THEME ARTICLE: RESPONSIBLE, EXPLAINABLE, AND EMOTIONAL AI

DeFakePro: Decentralized Deepfake Attacks Detection Using ENF Authentication

Deeraj Nagothu, Ronghua Xu, and Yu Chen, *Binghamton University, Binghamton, NY, 13902, USA*

Erik Blasch and Alexander Aved, *The U.S. Air Force Research Laboratory, Rome, NY, 13441, USA*

Advancements in generative models, such as deepfake, allow users to imitate a targeted person and manipulate online interactions. It has been recognized that disinformation may cause disturbance in society and ruin the foundation of trust. This article presents DeFakePro, a decentralized consensus mechanism-based deepfake detection technique in online video conferencing tools. Leveraging electrical network frequency (ENF), an environmental fingerprint embedded in digital media recording affords a consensus mechanism design called proof-of-ENF (PoENF) algorithm. The similarity in ENF signal fluctuations is utilized in the PoENF algorithm to authenticate the media broadcasted in conferencing tools. By utilizing the video conferencing setup with malicious participants to broadcast deepfake video recordings to other participants, the DeFakePro system verifies the authenticity of the incoming media in both audio and video channels.

The rise of the fifth-generation (5G) communication and the Internet of Video Things (IoVT) technologies enables a broader range of applications with megascale data (e.g., all conditions all time video), while COVID-19 forces more activities, such as meetings and conferences, migrated to the cyberspace. While these network-based applications become essential in the *new normal*, which highly depend on reliable, secure, real-time audio or/and video streaming (e.g., Zoom), they become a target for attackers.[1] Enhanced with such security features, users tend to rely on the communication channel for confidential conversations and have a higher trust factor on the information received through audio or video mediums. Hence, end-to-end multimedia attacks have a significant impact where the perpetrator is a trusted participant in the conference who can relay misinformation.[2]

Modern generative deep learning (DL) models have enabled forging audio and video recordings with another source and created false media called deepfakes.[3] The deepfakes are a more potent form of visual layer attacks since it involves manipulating the video and audio channels by imitating a targeted person's face and voice and creating a false recording to relay misinformation through a forged and trusted entity.[4] Generating such recordings is not difficult with the vast availability of source images and video recordings over the Internet.[5] Recent advancements in audio software called Descript allow a user to generate text-to-speech content with training data within 10 minutes.[6] deepfaked videos, audio, or photos in social media are highly disturbing and able to mislead the public, raising further challenges in policy, technology, social, and legal aspects.[7,8] Figure 1 shows an example of a deepfake attack on a celebrity mimicking the source actor.

Electrical network frequency (ENF) is a unique environmental fingerprinting technique for real-time distributed authentication.[9] ENF is an instantaneous frequency in the power distribution networks, and the fluctuations occur due to the load control variations. ENF is embedded in multimedia recording from different power sources, and the resultant media can be authenticated based on the time stamp and ENF fluctuation patterns recovered from it.[10] The similarity of the fluctuations along with its robust and random nature makes ENF a reliable source for authenticating digital media recordings. Existing ENF-exploited solutions rely on centralized architectures that can be a performance bottleneck and vulnerable to a single point of failure.[11]

1520-9202 © 2022 IEEE
Digital Object Identifier 10.1109/MITP.2022.3172653
Date of current version 30 November 2022.

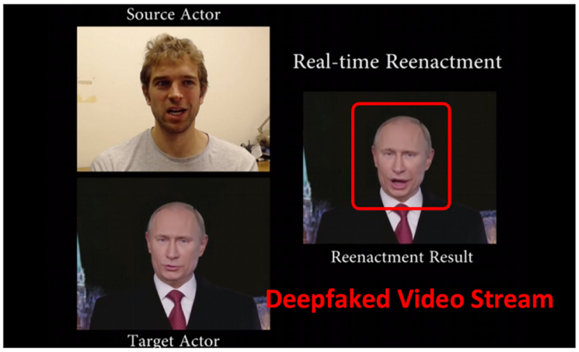

FIGURE 1. Facial reenactment deepfake attack.

Blockchain utilizes a decentralized architecture, which mitigates the problem of a single point of failure and allows for immutable data storage and verification. With the consensus protocol, blockchain executes transactions on a public distributed ledger, which allow for transparency, immutability, and auditability ensuring data authenticity among untrusted devices. Thus, blockchain is promising to enable a decentralized authentication mechanism for ENF-based deepfake detection.[12]

Inspired by spatio-temporal sensitive ENF contained in multimedia signals and decentralized consensus algorithms in blockchain, this article proposes DeFakePro, a novel decentralized ENF consensus-based deepfake detection in audio–video channels for online conferencing scenarios. DeFakePro contributes the following:

- the embedded ENF fingerprints of deepfaked audio and video streams are studied;
- DeFakePro, a secure deepfake authentication system is introduced along with details of key components and workflows;
- a partially decentralized PoENF consensus algorithm is designed to ensure the efficiency and security in distributed authentication of streaming media;
- a proof-of-concept prototype is tested with deepfaked audio and video authentication in an online video conferencing setup and verifies the feasibility of the DeFakePro system;
- an experimental evaluation of the proposed system with the current state-of-the-art technique shows that DeFakePro has similar accuracy performance, but comparatively faster making it suitable for online applications.

DEEPFAKE ATTACKS ON ONLINE CONFERENCING TOOLS

With advanced computation power and development in DL models, generative adversarial networks (GAN) can imprint the source facial landmarks or impressions on a targeted person to recreate similar content with a fake personality commonly referred to as deepfake.[3] Both audio and video recordings can be manipulated with enough training data available from the source. Current deepfake detection relies on DL models trained to detect visual artifacts introduced in the deepfake videos. However, with more training data and models, such artifacts can be removed, and deepfake videos' precision gets better.[5]

For online digital media in the context of conference tools, both audio and video are equally targeted to create the mirage of a fake digital presence. Authentication of both audio and video recording for forgery detection is eminent for information integrity[13] in all available channels. A detection technique void of training data and large-scale computational infrastructure, which depends on underlying fingerprints or multimedia artifacts to locate deepfake forgeries, enables reliable digital media authentication. Deepfakes can perform better with more training data and create visually perfect manipulation of the target, however, it results in high-frequency artifacts and shows poor performance in reconstructing spectral consistencies.[4,8]

Applications with simple video manipulation, such as Face-swap or Face-Shifting software, and audio manipulation, such as generating text-to-audio speech on the go using a modified voice, have become abundantly available for common users.[6,14] For online conferencing tools, the participating perpetrator has complete control over the audio and video broadcasted to other users. With such manipulation tools easily accessible, the perpetrator can imitate a targeted person and spread misinformation. Such attacks raise concerns over the virtual communication platforms, and along with the existing network-level security, online conferencing software also requires an authentication scheme for the information broadcasted.

A study in video deepfake detection using the spatial frequency inconsistencies caused by the up-sampling mechanism of most deepfake generator models confirms the frequency-level modifications.[8] The resulting frequency fingerprints are utilized to train neural networks to detect the GAN-based modifications.[4,15] However, the spatial frequency inconsistency still remains a trainable parameter and the resulting fingerprints could be minimized.[8] Leveraging the incapability of deepfake models to preserve spectral consistencies and the random nature of ENF consisting of spatio-temporal fingerprint information, we introduce DeFakePro channels as a distributed authentication system for online media broadcasts. The proposed DeFakePro identifies fingerprints sensitive to both spatial and temporal frequency manipulations.

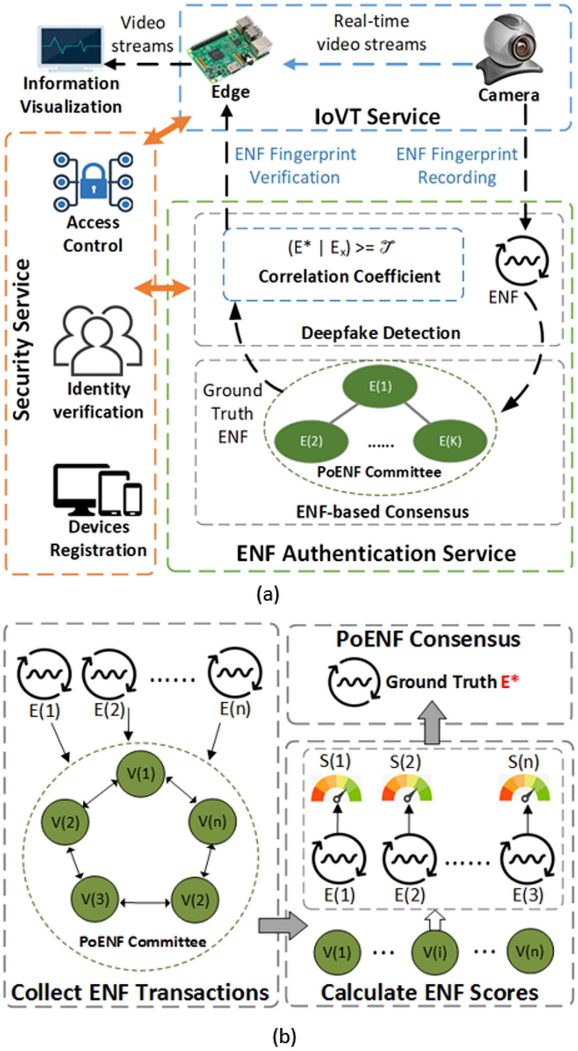

FIGURE 2. DeFakePro system. (a) DeFakePro System Architecture. (b) PoENF Consensus Workflow.

DEFAKEPRO SYSTEM

The DeFakePro system comprises a decentralized authentication system, where each participating node estimates ENF and broadcasts it for proof of authenticity. Figure 2(a) represents the system workflow. Each modules are discussed as follows.

ENF Estimation in Multimedia Recordings

The presence of an ENF signal in a multimedia recording depends on the location of the recording, where the estimation of the signal varies based on the type of multimedia source. DeFakePro tackles both audio and video streams for dual authentication and adds robustness to the system. For the online conferencing setup, we assume the recordings are made in an indoor environment.

ENF in Audio Recordings

The ENF is embedded in an audio recording through electromagnetic induction for devices powered by the electrical grid[10] or background hum for devices running on a battery-powered source.[16] Audio recordings are typically captured with a high sampling frequency (44.1 kHz), but for ENF estimation, a low frequency, such as 1–8 kHz, is sufficient.

ENF in Video Recordings

The source of ENF in video recordings is through the illumination frequency from light sources powered by the electrical grid. The illumination frequency, i.e., 120 Hz when the nominal frequency is 60 Hz, is captured by the recording devices based on the imaging sensor used.[17] The two most commonly used imaging sensors are complementary metal-oxide-semiconductor (CMOS) and charge-coupled device (CCD) sensors. Each sensor has its unique shutter mechanism associated with image capturing, and the total samples captured depend on the number of frames per second (FPS). CCD sensors utilize a global shutter mechanism where the whole sensor is exposed to light for each frame, resulting in lower samples. In the case of CMOS sensors, the frames are captured using a rolling shutter mechanism, where each row of the imaging sensor is sequentially exposed to light, and the number of ENF samples captured is increased by the frame height.[18] Among CCD and CMOS-based sensors, CMOS is most commonly used for general purposes due to its cost efficiency and broad applicability.

ENF Estimation

The ENF is estimated in the following steps:

1) *Power spectrum matrix* is computed using the spectrogram technique from the collected samples in audio and video recordings;
2) based on the *nominal frequency*, the weights are estimated from the harmonic frequencies in the power spectrum matrix;
3) the computed weights are used to *combine spectrum slices* resulting in a robust ENF estimation.

The detailed discussion of ENF estimation techniques for different multimedia recordings are described in our previous work.[11,19]

Security Service

DeFakePro leverages security services to provide basic cryptography security primitives for permission

IoVT, as shown on the left of Figure 2. All devices and users must complete registration to join the network, and DeFakePro assumes that a system administrator is a trusted oracle to manage the profiles of all registered entities. DeFakePro relies on container technology to implement security services that support resource isolation, data flexibility, and maintenance simplicity in a distributed network environment. Each service unit exposes a set of RESTful web service APIs for devices/users. Identity verification services rely on a virtual trust zone method to authenticate identities. Access control services use a capability-based access model to support decentralized access authorization and verification.[9]

Proof-of-ENF (PoENF) Consensus

To maintain ground-truth ENF benchmarks used for deepfake detection, the DeFakePro solution designed a byzantine-resistant PoENF consensus algorithm that is executed by a PoENF committee. Such a committee can be either preconfigured by a system administrator or randomly elected given a certain period of time. Figure 2(b) illustrates the PoENF consensus workflow consisting of three main procedures.

Collect ENF Transactions

At the beginning of the current consensus round, a validator $V(i)$ can broadcast an ENF transaction saving its ENF proof $E(i)$ among PoENF committee members. Then, other validators can verify a received ENF transaction given conditions that a) it should be sent by validators in committee; and b) it should be neither outdated nor existed in the local transactions pool. Finally, all valid transactions are locally buffered.

Calculate ENF Scores

Given a local ENF transaction pool, a validator $V(i)$ can extract ENF proofs from other committee members and build a global view of collected ENF proofs. To prevent against byzantine validators who send arbitrary or poisoned ENF proofs, DeFakePro adopts a byzantine resilient aggregation rule in the ENF score calculation. Finally, each validator has a global view of ENF scores, as shown by Figure 2(b).

PoENF Consensus

In the PoENF consensus stage, every validator can sort ENF scores and choose the minimum one as ground truth E^*. As all honest validators have an identical global view of ENF scores, they can generate the same E^*. The PoENF requires that a validator always uses E^* as the ground-truth ENF. Therefore, the PoENF consensus can make an agreement on E^* given an assumption that an adversary can only compromise at most f committee members.

Interested readers can refer to Xu *et al.*'s work[20] for details about PoENF consensus protocol.

Deepfake Detection Using PoENF Consensus

Once the PoENF consensus agrees on the ground-truth ENF E^* for the round, each node compares its local ENF with the ground-truth ENF using the correlation coefficient. The measure of similarity ranges from $[-1, 1]$, where 1 represents the highest similarity. Based on the experiments, we adopted a threshold of 0.8 to compare the ENF signals. For localization of the forgery, a sliding window protocol is used to compare the ENF signal.

PROTOTYPE AND EVALUATION

Experimental Setup

A proof-of-concept prototype of DeFakePro is implemented in Python. To emulate the participants in an online conferencing tool, we adopted Raspberry Pi-4 (RPi) as the nodes to cap the computation power requirements. For the performance of PoENF consensus, we compared the time latency on RPi and a Dell Optiplex-7010 desktop. The collected raw footage is processed in the devices and using the ENF estimation techniques, the ENF signal is broadcasted to the PoENF committee. For audio deepfakes, the *Descript* platform[6] is used, where the software can generate a text-to-speech synthesis in real time for any pretrained vocals. For video deepfakes, a live deepfake generator named *DeepFaceLive* is used.[14] The video deepfake generator uses the live webcam feed and synthesizes the targeted users' face on the source image.

Performance of PoENF Consensus Mechanism

Table 1 presents the cumulative time taken for a round of PoENF consensus, including ENF proof broadcast, verification, and PoENF algorithm execution. The time complexity of the PoENF consensus is $\mathcal{O}(K^2\ d)$, where K is the committee size, d is the ENF sample size, and the latency increases with the number of validators. A general conference scenario typically includes less than 50 participants and incurs delays up to 0.5 and 0.2 second on a desktop.

Detecting Audio and Video Deepfakes

The audio and the video deepfakes are generated independently to analyze the effects of ENF on each

TABLE 1. Poenf consensus latency (second) with different number of validators.

No. of validators	10	20	50	100	200	500
RPi-4	0.02	0.08	0.48	1.93	7.71	48.5
Desktop	0.01	0.02	0.15	0.59	2.4	15.4

Note: Comparative evaluations on platform benchmarks.

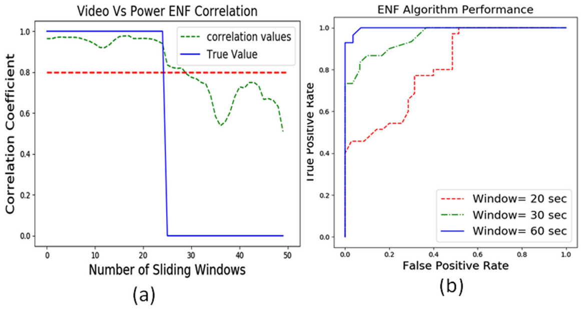

FIGURE 3. (a) Correlation coefficient values for deepfake localization. (b) ROC curve for optimal ENF window size for lower false positives and threshold selection.

recording. For audio recordings broadcasted, the text-to-speech modification is made in multiple locations throughout the recording.[6] After comparing the ENF from the ground-truth ENF and the audio deepfake ENF with multiple forgery locations, a low correlation coefficient indicates fake audio from which the modified section can be localized. Figure 4(a) represents the mismatch in the ENF, where the correlation coefficient is below the threshold.

For video recordings, the deepfakes are generated in realtime using a pretrained model face set in the DeepFaceLab tool.[14] For our experiment, multiple deepfake videos were generated using the DeepFaceLab tool. For the generated deepfake videos, a subset of frames from each video is analyzed for spatial frequency inconsistencies generated due to the deepfake model up-sampling mechanism.[8] Figure 4(c) represents the changes in spatial frequencies caused by most deepfake models, generated by analyzing the azimuthally averaged frequency spectrum of deepfake frames.[8] Along with the facial manipulations in the frame center, the spatial frequency inconsistencies represent that the deepfake model adds additional perturbations in the static background of the frame.

The video frames are buffered in online conferencing tools to collect enough samples for a reliable ENF estimation. With more samples, ENF estimation is more accurate. A sliding window approach is used for an online authentication system to buffer incoming frames and estimate ENF. We tested various window sizes and a fixed shift size of 5 seconds for our experiment since shift size has a low effect on ENF estimation. Figure 3(a) shows a clear separation between original frames and deepfake frames, while Figure 3(b) represents the accuracy of detecting deepfake videos as window sizes vary.

Using the appropriate window and shift sizes, ENF-based video authentication is presented in Figure 4(b). Given the ground- truth ENF, the measure of similarity of the incoming deepfake video ENF estimates is lower than the original video streams. In deepfake recordings, even though the facial landmark regions are forged, the pixel intensities through the frame are modified due to added perturbations, as seen in Figure 4(c).

Performance Evaluation

A comparison study is performed to analyze the effectiveness of ENF-based authentication compared to the spatial frequency-based GAN fingerprint. To the

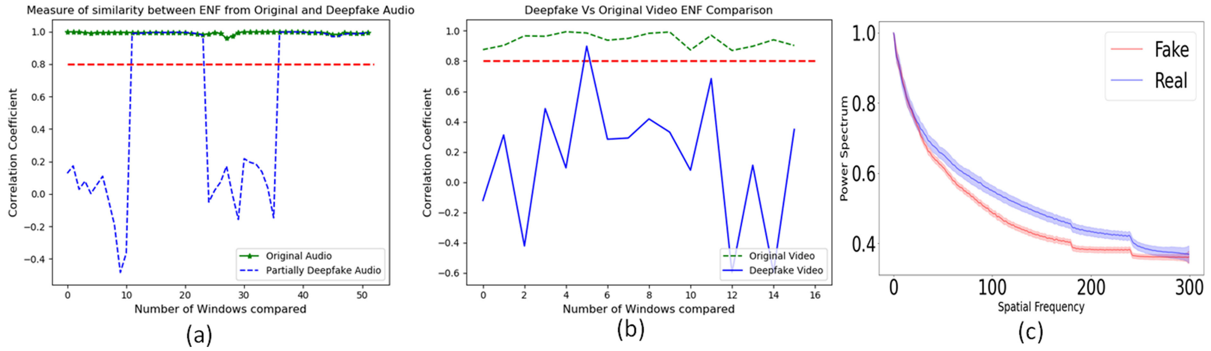

FIGURE 4. (a) Original and deepfaked audio recording correlated with ground-truth ENF. (b) Real-time Face-swap deepfake generated using DeepFaceLab software and compared with ground-truth ENF. (c) Spectral inconsistencies generated by deepfake model, generated using frequency transform and azimuthal averaging.

TABLE 2. Deepfake detection performance comparison.

Techniques	1080p		720p		480p	
	AUC	FPS	AUC	FPS	AUC	FPS
DeFakePro	0.95	19	0.96	25	0.95	33
UpConv	0.94	10	0.97	14	0.98	20

best of our knowledge, the presented approach is the only technique focused on online deepfake detection using a distributed backbone system. We collected ten 5-minutes deepfake videos, collected with ground-truth ENF, and evaluated the performance using ENF-technique and spatial inconsistencies-based detection *UpConv* in Durall *et al.*'s work.[8]

Table 2 represents the performance of each deepfake detection technique in multiple resolutions. For online conference scenarios, faster and more reliable techniques are more viable due to their time-sensitive nature. The area under the curve for both techniques is similar for all formats, however, the number of FPS for the proposed DeFakePro system is higher since there is minimal frame processing required and no feature training. DeFakePro is applicable to any input streams as long as it carries a background ENF signature, and the nominal frequency is known. The presented approach is effective against any kind of frame modification since the ENF fingerprint carries unique fluctuations and allows for a distributed authentication system enabling deepfake detection and byzantine nodes.

CONCLUSION

This article presents DeFakePro—a decentralized deepfake attack detection system leveraging embedded ENF signals in online video conferencing tools. The proposed DeFakePro adds resilience to byzantine nodes and verifies media integrity with minimal computational resources using the integrated PoENF consensus mechanism. The consensus mechanism establishes the ground-truth ENF in each round, and each participating node can verify the media authenticity using a correlation coefficient. Furthermore, the consensus mechanism is evaluated for time latency based on the number of participants in each round. However, the application of ENF-based authentication is limited to zones with passive ENF presence, such as indoor environments.

The experimental results show that the DeFakePro system can detect and localize the deepfake audio and video attacks using the estimated ENF signal. The DeFakePro system is evaluated against the current deepfake detection techniques, and the proposed system achieves similar performance and had a faster processing rate, which is a prerequisite for an online detection system.

ACKNOWLEDGMENTS

This work is supported in part by the U.S. National Science Foundation (NSF) under Grant CNS-2039342 and the U.S. Air Force Office of Scientific Research (AFOSR) Dynamic Data, and Information Processing Program (DDIP) under Grant FA9550-21-1-0229. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Air Force.

REFERENCES

1. V. Mehta, P. Gupta, R. Subramanian, and A. Dhall, "Fakebuster: A deepfakes detection tool for video conferencing scenarios," in *Proc. 26th Int. Conf. Intell. User Interfaces*, 2021, pp. 61–63.
2. D. Kagan, G. F. Alpert, and M. Fire, "Zooming into video conferencing privacy and security threats," 2020, *arXiv:2007.01059*.
3. I. Goodfellow *et al.*, "Generative adversarial nets," *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 1–9, 2014.
4. J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3247–3258.
5. L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, Aug. 2020.
6. "Descript create podcasts, videos, and transcripts." Accessed: Sep. 20, 2021. [Online]. Available: https://www.descript.com/
7. M. Westerlund, "The emergence of deepfake technology: A review," *Technol. Innov. Manage. Rev.*, vol. 9, no. 11, pp. 39–52, 2019.
8. R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7890–7899.
9. R. Xu, D. Nagothu, and Y. Chen, "Decentralized video input authentication as an edge service for smart cities," *IEEE Consum. Electron. Mag.*, vol. 10, no. 6, pp. 76–82, Nov. 2021.
10. C. Grigoras, "Applications of enf analysis in forensic authentication of digital audio and video recordings," *J. Audio Eng. Soc.*, vol. 57, no. 9, pp. 643–661, 2009.

11. D. Nagothu, Y. Chen, E. Blasch, A. Aved, and S. Zhu, “Detecting malicious false frame injection attacks on surveillance systems at the edge using electrical network frequency signals,” *Sensors*, vol. 19, no. 11, 2019, Art. no. 2424.
12. S. Y. Nikouei, R. Xu, D. Nagothu, Y. Chen, A. Aved, and E. Blasch, “Real-time index authentication for event-oriented surveillance video query using blockchain,” in *Proc. IEEE Int. Smart Cities Conf.*, 2018, pp. 1–8.
13. Y. Zhou and S.-N. Lim, “Joint audio-visual deepfake detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14800–14809.
14. I. Perov *et al.*, “Deepfacelab: A simple, flexible and extensible face swapping framework,” 2020, *arXiv:2005.05535*.
15. F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, “Do GANs leave artificial fingerprints?,” in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2019, pp. 506–511.
16. J. Chai, F. Liu, Z. Yuan, R. W. Conners, and Y. Liu, “Source of ENF in battery-powered digital recordings,” in *Audio Engineering Society Convention 135*. New York, NY, USA: Audio Engineering Society, 2013.
17. R. Garg, A. L. Varna, A. Hajj-Ahmad, and M. Wu, “‘Seeing’ ENF: Power-signature-based timestamp for digital multimedia via optical sensing and signal processing,” *IEEE Trans. Inf. Forensics Secur.*, vol. 8, no. 9, pp. 1417–1432, Sep. 2013.
18. S. Vatansever, A. E. Dirik, and N. Memon, “Analysis of rolling shutter effect on ENF-based video forensics,” *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 9, pp. 2262–2275, Sep. 2019.
19. D. Nagothu, Y. Chen, A. Aved, and E. Blasch, “Authenticating video feeds using electric network frequency estimation at the edge,” *EAI Endorsed Trans. Secur. Saf.*, vol. 7, no. 24, pp. 1–13, 2021.
20. R. Xu, D. Nagothu, and Y. Chen, “Econledger: A proof-of-ENF consensus based lightweight distributed ledger for IoVT networks,” *Future Internet, Special Issue Blockchain: Appl., Challenges, Solutions*, vol. 13, no. 10, pp. 1–24, 2021.

DEERAJ NAGOTHU is the Ph.D. candidate of electrical and computer engineering at the Binghamton University - SUNY, Binghamton, NY, 13902, USA. His research interests include multimedia forensics in Internet of Video Things (IoVT) and computer network security. Nagothu received his M.S. degree on electrical and computer engineering from Binghamton University. Contact him at dnagoth1@binghamton.edu.

RONGHUA XU is the Ph.D. candidate of electrical and computer engineering at Binghamton University, Binghamton, NY, 13902, USA. His research focuses on blockchain-based security solutions to Internet of Things (IoT). Xu received his M.S. degree in mechanical and electrical engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, China. He is a graduate student member of IEEE. Contact him at rxu22@binghamton.edu.

YU CHEN is an associate professor of electrical and computer engineering at Binghamton University, Binghamton, NY, 13902, USA. His research interests include edge-fog-cloud computing, IoTs, and smart cities. Chen received his Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA. He is a senior member of IEEE and SPIE, and a member of ACM. Contact him at ychen@binghamton.edu.

ERIK BLASCH is with the Air Force Research Laboratory (AFRL), Rome, NY, 13441, USA. His research interests include target tracking, image fusion, information fusion performance evaluation, and human–machine integration. Blasch received his Ph.D. degree from Wright State University, Dayton, OH, USA, in addition to seven master’s degrees. He is a fellow of IEEE, AIAA, SPIE, and MSS. Contact him at erik.blasch.1@us.af.mil.

ALEXANDER AVED is a technical advisor at the Air Force Research Laboratory Information Directorate, Rome, NY, 13441, USA. His research interests include multimedia databases, stream processing and dynamically executing models with feedback loops incorporating measurement and error data to improve the accuracy of the model. Aved received his Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA. Contact him at alexander.aved@us.af.mil.