struggle to achieve the levels of accuracy that would be needed to be incorporated into a practical video-conferencing application and run in real-time. On the other hand, new approaches based on *active* forensics, which interfere with the generation process to improve the detection efficiency and accuracy, *e.g.*, [6, 7], are gaining momentum recently. In particular, the work of [7] exploits the unique constrained environment afforded by a video-conferencing call to detect real-time DeepFakes by varying the lighting pattern on the screen and extracting the same lighting variation from the attendant's face. As the current real-time DeepFake synthesis methods are not sufficiently adaptable to capture such subtle changes, the lack of consistent lighting variation can be used as a telltale sign of synthesis and impersonation. However, controlling and estimating the subtle change of screen lighting may not be reliable as it can be affected by other environmental factors, such as the ambient light, room setting, and makeup.

In this work, we describe a new active forensic approach to exposing real-time DeepFakes. The main idea is illustrated in Fig. 1. This method can be initiated by a call participant or

directly integrated into the video-call client[1]. First, we briefly display a distinct pattern, which will be referred to as the *probing pattern*, on the shared screen during an ongoing video call. The image of the attendant's face will be captured by the camera, and we will focus on the cornea areas. As the attendant sits before the camera in a video call and the human cornea is mirror-like and highly reflective, the probing pattern on the screen should leave a reflective image on the cornea that can subsequently be extracted from the face image and compared with the probing pattern. We provide an automatic pipeline to display the probing pattern, capture the face image, extract the cornea reflections, and compare them with the original probing pattern. Our experiments with several state-of-the-art real-time DeepFake synthesis models show that they cannot recreate the probing pattern at the synthesized cornea region at all in a variety of real-world settings. Compared with the work in [7], our active detection method is less limited by the lighting environment. In addition, our method does not rely on complicated trained models, which allows use in a real-time video conferencing environment easily. On the other hand, our method can reliably extract the reflections and compare them with probing patterns to authenticate real persons under a range of imaging scenarios and validate this approach.

2. RELATED WORKS

Real-time DeepFake Synthesis. DeepFakes have been made for real-time synthesis in recent years. DeepFaceLive [2] was proposed to DeepFake in the real video-conferencing scenario. It obtains higher visual quality and real-time speed that could be used in practice. Using the DeepFaceLive, the users can swap their faces from a real webcam using trained face-swapping models in real-time. The generated fake screen in the DeepFaceLive software can be passed to the video-conferencing software via virtual camera software (*e.g.*, OBS-VirtualCam [2]). For example, in the Zoom software [8], the host can select a virtual camera instead of the actual camera to display the fake screen from the DeepFaceLive in the Zoom meeting. Examples of running DeepFaceLive in a Zoom meeting are shown in Fig. 2.

DeepFake Detection Using Eye Biometrics. Biometric cues from the eyes have been used for the detection of GAN-generated still images [9, 10, 11, 12, 13, 14]. The work [10] uses the inconsistency of corneal specular highlights in the two eyes to identify AI-synthesised faces. More recently, the work [11] spot the AI-synthesised faces by detecting the inconsistency of pupil shapes. These methods are further extended in [12] by using an attention-based robust deep network where the inconsistent components and artifacts in the iris region of the GAN-face are clearly highlighted in the attention maps. Although effective in exposing GAN-generated faces in high-resolution still images in a passive setting, these methods may not work to catch real-time DeepFake videos that are used in video conferences.

[1]We assume that a consensus can be obtained from the attendants to use their imagery for authentication purposes without privacy issues. This would be the same agreement required when the video call is live recorded.

Fig. 2: *Examples of video-conferencing DeepFake using DeepFaceLive [2]. For each pair,* ***Left:*** *The template Faces,* ***Right:*** *The DeepFakes.*

Active Detection of DeepFakes. The active detection for DeepFakes differs from the existing detection methods [15] in that it interferes with the generation process to make detection easier. Early work in [6] obstructs the DeepFake generation by attacking a key step of the generation pipeline, *i.e.*, facial landmark extraction. The method generates adversarial perturbations [16] to disrupt the facial landmark extraction, such that the DeepFake models cannot locate the real face to swap. Active illumination artifacts are studied for exploring the DeepFakes. For example, the work [17] shows that the correspondence of the brightness of the facial appearance in different active illumination can be used as a signal for active DeepFakes detection. Motivated by this work, [7] proposed a new active method for video-conferencing DeepFakes detection using active illumination.

3. METHOD

The overall process of our method is shown in Fig. 3. In a standard video conference setting, a person sits in front of a laptop computer, and her eyes are captured by the webcam, Fig. 3 (a). To verify if the attendant(s) is indeed a real person instead of a synthesis from real-time DeepFake models, the host will briefly put up the *probing pattern* on the shared screen. The probing pattern is a simple geometric shape on a white background to have good contrast. We expect the real attendants' eyes to have reflections of the *probing pattern*, while a real-time DeepFake will not. We first capture an image of the attendant's face and then run a face detector and extract facial landmarks using Dlib [18], Fig. 3 (b). From the facial landmarks, we localize the eye region, Fig. 3 (c), and then segment out the iris part using an edge detector and the Hough transform as in [10], Fig. 3 (d). The segmented iris images are then passed to the template matching steps for automatic DeepFake detection, Fig. 3 (e), where we compare the corneal reflection with the probing pattern. The matching of the two indicates a real person, and the lack of matching suggests a possible real-time DeepFake impersonation.

Our method is based on the assumption that a probing

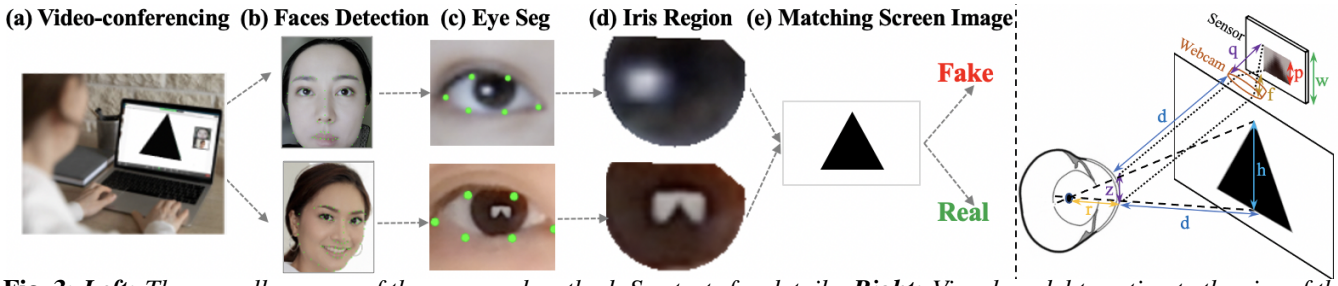

Fig. 3: ***Left:*** *The overall process of the proposed method. See texts for details.* ***Right:*** *Visual model to estimate the size of the probing pattern when it is shown on the sensor, i.e., variable p. We assume the probing pattern is square, and also estimate how many pixels are in the area of p^2 on the sensor.*

pattern on the screen in a video conference can be reliably captured. A key question is if we can have a sufficiently large image from the corneal reflection to match the probing pattern. In the following, we will give an estimation of the number of pixels using an idealized model of a real video conferencing scenario, Fig. 3 (Right).

Assuming the probing pattern is symmetric in both directions with size h (in centimeters). For simplicity, we only consider the vertical dimension. For a laptop display of dimension 30.41×21.24 centimeters (cms), we choose $h \approx 14.5$ cms, which accounts for 70% of the height on the laptop display. The attendant is assumed to sit at $d = 30$ cms away from the screen, which is approximately the distance from the center of the probing pattern to the center of the eye. We further denote r as the radius of the eyeball, which is approximately 1.25cm for a healthy adult [19]. The built-in webcam of the laptop computer has a sensor of height $w = 45 \times 10^{-2}$ cms. A vertical scan line of the sensor has $M = 720$ pixels. We also assume the focal length of the webcam is set to $f = 50 \times 10^{-2}$ cms.

We first compute the vertical height of the corneal reflection of the probing pattern (in cms), which is denoted as z. With a simple geometric relation, we have $\frac{z}{r} \approx \frac{h}{d} \Longrightarrow z = \frac{hr}{d}$. Next, using the focal length formula [20], we can estimate the horizontal distance from the lens of the camera to the sensor, q, as $\frac{1}{d} + \frac{1}{q} = \frac{1}{f} \Longrightarrow q = \frac{df}{d-f}$. With the geometrical relations between z, d, and q, we can have an estimation of the sensor image of the corneal reflection p, as $\frac{p}{q} = \frac{z}{d} \Longrightarrow p = \frac{qz}{d} = \frac{hrf}{d(d-f)}$. This corresponds to a total number of $\frac{pM}{w}$ pixels in the vertical direction for the sensor image of the corneal reflection. As the last step, assuming symmetry between the vertical and horizontal directions again, the number of pixels of the captured corneal reflections is $\left(\frac{hrfM}{wd(d-f)}\right)^2$. Plugging actual numbers for these variables, we can obtain the actual number of pixels of the corneal reflection. With the previous setting, the corneal reflection will have roughly 256 pixels in the image.

The previous derivation establishes the approximate size of the corneal reflection image of the probing pattern. The detection of real-time DeepFake impersonation in a video conference could be a straightforward search in the image of the probing pattern. However, there are some subtleties that motivate a more elaborated solution. First, because of the ambient light in the surrounding environment and the gamut difference between the screen and camera, directly comparing RGB images could lead to inaccurate matching. In Fig. 6 of Section 4 we experimentally demonstrate the effects of color on the matching performance. Since the shape of the probing pattern is more essential in this case, we binarize the patterns obtained from the corneal reflection images and use the binarized mask to compare with the probing pattern, Fig. 4. This is because we can use probing patterns with simple shapes and high contrast (*e.g.*, a triangle shape of saturated color on white background). The high contrast makes the binarization process easier, and we can use automatic thresholding algorithms [10] for that purpose.

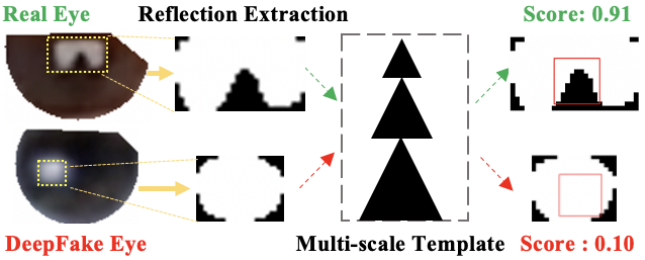

Fig. 4: *Overview of the probing pattern matching. The multi-scale templates are generated using the scaled probing pattern.*

In addition, to proceed with the matching, we need to make the probing pattern to have a similar size as the pattern in the corneal reflection image. This can be done using the analysis of the approximate size of the shape, as in the previous section. However, we need to adjust the matching scale over a range because (i) the estimation is only approximate and (ii) in practice, the attendant's face may move and lead to different distances to the camera and display. Therefore, we generate multi-scale templates and then search the templates in the iris images to identify the occurrence of the template in the image. To generate the multi-scale templates, we need to estimate the shown size of the probing pattern in the sensor (i.e., the reflected probing pattern) as the scaling parameter. With the estimated pixel number N as the scaling parameter, we can generate appropriate multi-scale probing patterns as templates for template matching. For each single matching step with one of the multi-scale templates, we use the method of normalized cross-correlation (NCC) [21]. We use the NCC

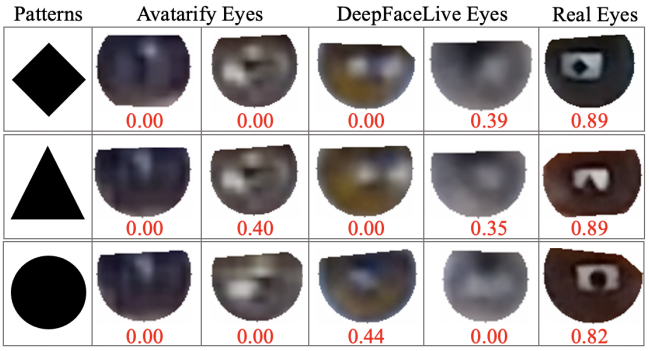

Fig. 5: ***Left:*** *The probing patterns used in our method.* ***Mid:*** *Eye reflection results using Avatar and DeepFaceLive.* ***Right:*** *Eye reflection results of real eyes. Due to the reflection, the probing patterns are clearly shown in the real eyes. But we can not find any probing patterns in the DeepFake eyes.*

implementation in the scikit-image library [22], and the red boxes in Fig. 4 indicate the predicted template location. As the last step, we identify videos as possible real-time DeepFakes if it has NCC lower than a preset threshold.

Limitations and Counter Measures. The effectiveness and robustness of our approach hinge on reliable detection and segmentation of face, eye/iris, and corneal reflections. Several factors may influence the final performance, such as occlusions, resolution requirements of the video-conferencing, etc. But as these components are actively studied as general Computer Vision topics, our approach will benefit from the improved techniques. If we use a fixed probing pattern, then a knowing adversary could predict the probing pattern and intentionally add it to the generated video with minimal temporal delay. We can counter this attack by randomizing the probing pattern and using more complex probing patterns beyond simple geometric shapes, for instance, texts representing the date and time of the meeting, so it is difficult to predict.

4. RESULTS

The efficacy of our method is evaluated on two datasets. The first includes video-conferencing videos of real attendants and their DeepFake synthesis. This dataset validates our method in a realistic video conferencing scenario. The second simulated dataset allows us to evaluate our method across a broad range of assumptions and environmental conditions.

Our real-world dataset was recorded from two users in a range of different environments. Users have placed approximately 30 cm away from the display and camera, with the probing patterns ranging in colors and shapes, such as simple diamond shapes with different colors. We use Zoom [8] as the video-conferencing environment with default settings.

From the real videos, we generate DeepFakes created using Avatarify [1], and DeepFaceLive [2]. The qualitative results are shown in Fig. 5, in which we test some geometrical signs with different shapes for the probing pattern. From the real videos, we can reliably extract the corneal reflections and

Fig. 6: *Evaluate the effectiveness of the colors of the probing pattern where the illumination is fixed.* ***Left:*** *Eye reflections. We can find that stronger global contrast of the sign image indicates better reflection on the iris of the real eyes.* ***Right:*** *A probing pattern with different colors. The red numbers are the corresponding NCC scores.*

match them with the input probing patterns. On the other hand, the DeepFakes do not incorporate environmental lighting and are therefore identifiable easily. Because in the presence of our active probing patterns, their corneal reflection patterns have a nearly zero correlation to the probing patterns. Our current method with unoptimized code has a running time of one frame per 4 seconds. It is certainly possible to improve the overall running efficiency of the algorithm by optimizing the code to be used in the video conferencing tools.

We evaluate aspects of the probing pattern and their influence on the efficacy of our approach.

Shape. We try several different shapes for the probing screen image, as shown in Fig. 5. We can find that these shapes can be reflected successfully in real eyes.

Color and Contrast. To understand what types of screen patterns are more effective, we synthesize common patterns with different colors for further evaluation. As shown in Fig. 6, we try different colors from light to dark for the probing pattern. We find that higher global contrast of the probing pattern indicates better reflection on the iris of the real eyes. Moreover, we can also see that the colors of the probing pattern are hard to show on the reflected iris. The reason may be due to the low video frame quality [23] of the video-conferencing software that uses video compressing technologies [24, 25] during the video-conferencing meeting.

5. CONCLUSION

Real-time DeepFakes pose new challenges to detection methods, which have to run in real-time as a video call is ongoing. In this paper, we describe a new active forensic method to detect real-time DeepFakes by displaying a distinct pattern on the screen and using the corneal reflection extracted from the images of the call participant's face. The direction of the biometric-based active forensic approach provides a promising alternative to the widely used passive forensic methods for DeepFake detection. With the increasing quality and ubiquity of synthetic media, [26, 27] (*e.g.*, online streaming video compression and VR-based metaverse), the active approach has more applications as it can also be used to expose *unauthorized* use of synthetic models. We plan to further explore this direction in the future. In addition, we will further improve the performance and robustness of this method's components to make it more practical.

Acknowledgement. This material is based upon work supported by the Center for Identification Technology Research and the National Science Foundation under Grant No.1822190 and No.2153112.

6. REFERENCES

[1] "Avatarify," https://github.com/alievk/avatarify-python.

[2] "Deepfacelive," github.com/iperov/DeepFacelive.

[3] Siwei Lyu, "DeepFake detection: Current challenges and next steps," in *International Workshop on Media-Rich Fake News (MedFake) in conjunction with ICME*, London, UK, 2020.

[4] "futurism," https://futurism.com/the-byte/researchers-simple-trick-unmask-deepfaker.

[5] "Avinteractive," https://www.avinteractive.com/news/collaboration/crypto-crooks-take-to-holographic-zoom-impersonation-24-08-2022/.

[6] Pu Sun, Yuezun Li, Honggang Qi, and Siwei Lyu, "Landmark breaker: Obstructing deepfake by disturbing landmark extraction," in *IEEE Workshop on Information Forensics and Security (WIFS)*, New York, NY, United States, 2020.

[7] Candice R. Gerstner and Hany Farid, "Detecting real-time deep-fake videos using active illumination," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022, pp. 53–60.

[8] "Zoom," https://zoom.us/.

[9] Falko Matern, Christian Riess, and Marc Stamminger, "Exploiting visual artifacts to expose Deepfakes and face manipulations," in *WACVW*, 2019.

[10] Shu Hu, Yuezun Li, and Siwei Lyu, "Exposing GAN-generated faces using inconsistent corneal specular highlights," in *ICASSP*, 2021.

[11] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu, "Eyes tell all: Irregular pupil shapes reveal gan-generated faces," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2904–2908.

[12] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu, "Robust attentive deep neural network for exposing gan-generated faces," *IEEE Access*, 2022.

[13] Xin Wang, Hui Guo, Shu Hu, Ming-Ching Chang, and Siwei Lyu, "Gan-generated faces detection: A survey and new perspectives," *arXiv preprint arXiv:2202.07145*, 2022.

[14] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu, "Open-eye: An open platform to study human performance on identifying ai-synthesized faces," *IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2022.

[15] Wenbo Pu, Jing Hu, Xin Wang, Yuezun Li, Shu Hu, Bin Zhu, Rui Song, Qi Song, Xi Wu, and Siwei Lyu, "Learning a deep dual-level network for robust deepfake detection," *Pattern Recognition*, vol. 130, pp. 108832, 2022.

[16] Shu Hu, Lipeng Ke, Xin Wang, and Siwei Lyu, "Tkml-ap: Adversarial attacks to top-k multi-label learning," in *ICCV*, 2021, pp. 7649–7657.

[17] Jiacheng Shang and Jie Wu, "Protecting real-time video chat against fake facial videos generated by face reenactment," in *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2020, pp. 689–699.

[18] Davis E King, "Dlib-ml: A machine learning toolkit," *JMLR*, vol. 10, pp. 1755–1758, 2009.

[19] "Human eye," https://en.wikipedia.org/wiki/Human_eye.

[20] "Focal length," https://en.wikipedia.org/wiki/Focal_length.

[21] Jae-Chern Yoo and Tae Hee Han, "Fast normalized cross-correlation," *Circuits, systems and signal processing*, vol. 28, no. 6, pp. 819–843, 2009.

[22] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu, "scikit-image: image processing in python," *PeerJ*, vol. 2, pp. e453, 2014.

[23] Munan Xu, Junming Chen, Haiqiang Wang, Shan Liu, Ge Li, and Zhiqiang Bai, "C3dvqa: Full-reference video quality assessment with 3d convolutional neural network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4447–4451.

[24] Jiahao Li, Bin Li, and Yan Lu, "Deep contextual video compression," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18114–18125, 2021.

[25] Jiahao Li, Bin Li, and Yan Lu, "Hybrid spatial-temporal entropy modelling for neural video compression," *ACM MM*, 2022.

[26] Fabian Mentzer, George Toderici, David Minnen, Sung-Jin Hwang, Sergi Caelles, Mario Lucic, and Eirikur Agustsson, "Vct: A video compression transformer," *Advances in Neural Information Processing Systems*, 2022.

[27] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10039–10049.