IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 31, 2021, accepted November 26, 2021, date of publication November 30, 2021,
date of current version December 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3131517

Real, Forged or Deep Fake? Enabling the Ground Truth on the Internet

MOHAMMAD A. HOQUE[1], MD SADEK FERDOUS[2,3], (Member, IEEE), MOHSIN KHAN[1], AND SASU TARKOMA[1], (Senior Member, IEEE)
[1]Department of Computer Science, University of Helsinki, 00560 Helsinki, Finland
[2]Department of Computer Science and Engineering, BRAC University, Dhaka 1212, Bangladesh
[3]Centre for Financial Technology, Imperial College Business School, London SW7 2AZ, U.K.

Corresponding author: Mohammad A. Hoque (mohammad.a.hoque@helsinki.fi)

This work was supported by the Academy of Finland IDEA-MILL research project.

ABSTRACT The proliferation of smartphones and mobile communication has enabled users to capture images or videos and share them immediately on social networking and messaging platforms. Unfortunately, these platforms are also used to manipulate the masses by performing social engineering attacks by sharing fabricated images (or videos). These attacks cause public shame, ethnic violence and claim lives. With the rise of advanced image processing tools, the deep fakes are automated, and their implications are boundless. In this article, we discuss different types of modification of images/videos and survey the corresponding methods and tools. We also highlight the ongoing efforts to detect fake images and videos using advanced machine learning tools and fact-checking. Along with these tools, we also need different complementary approaches discouraging the production and propagation of manipulative forged images and videos on the Internet. This paper further emphasizes that we desperately need socio-technological solutions that empower end-users with the right tools to make an informed moral decision while producing, uploading, and sharing media. Finally, supporting this, we discuss a holistic blockchain-based solution.

INDEX TERMS Deep fake, synthesis, reenactment, swapping, enhancement, authentication, deep learning, neural networks, image classification, blockchain, verification.

I. INTRODUCTION

We, the humans, perceive visual information such as images and videos with less effort than oral or textual information [2]. We are also better at recalling visual represents [3]. Visual perception is precise and intrigues us quickly. This is because a bit of familiarity with visual content gives us such a meta-cognitive experience that we perceive the content as credible [4]. The intriguing images and videos sometimes challenge one's deeply rooted beliefs and expectations. Consequently, one seeks emotional release and consolidation through sharing the content with others. The emotional expression of one begets the expression of many others, who have similar beliefs or views, that create a ripple effect [5], [6]. Moreover, easy access to the Internet through smart devices enables an unprecedented number of people to participate in social networks and messaging platforms, accelerating the ripple effect. Therefore, on-purpose manipulated visuals and audiovisuals mislead not only an individual but also a whole community.

In recent years, it has become easier to divert the attention of the masses by introducing fabricated or fake images and videos. For example, it is common to add/remove frames in a video or to replace a person's face with another in a photo with Photoshop-like tools is common. Although such alterations can have legitimate use cases, some media are engineered to be viral, manipulative, and harmful. Once on the Internet, they can ruin people's careers, claim human lives, or cause ethnic violence. For example, to raise awareness regarding child abduction, a campaign program in Pakistan produced a video demonstrating abduction play and how children should react in such circumstances. Unfortunately, some malicious actors in India removed the last few seconds and portrayed an actual abduction. This cleverly edited video created confusion, panic among people, and their outrage resulting in an eight-week-long riot that claimed several lives [7]. In the same year, morphed nude images of teenagers in India were shared on social media as a heinous act of revenge porn. The utter humiliation was unbearable, and the victims took their lives [8].

With the advent of modern AI, such unethical deformation of images has become smoother and undetectable to

The associate editor coordinating the review of this manuscript and approving it for publication was Chi-Tsun Cheng.

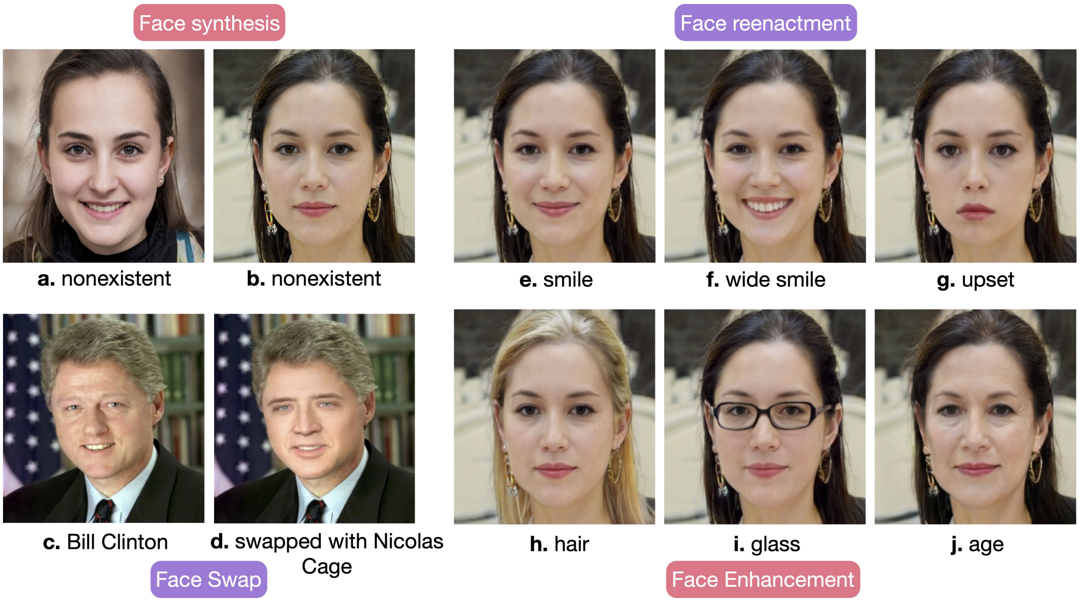


FIGURE 1. Types of image deformation or alteration. The first two nonexistent images are *synthesized* from numerous other images in https://thispersondoesnotexist.com [1]. Next, Bill Clinton's face (c) is *swapped* with Nicolas Cage in picture (d). The FaceApp application applies Facial *reenactments* on (b) and produces (e), (f), (g) on an iPhone 6. The FaceApp application also *enhanced* different facial attributes of face on image (b), i.e., from Asian hair to blonde (h), adding spectacles (i) and adding wrinkles (j).

human eyes. With sufficient training images or videos available, AI-based algorithms can generate new fake faces and reenact an individual's face gaze, facial expression, voice, and body expression. These algorithms are already being used to produce revenge porn, and the victims are mostly females. For example, an AI-powered Telegram bot recently transformed the images of more than one hundred thousand women into fake nudes [9]. To the worse, the news organizations also publish fabricated images [10]. These incidents highlight that any entities with the required technological know-how can be manipulators. These manipulators might engineer images and videos, which others may perceive as credible or believable. The people, in general, may lose trust in the existing establishment as the ground truth cannot be obtained or established [11]. Therefore, social media platforms are under constant scrutiny to examine and curb forged media.

This paper first explores different human visual alterations and surveys how images and videos are forged in the modern era. Next, we argue that detecting forged media after circulating on the Internet is insufficient to prevent their circulation. In addition to the detection by a third party, various involved entities should verify the authenticity of images or videos on the Internet whenever they are created → uploaded → viewed → shared on the Internet. Simultaneously, the relevant technological solutions should empower users to judge content credibility and make an informed moral decision in sharing media. Finally, we propose a more holistic solution that incorporates new socio-technological methods to prevent the introduction and circulation of unauthentic media on the Internet and social networks.

The rest of the article is organized as follows. Section II first presents different types and methods of human face alteration in images and videos and the tools to perform those modifications. Next, it discusses the mechanisms to detect such alterations. Section III, presents challenges in preventing the production and sharing of forged media and discusses a recent content authenticity initiative to detect forged media. Section IV presents a blockchain-based holistic system to prevent the production and sharing of forged media on the Internet. Section V discusses the additional technological challenges in realizing the system. Finally, we conclude in Section VI.

II. ALTERATION OF HUMAN VISUALS

In this section, we first present different types of alternation of images with their broad implications. Next, we briefly discuss the underlying mechanisms (tools and methods) and their use cases.

A. TYPES OF ALTERATION

From existing studies, we identify four types of alteration in the context of human visuals: (i) synthesis, (ii) reenactment, (iii) replacement, and (iv) enhancement.

1) SYNTHESIS

Synthesis allows creating a new arbitrary image from other images without any reference target. While there are no actual people with such synthesized faces, we might see people with such faces in the future through plastic surgery. These approaches allow to create new nonexistent images as shown in Figure 1 (a, b).

2) REENACTMENT

Reenactment is the transformation of a human visual (target) with respect to another human (source) without changing the identity of the target human. Expression reenactment may involve transferring the mouth, gaze, or pose and derived from the source image. Figure 1 (e,f,g) show three example expression reenactments of the target image in Figure 1 (b), where the source images are selected by the FaceApp [16] application on an iPhone. The famous Obama case is an example of mouth reenactment or lip-syncing in a video, where a source audio drives the mouth and facial reenactments of the target (Obama) according to an audio [17]. The sources of these reenactments could be the other images of Obama or other people. Facial reenactment is helpful to improve the performance of facial recognition algorithms. The other kind of facial reenactment is changing facial muscle expression, enabling us to understand the emotion. In the entertainment industry, this helps to re-create the facial expression of famous artists for post-editing. However, the most promising use case is very realistic dubbing movies or documentaries into different languages or even lip-reading speech therapy.

3) SWAPPING/TRANSFERRING

Swapping or Transferring is the most primitive method to tweak body, or facial expression. It happens by replacing the content of a target with that of the source with respect to the target. For example, Bill Clinton's facial attributes in Figure 1(c) are swapped with Nicolas Cage's on Clinton's image as shown in Figure 1 (d). Therefore, swapping allows a person to impersonate another, for example, by swapping an artist's identity with a popular person for entertainment. Such edits also can be applied for revenge porn and disseminating political opinions [18] by swapping the face of an artist with the victim's face.

4) ENHANCEMENT

Human visuals also can be enhanced by adding, removing, or modifying the attributes of the target images without any requirement from the source. For example, changing hair color, adding spectacles, and adding wrinkles in the face are shown in Figure 1(h-j). While the reenactment enables impersonation, enhancement impacts the persona of a person. The fashion industry can use such enhancements before actual changes in the hair or face. Plastic surgery is another example, where people can check prior facial rejuvenation [19] or breast augmentation. On the dark side, such enhancements also allow revenge porn by removing clothes digitally.

B. TOOLS AND METHODS FOR ALTERING IMAGES

Photoshop and other similar tools had been used for image editing, such as face transfer and swapping. Editing photos with such tools require various steps and take much time. Moreover, the artifacts due to editing with these tools are also visible to human eyes. There are numerous algorithmic approaches to change the facial attributes and transfer faces in images and videos other than manually editing. With the availability of a free large collection of images [12], a new set of computer vision and machine learning approaches, such as Face2Face [13], FaceSwapNet [14], FSGAN [15], are producing more smooth images. The media produced using these tools are called Deep Fake. Figure 1 shows fake images produced through synthesis, reenactment, swapping, and face editing with such modern tools.

1) FACE SYNTHESIS

The synthesized nonexistent photos presented in Figure 1 are generated by a Generative Adversarial Networks (GAN) called StyleGAN [20]. Figure 1(a, b) are two StyleGAN-generated synthetic photos. In GANs, two neural networks compete with each other to improve the synthetic media quality. The generator takes random noise as input and outputs synthetic data. In contrast, the discriminator takes both the true images and synthetic data as inputs and outputs the fakeness of the synthetic image. Multiple generators can be used either hierarchically or progressively to improve the performance or quality of images. Besides, the quality of the training images also affects the outcome of the synthesized photos. Karras *et al.* trained the GANs in a progressive fashion [20], which increased GAN's performance in generating more realistic human faces.

2) FACE SWAP AND REENACTMENT

Face Swap and reenactments go through several steps. The first step is to crop the face of the source and target images and derive the intermediary features, such as facial boundary, landmark points, and 3D Morphable Model (3DMM) of the human face. The next step is to generate the new face and stitch that on the target image. The primary method to create Deep Fakes is to use two encoder-decoder networks with a common encoder. This allows the encoder to learn the common features of the faces. The actual face transfer or swapping uses another pair with the same encoder with the decoder for the target image. Figure 2(b) shows that this approach affects the face in a rectangular fashion.

FSGAN [15] first generates a sequence of landmarks from the source image with small pose changes with a reenactment network. A segmentation network estimates the segmentation masks for the target image. An inpainting network reconstructs the missing parts of the source reenactments to complete the reenacted face according to the segmentation masks of the target image. Finally a blending network blends this reconstructed face with the target face using the segmentation mask of the target image. Figure 2(e) shows facial region affected by FSGAN. Another approach uses the landmark points and uses a convolutional neural network (CNN) to swap faces [21]. [22]

On the other hand, face reenactment methods transfer various facial expressions, such as gaze, to the target image with respect to the target. FSGAN relies on the sequence of the landmarks to drive the facial reenactments. In contrast, FaceswapNet enables reenactment from any arbitrary source

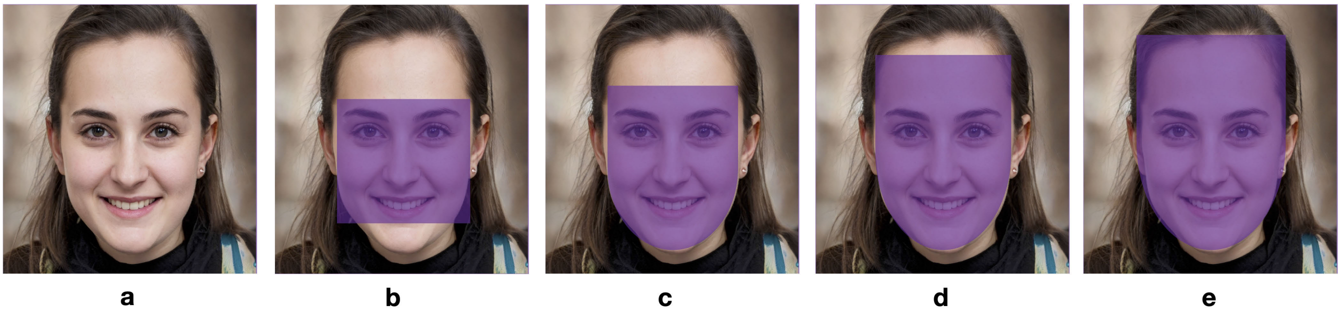


FIGURE 2. Different face swap/transfer methods affect facial region of the nonexistence image differently; (a) without any modification, (b) modification with FaceForensics++ [12] variants, (c) modification with Face2Face [13], (d) modification with FaceSwapNet [14], and (e) modification with FSGAN [15].

to any undefined target [14]. It relies on an encoder-decoder network to extract the landmark features from the sources and adapt to targets. Next, it leverages the geometry information from the landmarks and neutral expression of the target to generate realistic images. The training phase requires two pairs of encoder-decoder with two different decoders. Figure 2(e) shows that the method affects the middle of the forehead to the jaw of the face, excluding mouth internals. The Face2Face reenactment method manipulates the region corresponding to a 3DMM face model. Figure 2 shows how the technique affects the face similarly to FaceSwapNet, including the mouth internals.

3) FACE ENHANCEMENTS

Editing multiple face attributes may result in a face transfer or synthesis. Some approaches aim to change an individual attribute, as demonstrated in Figure 1 (h-j). In this way, it would require multiple methods to edit different attributes. Whereas the autoencoders or such architectures, such as AttGAN [23], allow to alter one or multiple attributes at a time. These approaches are already being used as augmented reality applications to check the new hairstyle or color [24].

C. APPROACHES TO ALTER VIDEOS

Alterations also can be done with the videos. The very general method for editing a video is to add and remove the frames. The recent deep learning tools apply face swapping and reenactments on real-time videos. In fact, Face2Face and FSGAN change the faces of target images with that of the source image while streaming. They superimpose the facial expressions of the target person on an online person. The information includes fiducial features of face features, such as nose, mouth, lips, head pose, alignment of the pose. The facial expression of the online user is continuously tracked to determine the head pose. Once determined, the user's face is compared with all the aligned faces of the target user, and the best match is returned.

D. DETECTING FAKE OR FORGED MEDIA

While the media deformation has excellent use cases, there are dark sides. Someone can easily use face-swapping methods to construct fake porn images and videos, and they can further use them as revenge porn. With face editing methods, such attackers can remove clothes on different body parts in the images and videos. Since the face transfer/swapping methods transfer user identities, these tools give an attacker control over others' identities [25], which can be used for blackmailing and shaming. The consequence can cause utter humiliation and claim lives when those are shared on Internet platforms like, for example, social networks. With some applications, such face alteration can happen in real-time. Even people can be made invisible in the videos, which can have severe consequences when video recordings are essential for evidence in a legal court. To combat such forged media and its effects, we need to detect the forgery. The state-of-the-art for detecting forged media can be classified as signal processing and deep learning approaches. The signal processing methods investigate the cues from the image sensors at the signal level [26], lighting, and shadow reflection [27] at the physical level or image metadata.

Deep Fake images are difficult to detect compared to the other types of manipulation. The Deep Fakes already threaten the signal processing methods by generating very realistic images. The detection of GAN-generated photos and videos is already challenging for the face recognition systems [28] and compression makes it more difficult. Towards this, Rossler *et al.* introduced a massive database of manipulated images called Forensics [29]. Later the dataset was extended to 1.5 Million Deep Fake images generated by Face2Face and FaceSwap [12]. Such large datasets assist in training the deep learning methods to detect Deep Fakes. FakeFaceDetect crops the facial areas from images and trains various CNNs with real and fake images [30].

Figure 2 demonstrates that different face swapping and reenactment methods affect the facial region differently. Another Deep Fake detection approach aims to automatically extract the features from such facial landmarks or areas from the GAN-generated images and use their similarities [31]. A simple SVM with these landmarks achieves good classification accuracy for GAN-synthesized faces [32]. However, there could be some artifacts on the algorithm-generated images as well, such as water droplets and asymmetries in the facial attributes [33]. For example, Figure 3 shows the presence of such asymmetries and artifacts generated

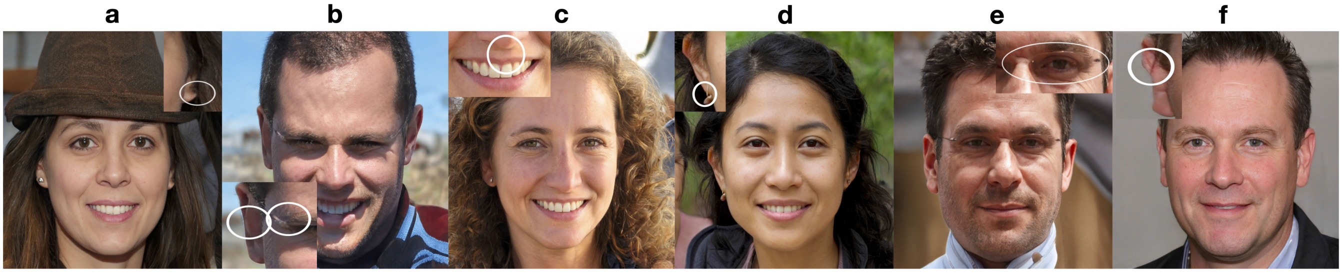


FIGURE 3. Artifacts in synthetic photos generated by StyleGAN. These images lack symmetry in ear pins, have visible marks in ears, and artifacts for spectacles, as overlayed at the corners of the images.

by StyleGAN. Zhang *et al.* investigated more specifically at the edges of ears, teeth, hair of such synthesized images and found distinct characteristics. They extracted such edge information and used a deep neural network to detect the GAN modified images [34]. Besides, the face synthesis methods may create a new face while keeping the facial expression unchanged, which results in mismatched facial artifacts. Although such artifacts might not be visible in the eyes, the head pose's differences may reveal those landmarks. Yang *et al.* showed that such a distinction could identify synthesized photos [35].

In synthesized videos, the research community overlooked intrinsic human physiological activities such as breathing or eye blinking. The apparent reason is that the training datasets do not include such information. Li *et al.* [36] used a convolutional neural network (CNN) with the recurrent neural network (RNN) to detect the presence of eye blinking in the videos. Another approach is to detect audio inconsistencies in videos [37]–[39]. Another CNN-based approach detects spoofed audio which also can be used to detect video Speech forgery [40]. Detecting inconsistencies in biological signals for different facial expressions in a video [41].

Figure 1 shows how different Deep Fake and Face2Face methods change faces. Afchar *et al.* [42] proposed two CNNs to detect the faces altered by these methods in the videos. Guera *et al.* used a CNN to learn the features from video frames and then applied an LSTM to detect common-encoder generated Deep Fakes in videos [43]. The CNN learns frame features, and the LSTM performs temporal sequence analysis on the feature vectors. In general, a local structural relationship exists between pixels in an image. In altered images, any changes in such relations should be detectable. Bayar *et al.* [44] added a new convolution layer that forces CNN to learn only the pixels' local structural relationships rather than content-specific features. The approach is robust against different manipulation techniques.

E. DISCUSSIONS

The Deep Fake generating tools are becoming more efficient in generating more realistic images and videos. Some deep learning approaches require subject-specific training for face swapping and reenactment. However, FSGAN and FaceSwapNet can work with arbitrary source images. FaceSwapNet even can transfer reenactment to an undefined target. Furthermore, computer vision-based approaches can be applied to faces in real-time. Previously, it would require attribute-specific systems to edit different parts of the face or body. For example, encoder-decoder networks allow editing more features. AttGAN allows changing the attributes on demand rather than a particular attribute. Similarly, new synthesized faces are even smoother. Similarly, the detection mechanisms are also advancing. ForensicTransfer [45] and Capsule networks [46] provide robust detection of modified images/videos against new modification techniques. Readers can dive into detail about Deep Fakes into these surveys [47], [48].

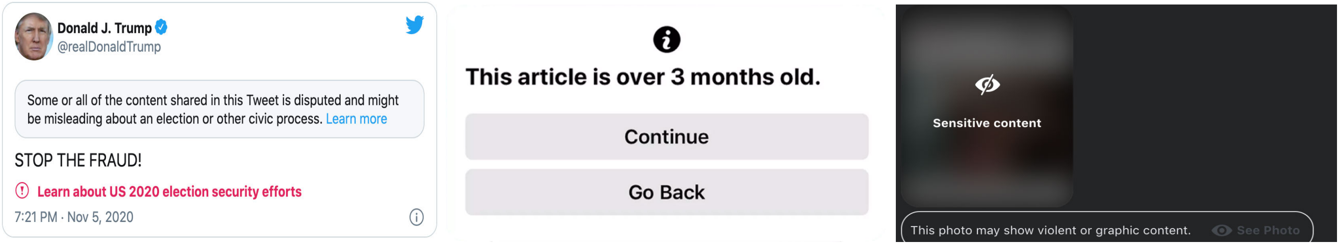
III. MEDIA VERIFICATION

Most popular media content is engineered to play with people's emotions such as pride, supremacy, fear, or anger across different groups. Popular media is transmitted on the Internet through likes and shares, as the users seek the reciprocation of similar feelings from the groups. With mobile devices and communication technologies, the replication of such media is fast and effortless. When an engineered media goes viral, the outcome can be devastating – from personal humiliation to ethnic violence. Therefore, we need verification measures to prevent the transmission of provocative forged media before they become viral. First, we need mechanisms to verify and label the media – for example, real, deep-faked, or forged – as they enter the Internet. Such verification and labeling mechanisms should lead to the ground truth, i.e., the original media. Detecting forged media is also essential when the media has escaped the verification and labeling mechanisms. This section presents the challenges in forged media verification and discusses a recent initiative that employs cryptographic verification methods.

A. CHALLENGES IN VERIFYING FORGED MEDIA

1) COGNITIVE BIASES

The existing practice in information sharing is to verify the source of information first, and the responsibility to verify the source falls upon consumers. The sheer volume of (mis)information shared in social networks and available on



Donald J. Trump
@realDonaldTrump

Some or all of the content shared in this Tweet is disputed and might be misleading about an election or other civic process. Learn more

STOP THE FRAUD!

Learn about US 2020 election security efforts

7:21 PM · Nov 5, 2020

This article is over 3 months old.

Sensitive content

This photo may show violent or graphic content.

(a) Twitter warning

(b) Facebook old news Alert

(c) Facebook violent content warning

FIGURE 4. Various warning messages on different social media platforms for assisting users to make decision before viewing and sharing content.

the Internet presents a cognitive challenge to credibility judgment. It is difficult for consumers to verify the authenticity of images and videos without assistance because we all do not have the same level of expertise or information route to verify an image or video's authenticity. While we all accept the same information, our cognitive bias or a little familiarity may leak some misinformation from us. Even the most educated/responsible users may share misinformation by relying on cognitive heuristics, such as endorsements, the number of shares/likes, the number of followers, website reputation, and information source. Indeed, the source or origin may not make an image/video authentic [49]. For example, recently, Fox News produced and published fabricated images of Seattle protests [10]. On top of that, credibility is not the objective quality of the images (or videos). It depends on the perceptions of individuals. Therefore, various Internet entities must assist users in credibility judgments while uploading, consuming, or sharing images (or videos) on the Internet.

2) LACK OF SUPPORTING TOOLS

There are several ways users share images or videos on the Internet from their computing devices. They can upload images or videos to the Internet via browser, social networking, or messaging applications. Fake image (or video) detection tools and fact-checkers are essential parts of the Internet to fight against forged media and misinformation. Social media platforms have been relying on them to fight against misinformation.

Figure 4 (a) shows that Twitter has been adding warning messages for Donald Trump's Tweets. Facebook recently has started adding alerts while sharing old news, as shown in Figure 4(b). This approach prevents contextual manipulation by reducing the propagation of old information. Facebook has been adding warning labels for violent photos or videos 4(c). Social networking platforms warn about age-restricted content. All these approaches should help users make informed decisions and thus prevent misinformation to some extent.

Recent studies have shown that users tend to be reserved in sharing news on social networks when the news articles are labeled as false [50], [51]. We believe that users would react similarly to tagged forged images and videos; however, the labeling could differ from the news. Furthermore, such explicit labeling should begin at the very early stage when any image (or video) enters the Internet for the first time. Because labeling late does not reduce the propagation of viral images or videos on the Internet and thus their consequences.

In addition, the fake-detection tools and fact-checkers play roles when something is already published. Whenever content is uploaded to the Internet, the Internet never forgets. Therefore, it is vital to prevent the uploading of fake or forged content in the first place. However, stopping to upload will appall the users using the services. Instead, they need to understand what they are going to share with the world.

B. CONTENT AUTHENTICITY INITIATIVE (CAI)

There have been tools like TinEye used by the journalists for fact-checking, which has an index of ≈ 47 billion images [52], [53]. In line with this discussion, a consortium of big companies such as Adobe, ARM, Qualcomm, New York Times, Microsoft has emerged recently [54]. Since our vision overlaps with this consortium's goals to some extent, we discuss how the CAI process works in light of an example and assert our viewpoints on the CAI process.

1) VERIFICATION PROCESS

Figure 5 illustrates the example operating process towards verifiable content publishing and sharing. According to the figure, a photojournalist takes a photo on the camera. The photo is digitally signed while storing (step 2). The photo is stored along with the metadata. The journalist can also edit the photo using Photoshop and then upload the photo to the content management server (steps 3, 4, 5). The edited photo also carries all the metadata, i.e., the history of editing. When an end-user views the content (step 6), the web page can directly verify while loading. The user can verify the images or videos separately using the consortium website (step 7). Alternately, a user can download an image shared in social media in the name of the consortium members and verify it by uploading it to the consortium verification site. At this moment, the service is only for the consortium parties. Nevertheless, the consortium is growing in size.

2) DISCUSSIONS

CAI aims to ensure that the media produced by the news organizations or consortium parties are not misused or manipulated by third parties. The verification page confirms whether

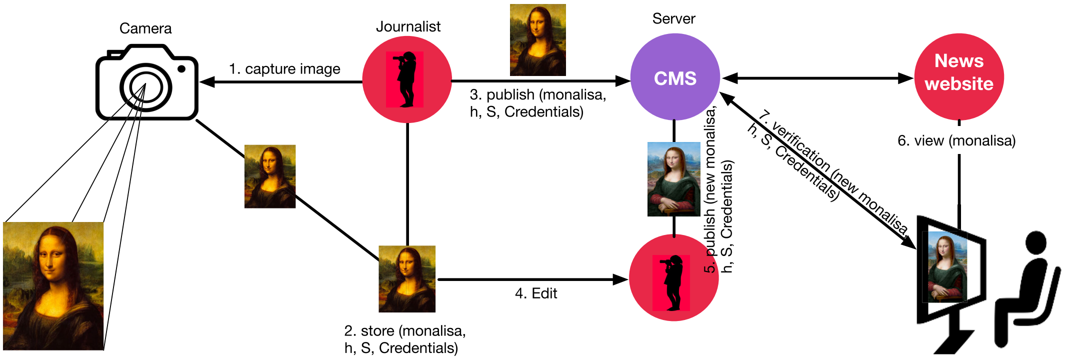

Camera
Journalist
1. capture image
3. publish (monalisa, h, S, Credentials)
Server
CMS
News website
6. view (monalisa)
7. verification (new monalisa, h, S, Credentials)
5. publish (new monalisa, h, S, Credentials)
4. Edit
2. store (monalisa, h, S, Credentials)

FIGURE 5. **The activities of different entities in producing and publishing authentic media in content authenticity initiative CAI.**

the consortium members did not produce a particular media or not. Therefore, CAI enables fact-checking for images, as it is common to modify an image/video from popular news outlets to manipulate the masses. Following, we summarize the limitations of CAI towards preventing forged media on the Internet.

- The news outlets publish fabricated media [10]. Consequently, the consortium members can edit a camera photo, sign with Adobe, and upload it without audit trails. In other words, they can neglect the third step in Figure 5. While the verification process confirms the ownership, the audit trail will not lead to the *original picture or ground truth*. Therefore, media from the members of the consortium does not make it more authentic.
- Only a CPU manufacturer is included in the consortium. The GPU manufacturers are not in the consortium yet. The consortium would also benefit by incorporating different open source tools AI frameworks, such as PyTorch [55], Scitool-learn, Tensorflow [56], Keras [57], Theano [58], MXNET [59], and CNTK [60].
- While the consortium aims to verify images and videos from influential publishers, smartphone users can produce and edit images and videos; however, they cannot upload media in the database of trusted images. There is an entry barrier for individuals with a camera or a smartphone or even for organizations.
- Individual interactions with the news websites are passive (consume content), whereas they can produce content and share in social media platforms. Therefore, the consortium misses the participation of Social Media Platforms (SMPs).

IV. ENABLING AUTHENTIC MEDIA PRODUCTION & SHARING

We envision a proactive approach that aims to prevent the production and sharing of unauthenticated media. The CAI consortium or social media platforms can adopt our approach. Similar to the CAI initiative, the heart of our proposal is the verification process. Our approach aims to satisfy the following requirements to realize this verification process.

1) To track the authenticity of any image/video, its source(s) must be identified. By source, we imply the hardware (e.g., camera) or software used to generate or edit the media.
2) The ownership of the respective image must be guaranteed to ensure that only the authenticated owner of the media can upload the media to a social networking service.
3) To ensure the integrity of an image or video, a historical update trail of the media must be recorded and made available so that *ground truth* can be reached when required. An implicit requirement for such a trail is to guarantee its immutability so that the trail data cannot be forged illicitly.

All in all, we would need to disrupt the current setting and envisage a holistic solution by which image (or video) is created/updated and then uploaded to a social networking service. This holistic solution needs to enable the participation of various entities, employ several cryptographic mechanisms, and manage and disseminate corresponding public keys for each entity. The proposed solution will utilize metadata accompanied by the media. For the solution's security, such metadata and public keys must be stored in a tamperproof manner. In the following, we first describe the involved entities and their envisioned roles. Next, we illustrate how these entities should interact on a blackchain-based system to realize the verifiable image or video on the Internet.

A. STAKEHOLDERS

The above approach requires the participation of various entities and fine-tuning their activities. Here, we present four entities below, along with a brief discussion illustrating the roles of each entity. The relations among these entities are illustrated in Figure 6.

1) OEM (Original EQUIPMENT Manufacturer)

Within the scope of this work, an OEM is the manufacturer of digital cameras with image processing units that can capture photos or videos. However, in contrast to the traditional digital cameras, we assume such a camera is

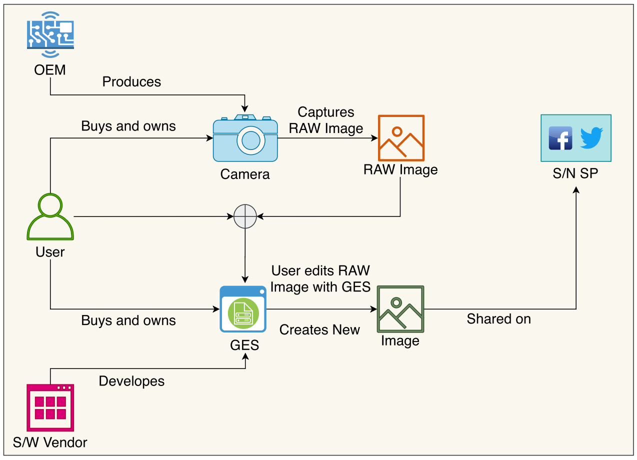


FIGURE 6. Relation among different entities involved in media production and sharing.

equipped with appropriate cryptographic credentials that the camera could use to compute cryptographic functions. For example, a standalone camera can have a trusted execution environment (TEE) to store the keys and perform different cryptographic functions utilizing these keys. TEE also provides boot integrity, secure storage, device identification, isolated execution, and device authentication capabilities [61], which safeguard the camera against various attacks. Modern smartphones are equipped with robust hardware to capture photos as well as videos. Many of them also have trusted environments such as Android Trusty TEE [62] and Security Enclave on iOS [63]. GPUs are extensively used in training deep learning models. For example, NVIDIA provides API support for Tensorflow. Such deep learning tools can use Intel SGX [64] support to sign the deep fakes. Alternatively, the TEEs of GPUs can be used [65].

2) SW VENDOR

A SW (Software) vendor is a software company that develops Graphical Editing Software (GES), which is an image/video producing/editing software, such as camera applications, Photoshop, and deep learning tools. Below we describe several GES and their requirements in creating verifiable authentic media.

a: CAMERA APPLICATIONS

On smartphones, operating systems, and third-party applications can access cameras and microphones based on user permission. Once permitted, all the social media platforms, such as FaceBook, WhatsApp, and Instagram, have camera access. Therefore, these applications should produce authentic images (or videos) with the help of on-device TEE systems. There could be public key infrastructure (PKI) systems for the OS and these SMPs to verify the images or videos whenever the users upload photos or videos. Alternatively, these social networking applications can receive key pairs from their platforms. Multiple platforms can verify each other's images or videos while uploading or sharing. In this case, a user can view or share the media produced by a social media application on other platforms.

b: SOFTWARE FOR CREATIVE ARTS

Apart from these sources, media can also be constructed using various tools or devices, such as Photoshop, Microsoft PowerPoint, augmented reality, virtual reality, and other animation software. These applications may use external media as the overlays on top of camera information or *local* objects. In this case, the local means the objects specifically belong to the environment these tools offer, such as various shapes. The example in Figure 7 shows that one can use local objects and tools to produce a new media object. The software might also incorporate objects from the outside world, such as storage and the Internet. In the case of ''external'' objects, the tool aims to verify the signature of those and performs similar steps as presented in Figure 8. The media record should also contain the metadata of such external objects.

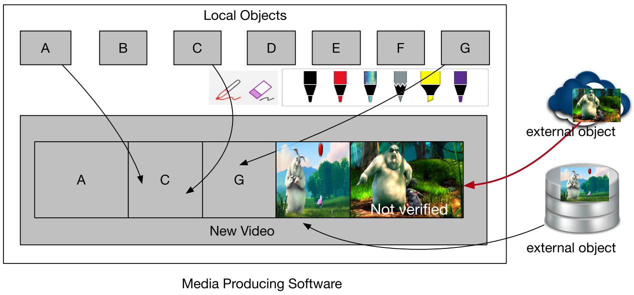


FIGURE 7. A software-based media editing software produces an image or video by using local & external objects.

c: DEEP LEARNING TOOLS

It is hard to verify that a deep fake is not modified after the image (or video) is created by a tool, as there is no deterministic way to confirm the authenticity of the deep fake images produced by different deep learning tools. DeepAttest [66] is an effort where the hardware specifically attests the output from the emerging hardware, such as GPU, FPGA, and ASIC, for the deep neural networks. A better solution would be that they are digitally signed by the respective tool, such as PyTorch, Scitool-learn, Tensorflow, and Keras. The signatures must be inserted as metadata to ensure their authenticity.

d: BROWSER/SOFTWARE FOR UPLOADING

While uploading, the verification process and the translated labels should intrigue a user. The outcomes of such methods should be carefully translated into human-readable messages so that the users can make an informed decision in uploading and sharing. Therefore, the dialog messages can also include ''Might be forged'' or ''Forged'' or ''Deep Fake'' labels. As a result, the user will give a second thought to uploading unauthenticated media. On top of that, if images

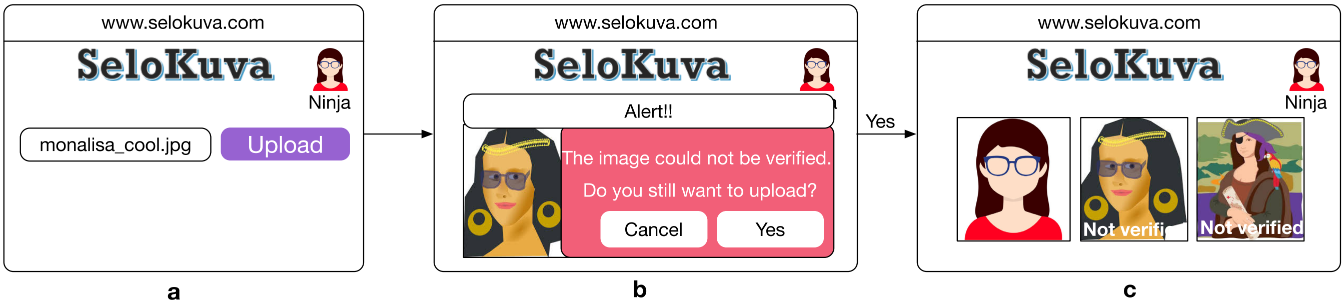


FIGURE 8. (a) Uploading an image to an website of social media platform. (b) When an image is to be uploaded, the browser/application verifies the image. (c) If the browser cannot verify the image, it adds a message, "Not Verified" to be shown by the viewer. The Ninja avatar is taken from freesvg.org and Monalisa images are from Wikimedia commons.

or videos are tagged with the proper messages and presented on social networks or messaging platforms, other users will be reluctant to like and share those as well. For example, these networking and messaging platforms can also add another alert message when the "share" button is clicked like those presented in Figure 4. The pop-up messages will remind users that the content may not be authentic to share. We are optimistic that the users will be reluctant to share, and so the uploaders when their unauthentic contents are not getting attention. Nevertheless, these approaches require behavioral study specific to this problem and the types of labeling to be used. A GES also incorporates the functionalities presented in Figure 8.

3) USER

A user is someone who buys and owns a camera or a GES. The user utilizes the digital camera to capture images. Furthermore, GES can be used by the user to edit the captured image or create new images or videos. Finally, the user can share their captured/created images and videos using SMPs (Social Media Platforms), discussed next.

4) SMP

An SMP (Social Media Platform) facilitates an online service that enables users, among other activities, to upload and share images or videos with others. Examples include WhatsApp, Facebook, Twitter, Snapchat, Instagram, and so on. As mentioned earlier that these SMPs have integrated audio/video/photo capturing functions on their mobile applications. Therefore, these services can have authentic image captures and verification mechanisms when end-users use their applications to produce and share media using a similar process by being adopted by the consortium (see Figure 5).

B. A BLOCKCHAIN-BASED APPROACH

Interestingly, blockchain exhibits several properties which coincide with the properties identified above for the novel system. For example, blockchain intrinsically provides mechanisms for secure tamper-proof storage of data (data immutability) and provenance [67]. Furthermore, a smart-contract supported blockchain can be utilized to deploy an immutable-autonomous program via a smart-contract which could be leveraged to provide the underlying logic for an irrefutable data audit trail. For these reasons, we envision that a blockchain system will be crucial to realize the holistic solution that we would like to propose. In the following, we discuss several approaches for the broader community.

The high-level architecture of the proposed solution is presented in Figure 9. In the center, we have a blockchain system. Three entities, i.e., OEM, SW Vendors, and SMPs, are integrated with the same blockchain system using a corresponding peer node. A user, on the other hand, utilizes the GES to interact with the blockchain system. Similarly, the GES facilitates any interaction between a camera and the blockchain system as required.

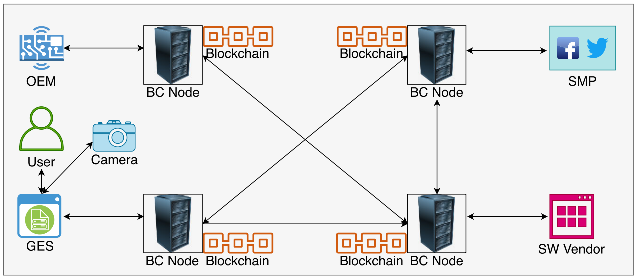


FIGURE 9. Architecture of the proposed solution consisting of a number of entities: OEM, SW Vendor, users and SMPs are interconnected via a blockchain.

In this section, we present three protocol flows illustrating the involvement of different entities. The first two flows demonstrate how a user can assert ownership of the camera and the GES. The third one explores how the proposal can ensure the authenticity and integrity of photos and videos while uploaded to and shared via SMPs.

1) CAMERA OWNERSHIP FLOW

In this flow, we demonstrate (Figure 10) how the camera ownership is assigned when a user buys a digital camera from an OEM or a retailer, which is discussed next. When the OEM produces a new digital camera, a new key pair (public and private key) and a unique identifier for the camera are created (Step 1 in Figure 10). The private key is stored in the TEE of

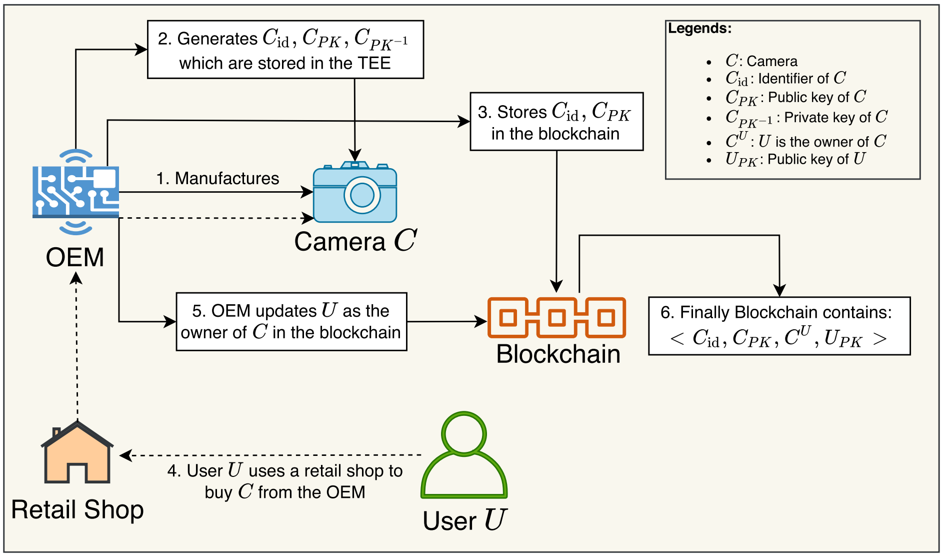

2. Generates $C_{\text{id}}, C_{PK}, C_{PK^{-1}}$ which are stored in the TEE
3. Stores C_{id}, C_{PK} in the blockchain
Legends:
• C: Camera
• C_{id}: Identifier of C
• C_{PK}: Public key of C
• $C_{PK^{-1}}$: Private key of C
• C^U: U is the owner of C
• U_{PK}: Public key of U
1. Manufactures
OEM
Camera C
5. OEM updates U as the owner of C in the blockchain
Blockchain
6. Finally Blockchain contains: $< C_{\text{id}}, C_{PK}, C^U, U_{PK} >$
4. User U uses a retail shop to buy C from the OEM
Retail Shop
User U

FIGURE 10. Flow for storing camera ownership on a blockchain. The number represents the sequence of steps.

the camera (Step 2 in Figure 10). The OEM stores identifier and public key of the camera on the blockchain (Step 3 in Figure 10). When a user buys the camera from the OEM or retailer, the user is assigned as the camera owner, and this information is stored on the blockchain (Steps 4 and 5 in Figure 10). To facilitate this, we assume that the user utilizes a mobile wallet app (or other web interfaces), either provided by the OEM/Retailer or by any open source community. This wallet app also can generate key pairs and the corresponding addresses for the user, which act as identifiers. The ownership information binds the camera identifier with a respective address as provided by the wallet app. We also assume that the user has the provision of exporting their addresses to other compatible software. Once the ownership information is stored in the blockchain, the camera is considered sold to the user (Step 6 in Figure 10).

2) GES OWNERSHIP FLOW

In this flow, we examine the use-cases of buying the GES from an SW Vendor or a retailer similar to the previous use case. First, the user purchases a new GES by providing the vendor/retailer with their address. Like before, the user can utilize a wallet app for this purpose. Next, the SW Vendor creates a new copy of the GES and generates a new key-pair for the respective copy along with a unique identifier. Finally, the vendor stores the identifier and the public key of the GES and ownership information (an association between the GES and the user utilized the GES identifier and the user address) in the blockchain. Once the information is stored in the blockchain, the GES is considered sold to the user.

3) MEDIA OWNERSHIP & SHARING FLOW

Finally, in this section, we present the final scenario, which illustrates (Figures 11 & 12.) what happens when the user utilizes the camera to capture a photo and/or uses the GES to edit (or create) the captured (or a new) photo and consequently uploads the media on an SMP. The protocol flow for sharing media satisfying ownership, authenticity, and integrity camera ownership is discussed next. The flow can be easily modified to create an image with the GES and then share it. It is also assumed that the GES provides an interface for importing key pairs as provided by the users.

At first, we explore the flow for creating a new image which is illustrated in Figure 11. The user captures a photo with the camera (Step 1 in Figure 11). After capturing the image, the camera generates a hash of the picture and signs with the private key. The signature is added as metadata within the photo and the camera information (e.g., identifier) (Steps 2.1, 2.2, and 2.3 in Figure 11). When the user connects the camera to the GES (Step 3.1 in Figure 11), the GES

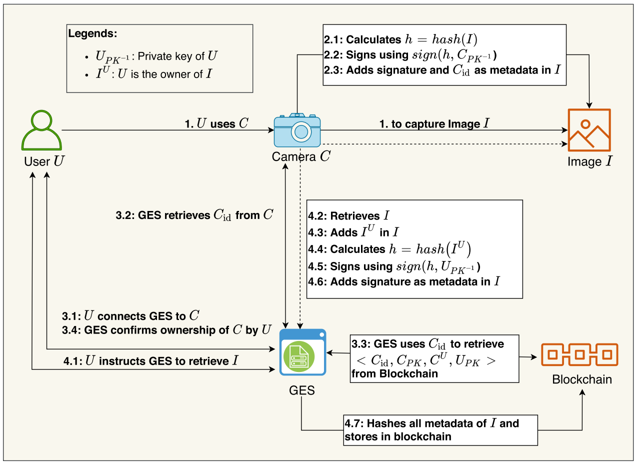

Legends:
• $U_{PK^{-1}}$: Private key of U
• I^U: U is the owner of I
2.1: Calculates $h = hash(I)$
2.2: Signs using $sign(h, C_{PK^{-1}})$
2.3: Adds signature and C_{id} as metadata in I
1. U uses C
1. to capture Image I
User U
Camera C
Image I
3.2: GES retrieves C_{id} from C
4.2: Retrieves I
4.3: Adds I^U in I
4.4: Calculates $h = hash(I^U)$
4.5: Signs using $sign(h, U_{PK^{-1}})$
4.6: Adds signature as metadata in I
3.1: U connects GES to C
3.4: GES confirms ownership of C by U
4.1: U instructs GES to retrieve I
3.3: GES uses C_{id} to retrieve $< C_{\text{id}}, C_{PK}, C^U, U_{PK} >$ from Blockchain
GES
Blockchain
4.7: Hashes all metadata of I and stores in blockchain

FIGURE 11. Flow for creating a new image.

retrieves the camera identifier from the camera (Step 3.2 in Figure 11) and uses that identifier to retrieve the corresponding public key and ownership information from the blockchain (Step 3.3 in Figure 11). The GES then verifies the authenticity and ownership of the camera for the user (Step 3.4 in Figure 11). Once the ownership of the camera is confirmed, the user utilizes the GES to retrieve the captured image (Step 4.1 in Figure 11). The GES generates an ownership signature on behalf of the user and other relevant information, which are then embedded within the image as metadata (Steps 4.2 to 4.6 in Figure 11). Then, the GES generates a hash combining image and the metadata and stores this hash on the blockchain (Step 4.7 in Figure 11).

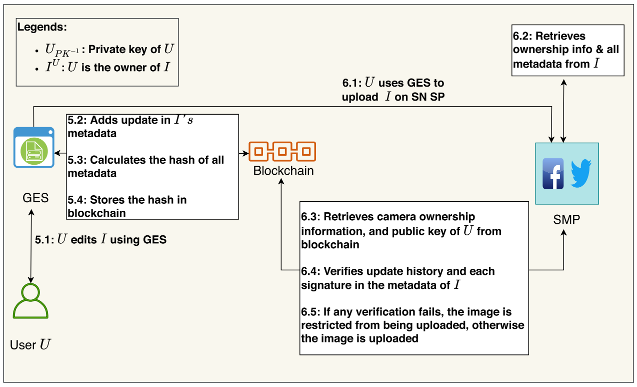

Legends:
• $U_{PK^{-1}}$: Private key of U
• I^U: U is the owner of I
6.2: Retrieves ownership info & all metadata from I
6.1: U uses GES to upload I on SN SP
5.2: Adds update in I's metadata
5.3: Calculates the hash of all metadata
5.4: Stores the hash in blockchain
Blockchain
GES
5.1: U edits I using GES
User U
SMP
6.3: Retrieves camera ownership information, and public key of U from blockchain
6.4: Verifies update history and each signature in the metadata of I
6.5: If any verification fails, the image is restricted from being uploaded, otherwise the image is uploaded

FIGURE 12. **Flow for updating an old image and sharing an image to the SMP.**

Let us consider the scenario when the user edits an image with the GES. The flow is illustrated in Figure 12. In this regard, we assume that the user edits the image with the GES (Step 5.1 in Figure 12). Once the editing is completed, the GES generates another hash by signing the edited image. Next, this signature is used to update the previous signature metadata. The GES generates a hash combining all the metadata and the image. This hash is then stored in the blockchain. This step is repeated whenever the user updates the image using the GES and is presented in Steps 5.2, 5.3, and 5.4 in Figure 12.

Let us consider the scenario when the user uploads the (edited) image to the SMP (Step 6.1 in Figure 12). The platform retrieves the camera identifier, ownership information, and all signature metadata from the image (Step 6.2 in Figure 12). The SMP also retrieves the ownership information regarding the camera and the owner's public key from the blockchain (Step 6.3 in Figure 12). It then verifies the signatures using the public keys of the camera and the owner (Step 6.4 in Figure 12). The platform may also validate the editing history of the image as stored in the blockchain. Suppose any signature verification fails or any discrepancy in the update history stored in the metadata of the image and the blockchain. In that case, the platform can discard the upload by displaying an appropriate message to the user or uploading the image with a proper label, as discussed previously (Step 6.5 in Figure 12).

Let us explore the situation when a user tries to act maliciously, i.e., sharing someone else's image as their image. We assume that the user downloads an image from the Internet, deletes all its metadata, and then uses the GES to create new metadata to confirm their ownership. To mitigate this attack, the GES can be equipped with the capability not to allow a user to import any image without any ownership metadata. Suppose the user uses a camera to capture an image or another type of GES to create an image (i.e., digital art). In that case, ownership data should be there, and the GES would verify the ownership (as illustrated previously). In this way, it can be this situation can be effectively addressed.

V. DISCUSSIONS

We believe that our propositions will gear us forward in this fight against fake or forged media. However, the underlying reason for not adopting such approaches could be the lack of appropriate partners, and the user can make authentic images with numerous tools. Our propositions should be straightforward to implement on modern smartphones, other computing architectures, and creative tools. The additional technological challenges are the following.

4) PRIVACY

Journalist's or artist's privacy is essential for investigative journalism, where confidentiality might be desired in some scenarios. The metadata presented in the media might contain unique identifiers which could essentially identify a user and can be abused by a repressive government or a powerful entity. Therefore, to ensure the privacy of the users, the media producing hardware or software should have mechanisms in place to incorporate privacy mechanisms in a programmable fashion. Simply removing metadata might preserve privacy but might compromise the authenticity of the media. To facilitate this, we envision that an approach could be adapted which would allow the user to choose if the ownership information is removed while uploading the image to the SMP, while preserving the integrity information. Alternatively, a more privacy-preserving option using Zero-knowledge Proof (ZKP) [68] could be adopted, which would eliminate the need to attach any unique identifier to assert ownership over a media. In our future work, we will explore different privacy-preserving approaches in this regard.

5) BLOCKCHAIN PLATFORMS

There are a number of smart-contract supporting public and private blockchain platforms currently available. Examples of public blockchain platforms are Ethereum [69], Cardano [70], Algorand [71], Polkadot [72] and others while examples of private blockchain platforms include Hyperledger Fabric [73], Hyperledger Sawtooth [74], Hyperledger Burrow [75], R3 Corda [76] and others. Even though public blockchain platforms are very secure, they still are quite slow, open to all, and require a high cost to store and process data within their platform [67]. On the other hand, private blockchain platforms are fast, provide reasonable security, and there is no cost associated with them. In addition, the management of a consortium involving different entities can be handled well using a private blockchain platform. That is why we envision the usage of a private blockchain platform utilized by the consortium entities.

6) CONTEXT

While taking photos, it is vital to consider the time of the capture, as the images could have evidentiary value. Only taking system timestamp is not appropriate. Users can configure the local camera or mobile system at any old date, take pictures or videos, and use them as evidence or use such photos as out of context. Therefore such devices or applications need time synchronization either using NTP or GPS clock. If such a facility is absent, the camera/GES does not sign the photo/video. Another critical challenge is that an attacker can use cameras to take very realistic high-resolution images from screens and then claim the captured images as their own. Therefore, we need to have mechanisms inside the camera to detect such screens in real-time and not add signature or ownership metadata within such photos.

VI. CONCLUSION

The threat of fake/edited images or videos is genuine. Apart from human editing, fake image/video generation are real-time and automated. They are already being weaponized against females for target porn. Such activities also allow the perpetrator to deny their actions, merely citing that the image or content is manipulated. The damages already outweigh the benefits. We might have a more challenging future as the existing deep learning approaches heavily relies on deep fake and actual image datasets. An attacker can mix the datasets and produce and disseminate polluted models for more realistic or undetectable deep fakes. Another critical issue that has been neglected is people's privacy, whose faces are initially used to create deep fakes.

Nevertheless, only using digital signatures may not stop the forged media on the Internet unless different entities verify them and assist the users in making an informed moral decision, and the users are not responsible. Similarly, detecting altered images and videos after being viral is less effective in fighting against forged images and videos. Therefore we need a combination of various approaches guiding users in producing and sharing authentic content. Our propositions aim to reduce the entry and transmission of forged media for the Internet, social networking, and messaging platforms.

REFERENCES

[1] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," 2019, *arXiv:1912.04958*.

[2] C. L. Ulloa, M. C. Mora, C. R. Pros, and C. A. Tarrida, "News photography for Facebook: Effects of images on the visual behaviour of readers in three simulated newspaper formats," *Inf. Res.*, vol. 20, p. 660, Mar. 2015.

[3] M. Prior, "Visual political knowledge: A different road to competence?" *J. Politics*, vol. 76, no. 1, pp. 41–57, Jan. 2014.

[4] A. J. Berinsky, "Rumors and health care reform: Experiments in political misinformation," *Brit. J. Political Sci.*, vol. 47, no. 2, pp. 241–262, 2017.

[5] D. Choi, S. Chun, H. Oh, J. Han, and T. Kwon, "Rumor propagation is amplified by echo chambers in social media," *Sci. Rep.*, vol. 10, no. 1, p. 310, Jan. 2020.

[6] L. Coviello, Y. Sohn, A. D. I. Kramer, C. Marlow, M. Franceschetti, N. A. Christakis, and J. H. Fowler, "Detecting emotional contagion in massive social networks," *PLoS ONE*, vol. 9, no. 3, Mar. 2014, Art. no. e90315.

[7] The Logical Indian. (2018). *Fact Check: An Edited Version of This Video From Pakistan Is Driving People To Lynch*. [Online]. Available: https://thelogicalindian.com/fact-check/child-abduction-video-pakistan/?infinitescroll=1

[8] G. Emmanuel. (Jun. 2019). *23-Year-Old Commits Suicide After Her Morphed Pictures Were Circulated on Social Media*. [Online]. Available: https://punemirror.indiatimes.com/news/india/congress-mlas-defection-te%langana-hc-serves-notice-to-assembly-speaker-election-commission/articleshow/6%9741097.cms

[9] M. Burgess. *The AI Telegram BOT That Abused Women is Still Out of Control*. [Online]. Available: https://www.wired.co.UK/article/porn-bots-in-telegram-deepfake

[10] J. Drucker. (Jun. 2020). *Fox News Removes a Digitally Altered Image of Seattle Protests*. [Online]. Available: https://www.nytimes.com/2020/06/13/business/media/fox-news-george-floyd-protests-seattle.html

[11] C. Vaccari and A. Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news," *Social Media Soc.*, vol. 6, no. 1, Jan. 2020, Art. no. 205630512090340, doi: 10.1177/2056305120903408.

[12] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," 2019, *arXiv:1901.08971*.

[13] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2387–2395.

[14] J. Zhang, X. Zeng, M. Wang, Y. Pan, L. Liu, Y. Liu, Y. Ding, and C. Fan, "FReeNet: Multi-identity face reenactment," 2019, *arXiv:1905.11805*.

[15] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7184–7193.

[16] FaceApp Technology Limited. *Faceapp—AI Face Editor*. [Online]. Available: https://apps.apple.com/us/app/faceapp-ai-face-editor/id1180884341

[17] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, Jul. 2017, doi: 10.1145/3072959.3073640.

[18] O. Schwartz. (May 2018). *You Thought Fake News Was Bad?—The Guardian*. [Online]. Available: https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth

[19] B. Kim, J. Choi, and Y. Lee, "Development of facial rejuvenation procedures: Thirty years of clinical experience with face lifts," *Arch. Plastic Surg.*, vol. 42, pp. 521–531, Oct. 2015.

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2018. [Online]. Available: https://openreview.net/forum?id=Hk99zCeAb

[21] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," 2016, *arXiv:1611.09577*.

[22] R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap: An efficient framework for high fidelity face swapping," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2003–2011, doi: 10.1145/3394171.3413630.

[23] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5464–5478, Nov. 2019.

[24] *ModiFace New Video App. Uses Machine Learning To Change Hair Color in Real Time*. Accessed: Oct. 17, 2021. [Online]. Available: https://techthelead.com/modiface-machine-learning-change-hair-color/

[25] R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," California Law Rev., U Texas Law, Public Law Res. Paper No. 692, U Maryland Legal Studies Research, College Park, MD, USA, White Paper2018-21, 2018.

[26] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in *Proc. IEEE Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2017, pp. 1–6.

[27] E. Kee, J. F. O'brien, and H. Farid, "Exposing photo manipulation from shading and shadows," *ACM Trans. Graph.*, vol. 33, no. 5, pp. 1–21, Sep. 2014, doi: 10.1145/2629646.

[28] P. Korshunov and S. Marcel, "DeepFakes: A new threat to face recognition? Assessment and detection," 2018, *arXiv:1812.08685*.

[29] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A large-scale video dataset for forgery detection in human faces," 2018, *arXiv:1803.09179*.

[30] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo, "GAN is a friend or foe?: A framework to detect various fake face images," in *Proc. 34th ACM/SIGAPP Symp. Appl. Comput.*, Apr. 2019, pp. 1296–1303, doi: 10.1145/3297280.3297410.

[31] M. Tarasiou and S. Zafeiriou, "Extracting deep local features to detect manipulated images of human faces," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1821–1825.

[32] X. Yang, Y. Li, H. Qi, and S. Lyu, "Exposing GAN-synthesized faces using landmark locations," in *Proc. ACM Workshop Inf. Hiding Multimedia Secur.*, Jul. 2019, pp. 113–118, doi: 10.1145/3335203.3335724.

[33] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of styleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.

[34] K. Zhang, Y. Liang, J. Zhang, Z. Wang, and X. Li, "No one can escape: A general approach to detect tampered and generated image," *IEEE Access*, vol. 7, pp. 129494–129503, 2019.

[35] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," 2018, *arXiv:1811.00661*.

[36] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu oculi: Exposing AI generated fake face videos by detecting eye blinking," 2018, *arXiv:1806.02877*.

[37] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019, pp. 1–8.

[38] P. Korshunov and S. Marcel, "Speaker inconsistency detection in tampered video," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2375–2379.

[39] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep features perspective on ASVspoof 2019," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 1018–1025.

[40] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," 2019, *arXiv:1907.00501*.

[41] U. A. Ciftci and I. Demir, "FakeCatcher: Detection of synthetic portrait videos using biological signals," 2019, *arXiv:1901.02212*.

[42] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," 2018, *arXiv:1809.00888*.

[43] D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.

[44] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proc. 4th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2016, pp. 5–10.

[45] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "ForensicTransfer: Weakly-supervised domain adaptation for forgery detection," 2018, *arXiv:1812.02510*.

[46] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *Proc. IEEE 10th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2019, pp. 1–8.

[47] D. Yadav and S. Salmani, "Deepfake: A survey on facial forgery technique using generative adversarial network," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICCS)*, May 2019, pp. 852–857.

[48] Y. Mirsky, T. Mahler, I. Shelef, and Y. Elovici, "CT-GAN: Malicious tampering of 3D medical imagery using deep learning," in *Proc. USENIX Secur. Symp.*, 2019, pp. 461–478.

[49] N. Dias, G. Pennycook, and D. G. Rand, "Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media," *Harvard Kennedy School Misinformation Rev.*, to be published. [Online]. Available: https://misinforeview.hks.harvard.edu/article/emphasizing-publishers-does-not-reduce-misinformation/

[50] G. Pennycook, A. Bear, E. T. Collins, and D. G. Rand, "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings," *Manage. Sci.*, vol. 66, no. 11, pp. 4944–4957, Nov. 2020.

[51] P. Mena, "Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook," *Policy Internet*, vol. 12, no. 2, pp. 165–183, Jun. 2020.

[52] P. B. Brandtzaeg, A. Følstad, and M. Á. C. Domínguez, "How journalists and social media users perceive online fact-checking and verification services," *Journalism Pract.*, vol. 12, no. 9, pp. 1109–1129, Oct. 2018, doi: 10.1080/17512786.2017.1363657.

[53] P. B. Brandtzaeg, M. Lüders, J. Spangenberg, L. Rath-Wiggins, and A. Følstad, "Emerging journalistic verification practices concerning social media," *Journalism Pract.*, vol. 10, no. 3, pp. 323–342, Apr. 2016, doi: 10.1080/17512786.2015.1020331.

[54] L. Rosenthol, A. Parsons, E. Scouten, J. Aythora, B. MacCormack, P. England, M. Levallee, J. Dotan, S. Hanna, H. Farid, and S. Gregory. (2020). *The Content Authenticity Initiative Setting the Standard for Digital Content Attribution*. [Online]. Available: https://documentcloud.adobe.com/link/track?uri=urn%3Aaaid%3Ascds%3AUS%3A2c6361d5-b8da-4aca-89bd-1ed66cd22d19#pageNum=2

[55] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035.

[56] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, and S. Ghemawat. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software. [Online]. Available: https://www.tensorflow.org/

[57] François Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[58] R. Al-Rfou, "Theano: A Python framework for fast computation of mathematical expressions," 2016, *arXiv:1605.02688*.

[59] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems," 2015, *arXiv:1512.01274*.

[60] F. Seide and A. Agarwal, "CNTK: Microsoft's open-source deep-learning toolkit," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, p. 2135, doi: 10.1145/2939672.2945397.

[61] J.-E. Ekberg, K. Kostiainen, and N. Asokan, "The untapped potential of trusted execution environments on mobile devices," *IEEE Secur. Privacy*, vol. 12, no. 4, pp. 29–37, Jul. 2014.

[62] *Android TEE*. Accessed: Oct. 17, 2021. [Online]. Available: https://source.android.com/security/trusty

[63] *Secure Enclave*. Accessed: Oct. 17, 2021. [Online]. Available: https://support.apple.com/en-gb/guide/security/sec59b0b31ff/web

[64] V. Costan and S. Devadas, "Intel SGX explained," Cryptol. ePrint Arch., Comput. Sci. Artif. Intell. Lab., Massachusetts Inst. Technol., Tech. Rep. 2016/086, 2016. [Online]. Available: https://eprint.iacr.org/2016/086

[65] S. Volos, K. Vaswani, and R. Bruno, "Graviton: Trusted execution environments on GPUs," in *Proc. 13th USENIX Conf. Operating Syst. Design Implement. (OSDI)*. Berkeley, CA, USA: USENIX Association, 2018, pp. 681–696.

[66] H. Chen, B. D. Rouhani, J. Zhao, and F. Koushanfar, "Deepattest: An end-to-end attestation framework for deep neural networks," in *Proc. ACM/IEEE 46th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2019, pp. 487–498.

[67] M. J. M. Chowdhury, M. S. Ferdous, K. Biswas, N. Chowdhury, A. S. M. Kayes, M. Alazab, and P. Watters, "A comparative analysis of distributed ledger technology platforms," *IEEE Access*, vol. 7, pp. 167930–167943, 2019.

[68] O. Goldreich and Y. Oren, "Definitions and properties of zero-knowledge proof systems," *J. Cryptol.*, vol. 7, no. 1, pp. 1–32, Dec. 1994.

[69] *Ethereum*. Accessed: Oct. 17, 2021. [Online]. Available: https://ethereum.org/en/

[70] *Cardano*. Accessed: Oct. 17, 2021. [Online]. Available: https://cardano.org/

[71] *Algorand*. Accessed: Oct. 17, 2021. [Online]. Available: https://www.algorand.com/futurefi/

[72] *Polkadot*. Accessed: Oct. 17, 2021. [Online]. Available: https://polkadot.network/

[73] *Hyperledger Fabric*. Accessed: Oct. 17, 2021. [Online]. Available: https://www.hyperledger.org/use/fabric

[74] *Hyperledger Sawtooth*. Accessed: Oct. 17, 2021. [Online]. Available: https://www.hyperledger.org/use/sawtooth

[75] *Hyperledger Burrow*. Accessed: Oct. 17, 2021. [Online]. Available: https://www.hyperledger.org/use/hyperledger-burrow

[76] *R3 Corda*. Accessed: Oct. 17, 2021. [Online]. Available: https://www.r3.com/corda-enterprise/

MOHAMMAD A. HOQUE received the M.Sc. degree in computer science and engineering, in 2010, and the Ph.D. degree from Aalto University, in 2013. He is currently a Senior Researcher with the University of Helsinki. His research interests include user-centric pervasive computing and communications, data-driven analysis and optimizations of cyber-physical systems, distributed, and mobile systems.

MD SADEK FERDOUS (Member, IEEE) received the Ph.D. degree in identity management from the University of Glasgow. He is currently an Associate Professor with BRAC University, Dhaka, Bangladesh. He is also a Research Associate with the Centre for Global Finance and Technology, Imperial College Business School, investigating how the blockchain technology can be leveraged to mitigate different attacks. He has several years of experience of working as a Postdoctoral Researcher in different universities in different European and U.K.-funded research projects. He has coauthored more than 60 research papers published in top conferences, journals and symposiums.

MOHSIN KHAN received the B.Sc. degree in computer science and information technology from the Islamic University of Technology, Gazipur, Bangladesh, in 2006, the M.Sc. degree in foundations of advanced computing from Aalto University, Espoo, Finland, in 2015, and the Ph.D. degree in computer science from the University of Helsinki, Helsinki, Finland, in 2021.

From 2007 to 2012, he worked in several roles, such as a System Analyst and Project Manager, at two subsidiaries of Telenor, Bangladesh. From 2013 to 2016, he worked as a Teaching Assistant in various graduate-level courses and as a Research Assistant in several short-term internship programs at Aalto University, University of Helsinki, and Nokia Bell Labs. In the autumn of 2021, he joined Ericsson, Sweden, as an Experienced Researcher in security standardization. He was with the University of Helsinki while contributing to this paper. He is the author of nine articles. His research interests include cellular networks security, applied cryptography, and statistical cryptanalysis.

SASU TARKOMA (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in computer science from the Department of Computer Science, University of Helsinki. He is currently a Professor in computer science with the University of Helsinki and the Head of the Department of Computer Science. He has authored four textbooks and has published over 200 scientific articles. He has nine granted U.S. patents. His research interests include internet technology, distributed systems, data analytics, and mobile and ubiquitous computing. He is a fellow of IET and EAI. His research has received several Best Paper awards and mentions, for example at IEEE PerCom, IEEE ICDCS, ACM CCR, and ACM OSR.