IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 August 2023, accepted 24 August 2023, date of publication 4 September 2023, date of current version 28 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3311461

RESEARCH ARTICLE

Cyber Vaccine for Deepfake Immunity

CHING-CHUN CHANG, HUY H. NGUYEN, JUNICHI YAMAGISHI, AND ISAO ECHIZEN
National Institute of Informatics, Tokyo 101-8430, Japan

Corresponding author: Ching-Chun Chang (ccchang@nii.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP16H06302, Grant JP18H04120, Grant JP20K23355, Grant JP21H04907, and Grant JP21K18023; and in part by the Japan Science and Technology Agency (JST) CREST under Grant JPMJCR18A6 and Grant JPMJCR20D3.

ABSTRACT Deepfakes pose an evolving cybersecurity threat that calls for the development of automated countermeasures. While considerable forensic research has been devoted to the detection and localisation of deepfakes, solutions for 'fake-to-real' reversal are yet to be developed. In this study, we introduce the concept of cyber vaccination for conferring immunity to deepfakes. In other words, we aim to impart a self-healing ability to the face-containing media so that the original content can be recovered after manipulation by AI-based deepfake technology. Analogous to biological vaccination which uses injected antigens to induce immunity prior to infection by an actual pathogen, cyber vaccination simulates deepfakes and performs adversarial training to build a defensive immune system. Aiming to build up attack-agnostic immunity with limited computational resources, we propose simulating various deepfakes with one single overpowering attack: face masking. The proposed immune system consists of a vaccinator for inducing immunity and a neutraliser for recovering facial content. Experimental evaluations demonstrate effective immunity to face replacement and various types of corruption.

INDEX TERMS Cybersecurity, deepfakes, forensics, immunity, vaccine.

I. INTRODUCTION

Deepfakes, as an emergent cyber-security threat, leverage artificial intelligence and machine learning to create synthetic media. This technology is proving a double-edged sword: facilitating innocent entertainment, but also entailing insidious ramifications to the economy, politics and society, including but not limited to market manipulation, electoral influence, nonconsensual pornography, defamatory accusation, evidence fabrication and identity fraud. As the saying goes, 'seeing is believing', and this natural human tendency fuels the spread of disinformation posed by deepfakes. The widespread use and rapid advancement of deepfakes present an evolving challenge to develop countermeasures to enable us to tell fact from fiction [1].

While the term deepfake has been generalised to refer to a broad range of synthetic media nowadays, we focus on one main category that arouses major public concern—facial manipulation [2]. In general, the defence against manipulation centres around three fundamental (step-by-step) tasks: detection, localisation and restoration, as illustrated in Figure 1. Computational forensics has offered fruitful outcomes for detecting whether a given image has been tampered with by deepfake algorithms [3], [4], [5], [6] and for further localising the fake parts [7], [8], [9], [10]. However, solutions for reversing the fraudulent content are yet to be developed. Unlike detection and localisation which can rely on passive diagnostics (observation of abnormality in media after attack), reliable restoration often requires active precautions before the media becomes exposed to public cyberspace, as illustrated in Figure 2. In digital communications, for instance, error correction codes are used as an active precaution to protect a message prior to transmission over a noisy channel. Vaccination, as another example from medicine, injects antigens to trigger an immune response within the body, thereby inducing protective immunity prior to infection by an actual pathogen. These precautionary measures make it possible to restore the original state after the attacks take place. On top of this, the restored content can also serve as a reference for deepfake detection and localisation.

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

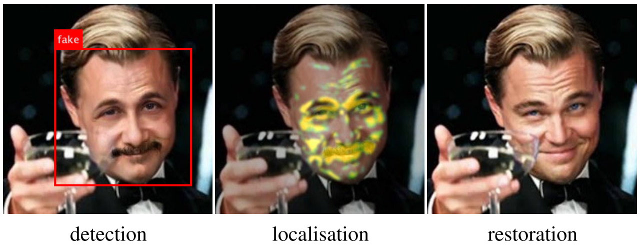


FIGURE 1. Aims of cyber forensics: detection, localisation, restoration.

Adversarial training, similar to the notion of pathogen-specific vaccination, is an intuitive approach combating security vulnerabilities [11], [12], [13]. It is performed by training models in the presence of simulated adversaries so that the computational models, similar to a biological immune system, learn to defend themselves against future attack. While adversarial training can be a potential solution, there are a number of inherent limitations. First of all, preparation and execution of a wide variety of deepfake algorithms can be expensive in terms of (computational) time and resources. On top of this, a vaccine against one specific pathogen typically cannot protect against another, except when the two are very similar. However, deepfakes, as a type of cyber virus, evolve over time, and it is virtually impossible to take every possibility into consideration in practice. For real-world applications, it is important to pursue the capability to defend against various (even unforeseen) attacks, sometimes referred to as attack agnosticism.

The contributions of this study are summarised as follows.

- As far as we are aware, this is one of the first studies to address the problem of deepfake restoration. Revealing the original media content can provide useful forensic clues to combat cyber crimes.
- A novel framework of a cyber immune system based on neural networks is proposed for automatic immunity acquirement. It consists of a vaccinator for inducing immunity, a neutraliser for recovering facial content, a validator for distinguishing between vaccinated and unvaccinated media, and an adversary for simulating deepfakes.
- An attack-agnostic method is developed based on face masking. Instead of exhausting every possible kind of deepfake, we consider a single overwhelming adversary model, the masked-face model, in an attempt to build up attack-agnostic capability with limited computational resources.

II. PRELIMINARIES

The battle between the attacker and the defender is never-ending. To pave the way for the development of further countermeasures, we introduce basic concepts and review relevant literature regarding deepfakes.

A. ATTACK

The term 'deepfake', a portmanteau of deep learning and fake media, is considered to have originated from an anonymous user named 'deepfakes' who posted face-swap videos on a social media platform. Face replacement, as the very first example of deepfakes, creates convincing face-swap videos using an auto-encoder model, which consists of a shared encoder and two decoders for two respective identities (source and target), as illustrated in Figure 3. An auto-encoder is a class of neural network model that uses an encoder to reduce data dimensionality, or to project data into a compact latent space, and a decoder to reconstruct data from the latent features. This process mimics squeezing information through a bottleneck, thereby retaining useful information for prediction. A heuristic idea of the face-swap auto-encoder is that the shared encoder learns to extract features such as facial expressions and poses, whereas each decoder learns to use such features along with invariant features of the corresponding identity to reconstruct the video frame. Once trained, the synthetic source frames are generated by passing the target frames through the shared encoder, while reconstructing the video frames with the decoder of the source identity. A face-swap video is then produced by blending the face region in the synthetic source frames with non-face region in the target frames.

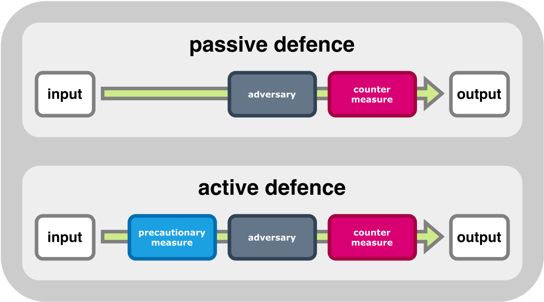


FIGURE 2. Types of defence: passive and active. Passive defence responds solely after the occurrence of attacks, whereas active defence takes precautions before being compromised.

Another example of deepfake is referred to as face reenactment, which turns an identity into a virtual puppet. As the name suggests, this technique manipulates a target individual's facial movements such as expression [14], [15], [16], gaze [17], [18], [19], and head-pose [20], [21], [22], [23], [24]. Mouth reenactment, also known as lip synchronisation, matches a target's lip movements with a vocal audio track [25], [26], [27], [28]. Facial editing or retouching is also a common form of deepfake manipulation that alters the appearance of a target, usually for entertainment [29], [30], [31].

B. DEFENCE

Deepfake artefacts may appear subtle to the human eye, but can often be detected by forensic analysis. For instance, the blending boundary may leave detectable imperfections [32], [33], [34], [35], [36]. Generative deepfake models may also leave statistically identifiable patterns [37], [38], [39]. For deepfake video sequences, temporal inconsistencies are

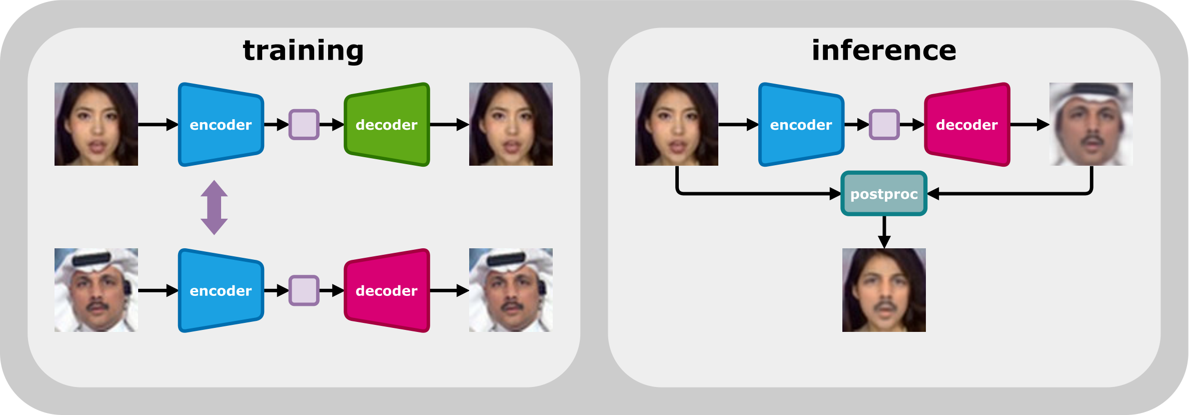


FIGURE 3. **Deepfake autoencoder for face replacement.**

recognisable features [40], [41], [42]. In addition to this, the mismatch of facial landmarks in the face synthesis process can result in an inconsistency in head poses (orientation and position) [43]. Camera fingerprints can also provide clues for distinguishing deepfakes from authentic videos [44]. For audio-to-video synchronisation, there may be anomalous mismatches between the visemes (mouth shapes) and the spoken phonemes (utterances) [45], [46], [47]. Biometric liveness detection is another approach that monitors, for example, irregular eye blinking patterns [48]. To yield further forensic clues, certain methods involve training neural networks to localise the tampered-with areas [49], [50], [51].

III. METHODOLOGY

We begin with a fundamental conceptual model of cyber vaccination and discuss the limitations of a natural solution. We then present the proposed solution as well as the procedures for building an immune system.

A. DEEPFAKE SAMPLING

Cyber vaccination can be viewed as a form of communications. A communication system typically consists of an encoder (at the source) and a decoder (at the destination) [52]. The goal is to accurately transmit a message from the source to the destination over a noisy channel with the help of an encoder/decoder pair. As the communication system employs this pair consisting of an encoder and decoder for error correction, the cyber vaccination system has a corresponding pair consisting of a vaccinator and neutraliser for manipulation reversal. It is possible to train a pair of neural networks jointly with an attack model in the middle. Random sampling of attack models during the training process may lead to adaptability to various hostile conditions. However, there are a number of challenges that may limit the practicality of this approach.

From an engineering standpoint, it is difficult to construct a universal deepfake toolkit containing diverse attack models. Impediments to constructing a universal toolkit include, but are not limited to, different input requirements (e.g. photo size, face position, portrait composition, auxiliary data), different levels of generalisability (i.e. identity-specific or identity-agnostic) and different pre/post-processing procedures. In addition to this, different attack models would cause very different degrees and forms of distortion (e.g. face reenactment and face replacement) and therefore it is arguable whether the training loss can converge within a reasonable time frame. Furthermore, it is a formidable challenge to prove that in-the-lab vaccines can reliably protect against in-the-wild virus variants.

B. FACE MASKING

To overcome these issues, neither restricting ourselves by focusing on a limited number of deepfake algorithms nor exhausting all possibilities, we attempt to consider a single attack model. It should be identity-agnostic and able to be executed in real time. Most importantly, addressing this fatal type of attack implies addressing a variety of types of deepfake attack. A common factor for nearly all deepfake manipulations is that the face region becomes untrustworthy to a greater or lesser extent. As an extreme case, face masking can be an ideal attack model that satisfies all the aforementioned requirements.

If we mask out the face region and attempt to reconstruct it based on the rest of the context information, this becomes an image inpainting problem [53], [54], [55]. Although one could expect a plausible and realistic reconstruction of the missing parts, the original content cannot be recovered with absolute certainty in most cases. If we permit imperceptible modifications prior to face masking, we can apply steganographic algorithms to embed the reconstruction information about the face region into the non-face region. Nonetheless, this steganographic solution (also referred to as self-embedding) often requires non-trivial manual adjustments of parameters when put into practice [56], [57], [58], [59], [60], [61]. Consider that steganographic capacity is limited under an embedding distortion constraint. For

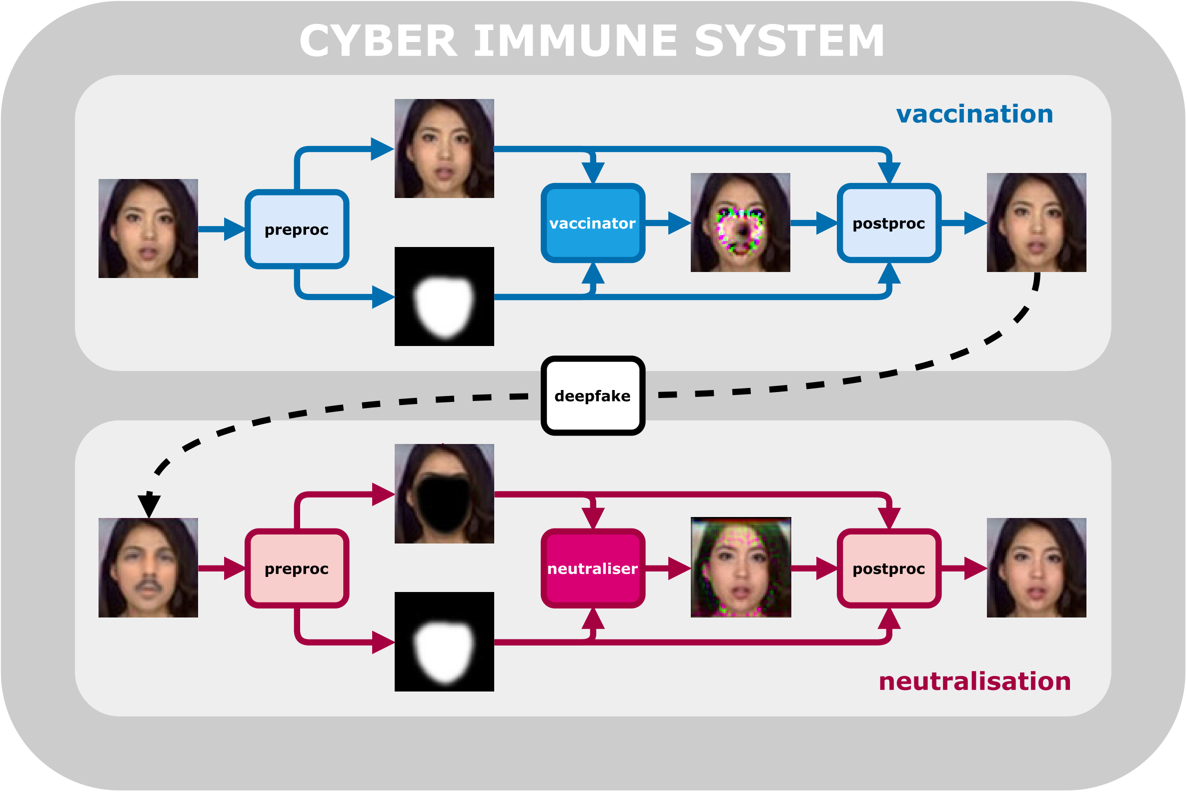


FIGURE 4. Cyber immune system using a vaccinator and neutraliser pair for deepfake restoration.

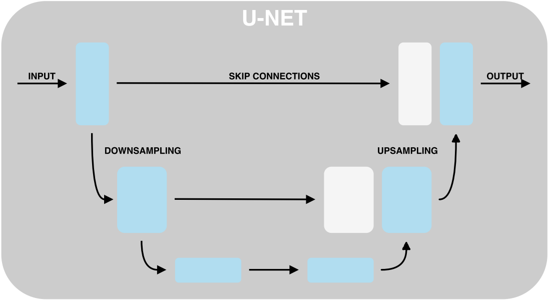
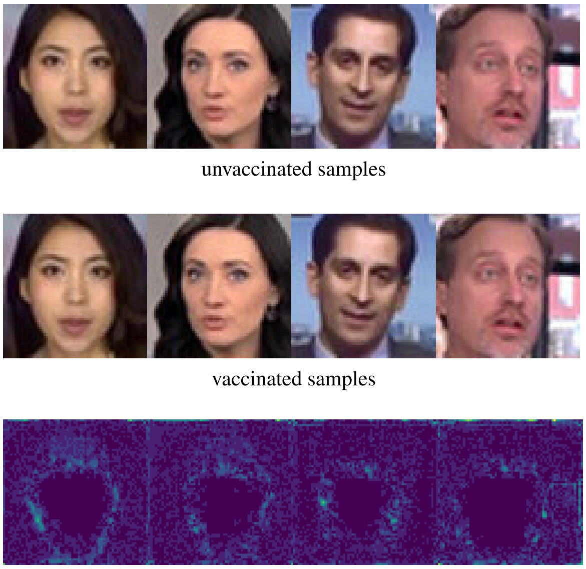


FIGURE 5. U-shaped backbone network architecture for vaccinator and neutraliser.

controlling the amount of data to be embedded below the capacity limit, one may apply source coding to compress the data, which involves a trade-off between code efficiency and reconstruction quality. Consider further that digital images may be processed by a wide range of transformations in blurriness, brightness, contrast, saturation, hue, etc. Due to the fragility of steganography, one may apply channel coding to correct errors to ensure reliable data extraction, which involves a trade-off between code redundancy and correction capability. Moreover, some content-dependent algorithmic parameters may need to be optimised for each individual image. While there are learning-based steganographic methods aiming to embed messages in an automatic and robust manner [62], [63], [64], the message is usually assumed to be a sequence of random binary digits or a secret image and the location of the hidden information cannot be specified. In our context, the message is the face region of the portrait image itself and the location of hidden information is the non-face region. Synchronisation may also be a problem if the face region detected at the encoder side is inconsistent with that at the decoder side.

C. CYBER IMMUNE SYSTEM

Machine learning forges a path to be (mostly) free from manual intervention in parameter configurations. Consider the transmission of a portrait image between a vaccinator and a neutraliser over a face masking channel. Both vaccinator and neutraliser are neural networks and the goal is to preserve the quality of the vaccinated image while ensuring the quality of the neutralised image. Our idea is to embed the information about the face region into the non-face region through imperceptible modification, similar to the steganographic approach. Nonetheless, we do not need to explicitly specify the information to be embedded, nor its location. In addition to this, we need to be able to distinguish between the vaccinated and unvaccinated objects. Furthermore, robustness against common image processing operations would be an appealing feature in practice. Both vaccination and neutralisation are composed



unvaccinated samples

vaccinated samples

residuals

FIGURE 6. Visualisation of residuals between unvaccinated and vaccinated image samples.

```
Algorithm 1 Training (Cyber Immune System)
Input: $x_\circ \in \mathcal{D}$                                   ▷ image from dataset
Output: [Vaccinator, Neutraliser]
  ▷ vaccination
  $m$ = MaskDetector($x_\circ$)                                     ▷ pre-proc.
  $x_\bullet^{\mathrm{raw}}$ = Vaccinator($x_\circ, m$)             ▷ mid-proc.
  $x_\bullet = x_\bullet^{\mathrm{raw}} \cdot m + x_\circ \cdot \bar{m}$      ▷ post-proc.
  $\mathcal{L}_{\mathrm{imp}}$ = Distance($x_\bullet, x_\circ$)     ▷ loss

  ▷ neutralisation
  $m^{\mathrm{rnd}}$ = RandomAffine($m$)                            ▷ pre-proc.
  $x_\bullet^{\mathrm{rnd}}$ = RandomTransform($x_\bullet$)         ▷ augmentation
  ▷ vaccinated case
  $y_\bullet^{\mathrm{raw}}$ = Neutraliser($x_\bullet^{\mathrm{rnd}}, m^{\mathrm{rnd}}$)   ▷ mid-proc.
  $y_\bullet = y_\bullet^{\mathrm{raw}} \cdot m^{\mathrm{rnd}} + x_\bullet \cdot \bar{m}^{\mathrm{rnd}}$   ▷ post-proc.
  $\mathcal{L}_{\mathrm{rev}}$ = Distance($y_\bullet, x_\circ$)     ▷ loss
  ▷ unvaccinated case
  $y_\circ^{\mathrm{raw}}$ = Neutraliser($x_\circ, m^{\mathrm{rnd}}$)   ▷ mid-proc.
  $y_\circ = y_\circ^{\mathrm{raw}} \cdot m^{\mathrm{rnd}} + x_\circ \cdot \bar{m}^{\mathrm{rnd}}$   ▷ post-proc.
  $\mathcal{L}_{\mathrm{val}}$ = Distance($y_\circ, x_\circ \cdot \bar{m}^{\mathrm{rnd}}$)   ▷ loss

  ▷ back-propagation
  $\mathcal{L} = \mathcal{L}_{\mathrm{imp}} + \mathcal{L}_{\mathrm{rev}} + \mathcal{L}_{\mathrm{val}}$   ▷ loss
  Backprop($\mathcal{L}$, [Vaccinator, Neutraliser])                ▷ update
```

```
Algorithm 2 Inference (Vaccination)
Input: $x_\circ$                                                    ▷ image (unvaccinated)
Output: $x_\bullet$                                                 ▷ image (vaccinated)
  $m$ = MaskDetector($x_\circ$)                                     ▷ pre-proc.
  $x_\bullet^{\mathrm{raw}}$ = Vaccinator($x_\circ, m$)             ▷ mid-proc.
  $x_\bullet = x_\bullet^{\mathrm{raw}} \cdot m + x_\circ \cdot \bar{m}$      ▷ post-proc.
```

```
Algorithm 3 Inference (Neutralisation)
Input: $x$                                      ▷ image (attacked/not; vaccinated/not)
Output: $y$                                                         ▷ image (neutralised)
  $m$ = MaskDetector($x$)                                           ▷ pre-proc.
  $y^{\mathrm{raw}}$ = Neutraliser($x \cdot \bar{m}, m$)            ▷ mid-proc.
  $y = y^{\mathrm{raw}} \cdot m + x \cdot \bar{m}$                  ▷ post-proc.
```

of pre-processing, mid-processing and post-processing steps, as illustrated in Figure 4. In our implementation, we use the U-Net as the backbone network architecture [65], as illustrated in Figure 5.

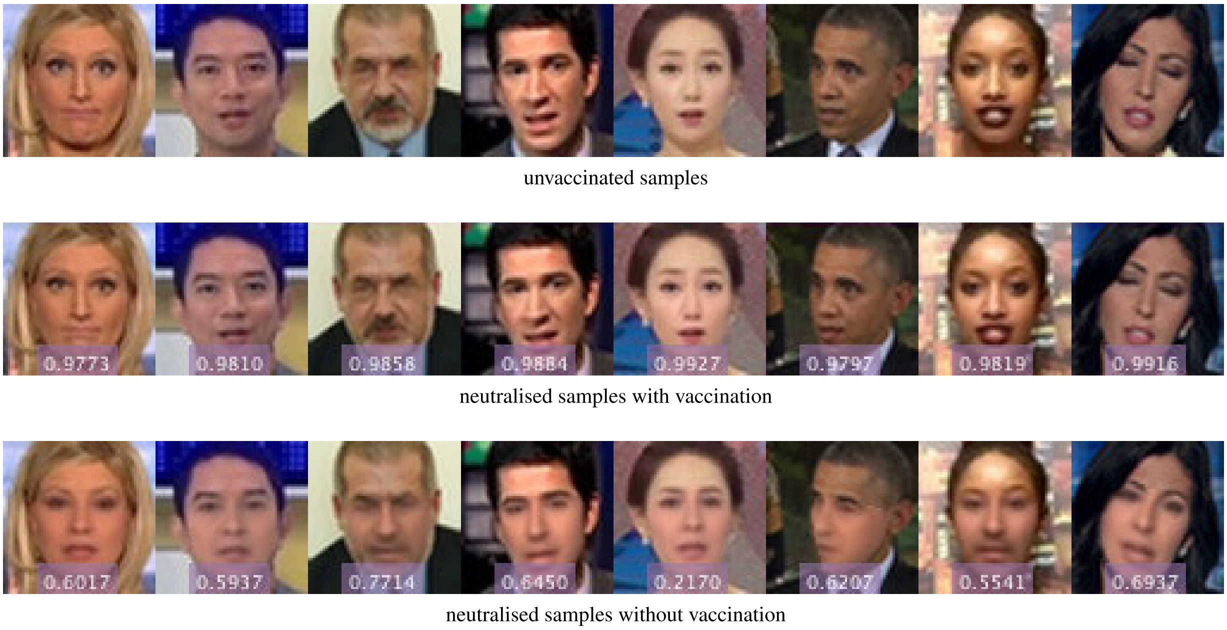
1) VACCINATION

Given video footage or a still image, a preliminary procedure is to detect the face region (with any available face detection algorithms), crop a portrait (for each video frame) and resize it to meet the input resolution specification (e.g. 64 × 64 pixels), leaving aside the case of multiple faces per image for simplicity. The pre-processing step of vaccination prepares a mask by detecting facial landmarks with off-the-shelf face alignment algorithms and assigning binary digits to pixels inside/outside the face contour. The mid-processing step inputs both portrait and mask into the vaccinator and obtains a raw output. The post-processing step produces the vaccinated portrait by substituting the face region of the raw output with that of the original portrait according to the mask. The vaccinated portrait is required to be similar to the original portrait. Since the face region is replaced in the post-processing step, the vaccination process is allowed to introduce imperceptible perturbations only to the non-face region. In this way, the vaccinator can automatically learn to embed the information about the face region into the non-face region. Figure 6 visualises the residuals (differences) between the unvaccinated and vaccinated portraits.

2) NEUTRALISATION

The pre-processing step of neutralisation process prepares a mask and uses it to mask out the face region of the given portrait, which can be either vaccinated or unvaccinated.

In the actual inference stage, the face region of the given portrait may be manipulated by arbitrary deepfake algorithms, which will cause slight misalignment of facial landmarks in the neutralisation process. During the training stage, it is unnecessary to involve deepfake algorithms because this kind of misalignment can be simulated by applying random affine (geometric) transformations to the mask generated in the vaccination process. We also apply random colour transformations to reinforce robustness against common image distortions. The mid-processing step inputs the masked portrait along with the mask into the neutraliser and yields a raw output. The post-processing step merges the face region of the raw output into the masked portrait, resulting in the neutralised portrait. We require the neutralised portrait to


unvaccinated samples

neutralised samples with vaccination

neutralised samples without vaccination

FIGURE 7. Comparison between neutralised image samples with and without vaccination. Numerical data denotes latent-space cosine similarity.

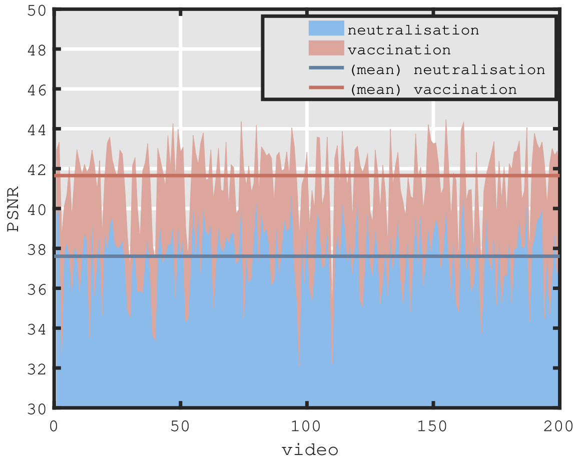


FIGURE 8. Evaluation of imperceptibility and reversibility.

be similar to the original portrait. Since the vaccinator and neutraliser neural networks are trained in a joint manner, the neutraliser learns to use the information embedded imperceptibly in the non-face region to reconstruct the face region.

3) LOSS FUNCTIONS AND VALIDATION

Loss functions are essential for training neural networks. For cyber vaccination, we aim to achieve imperceptibility, reversibility and validatability. We evaluate imperceptibility by the similarity between the non-face region of the vaccinated image and that of the original image, and reversibility by the similarity between the face region of the neutralised image and that of the original image. There are several feasible ways of imparting validatability to the system and one viable option is to make the neutraliser unresponsive to the unvaccinated input. In other words, the output is expected to remain a masked portrait when unvaccinated. It can then be identified effortlessly with the naked eye or automatically with a data-driven visual classifier. Let $\boldsymbol{x}_\circ$ and $\boldsymbol{x}_\bullet$ denote unvaccinated and vaccinated images, and $\boldsymbol{y}_\circ$ and $\boldsymbol{y}_\bullet$ their neutralised counterparts respectively. Also let $\boldsymbol{m}$ denote a mask where the face region is assigned as 1 and the non-face region as 0. Its inverse is denoted by $\bar{\boldsymbol{m}}$ where the binary assignment is opposite. Note that in practice it is not necessary that the mask be composed of binary digits but rather of real numbers with soft edges to provide a smoother blending effect. The loss function for optimising both vaccinator and neutraliser is the (weighted) sum of three loss terms:

$$\mathcal{L} = \mathcal{L}_{\text{imp}}(\boldsymbol{x}_\bullet, \boldsymbol{x}_\circ) + \mathcal{L}_{\text{rev}}(\boldsymbol{y}_\bullet, \boldsymbol{x}_\circ) + \mathcal{L}_{\text{val}}(\boldsymbol{y}_\circ, \boldsymbol{x}_\circ \cdot \bar{\boldsymbol{m}}). \tag{1}$$

In our implementation, each loss term is composed of mean absolute error (MAE) and structural similarity index measure (SSIM). The algorithmic procedures for training and inference (vaccination and neutralisation) are shown in Algorithms 1, 2 and 3. To train an automatic validator to distinguish between vaccinated and unvaccinated images, we use neutralised images as the inputs and binary cross entropy as the loss function.

IV. EVALUATION

We evaluate the proposed cyber vaccination system with respect to imperceptibility, reversibility, robustness and validatability. We also carry out case studies for demonstrating deepfake immunity and discuss potential limitations.

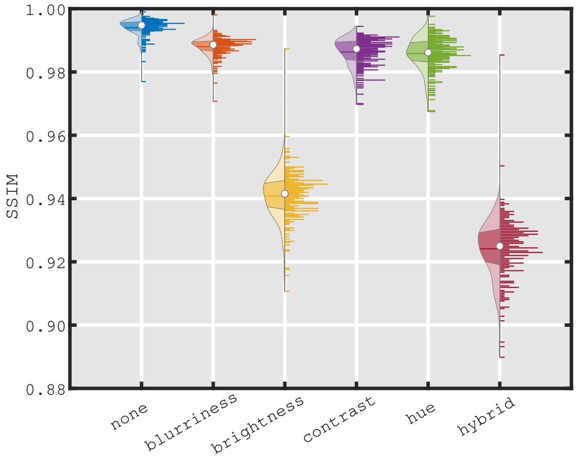


FIGURE 9. Evaluation of robustness with respect to different types of corruption.

A. EXPERIMENTAL SETUP

The models are trained and tested on the FaceForensics++ dataset, consisting of a thousand video clips with trackable faces [66]. For both vaccinator and neutraliser, we use the U-Net as the backbone architecture and apply a state-of-the-art version by OpenAI with residual connection and multi-head attention mechanisms [67], which has been widely used as a diffusion probabilistic model for various computer vision tasks [68], [69], [70]. We specify the number of residual blocks as 3 and the attention resolutions as 4, 8 and 16. The number of parameters is about 83 million for the above specifications. In terms of computational time complexity, the reported running time of the model is approximately 0.15 seconds per frame on the Apple M1 CPU (consumer electronics) without batch processing. For the validator, we test several representative classification models from pioneering to contemporary architectures, including the multi-layer perceptron (MLP) [71], the neural network by Lecun et al. (LeNet) [72], the residual neural network (ResNet) [73], the vision transformer (ViT) [74] and the next-generation convolutional neural network (ConvNeXT) [75]. We test immunity to face replacement with both mask-dependent and mask-independent deepfakes. For the former, we use the original deepfake method and train several pairs of autoencoders to swap between different pairs of identities. For the latter, we employ a pre-trained identity-agnostic SimSwap model [76]. We also test immunity to face reenactment by using a pre-trained X2Face model [77].

B. ABLATION STUDY ON CYBER VACCINE

We evaluate the effects of the cyber vaccine by making comparisons with a variant system without the vaccinator. This variant system consists only of a neutraliser trained to fill up the masked face area in a way similar to image inpainting. To compare the neutralised results with and without vaccination, we measure identity similarity. This is performed by projecting the images into a latent space using the FaceNet and computing the cosine similarity between the latent-space vectors [78]. It can be seen from Figure 7 that although the inpainting-based approach can reconstruct plausible portraits, the results are visually different from the original samples and the identity similarity is much lower than that achieved with the vaccination-based approach. For 200 test samples, the average identity similarity with the vaccine is 0.99 and that without the vaccine is 0.57, suggesting that vaccination generally leads to a close identity similarity.

C. IMPERCEPTIBILITY AND REVERSIBILITY

We require the vaccinator to keep distortion imperceptible and the neutraliser to reconstruct the original content with high fidelity. Imperceptibility and reversibility refer to the visual qualities of vaccinated images and neutralised images respectively in comparison to the original images. Peak signal-to-noise ratio (PSNR) is a common objective assessment of image quality. Figure 8 presents the PSNR values of 200 video samples for which each value is averaged over all the frames. Typical PSNR values in lossy image/video compression are between 30 and 50 decibels (dB) under the condition of 8 bits per colour channel. The mean PSNR values of vaccinated and neutralised videos are above 40 and 35, respectively, indicating acceptable imperceptibility and reversibility.

D. ROBUSTNESS

Robustness is an important consideration when translating a system into practical applications. It concerns the extent to which a system can continue to function despite faults, disruptions and varying conditions. We measure the neutralisation performance by the SSIM of neutralised videos with regard to changes in blurriness, brightness, contrast and hue. Figure 9 compares each case in terms of the distribution of SSIM values. The degradation parameters are randomly sampled within a limited range. The average SSIM values for the cases of no and hybrid corruptions, as expected, are at opposite ends. Among individual cases, brightness adjustment appears to have the most negative effect on the neutralisation performance, whereas other adjustments cause minor fluctuations compared with the no-corruption case. Figure 10 compares each case over the full spectrum of degradation. It can be seen that the system is capable of withstanding major changes in brightness and contrast, while sudden performance drops are observed in the presence of extreme adjustments of blurriness and hue. Overall, the evaluations show that the system can be considered reasonably resilient within a certain range of degradation and reliable for certain types of corruption.

E. VALIDATABILITY

To demonstrate that the vaccinated videos can be automatically validated and readily distinguished from the unvaccinated ones, we apply several neural network classifiers with

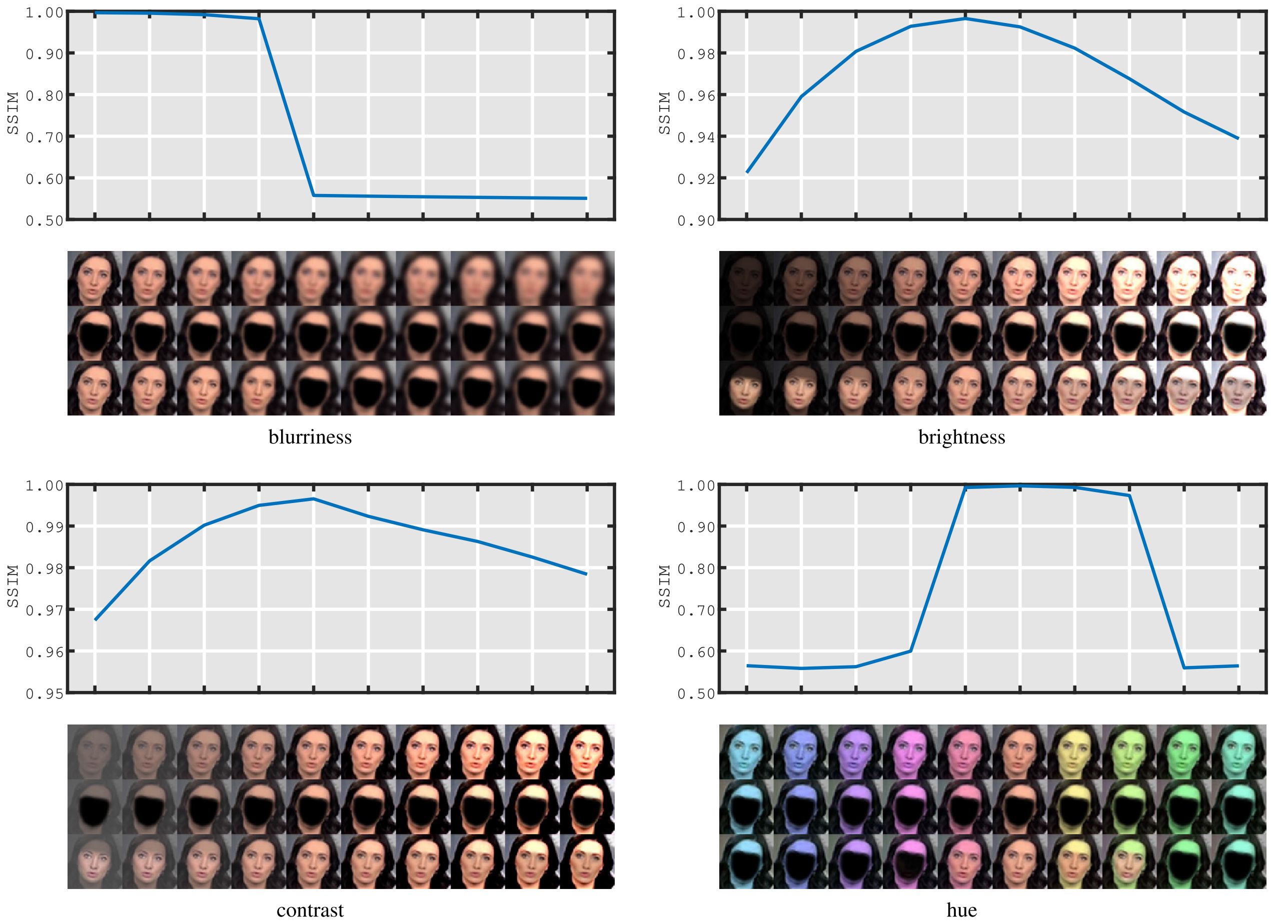


FIGURE 10. **Evaluation of robustness with respect to full spectrum of degradation.**

a wide variety of model size and architectural complexity and evaluate their classification performance across different degradation conditions. As shown in Figure 11, the average accuracy of each classifier is around 99 percent for every condition and the true positive rate is only slightly lower than the true negative rate. In general, it appears that an advanced model with a larger number of parameters and a more complex connection of neurones tends to achieve higher accuracy. Nevertheless, it is remarkable that even a most rudimentary perceptron model can demonstrate classification performance comparable to state-of-the-art models, suggesting that the neutralised results from vaccinated and unvaccinated videos are readily distinguishable.

F. DEEPFAKE IMMUNITY AND LIMITATIONS

We demonstrate immunity to face replacement and face reenactment, which are the types of deepfake that could lead to serious consequences. We evaluate the immunity to both mask-dependent and mask-independent deepfake methods. In general, mask-dependent deepfakes would be easier to cope with since the non-face region is more likely to be kept intact. For mask-dependent face replacement, we employ the classic autoencoder-based method. From Figure 12, we can see that the facial area is restored with high fidelity, providing that the videos are vaccinated prior to attacks. By contrast, an empty facial area is observed for the unvaccinated cases, suggesting that the system shows no immune response to unvaccinated videos. The bounding box annotations on the neutralised frames show the validity of vaccination determined by a validator model. It is worth noting that the system also shows adaptability to head positions and occlusive objects (e.g. eyeglasses). There are nonetheless some limitations such as inaccurate skin tones and mismatched make-up colours, as shown in Figure 13. The former problem may be improved through post-processing by using a more delicate blending mechanism. The latter phenomenon is likely due to unusual colours, namely out-of-distribution data, and hence a possible improvement is to train the models with a greater diversity of data collected from real sources or created artificially. For mask-independent face replacement, we use a pre-trained SimSwap model, which is an identity-agnostic model, being able to swap arbitrary identities with a single model rather than requiring a model for each pair of identities. Figure 14 shows an increase of SSIM scores between the infected and neutralised videos. From the visual examples provided in Figure 15, it can

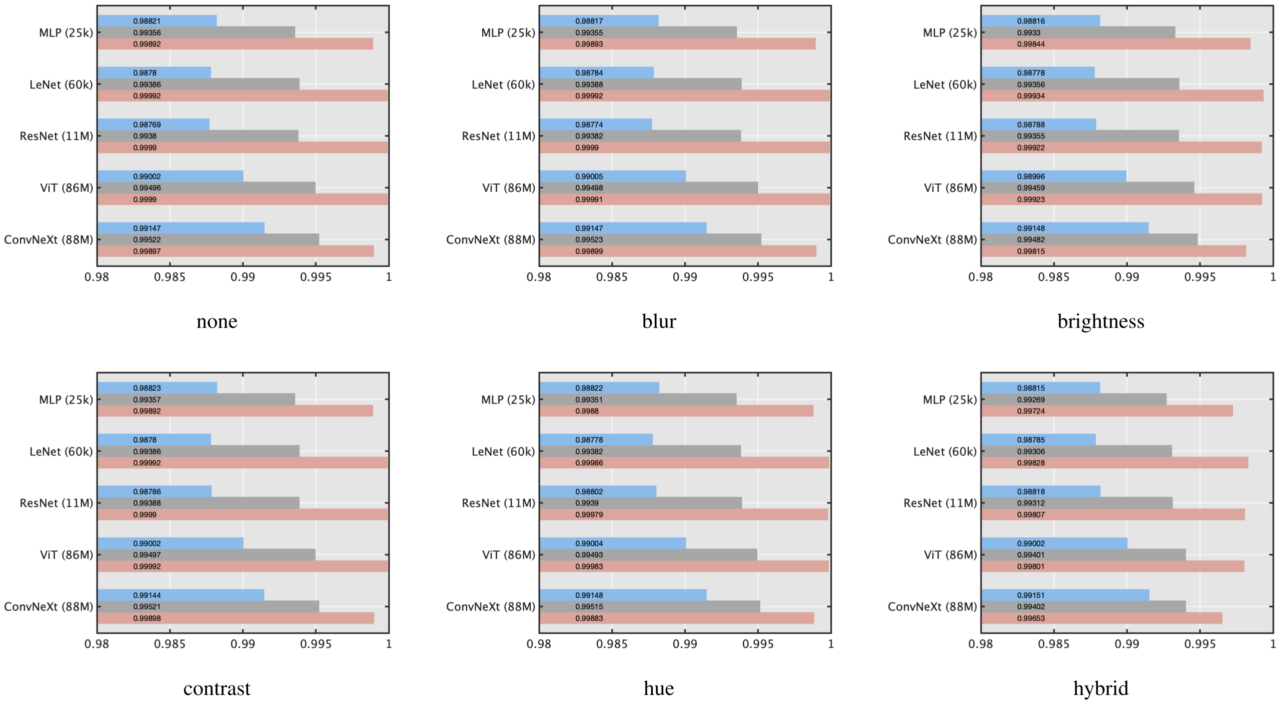


FIGURE 11. **Validatability with different neural network classifiers in terms of true positive rate (top bars), accuracy (middle bars) and true negative rate (bottom bars).**

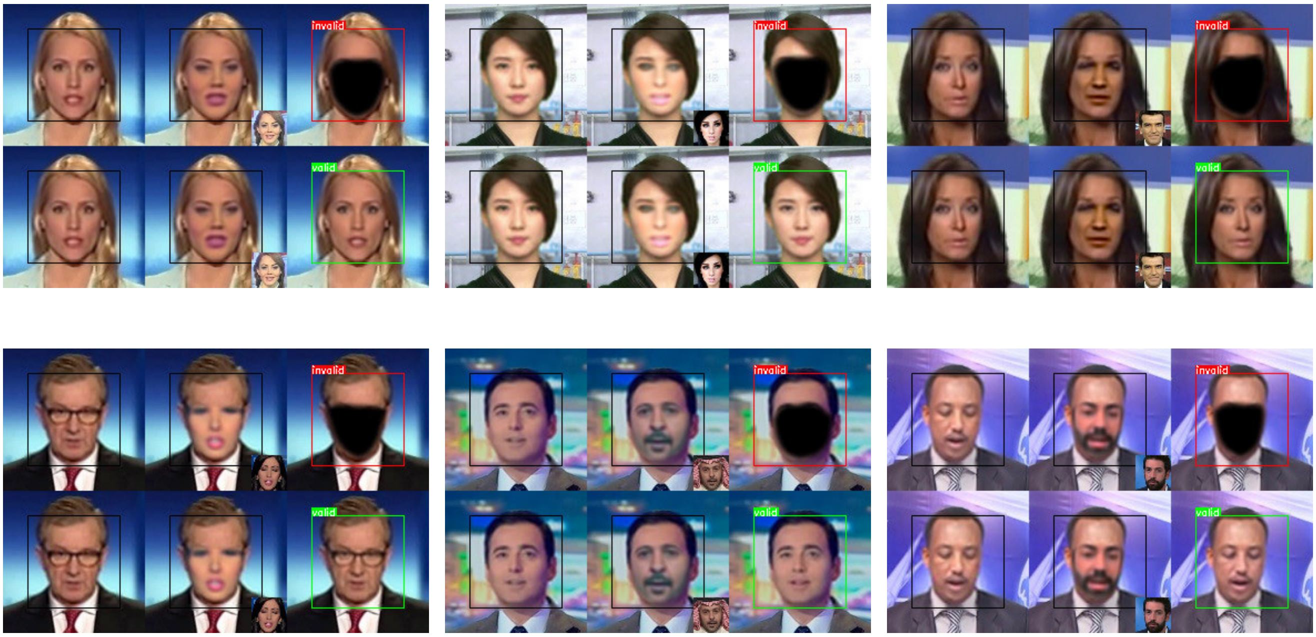


FIGURE 12. **Demonstration of immunity to mask-dependent face replacement by the original deepfakes. Top (from left to right): unvaccinated videos, infected videos and neutralised videos. Bottom (from left to right): vaccinated videos, infected videos and neutralised videos.**

be seen that while the neutralisation is functioning, the reversibility in the mask-independent case is inferior to that in the mask-dependent case, especially in the boundary of the face area (e.g. eyebrows). We also test the performance against face reenactment with a pre-trained X2Face model. In particular, it is interesting to explore the impact of pose changes. The results from Figure 16 suggests that neutralisation would be viable in the case of minor pose changes and yet infections are irreversible when major pose changes are presented.

G. DETECTION AND LOCALISATION

The restored content can also serve as a reference for determining whether the given video contains inauthentic content (deepfake detection) and which parts are likely to be forged (deepfake localisation). Compared with other detection and

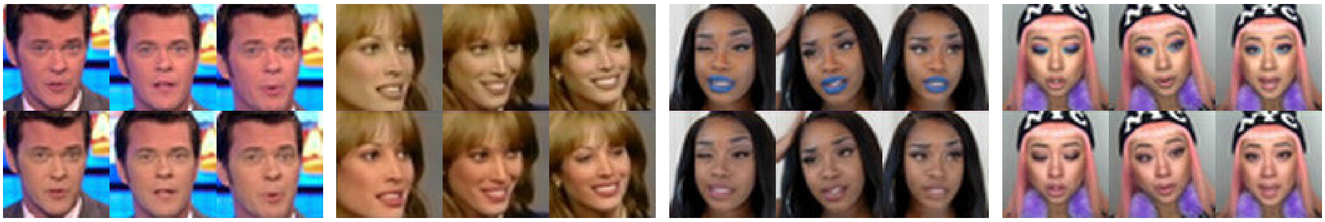


FIGURE 13. Case study on colour misalignment between original videos (top row) and neutralised videos (bottom row).

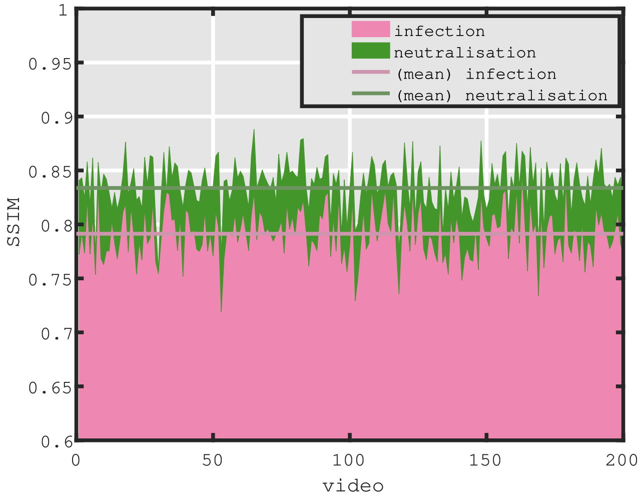


FIGURE 14. Evaluation of reversibility of infections by SimSwap.

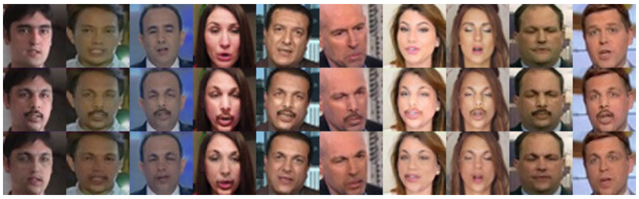


FIGURE 15. Demonstration of immunity to mask-independent face replacement by SwimSwap. Top: original images. Middle: infected images. Bottom: neutralised images.

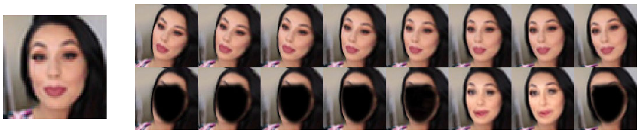
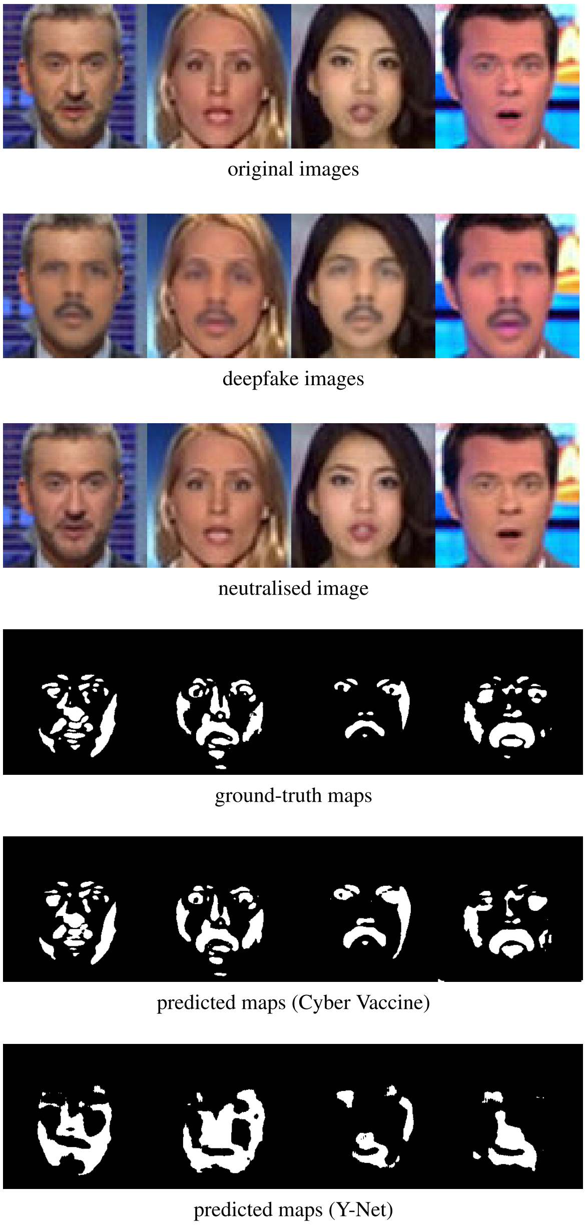


FIGURE 16. Demonstration of immunity to face reenactment by X2Face. Left: vaccinated image. Right: infected images (top row) and neutralised images (bottom row).



original images

deepfake images

neutralised image

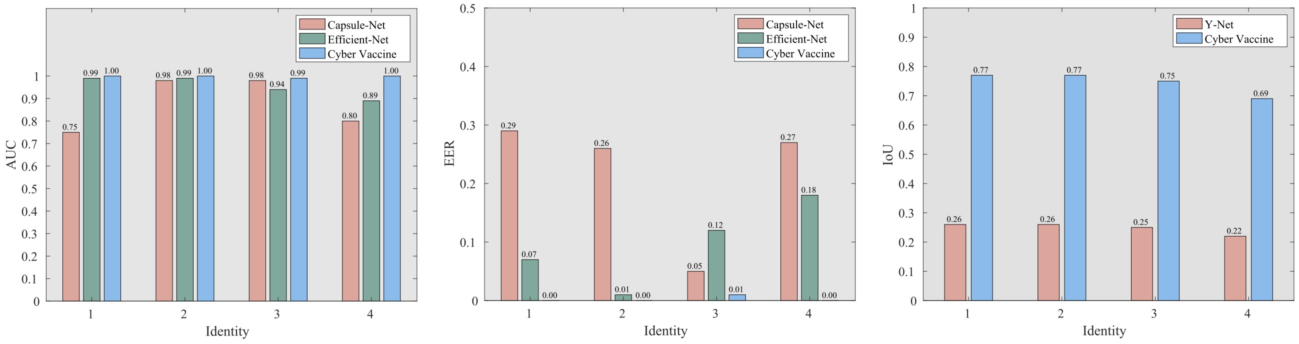
ground-truth maps

predicted maps (Cyber Vaccine)

predicted maps (Y-Net)

FIGURE 17. Demonstration of deepfake localisation using cyber-vaccine method (active) alongside another forensic method (passive).

localisation methods, restoration-based approach is more interpretable (because the inconsistent parts are observable), albeit at the cost of requiring active precautions. For instance, in Figure 17, we can generate a binary map that segments between the real and fake areas by applying a threshold on the differences between the test and neutralised images. In our experiments, the threshold value is derived from the 90th percentile of the difference map. It can be observed that our active approach can generate a map more similar to the ground truth than a passive approach. In Figure 18, we assess the performance of our method in terms of detection and localisation compared to the Capsule-Net [6], Efficient-Net [79] and Y-Net [10]. We use the area under the curve (AUC) and equal error rate (EER) for the detection performance, and the intersection over union (IoU) between the predicted and ground-truth maps for the localisation performance. Our performance testing involves 4 videos,



Detection (AUC) Detection (EER) Localisation (IoU)

FIGURE 18. A comparison of detection and localisation performance between cyber-vaccine method and other forensic methods.

each containing around 200 frames. The forged videos are created using the classic autoencoder-based deepfake algorithm. For our method, the latent-space cosine similarity between the test and neutralised images serves as the indicator of non-deepfakes. Evaluation results indicate that our method surpasses state-of-the-art detection methods, achieving AUC values ranging from 0.99 to 1.00 and EER values ranging from 0.00 to 0.01. For localisation, we select the deepfake video frames on which the Y-Net can successfully detect fake content. Our method achieves an IoU score ranging approximately from 0.7 to 0.8, whereas that of the Y-Net remains below 0.3.

V. CONCLUSION

In this study, we propose a cyber vaccination mechanism for conferring immunity to deepfakes. It is shown that the distortion caused by vaccination is generally imperceptible and effective neutralisation is achieved under various corruption conditions. Furthermore, the validity of vaccination can be readily verified with a wide range of neural network classifiers. There is nonetheless colour misalignment in certain cases, which may be improved through post-processing and data augmentation. There is also room for improvement in the recovery performance on mask-independent face replacement. For face reenactment that causes changes in pose and thus major alterations in the non-face region, novel mechanisms with greater robustness are worth further investigation. We envisage further progress in cyber vaccines for addressing more threats posed by deepfakes 'in the wild'.

REFERENCES

[1] S. Greengard, "Will deepfakes do deep damage?" *Commun. ACM*, vol. 63, no. 1, pp. 17–19, Dec. 2019.

[2] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–41, 2021.

[3] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of GAN-generated fake images over social networks," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Miami, FL, USA, Apr. 2018, pp. 384–389.

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Hong Kong, Dec. 2018, pp. 1–7.

[5] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Hong Kong, Dec. 2018, pp. 1–7.

[6] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 2307–2311.

[7] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and B. S. Manjunath, "Exploiting spatial structure for localizing manipulated image regions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4980–4989.

[8] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1053–1061.

[9] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath, and A. K. Roy-Chowdhury, "Hybrid LSTM and encoder–decoder architecture for detection of image forgeries," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3286–3300, Jul. 2019.

[10] 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)*, Tampa, FL, USA, Sep. 2019, pp. 1–8.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, 2018, pp. 1–23.

[12] F. Tramèr, A. Kurakin, N. Papernot, I. J. Goodfellow, D. Boneh, and P. D. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, 2018, pp. 1–20.

[13] F. Tramèr and D. Boneh, "Adversarial training and robustness for multiple perturbations," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, 2019, pp. 5866–5876.

[14] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–14, Nov. 2015.

[15] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–14, Aug. 2018.

[16] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," *Commun. ACM*, vol. 62, no. 1, pp. 96–104, 2018.

[17] D. Kononenko, Y. Ganin, D. Sungatullina, and V. Lempitsky, "Photorealistic monocular gaze redirection using machine learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2696–2710, Nov. 2018.

[18] Y. Yu, G. Liu, and J.-M. Odobez, "Improving few-shot user-specific gaze adaptation via gaze redirection synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 11929–11938.

[19] Y. Ganin, D. Kononenko, D. Sungatullina, and V. Lempitsky, "DeepWarp: Photorealistic image resynthesis for gaze manipulation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 311–326.

[20] L. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 3007–3021, Dec. 2019.

[21] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable GANs for pose-based human image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 3408–3416.

[22] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing images of humans in unseen poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8340–8348.

[23] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8398–8406.

[24] N. Neverova, R. A. Güler, and I. Kokkinos, "Dense pose transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 128–143.

[25] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, United Kingdom, 2020, pp. 716–731.

[26] 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, Aug. 2017.

[27] A. Jamaludin, J. S. Chung, and A. Zisserman, "You said that: Synthesising talking faces from audio," *Int. J. Comput. Vis.*, vol. 127, nos. 11–12, pp. 1767–1779, Dec. 2019.

[28] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with GANs," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1398–1413, May 2020.

[29] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Neural photo editing with introspective adversarial networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, 2017, pp. 1–15.

[30] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5444–5453.

[31] 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.

[32] Y. Zhang, L. Zheng, and V. L. L. Thing, "Automated face swapping and its detection," in *Proc. IEEE 2nd Int. Conf. Signal Image Process. (ICSIP)*, Singapore, Aug. 2017, pp. 15–19.

[33] A. Agarwal, R. Singh, M. Vatsa, and A. Noore, "SWAPPED! Digital face presentation attack detection via weighted local magnitude pattern," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Denver, CO, USA, Oct. 2017, pp. 659–665.

[34] H. Mo, B. Chen, and W. Luo, "Fake faces identification via convolutional neural network," in *Proc. 6th ACM Workshop Inf. Hiding Multimedia Secur.*, Innsbruck, Austria, Jun. 2018, pp. 43–47.

[35] Z. Akhtar and D. Dasgupta, "A comparative evaluation of local feature descriptors for DeepFakes detection," in *Proc. IEEE Int. Symp. Technol. Homeland Secur. (HST)*, Woburn, MA, USA, Nov. 2019, pp. 1–5.

[36] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 5000–5009.

[37] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs leave artificial fingerprints?" in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, San Jose, CA, USA, Mar. 2019, pp. 506–511.

[38] N. Yu, L. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 7555–7565.

[39] 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. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 7887–7896.

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

[41] A. Chintha, B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright, and R. Ptucha, "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 1024–1037, Aug. 2020.

[42] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K., 2020, pp. 667–684.

[43] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 8261–8265.

[44] F. Lugstein, S. Baier, G. Bachinger, and A. Uhl, "PRNU-based deepfake detection," in *Proc. ACM Workshop Inf. Hiding Multimedia Sec. (IH&MMSec)*, Virtual Event, Belgium, 2021, pp. 7–12.

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

[46] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting deep-fake videos from phoneme-viseme mismatches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 2814–2822.

[47] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 5037–5047.

[48] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[49] M. Du, S. Pentyala, Y. Li, and X. Hu, "Towards generalizable deepfake detection with locality-aware autoencoder," in *Proc. ACM Int. Conf. Inf. Knowl. Manag. (CIKM)*, Virtual Event, Ireland, 2020, pp. 325–334.

[50] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 5780–5789.

[51] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Face forensics in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 5774–5784.

[52] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[53] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. Annu. Conf. Comput. Graph. Interact. Tech. (SIGGRAPH)*, New Orleans, LA, USA, 2000, pp. 417–424.

[54] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, 2017, Art. no. 107.

[55] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5892–5900.

[56] J. Fridrich and M. Goljan, "Images with self-correcting capabilities," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Kobe, Japan, vol. 3, Oct. 1999, pp. 792–796.

[57] X. Zhang and S. Wang, "Fragile watermarking with error-free restoration capability," *IEEE Trans. Multimedia*, vol. 10, no. 8, pp. 1490–1499, Dec. 2008.

[58] H. He, F. Chen, H.-M. Tai, T. Kalker, and J. Zhang, "Performance analysis of a block-neighborhood-based self-recovery fragile watermarking scheme," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 1, pp. 185–196, Feb. 2012.

[59] C. Qin, C.-C. Chang, and P.-Y. Chen, "Self-embedding fragile watermarking with restoration capability based on adaptive bit allocation mechanism," *Signal Process.*, vol. 92, no. 4, pp. 1137–1150, Apr. 2012.

[60] P. Korus and A. Dziech, "Efficient method for content reconstruction with self-embedding," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1134–1147, Mar. 2013.

[61] C. Qin, P. Ji, C.-C. Chang, J. Dong, and X. Sun, "Non-uniform watermark sharing based on optimal iterative BTC for image tampering recovery," *IEEE MultimediaMag.*, vol. 25, no. 3, pp. 36–48, Jul. 2018.

[62] S. Baluja, "Hiding images in plain sight: Deep steganography," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, pp. 2066–2076.

[63] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "HiDDeN: Hiding data with deep networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 682–697.

[64] M. Tancik, B. Mildenhall, and R. Ng, "StegaStamp: Invisible hyperlinks in physical photographs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 2114–2123.

[65] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Cham, Switzerland, 2015, pp. 234–241.

[66] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 1–11.

[67] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 8162–8171.

[68] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, 2020, pp. 6840–6851.

[69] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 10674–10685.

[70] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2023.

[71] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362.

[72] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[74] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–21.

[75] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 11966–11976.

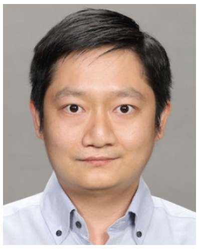
[76] R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap: An efficient framework for high fidelity face swapping," in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, Oct. 2020, pp. 2003–2011.

[77] O. Wiles, A. S. Koepke, and A. Zisserman, "X2Face: A network for controlling face generation using images, audio, and pose codes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 690–706.

[78] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[79] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 18699–18708.

CHING-CHUN CHANG received his PhD in Computer Science from the University of Warwick, UK, in 2019. He participated in a short-term scientific mission supported by European Cooperation in Science and Technology Actions at the Faculty of Computer Science, Otto-von-Guericke-Universität Magdeburg, Germany, in 2016. He was granted the Marie-Curie fellowship and participated in a research and innovation staff exchange scheme supported by Marie Skłodowska-Curie Actions at the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, USA, in 2017. He was a Visiting Scholar with the School of Computer and Mathematics, Charles Sturt University, Australia, in 2018, and with the School of Information Technology, Deakin University, Australia, in 2019. He was a Research Fellow with the Department of Electronic Engineering, Tsinghua University, China, in 2020. He is currently a Postdoctoral Research Fellow with the National Institute of Informatics, Japan. His research interests include steganography, forensics, biometrics, cybersecurity, computer vision, computational linguistics and artificial intelligence.



HUY H. NGUYEN received his PhD from the Graduate University for Advanced Studies (SOKENDAI) Japan, in 2022. He is currently a Specially Appointed Assistant Professor with the National Institute of Informatics, Tokyo, Japan. His research interests include security and privacy in biometrics and machine learning.

JUNICHI YAMAGISHI received his PhD from the Tokyo Institute of Technology (Tokyo Tech), Tokyo, Japan, in 2006. From 2007 to 2013, he was a research fellow in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, UK. He was appointed Associate Professor at the National Institute of Informatics, Japan in 2013 and is currently a Professor at the same institution. His research topics include speech processing, machine learning, signal processing, biometrics, digital media cloning, and media forensics. He served previously as co-organiser for the bi-annual ASVspoof Challenge and the bi-annual Voice Conversion Challenge. He also served as a member of the IEEE Speech and Language Technical Committee (2013–2019), as an Associate Editor of the *IEEE/ACM Transactions on Audio Speech and Language Processing* (2014–2017), and as a Chairperson of International Speech Communication Association (ISCA) SynSIG (2017–2021). He is currently a Principal Investigator of the CREST VoicePersonae project supported by the Japan Science and Technology Agency (JST) and Agence Nationale de la Recherche (ANR), and a Senior Area Editor of the *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

ISAO ECHIZEN received BS, MS, and DE degrees from the Tokyo Institute of Technology, Japan, in 1995, 1997 and 2003, respectively. He joined Hitachi, Ltd. in 1997 and until 2007 was a Research Engineer in the company's systems development laboratory. He is currently a Director and Professor of the Information and Society Research Division, as well as a Director of the Global Research Center for Synthetic Media, at the National Institute of Informatics; a Professor in the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, the University of Tokyo; and a Professor in the Graduate Institute for Advanced Studies, the Graduate University For Advanced Studies (SOKENDAI), Japan. He was a Visiting Professor at the Tsuda University, Japan; at the University of Freiburg, Germany; and at the University of Halle-Wittenberg, Germany. He is currently engaged in research on multimedia security and multimedia forensics, serving as a Research Director in the CREST FakeMedia project, Japan Science and Technology Agency (JST). He received the Best Paper Award from the IPSJ in 2005 and 2014; the Fujio Frontier Award and the Image Electronics Technology Award in 2010; the One of the Best Papers Award from the Information Security and Privacy Conference in 2011; the IPSJ Nagao Special Researcher Award in 2011; the DOCOMO Mobile Science Award in 2014; the Information Security Cultural Award in 2016; the IEEE Workshop on Information Forensics and Security Best Paper Award in 2017; and the Best Paper Award from the IEICE in 2023. He was a Member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society. He is an IEICE Fellow, IEEE Senior Member, and Japanese Representative on IFIP TC11 (Security and Privacy Protection in Information Processing Systems), Member-at-Large of the Board of Governors of APSIPA, and Editorial Board Member of the *IEEE Transactions on Dependable and Secure Computing*, *EURASIP Journal on Image and Video Processing*, and *Elsevier Journal of Information Security and Applications*.