Deep Learning for Deepfakes Creation and Detection: A Survey

Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Saeid Nahavandi, *Fellow, IEEE*

***Abstract*—Deep learning has been successfully applied to solve various complex problems ranging from big data analytics to computer vision and human-level control. Deep learning advances however have also been employed to create software that can cause threats to privacy, democracy and national security. One of those deep learning-powered applications recently emerged is "deepfake". Deepfake algorithms can create fake images and videos that humans cannot distinguish them from authentic ones. The proposal of technologies that can automatically detect and assess the integrity of digital visual media is therefore indispensable. This paper presents a survey of algorithms used to create deepfakes and, more importantly, methods proposed to detect deepfakes in the literature to date. We present extensive discussions on challenges, research trends and directions related to deepfake technologies. By reviewing the background of deepfakes and state-of-the-art deepfake detection methods, this study provides a comprehensive overview of deepfake techniques and facilitates the development of new and more robust methods to deal with the increasingly challenging deepfakes.**

***Index Terms*—survey, review, deepfakes, artificial intelligence, deep learning, computer vision, autoencoders, forensics, GAN, generative adversarial networks.**

I. Introduction

Deepfake (stemming from "deep learning" and "fake") is a technique that can superimpose face images of a target person to a video of a source person to create a video of the target person doing or saying things the source person does. The underlying mechanism for deepfake creation is deep learning models such as autoencoders and generative adversarial networks, which have been applied widely in the computer vision domain [1]–[7]. These models are used to examine facial expressions and movements of a person and synthesize facial images of another person making analogous expressions and movements [8]. Deepfake methods normally require a large amount of image and video data to train models to create photo-realistic images and videos. As public figures such as celebrities and politicians may have a large number of videos and images available online, they are initial targets of deepfakes. Deepfakes were used to swap faces of celebrities or politicians to bodies in porn images and videos. The first deepfake video emerged in 2017 where face of a celebrity was swapped to face of a porn actor. It is threatening to world security when deepfake methods can be employed to create videos of world leaders with fake speeches for falsification purposes [9], [10]. Deepfakes therefore can be abused to cause political or religion tensions between countries, to fool public and affect results in election campaigns, or create chaos in financial markets by creating fake news [11]. It can be even used to generate fake satellite images of the Earth to contain objects that do not really exist to confuse military analysts, e.g., creating a fake bridge across a river although there is no such a bridge in reality. This can mislead a troop who have been guided to cross the bridge in a battle [12], [13].

There is also positive use of deepfakes such as creating voices of those who have lost theirs or updating episodes of movies without reshooting them [14]. However, the number of malicious uses of deepfakes largely dominates that of the positive ones. The development of advanced deep networks and the availability of large amount of data have made the forged images and videos almost indistinguishable to humans and even to sophisticated computer algorithms. The process of creating those manipulated images and videos is also much simpler today as it needs as little as an identity photo or a short video of a target individual. Less and less effort is required to produce a stunningly convincing tempered footage. Recent advances can even create a deepfake with just a still image [15]. Deepfakes therefore can be a threat affecting not only public figures but also ordinary people. For example, a voice deepfake was used to scam a CEO out of $243,000 [16]. A recent release of a software called DeepNude shows more disturbing threats as it can transform a person to a non-consensual porn [17]. Likewise, the Chinese app Zao has gone viral lately as less-skilled users can swap their faces onto bodies of movie stars and insert themselves into well-known movies and TV clips [18]. These forms of falsification create a huge threat to violation of privacy and identity, and affect many aspects of human lives.

Finding the truth in digital domain therefore has become increasingly critical. It is even more challenging when dealing with deepfakes as they are majorly used to serve malicious purposes and almost anyone can create deepfakes these days using existing deepfake tools. Thus far, there have been numerous methods proposed to detect deepfakes [19]–[23]. Most of them are based on deep learning, and thus a battle between malicious and positive uses of deep learning methods has been arising. To address the threat of face-swapping technology or deepfakes, the United States Defense Advanced Research Projects Agency (DARPA) initiated a research scheme in media forensics (named Media Forensics or MediFor) to accelerate the development of fake digital visual media detection methods [24]. Recently, Facebook Inc.

T. T. Nguyen, D. T. Nguyen and D. T. Nguyen are with the School of Information Technology, Deakin University, Victoria, Australia, 3125.

C. M. Nguyen is with the School of Engineering, Deakin University, Victoria, Australia, 3216.

S. Nahavandi is with the Institute for Intelligent Systems Research and Innovation, Deakin University, Victoria, Australia, 3216.

Corresponding e-mail: thanh.nguyen@deakin.edu.au (T. T. Nguyen).

teaming up with Microsoft Corp and the Partnership on AI coalition have launched the Deepfake Detection Challenge to catalyse more research and development in detecting and preventing deepfakes from being used to mislead viewers [25]. Data obtained from https://app.dimensions.ai at the end of July 2020 show that the number of deepfake papers has increased significantly in recent years (Fig. 1). Although the obtained numbers of deepfake papers may be lower than actual numbers but the research trend of this topic is obviously increasing.

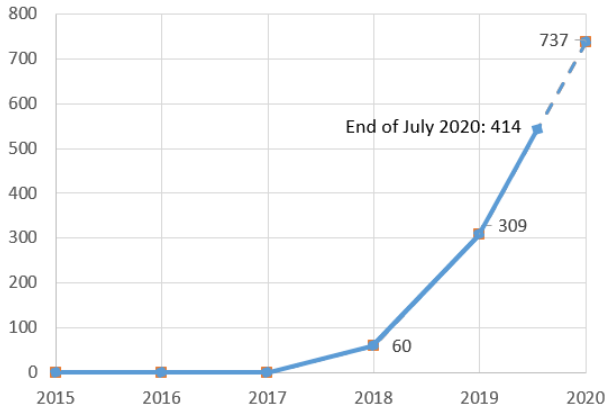

Fig. 1. Number of papers related to deepfakes in years from 2015 to 2020, obtained from https://app.dimensions.ai on 24 July 2020 with the search keyword "deepfake" applied to full text of scholarly papers. The number of such papers in 2018 and 2019 are 60 and 309, respectively. From the beginning of 2020 to near the end of July 2020, there are 414 papers about deepfakes and we linearly estimate that this number will be rising to more than 730 until the end of 2020.

This paper presents a survey of methods for creating as well as detecting deepfakes. In Section II, we present the principles of deepfake algorithms and how deep learning has been used to enable such disruptive technologies. Section III reviews different methods for detecting deepfakes as well as their advantages and disadvantages. We discuss challenges, research trends and directions on deepfake detection and multimedia forensics problems in Section IV.

II. Deepfake Creation

Deepfakes have become popular due to the quality of tampered videos and also the easy-to-use ability of their applications to a wide range of users with various computer skills from professional to novice. These applications are mostly developed based on deep learning techniques. Deep learning is well known for its capability of representing complex and high-dimensional data. One variant of the deep networks with that capability is deep autoencoders, which have been widely applied for dimensionality reduction and image compression [26]–[28]. The first attempt of deepfake creation was FakeApp, developed by a Reddit user using autoencoder-decoder pairing structure [29], [30]. In that method, the autoencoder extracts latent features of face images and the decoder is used to reconstruct the face images. To swap faces between source images and target images, there is a need of two encoder-decoder pairs where each pair is used to train on an image set, and the encoder's parameters are shared between two network pairs. In other words, two pairs have the same encoder network. This strategy enables the common encoder to find and learn the similarity between two sets of face images, which are relatively unchallenging because faces normally have similar features such as eyes, nose, mouth positions. Fig. 2 shows a deepfake creation process where the feature set of face A is connected with the decoder B to reconstruct face B from the original face A. This approach is applied in several works such as DeepFaceLab [31], DFaker [32], DeepFake_tf (tensorflow-based deepfakes) [33].

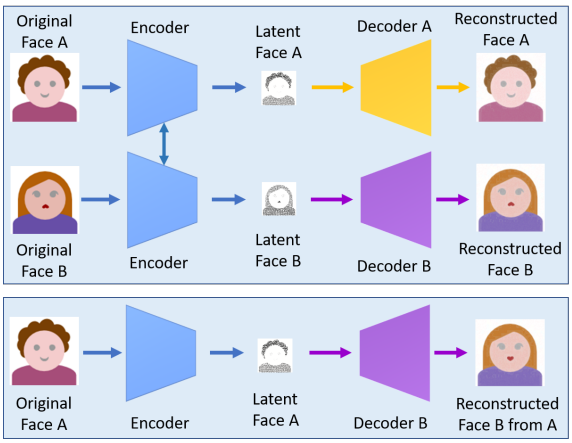

Fig. 2. A deepfake creation model using two encoder-decoder pairs. Two networks use the same encoder but different decoders for training process (top). An image of face A is encoded with the common encoder and decoded with decoder B to create a deepfake (bottom).

By adding adversarial loss and perceptual loss implemented in VGGFace [34] to the encoder-decoder architecture, an improved version of deepfakes based on the generative adversarial network (GAN) [35], i.e. faceswap-GAN, was proposed in [36]. The VGGFace perceptual loss is added to make eye movements to be more realistic and consistent with input faces and help to smooth out artifacts in segmentation mask, leading to higher quality output videos. This model facilitates the creation of outputs with 64x64, 128x128, and 256x256 resolutions. In addition, the multi-task convolutional neural network (CNN) from the FaceNet implementation [37] is introduced to make face detection more stable and face alignment more reliable. The CycleGAN [38] is utilized for generative network implementation. Popular deepfake tools and their features are summarized in Table I.

III. Deepfake Detection

Deepfakes are increasingly detrimental to privacy, society security and democracy [43]. Methods for detecting deepfakes have been proposed as soon as this threat was introduced. Early attempts were based on handcrafted features obtained from artifacts and inconsistencies of the fake video synthesis process. Recent methods, on the other hand, applied deep learning to automatically extract salient and discriminative features to detect deepfakes [44], [45].

Deepfake detection is normally deemed a binary classification problem where classifiers are used to classify between authentic videos and tampered ones. This kind of methods

TABLE I
SUMMARY OF NOTABLE DEEPFAKE TOOLS

Tools	Links	Key Features
Faceswap	https://github.com/deepfakes/faceswap	- Using two encoder-decoder pairs. - Parameters of the encoder are shared.
Faceswap-GAN	https://github.com/shaoanlu/faceswap-GAN	Adversarial loss and perceptual loss (VGGface) are added to an auto-encoder architecture.
Few-Shot Face Translation GAN	https://github.com/shaoanlu/fewshot-face-translation-GAN	- Use a pre-trained face recognition model to extract latent embeddings for GAN processing. - Incorporate semantic priors obtained by modules from FUNIT [39] and SPADE [40].
DeepFaceLab	https://github.com/iperov/DeepFaceLab	- Expand from the Faceswap method with new models, e.g. H64, H128, LIAEF128, SAE [41]. - Support multiple face extraction modes, e.g. S3FD, MTCNN, dlib, or manual [41].
DFaker	https://github.com/dfaker/df	- DSSIM loss function [42] is used to reconstruct face. - Implemented based on Keras library.
DeepFake_tf	https://github.com/StromWine/DeepFake_tf	Similar to DFaker but implemented based on tensorflow.
Deepfakes web β	https://deepfakesweb.com/	Commercial website for face swapping using deep learning algorithms.

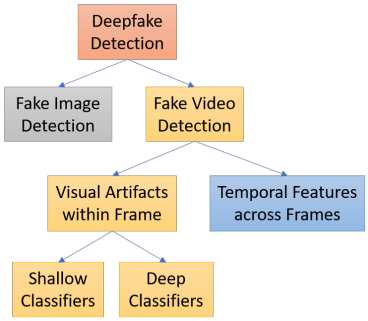

Fig. 3. Categories of reviewed papers relevant to deepfake detection methods where we divide papers into two major groups, i.e. fake image detection and face video detection.

requires a large database of real and fake videos to train classification models. The number of fake videos is increasingly available, but it is still limited in terms of setting a benchmark for validating various detection methods. To address this issue, Korshunov and Marcel [46] produced a notable deepfake data set consisting of 620 videos based on the GAN model using the open source code Faceswap-GAN [36]. Videos from the publicly available VidTIMIT database [47] were used to generate low and high quality deepfake videos, which can effectively mimic the facial expressions, mouth movements, and eye blinking. These videos were then used to test various deepfake detection methods. Test results show that the popular face recognition systems based on VGG [48] and Facenet [37], [49] are unable to detect deepfakes effectively. Other methods such as lip-syncing approaches [50]–[52] and image quality metrics with support vector machine (SVM) [53], [54] produce very high error rate when applied to detect deepfake videos from this newly produced data set. This raises concerns about the critical need of future development of more robust methods that can detect deepfakes from genuine.

This section presents a survey of deepfake detection methods where we group them into two major categories: fake image detection methods and fake video detection ones (see Fig. 3). The latter is distinguished into two groups: visual artifacts within single video frame-based methods and temporal features across frames-based ones. Whilst most of the methods based on temporal features use deep learning recurrent classification models, the methods use visual artifacts within video frame can be implemented by either deep or shallow classifiers.

A. Fake Image Detection

Face swapping has a number of compelling applications in video compositing, transfiguration in portraits, and especially in identity protection as it can replace faces in photographs by ones from a collection of stock images. However, it is also one of the techniques that cyber attackers employ to penetrate identification or authentication systems to gain illegitimate access. The use of deep learning such as CNN and GAN has made swapped face images more challenging for forensics models as it can preserve pose, facial expression and lighting of the photographs [55]. Zhang et al. [56] used the bag of words method to extract a set of compact features and fed it into various classifiers such as SVM [57], random forest (RF) [58] and multi-layer perceptrons (MLP) [59] for discriminating swapped face images from the genuine. Among deep learning-generated images, those synthesised by GAN models are probably most difficult to detect as they are realistic and high-quality based on GAN's capability to learn distribution of the complex input data and generate new outputs with similar input distribution.

Most works on detection of GAN generated images however do not consider the generalization capability of the detection models although the development of GAN is ongoing, and many new extensions of GAN are frequently introduced. Xuan et al. [60] used an image preprocessing step, e.g. Gaussian blur and Gaussian noise, to remove low level high frequency clues of GAN images. This increases the pixel level statistical similarity between real images and fake images and requires the forensic classifier to learn more intrinsic and meaningful features, which has better generalization capability than previous image forensics methods [61], [62] or image steganalysis networks [63].

On the other hand, Agarwal and Varshney [64] cast the GAN-based deepfake detection as a hypothesis testing problem where a statistical framework was introduced using the information-theoretic study of authentication [65]. The minimum distance between distributions of legitimate images and images generated by a particular GAN is defined, namely the oracle error. The analytic results show that this distance

increases when the GAN is less accurate, and in this case, it is easier to detect deepfakes. In case of high-resolution image inputs, an extremely accurate GAN is required to generate fake images that are hard to detect.

Recently, Hsu et al. [66] introduced a two-phase deep learning method for detection of deepfake images. The first phase is a feature extractor based on the common fake feature network (CFFN) where the Siamese network architecture presented in [67] is used. The CFFN encompasses several dense units with each unit including different numbers of dense blocks [68] to improve the representative capability for the fake images. The number of dense units is three or five depending on the validation data being face or general images, and the number of channels in each unit is varied up to a few hundreds. Discriminative features between the fake and real images, i.e. pairwise information, are extracted through CFFN learning process. These features are then fed into the second phase, which is a small CNN concatenated to the last convolutional layer of CFFN to distinguish deceptive images from genuine. The proposed method is validated for both fake face and fake general image detection. On the one hand, the face data set is obtained from CelebA [69], containing 10,177 identities and 202,599 aligned face images of various poses and background clutter. Five GAN variants are used to generate fake images with size of 64x64, including deep convolutional GAN (DCGAN) [70], Wasserstein GAN (WGAN) [71], WGAN with gradient penalty (WGAN-GP) [72], least squares GAN [73], and progressive growth of GAN (PGGAN) [74]. A total of 385,198 training images and 10,000 test images of both real and fake ones are obtained for validating the proposed method. On the other hand, the general data set is extracted from the ILSVRC12 [75]. The large scale GAN training model for high fidelity natural image synthesis (BIGGAN) [76], self-attention GAN [77] and spectral normalization GAN [78] are used to generate fake images with size of 128x128. The training set consists of 600,000 fake and real images whilst the test set includes 10,000 images of both types. Experimental results show the superior performance of the proposed method against its competing methods such as those introduced in [79]–[82].

B. Fake Video Detection

Most image detection methods cannot be used for videos because of the strong degradation of the frame data after video compression [83]. Furthermore, videos have temporal characteristics that are varied among sets of frames and thus challenging for methods designed to detect only still fake images. This subsection focuses on deepfake video detection methods and categorizes them into two groups: methods that employ temporal features and those that explore visual artifacts within frames.

1) Temporal Features across Video Frames: Based on the observation that temporal coherence is not enforced effectively in the synthesis process of deepfakes, Sabir et al. [84] leveraged the use of spatio-temporal features of video streams to detect deepfakes. Video manipulation is carried out on a frame-by-frame basis so that low level artifacts produced by face manipulations are believed to further manifest themselves as temporal artifacts with inconsistencies across frames. A recurrent convolutional model (RCN) was proposed based on the integration of the convolutional network DenseNet [68] and the gated recurrent unit cells [85] to exploit temporal discrepancies across frames (see Fig. 4). The proposed method is tested on the FaceForensics++ data set, which includes 1,000 videos [86], and shows promising results.

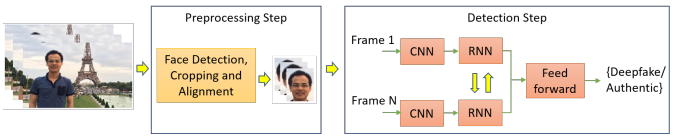

Fig. 4. A two-step process for face manipulation detection where the preprocessing step aims to detect, crop and align faces on a sequence of frames and the second step distinguishes manipulated and authentic face images by combining convolutional neural network (CNN) and recurrent neural network (RNN) [84].

Likewise, Guera and Delp [87] highlighted that deepfake videos contain intra-frame inconsistencies and temporal inconsistencies between frames. They then proposed the temporal-aware pipeline method that uses CNN and long short term memory (LSTM) to detect deepfake videos. CNN is employed to extract frame-level features, which are then fed into the LSTM to create a temporal sequence descriptor. A fully-connected network is finally used for classifying doctored videos from real ones based on the sequence descriptor as illustrated in Fig. 5.

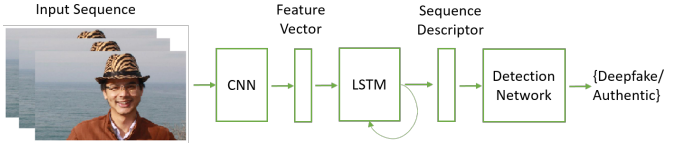

Fig. 5. A deepfake detection method using convolutional neural network (CNN) and long short term memory (LSTM) to extract temporal features of a given video sequence, which are represented via the sequence descriptor. The detection network consisting of fully-connected layers is employed to take the sequence descriptor as input and calculate probabilities of the frame sequence belonging to either authentic or deepfake class [87].

On the other hand, the use of a physiological signal, eye blinking, to detect deepfakes was proposed in [88] based on the observation that a person in deepfakes has a lot less frequent blinking than that in untampered videos. A healthy adult human would normally blink somewhere between 2 to 10 seconds, and each blink would take 0.1 and 0.4 seconds. Deepfake algorithms, however, often use face images available online for training, which normally show people with open eyes, i.e. very few images published on the internet show people with closed eyes. Thus, without having access to images of people blinking, deepfake algorithms do not have the capability to generate fake faces that can blink normally. In other words, blinking rates in deepfakes are much lower than those in normal videos. To discriminate real and fake videos, Li et al. [88] first decompose the videos into frames where face regions and then eye areas are extracted based on six eye landmarks. After few steps of pre-processing such as aligning faces, extracting and scaling the bounding boxes

of eye landmark points to create new sequences of frames, these cropped eye area sequences are distributed into long-term recurrent convolutional networks (LRCN) [89] for dynamic state prediction. The LRCN consists of a feature extractor based on CNN, a sequence learning based on long short term memory (LSTM), and a state prediction based on a fully connected layer to predict probability of eye open and close state. The eye blinking shows strong temporal dependencies and thus the implementation of LSTM helps to capture these temporal patterns effectively. The blinking rate is calculated based on the prediction results where a blink is defined as a peak above the threshold of 0.5 with duration less than 7 frames. This method is evaluated on a data set collected from the web consisting of 49 interview and presentation videos and their corresponding fake videos generated by the deepfake algorithms. The experimental results indicate promising performance of the proposed method in detecting fake videos, which can be further improved by considering dynamic pattern of blinking, e.g. highly frequent blinking may also be a sign of tampering.

2) Visual Artifacts within Video Frame: As can be noticed in the previous subsection, the methods using temporal patterns across video frames are mostly based on deep recurrent network models to detect deepfake videos. This subsection investigates the other approach that normally decomposes videos into frames and explores visual artifacts within single frames to obtain discriminant features. These features are then distributed into either a deep or shallow classifier to differentiate between fake and authentic videos. We thus group methods in this subsection based on the types of classifiers, i.e. either deep or shallow.

a) Deep classifiers: Deepfake videos are normally created with limited resolutions, which require an affine face warping approach (i.e., scaling, rotation and shearing) to match the configuration of the original ones. Because of the resolution inconsistency between the warped face area and the surrounding context, this process leaves artifacts that can be detected by CNN models such as VGG16 [90], ResNet50, ResNet101 and ResNet152 [91]. A deep learning method to detect deepfakes based on the artifacts observed during the face warping step of the deepfake generation algorithms was proposed in [92]. The proposed method is evaluated on two deepfake data sets, namely the UADFV and DeepfakeTIMIT. The UADFV data set [93] contains 49 real videos and 49 fake videos with 32,752 frames in total. The DeepfakeTIMIT data set [52] includes a set of low quality videos of 64 x 64 size and another set of high quality videos of 128 x 128 with totally 10,537 pristine images and 34,023 fabricated images extracted from 320 videos for each quality set. Performance of the proposed method is compared with other prevalent methods such as two deepfake detection MesoNet methods, i.e. Meso-4 and MesoInception-4 [83], HeadPose [93], and the face tampering detection method two-stream NN [94]. Advantage of the proposed method is that it needs not to generate deepfake videos as negative examples before training the detection models. Instead, the negative examples are generated dynamically by extracting the face region of the original image and aligning it into multiple scales before applying Gaussian blur to a scaled image of random pick and warping back to the original image. This reduces a large amount of time and computational resources compared to other methods, which require deepfakes are generated in advance.

Recently, Nguyen et al. [95] proposed the use of capsule networks for detecting manipulated images and videos. The capsule network was initially proposed to address limitations of CNNs when applied to inverse graphics tasks, which aim to find physical processes used to produce images of the world [96]. The recent development of capsule network based on dynamic routing algorithm [97] demonstrates its ability to describe the hierarchical pose relationships between object parts. This development is employed as a component in a pipeline for detecting fabricated images and videos as illustrated in Fig. 6. A dynamic routing algorithm is deployed to route the outputs of the three capsules to the output capsules through a number of iterations to separate between fake and real images. The method is evaluated through four data sets covering a wide range of forged image and video attacks. They include the well-known Idiap Research Institute replay-attack data set [98], the deepfake face swapping data set created by Afchar et al. [83], the facial reenactment FaceForensics data set [99], produced by the Face2Face method [100], and the fully computer-generated image data set generated by Rahmouni et al. [101]. The proposed method yields the best performance compared to its competing methods in all of these data sets. This shows the potential of the capsule network in building a general detection system that can work effectively for various forged image and video attacks.

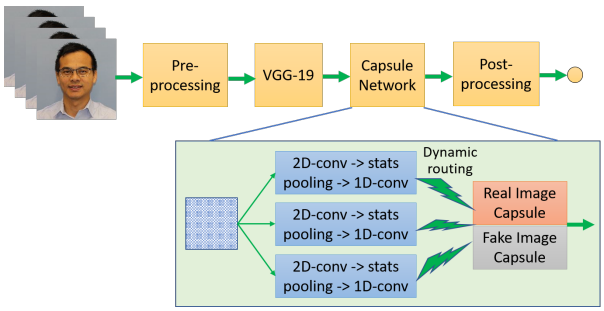

Fig. 6. Capsule network takes features obtained from the VGG-19 network [90] to distinguish fake images or videos from the real ones (top). The pre-processing step detects face region and scales it to the size of 128x128 before VGG-19 is used to extract latent features for the capsule network, which comprises three primary capsules and two output capsules, one for real and one for fake images (bottom). The statistical pooling constitutes an important part of capsule network that deals with forgery detection [95].

b) Shallow classifiers: Deepfake detection methods mostly rely on the artifacts or inconsistency of intrinsic features between fake and real images or videos. Yang et al. [93] proposed a detection method by observing the differences between 3D head poses comprising head orientation and position, which are estimated based on 68 facial landmarks of the central face region. The 3D head poses are examined because there is a shortcoming in the deepfake face generation pipeline. The extracted features are fed into an SVM classifier to obtain the detection results. Experiments on two data sets show the great performance of the proposed approach against

its competing methods. The first data set, namely UADFV, consists of 49 deep fake videos and their respective real videos [93]. The second data set comprises 241 real images and 252 deep fake images, which is a subset of data used in the DARPA MediFor GAN Image/Video Challenge [102]. Likewise, a method to exploit artifacts of deepfakes and face manipulations based on visual features of eyes, teeth and facial contours was studied in [103]. The visual artifacts arise from lacking global consistency, wrong or imprecise estimation of the incident illumination, or imprecise estimation of the underlying geometry. For deepfakes detection, missing reflections and missing details in the eye and teeth areas are exploited as well as texture features extracted from the facial region based on facial landmarks. Accordingly, the eye feature vector, teeth feature vector and features extracted from the full-face crop are used. After extracting the features, two classifiers including logistic regression and small neural network are employed to classify the deepfakes from real videos. Experiments carried out on a video data set downloaded from YouTube show the best result of 0.851 in terms of the area under the receiver operating characteristics curve. The proposed method however has a disadvantage that requires images meeting certain prerequisite such as open eyes or visual teeth.

The use of photo response non uniformity (PRNU) analysis was proposed in [104] to detect deepfakes from authentic ones. PRNU is a component of sensor pattern noise, which is attributed to the manufacturing imperfection of silicon wafers and the inconsistent sensitivity of pixels to light because of the variation of the physical characteristics of the silicon wafers. When a photo is taken, the sensor imperfection is introduced into the high-frequency bands of the content in the form of invisible noise. Because the imperfection is not uniform across the silicon wafer, even sensors made from the silicon wafer produce unique PRNU. Therefore, PRNU is often considered as the fingerprint of digital cameras left in the images by the cameras [105], [106]. The analysis is widely used in image forensics [107]–[110] and advocated to use in [104] because the swapped face is supposed to alter the local PRNU pattern in the facial area of video frames. The videos are converted into frames, which are cropped to the questioned facial region. The cropped frames are then separated sequentially into eight groups where an average PRNU pattern is computed for each group. Normalised cross correlation scores are calculated for comparisons of PRNU patterns among these groups. The authors in [104] created a test data set consisting of 10 authentic videos and 16 manipulated videos, where the fake videos were produced from the genuine ones by the DeepFaceLab tool [31]. The analysis shows a significant statistical difference in terms of mean normalised cross correlation scores between deepfakes and the genuine. This analysis therefore suggests that PRNU has a potential in deepfake detection although a larger data set would need to be tested.

When seeing a video or image with suspicion, users normally want to search for its origin. However, there is currently no feasibility for such a tool. Hasan and Salah [111] proposed the use of blockchain and smart contracts to help users detect deepfake videos based on the assumption that videos are only real when their sources are traceable. Each video is associated with a smart contract that links to its parent video and each parent video has a link to its child in a hierarchical structure. Through this chain, users can credibly trace back to the original smart contract associated with pristine video even if the video has been copied multiple times. An important attribute of the smart contract is the unique hashes of the interplanetary file system (IPFS), which is used to store video and its metadata in a decentralized and content-addressable manner [112]. The smart contract's key features and functionalities are tested against several common security challenges such as distributed denial of services, replay and man in the middle attacks to ensure the proposed solution meeting security requirements. This approach is generic, and it can be extended to other types of digital content such as images, audios and manuscripts.

IV. Discussions, Conclusions and Future Research Directions

Deepfakes have begun to erode trust of people in media contents as seeing them is no longer commensurate with believing in them. They could cause distress and negative effects to those targeted, heighten disinformation and hate speech, and even could stimulate political tension, inflame the public, violence or war. This is especially critical nowadays as the technologies for creating deepfakes are increasingly approachable and social media platforms can spread those fake contents quickly. Sometimes deepfakes do not need to be spread to massive audience to cause detrimental effects. People who create deepfakes with malicious purpose only need to deliver them to target audiences as part of their sabotage strategy without using social media. For example, this approach can be utilized by intelligence services trying to influence decisions made by important people such as politicians, leading to national and international security threats [113]. Catching the deepfake alarming problem, research community has focused on developing deepfake detection algorithms and numerous results have been reported. This paper has reviewed the state-of-the-art methods and a summary of typical approaches is provided in Table II. It is noticeable that a battle between those who use advanced machine learning to create deepfakes with those who make effort to detect deepfakes is growing.

Deepfakes' quality has been increasing and the performance of detection methods needs to be improved accordingly. The inspiration is that what AI has broken can be fixed by AI as well [114]. Detection methods are still in their early stage and various methods have been proposed and evaluated but using fragmented data sets. An approach to improve performance of detection methods is to create a growing updated benchmark data set of deepfakes to validate the ongoing development of detection methods. This will facilitate the training process of detection models, especially those based on deep learning, which requires a large training set [115].

On the other hand, current detection methods mostly focus on drawbacks of the deepfake generation pipelines, i.e. finding weakness of the competitors to attack them. This kind of information and knowledge is not always available in adversarial environments where attackers commonly attempt not to

reveal such deepfake creation technologies. Recent works on adversarial perturbation attacks to fool DNN-based detectors make the deepfake detection task more difficult [116]–[120]. These are real challenges for detection method development and a future research needs to focus on introducing more robust, scalable and generalizable methods.

Another research direction is to integrate detection methods into distribution platforms such as social media to increase its effectiveness in dealing with the widespread impact of deepfakes. The screening or filtering mechanism using effective detection methods can be implemented on these platforms to ease the deepfakes detection [113]. Legal requirements can be made for tech companies who own these platforms to remove deepfakes quickly to reduce its impacts. In addition, watermarking tools can also be integrated into devices that people use to make digital contents to create immutable metadata for storing originality details such as time and location of multimedia contents as well as their untampered attestment [113]. This integration is difficult to implement but a solution for this could be the use of the disruptive blockchain technology. The blockchain has been used effectively in many areas and there are very few studies so far addressing the deepfake detection problems based on this technology. As it can create a chain of unique unchangeable blocks of metadata, it is a great tool for digital provenance solution. The integration of blockchain technologies to this problem has demonstrated certain results [111] but this research direction is far from mature.

Using detection methods to spot deepfakes is crucial, but understanding the real intent of people publishing deepfakes is even more important. This requires the judgement of users based on social context in which deepfake is discovered, e.g. who distributed it and what they said about it [121]. This is critical as deepfakes are getting more and more photorealistic and it is highly anticipated that detection software will be lagging behind deepfake creation technology. A study on social context of deepfakes to assist users in such judgement is thus worth performing.

Videos and photographics have been widely used as evidences in police investigation and justice cases. They may be introduced as evidences in a court of law by digital media forensics experts who have background in computer or law enforcement and experience in collecting, examining and analysing digital information. The development of machine

TABLE II
SUMMARY OF PROMINENT DEEPFAKE DETECTION METHODS

Methods	**Classifiers/ Techniques**	**Key Features**	**Dealing with**	**Data Sets Used**
Eye blinking [88]	LRCN	- Use LRCN to learn the temporal patterns of eye blinking. - Based on the observation that blinking frequency of deepfakes is much smaller than normal.	Videos	Consist of 49 interview and presentation videos, and their corresponding generated deepfakes.
Using spatio-temporal features [84]	RCN	Temporal discrepancies across frames are explored using RCN that integrates convolutional network DenseNet [68] and the gated recurrent unit cells [85]	Videos	FaceForensics++ data set, including 1,000 videos [86].
Intra-frame and temporal inconsistencies [87]	CNN and LSTM	CNN is employed to extract frame-level features, which are distributed to LSTM to construct sequence descriptor useful for classification.	Videos	A collection of 600 videos obtained from multiple websites.
Using face warping artifacts [92]	VGG16 [90] ResNet50, 101 or 152 [91]	Artifacts are discovered using CNN models based on resolution inconsistency between the warped face area and the surrounding context.	Videos	- UADFV [93], containing 49 real videos and 49 fake videos with 32752 frames in total. - DeepfakeTIMIT [52]
MesoNet [83]	CNN	- Two deep networks, i.e. Meso-4 and MesoInception-4 are introduced to examine deepfake videos at the mesoscopic analysis level. - Accuracy obtained on deepfake and FaceForensics data sets are 98	Videos	Two data sets: deepfake one constituted from online videos and the FaceForensics one created by the Face2Face approach [100].
Capsule-forensics [95]	Capsule networks	- Latent features extracted by VGG-19 network [90] are fed into the capsule network for classification. - A dynamic routing algorithm [97] is used to route the outputs of three convolutional capsules to two output capsules, one for fake and another for real images, through a number of iterations.	Videos/ Images	Four data sets: the Idiap Research Institute replay-attack [98], deepfake face swapping by [83], facial reenactment FaceForensics [99], and fully computer-generated image set using [101].
Head poses [93]	SVM	- Features are extracted using 68 landmarks of the face region. - Use SVM to classify using the extracted features.	Videos/ Images	- UADFV consists of 49 deep fake videos and their respective real videos. - 241 real images and 252 deep fake images from DARPA MediFor GAN Image/Video Challenge.
Eye, teach and facial texture [103]	Logistic regression and neural network	- Exploit facial texture differences, and missing reflections and details in eye and teeth areas of deepfakes. - Logistic regression and neural network are used for classifying.	Videos	A video data set downloaded from YouTube.
PRNU Analysis [104]	PRNU	- Analysis of noise patterns of light sensitive sensors of digital cameras due to their factory defects. - Explore the differences of PRNU patterns between the authentic and deepfake videos because face swapping is believed to alter the local PRNU patterns.	Videos	Created by the authors, including 10 authentic and 16 deepfake videos using DeepFaceLab [31].
Using phoneme-viseme mismatches [122]	CNN	- Exploit the mismatches between the dynamics of the mouth shape, i.e. visemes, with a spoken phoneme. - Focus on sounds associated with the M, B and P phonemes as they require complete mouth closure while deepfakes often incorrectly synthesize it.	Videos	Four in-the-wild lip-sync deepfakes from Instagram and YouTube (www.instagram.com/bill_posters_uk and youtu.be/VWMEDacz3L4) and others are created using synthesis techniques, i.e. Audio-to-Video (A2V) [51] and Text-to-Video (T2V) [123].

Methods	Classifiers/ Techniques	Key Features	Dealing with	Data Sets Used
Using attribution-based confidence (ABC) metric [124]	ResNet50 model [91], pre-trained on VGGFace2 [125]	- The ABC metric [126] is used to detect deepfake videos without accessing to training data. - ABC values obtained for original videos are greater than 0.94 while those of deepfakes have low ABC values.	Videos	VidTIMIT and two other original datasets obtained from the COHFACE (https://www.idiap.ch/dataset/cohface) and from YouTube. Datasets from COHFACE [127] and YouTube are used to generate two deepfake datasets by commercial website https://deepfakesweb.com and another deepfake dataset is DeepfakeTIMIT [128].
Using spatial and temporal signatures [129]	Convolutional bidirectional recurrent LSTM network	- An XceptionNet CNN is used for facial feature extraction while audio embeddings are obtained by stacking multiple convolution modules. - Two loss functions, i.e. cross-entropy and Kullback-Leibler divergence, are used.	Videos	FaceForensics++ [86] and Celeb-DF (5,639 deepfake videos) [130] datasets and the ASVSpoof 2019 Logical Access audio dataset [131].
Using emotion audio-visual affective cues [132]	Siamese network architecture [67]	Modality and emotion embedding vectors for the face and speech are extracted for deepfake detection.	Videos	DeepfakeTIMIT [128] and DFDC [133].
Using appearance and behaviour [134]	Rules defined based on facial and behavioural features.	Temporal, behavioral biometric based on facial expressions and head movements are learned using ResNet-101 [91] while static facial biometric is obtained using VGG [48].	Videos	The world leaders dataset [135], FaceForensics++ [86], Google/Jigsaw deepfake detection dataset [136], DFDC [133] and Celeb-DF [130].
Preprocessing combined with deep network [60]	DCGAN, WGAN-GP and PGGAN.	- Enhance generalization ability of models to detect GAN generated images. - Remove low level features of fake images. - Force deep networks to focus more on pixel level similarity between fake and real images to improve generalization ability.	Images	- Real data set: CelebA-HQ [74], including high quality face images of 1024x1024 resolution. - Fake data sets: generated by DCGAN [70], WGAN-GP [72] and PGGAN [74].
Bag of words and shallow classifiers [56]	SVM, RF, MLP	Extract discriminant features using bag of words method and feed these features into SVM, RF and MLP for binary classification: innocent vs fabricated.	Images	The well-known LFW face database [137], containing 13,223 images with resolution of 250x250.
Pairwise learning [66]	CNN concatenated to CFFN	Two-phase procedure: feature extraction using CFFN based on the Siamese network architecture [67] and classification using CNN.	Images	- Face images: real ones from CelebA [69], and fake ones generated by DCGAN [70], WGAN [71], WGAN-GP [72], least squares GAN [73], and PGGAN [74]. - General images: real ones from ILSVRC12 [75], and fake ones generated by BIGGAN [76], self-attention GAN [77] and spectral normalization GAN [78].
Defenses against adversarial perturbations introduced to deepfakes [116]	VGG [48] and ResNet [91]	- Introduce adversarial perturbations to enhance deepfakes and fool deepfake detectors. - Improve accuracy of deepfake detectors using Lipschitz regularization and deep image prior techniques.	Images	5,000 real images from CelebA [69] and 5,000 fake images created by the Few-Shot Face Translation GAN method [138].
Analyzing convolutional traces [139]	K-nearest neighbors, SVM, and linear discriminant analysis	Using expectation maximization algorithm to extract local features pertaining to convolutional generative process of GAN-based image deepfake generators.	Images	Authentic images from CelebA and corresponding deepfakes are created by five different GANs (group-wise deep whitening-and-coloring transformation GDWCT [140], StarGAN [141], AttGAN [142], StyleGAN [143], StyleGAN2 [144]).

learning and AI technologies might have been used to modify these digital contents and thus the experts' opinions may not be enough to authenticate these evidences because even experts are unable to discern manipulated contents. This aspect needs to take into account in courtrooms nowadays when images and videos are used as evidences to convict perpetrators because of the existence of a wide range of digital manipulation methods [145]. The digital media forensics results therefore must be proved to be valid and reliable before they can be used in courts. This requires careful documentation for each step of the forensics process and how the results are reached. Machine learning and AI algorithms can be used to support the determination of the authenticity of digital media and have obtained accurate and reliable results, e.g. [146]–[148], but most of these algorithms are unexplainable. This creates a huge hurdle for the applications of AI in forensics problems because not only the forensics experts oftentimes do not have expertise in computer algorithms, but the computer professionals also cannot explain the results properly as most of these algorithms are black box models [149]. This is more critical as the most recent models with the most accurate results are based on deep learning methods consisting of many neural network parameters. Explainable AI in computer vision therefore is a research direction that is needed to promote and utilize the advances and advantages of AI and machine learning in digital media forensics.

REFERENCES

[1] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481-2495.

[2] Yang, W., Hui, C., Chen, Z., Xue, J. H., and Liao, Q. (2019). FV-GAN: Finger vein representation using generative adversarial networks. *IEEE Transactions on Information Forensics and Security*, 14(9), 2512-2524.

[3] Tewari, A., Zollhoefer, M., Bernard, F., Garrido, P., Kim, H., Perez, P., and Theobalt, C. (2020). High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI: 10.1109/TPAMI.2018.2876842.

[4] Guo, Y., Jiao, L., Wang, S., Wang, S., and Liu, F. (2018). Fuzzy sparse autoencoder framework for single image per person face recognition. *IEEE Transactions on Cybernetics*, 48(8), 2402-2415.

[5] Liu, F., Jiao, L., and Tang, X. (2019). Task-oriented GAN for PolSAR image classification and clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9), 2707-2719.

[6] Cao, J., Hu, Y., Yu, B., He, R., and Sun, Z. (2019). 3D aided duet GANs for multi-view face image synthesis. *IEEE Transactions on Information Forensics and Security*, 14(8), 2028-2042.

[7] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2019). StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8), 1947-1962.

[8] Lyu, S. (2018, August 29). Detecting deepfake videos in the blink of an eye. Available at http://theconversation.com/detecting-deepfake-videos-in-the-blink-of-an-eye-101072

[9] Bloomberg (2018, September 11). How faking videos became easy and why that's so scary. Available at https://fortune.com/2018/09/11/deep-fakes-obama-video/

[10] Chesney, R., and Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. Foreign Affairs, 98, 147.

[11] Kaliyar, R. K., Goswami, A., and Narang, P. (2020). Deepfake: improving fake news detection using tensor decomposition based deep neural network. *Journal of Supercomputing*, doi: https://doi.org/10.1007/s11227-020-03294-y.

[12] Tucker, P. (2019, March 31). The newest AI-enabled weapon: Deep-Faking photos of the earth. Available at https://www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/

[13] Fish, T. (2019, April 4). Deep fakes: AI-manipulated media will be weaponised to trick military. Available at https://www.express.co.uk/news/science/1109783/deep-fakes-ai-artificial-intelligence-photos-video-weaponised-china

[14] Marr, B. (2019, July 22). The best (and scariest) examples of AI-enabled deepfakes. Available at https://www.forbes.com/sites/bernardmarr/2019/07/22/the-best-and-scariest-examples-of-ai-enabled-deepfakes/

[15] Zakharov, E., Shysheya, A., Burkov, E., and Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. *arXiv preprint* arXiv:1905.08233.

[16] Damiani, J. (2019, September 3). A voice deepfake was used to scam a CEO out of $243,000. Available at https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/

[17] Samuel, S. (2019, June 27). A guy made a deepfake app to turn photos of women into nudes. It didn't go well. Available at https://www.vox.com/2019/6/27/18761639/ai-deepfake-deepnude-app-nude-women-porn

[18] The Guardian (2019, September 2). Chinese deepfake app Zao sparks privacy row after going viral. Available at https://www.theguardian.com/technology/2019/sep/02/chinese-face-swap-app-zao-triggers-privacy-fears-viral

[19] Lyu, S. (2020, July). Deepfake detection: current challenges and next steps. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (pp. 1-6). IEEE.

[20] Guarnera, L., Giudice, O., Nastasi, C., and Battiato, S. (2020). Preliminary forensics analysis of deepfake images. *arXiv preprint* arXiv:2004.12626.

[21] Jafar, M. T., Ababneh, M., Al-Zoube, M., and Elhassan, A. (2020, April). Forensics and analysis of deepfake videos. In *The 11th International Conference on Information and Communication Systems (ICICS)* (pp. 053-058). IEEE.

[22] Trinh, L., Tsang, M., Rambhatla, S., and Liu, Y. (2020). Interpretable deepfake detection via dynamic prototypes. *arXiv preprint* arXiv:2006.15473.

[23] Younus, M. A., and Hasan, T. M. (2020, April). Effective and fast deepfake detection method based on Haar wavelet transform. In *2020 International Conference on Computer Science and Software Engineering (CSASE)* (pp. 186-190). IEEE.

[24] Turek, M. (2019). Media Forensics (MediFor). Available at https://www.darpa.mil/program/media-forensics

[25] Schroepfer, M. (2019, September 5). Creating a data set and a challenge for deepfakes. Available at https://ai.facebook.com/blog/deepfake-detection-challenge

[26] Punnappurath, A., and Brown, M. S. (2019). Learning raw image reconstruction-aware deep image compressors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI: 10.1109/TPAMI.2019.2903062.

[27] Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. (2019). Energy compaction-based image compression using convolutional autoencoder. *IEEE Transactions on Multimedia*. DOI: 10.1109/TMM.2019.2938345.

[28] Chorowski, J., Weiss, R. J., Bengio, S., and Oord, A. V. D. (2019). Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 27(12), pp. 2041-2053.

[29] Faceswap: Deepfakes software for all. Available at https://github.com/deepfakes/faceswap

[30] FakeApp 2.2.0. Available at https://www.malavida.com/en/soft/fakeapp/

[31] DeepFaceLab. Available at https://github.com/iperov/DeepFaceLab

[32] DFaker. Available at https://github.com/dfaker/df

[33] DeepFake_tf: Deepfake based on tensorflow. Available at https://github.com/StromWine/DeepFake_tf

[34] Keras-VGGFace: VGGFace implementation with Keras framework. Available at https://github.com/rcmalli/keras-vggface

[35] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (pp. 2672-2680).

[36] Faceswap-GAN. Available at https://github.com/shaoanlu/faceswap-GAN.

[37] FaceNet. Available at https://github.com/davidsandberg/facenet.

[38] CycleGAN. Available at https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.

[39] Liu, M. Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., and Kautz, J. (2019). Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 10551-10560).

[40] Park, T., Liu, M. Y., Wang, T. C., and Zhu, J. Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2337-2346).

[41] DeepFaceLab: Explained and usage tutorial. Available at https://mrdeepfakes.com/forums/thread-deepfacelab-explained-and-usage-tutorial.

[42] DSSIM. Available at https://github.com/keras-team/keras-contrib/blob/master/keras_contrib/losses/dssim.py.

[43] Chesney, R., and Citron, D. K. (2018). Deep fakes: a looming challenge for privacy, democracy, and national security. https://dx.doi.org/10.2139/ssrn.3213954.

[44] de Lima, O., Franklin, S., Basu, S., Karwoski, B., and George, A. (2020). Deepfake detection using spatiotemporal convolutional networks. *arXiv preprint* arXiv:2006.14749.

[45] Amerini, I., and Caldelli, R. (2020, June). Exploiting prediction error inconsistencies through LSTM-based classifiers to detect deepfake videos. In *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security* (pp. 97-102).

[46] Korshunov, P., and Marcel, S. (2019). Vulnerability assessment and detection of deepfake videos. In *The 12th IAPR International Conference on Biometrics (ICB)*, pp. 1-6.

[47] VidTIMIT database. Available at http://conradsanderson.id.au/vidtimit/

[48] Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015, September). Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)* (pp. 41.1-41.12).

[49] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 815-823).

[50] Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2017, July). Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3444-3453).

[51] Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4), 113.

[52] Korshunov, P., and Marcel, S. (2018, September). Speaker inconsistency detection in tampered video. In *2018 26th European Signal Processing Conference (EUSIPCO)* (pp. 2375-2379). IEEE.

[53] Galbally, J., and Marcel, S. (2014, August). Face anti-spoofing based on general image quality assessment. In *2014 22nd International Conference on Pattern Recognition* (pp. 1173-1178). IEEE.

[54] Wen, D., Han, H., and Jain, A. K. (2015). Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4), 746-761.

[55] Korshunova, I., Shi, W., Dambre, J., and Theis, L. (2017). Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3677-3685).

[56] Zhang, Y., Zheng, L., and Thing, V. L. (2017, August). Automated face swapping and its detection. In *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)* (pp. 15-19). IEEE.

[57] Wang, X., Thome, N., and Cord, M. (2017). Gaze latent support vector machine for image classification improved by weakly supervised region selection. *Pattern Recognition*, 72, 59-71.

[58] Bai, S. (2017). Growing random forest on deep convolutional neural networks for scene categorization. *Expert Systems with Applications*, 71, 279-287.

[59] Zheng, L., Duffner, S., Idrissi, K., Garcia, C., and Baskurt, A. (2016). Siamese multi-layer perceptrons for dimensionality reduction and face identification. *Multimedia Tools and Applications*, 75(9), 5055-5073.

[60] Xuan, X., Peng, B., Dong, J., and Wang, W. (2019). On the generalization of GAN image forensics. *arXiv preprint* arXiv:1902.11153.

[61] Yang, P., Ni, R., and Zhao, Y. (2016, September). Recapture image forensics based on Laplacian convolutional neural networks. In *International Workshop on Digital Watermarking* (pp. 119-128).

[62] Bayar, B., and Stamm, M. C. (2016, June). A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security* (pp. 5-10). ACM.

[63] Qian, Y., Dong, J., Wang, W., and Tan, T. (2015, March). Deep learning for steganalysis via convolutional neural networks. In *Media Watermarking, Security, and Forensics* 2015 (Vol. 9409, p. 94090J).

[64] Agarwal, S., and Varshney, L. R. (2019). Limits of deepfake detection: A robust estimation viewpoint. *arXiv preprint* arXiv:1905.03493.

[65] Maurer, U. M. (2000). Authentication theory and hypothesis testing. *IEEE Transactions on Information Theory*, 46(4), 1350-1356.

[66] Hsu, C. C., Zhuang, Y. X., and Lee, C. Y. (2020). Deep fake image detection based on pairwise learning. *Applied Sciences*, 10(1), 370.

[67] Chopra, S. (2005). Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Compter Vision and Pattern Recognition* (pp. 539-546).

[68] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4700-4708).

[69] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3730-3738).

[70] Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint* arXiv:1511.06434.

[71] Arjovsky, M., Chintala, S., and Bottou, L. (2017, July). Wasserstein generative adversarial networks. In *International Conference on Machine Learning* (pp. 214-223).

[72] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems* (pp. 5767-5777).

[73] Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2794-2802).

[74] Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint* arXiv:1710.10196.

[75] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... and Berg, A. C. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211-252.

[76] Brock, A., Donahue, J., and Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint* arXiv:1809.11096.

[77] Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2018). Self-attention generative adversarial networks. *arXiv preprint* arXiv:1805.08318.

[78] Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint* arXiv:1802.05957.

[79] Farid, H. (2009). Image forgery detection. *IEEE Signal Processing Magazine*, 26(2), 16-25.

[80] Mo, H., Chen, B., and Luo, W. (2018, June). Fake faces identification via convolutional neural network. In *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security* (pp. 43-47).

[81] Marra, F., Gragnaniello, D., Cozzolino, D., and Verdoliva, L. (2018, April). Detection of GAN-generated fake images over social networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp. 384-389). IEEE.

[82] Hsu, C. C., Lee, C. Y., and Zhuang, Y. X. (2018, December). Learning to detect fake face images in the wild. In *2018 International Symposium on Computer, Consumer and Control (IS3C)* (pp. 388-391). IEEE.

[83] Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I. (2018, December). MesoNet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1-7). IEEE.

[84] Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., and Natarajan, P. (2019). Recurrent convolutional strategies for face manipulation detection in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 80-87).

[85] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014, October). Learning phrase representations using RNN encoderdecoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724-1734).

[86] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Niener, M. (2019). FaceForensics++: Learning to detect manipulated facial images. *arXiv preprint* arXiv:1901.08971.

[87] Guera, D., and Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1-6). IEEE.

[88] Li, Y., Chang, M. C., and Lyu, S. (2018, December). In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1-7). IEEE.

[89] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2625-2634).

[90] Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556.

[91] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

[92] Li, Y., and Lyu, S. (2019). Exposing deepfake videos by detecting face warping artifacts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 46-52).

[93] Yang, X., Li, Y., and Lyu, S. (2019, May). Exposing deep fakes using inconsistent head poses. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8261-8265). IEEE.

[94] Zhou, P., Han, X., Morariu, V. I., and Davis, L. S. (2017, July). Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1831-1839). IEEE.

[95] Nguyen, H. H., Yamagishi, J., and Echizen, I. (2019, May). Capsule-forensics: Using capsule networks to detect forged images and videos. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2307-2311). IEEE.

[96] Hinton, G. E., Krizhevsky, A., and Wang, S. D. (2011, June). Transforming auto-encoders. In *International Conference on Artificial Neural Networks* (pp. 44-51). Springer, Berlin, Heidelberg.

[97] Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. In *Advances in Neural Information Processing Systems* (pp. 3856-3866).

[98] Chingovska, I., Anjos, A., and Marcel, S. (2012, September). On the effectiveness of local binary patterns in face anti-spoofing. In *Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)* (pp. 1-7). IEEE.

[99] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Niener, M. (2018). FaceForensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint* arXiv:1803.09179.

[100] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Niener, M. (2016). Face2Face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2387-2395).

[101] Rahmouni, N., Nozick, V., Yamagishi, J., and Echizen, I. (2017, December). Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)* (pp. 1-6). IEEE.

[102] Guan, H., Kozak, M., Robertson, E., Lee, Y., Yates, A. N., Delgado, A., ... and Fiscus, J. (2019, January). MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)* (pp. 63-72).

[103] Matern, F., Riess, C., and Stamminger, M. (2019, January). Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)* (pp. 83-92). IEEE.

[104] Koopman, M., Rodriguez, A. M., and Geradts, Z. (2018). Detection of deepfake video manipulation. In *The 20th Irish Machine Vision and Image Processing Conference (IMVIP)* (pp. 133-136).

[105] Lukas, J., Fridrich, J., and Goljan, M. (2006). Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2), 205-214.

[106] Rosenfeld, K., and Sencar, H. T. (2009, February). A study of the robustness of PRNU-based camera identification. In *Media Forensics and Security* (Vol. 7254, p. 72540M). International Society for Optics and Photonics.

[107] Li, C. T., and Li, Y. (2012). Color-decoupled photo response non-uniformity for digital image forensics. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(2), 260-271.

[108] Lin, X., and Li, C. T. (2017). Large-scale image clustering based on camera fingerprints. *IEEE Transactions on Information Forensics and Security*, 12(4), 793-808.

[109] Scherhag, U., Debiasi, L., Rathgeb, C., Busch, C., and Uhl, A. (2019). Detection of face morphing attacks based on PRNU analysis. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(4), 302-317.

[110] Phan, Q. T., Boato, G., and De Natale, F. G. (2019). Accurate and scalable image clustering based on sparse representation of camera fingerprint. *IEEE Transactions on Information Forensics and Security*, 14(7), 1902-1916.

[111] Hasan, H. R., and Salah, K. (2019). Combating deepfake videos using blockchain and smart contracts. *IEEE Access*, 7, 41596-41606.

[112] IPFS powers the Distributed Web. Available at https://ipfs.io/

[113] Chesney, R. and Citron, D. K. (2018, October 16). Disinformation on steroids: The threat of deep fakes. Available at https://www.cfr.org/report/deep-fake-disinformation-steroids.

[114] Floridi, L. (2018). Artificial intelligence, deepfakes and a future of ectypes. *Philosophy and Technology*, 31(3), 317-321.

[115] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C. (2020). The deepfake detection challenge dataset. *arXiv preprint* arXiv:2006.07397.

[116] Gandhi, A., and Jain, S. (2020). Adversarial perturbations fool deepfake detectors. *arXiv preprint* arXiv:2003.10596.

[117] Neekhara, P., Hussain, S., Jere, M., Koushanfar, F., and McAuley, J. (2020). Adversarial deepfakes: evaluating vulnerability of deepfake detectors to adversarial examples. *arXiv preprint* arXiv:2002.12749.

[118] Carlini, N., and Farid, H. (2020). Evading deepfake-image detectors with white-and black-box attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 658-659).

[119] Yang, C., Ding, L., Chen, Y., and Li, H. (2020). Defending against GAN-based deepfake attacks via transformation-aware adversarial faces. *arXiv preprint* arXiv:2006.07421.

[120] Yeh, C. Y., Chen, H. W., Tsai, S. L., and Wang, S. D. (2020). Disrupting image-translation-based deepfake algorithms with adversarial attacks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops* (pp. 53-62).

[121] Read, M. (2019, June 27). Can you spot a deepfake? Does it matter? Available at http://nymag.com/intelligencer/2019/06/how-do-you-spot-a-deepfake-it-might-not-matter.html.

[122] Agarwal, S., Farid, H., Fried, O., and Agrawala, M. (2020). Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 660-661).

[123] Fried, O., Tewari, A., Zollhfer, M., Finkelstein, A., Shechtman, E., Goldman, D. B., ... and Agrawala, M. (2019). Text-based editing of talking-head video. *ACM Transactions on Graphics*, 38(4), 1-14.

[124] Fernandes, S., Raj, S., Ewetz, R., Singh Pannu, J., Kumar Jha, S., Ortiz, E., ... and Salter, M. (2020). Detecting deepfake videos using attribution-based confidence metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 308-309).

[125] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018, May). VGGFace2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 67-74). IEEE.

[126] Jha, S., Raj, S., Fernandes, S., Jha, S. K., Jha, S., Jalaian, B., ... and Swami, A. (2019). Attribution-based confidence metric for deep neural networks. In *Advances in Neural Information Processing Systems* (pp. 11826-11837).

[127] Fernandes, S., Raj, S., Ortiz, E., Vintila, I., Salter, M., Urosevic, G., and Jha, S. (2019, October). Predicting heart rate variations of deepfake videos using neural ODE. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (pp. 1721-1729). IEEE.

[128] Korshunov, P., and Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint* arXiv:1812.08685.

[129] Chintha, A., Thai, B., Sohrawardi, S. J., Bhatt, K. M., Hickerson, A., Wright, M., and Ptucha, R. (2020). Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE Journal of Selected Topics in Signal Processing*, doi: 10.1109/JSTSP.2020.2999185.

[130] Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3207-3216).

[131] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., ... and Lee, K. A. (2019). ASVspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint* arXiv:1904.05441.

[132] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., and Manocha, D. (2020). Emotions don't lie: A deepfake detection method using audio-visual affective cues. *arXiv preprint* arXiv:2003.06711.

[133] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., and Ferrer, C. C. (2019). The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint* arXiv:1910.08854.

[134] Agarwal, S., El-Gaaly, T., Farid, H., and Lim, S. N. (2020). Detecting deep-fake videos from appearance and behavior. *arXiv preprint* arXiv:2004.14491.

[135] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., and Li, H. (2019, June). Protecting world leaders against deep fakes. In *Computer Vision and Pattern Recognition Workshops* (pp. 38-45).

[136] Dufour, N., and Gully, A. (2019). Contributing Data to Deepfake Detection Research. Available at: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

[137] Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2007, October). Labelled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, http://vis-www.cs.umass.edu/lfw/.

[138] Shaoanlus GitHub. (2019). Few-Shot Face Translation GAN. Available at https://github.com/shaoanlu/fewshot-face-translation-GAN.

[139] Guarnera, L., Giudice, O., and Battiato, S. (2020). Deepfake detection by analyzing convolutional traces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 666-667).

[140] Cho, W., Choi, S., Park, D. K., Shin, I., and Choo, J. (2019). Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 10639-10647).

[141] Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., and Choo, J. (2018). StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8789-8797).

[142] He, Z., Zuo, W., Kan, M., Shan, S., and Chen, X. (2019). AttGAN: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11), 5464-5478.

[143] Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4401-4410).

[144] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8110-8119).

[145] Maras, M. H., and Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. *The International Journal of Evidence and Proof*, 23(3), 255-262.

[146] Fridrich, J., and Kodovsky, J. (2012). Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3), 868-882.

[147] Su, L., Li, C., Lai, Y., and Yang, J. (2017). A fast forgery detection algorithm based on exponential-Fourier moments for video region duplication. *IEEE Transactions on Multimedia*, 20(4), 825-840.

[148] Iuliani, M., Shullani, D., Fontani, M., Meucci, S., and Piva, A. (2018). A video forensic framework for the unsupervised analysis of MP4-like file container. *IEEE Transactions on Information Forensics and Security*, 14(3), 635-645.

[149] Malolan, B., Parekh, A., and Kazi, F. (2020, March). Explainable deep-fake detection using visual interpretability methods. In *The 3rd International Conference on Information and Computer Technologies (ICICT)* (pp. 289-293). IEEE.

Thanh Thi Nguyen was a Visiting Scholar with the Computer Science Department at Stanford University, California, USA in 2015 and the Edge Computing Lab, John A. Paulson School of Engineering and Applied Sciences, Harvard University, Massachusetts, USA in 2019. He received an Alfred Deakin Postdoctoral Research Fellowship in 2016, a European-Pacific Partnership for ICT Expert Exchange Program Award from European Commission in 2018, and an AustraliaIndia Strategic Research Fund Early- and Mid-Career Fellowship Awarded by the Australian Academy of Science in 2020. Dr. Nguyen obtained a PhD in Mathematics and Statistics from Monash University, Australia in 2013 and has expertise in various areas, including artificial intelligence, deep learning, reinforcement learning, computer vision, cyber security, IoT, and data science.

Dr. Nguyen has been recognized as a leading researcher in Australia in the field of Artificial Intelligence by The Australian Newspaper in a report published in 2018. He is currently a Senior Lecturer in the School of Information Technology, Deakin University, Victoria, Australia.

Cuong M. Nguyen received the B.Sc. and M.Sc. degrees in Mathematics from Vietnam National University, Hanoi, Vietnam. In 2017, he received the Ph.D. degree from School of Engineering, Deakin University, Victoria, Australia, where he is currently a postdoctoral researcher. His research interests lie in the areas of Optimization, Machine Learning, and Control Systems.

Dung Tien Nguyen received the B.Eng. and M.Eng. degrees from the People Securiry Academy and the Vietnam National University University of Enginerring and Technology in 2006 and 2013, respectively. He received his PhD degree in the area of multimodal emotion recognition using deep learning techniques from Queensland University of Technology in Brisbane, Australia in 2019. He is currently working as a research fellow at Deakin University in computer vision, machine learning, deep learning, image processing, and affective computing.

Duc Thanh Nguyen was awarded a PhD in Computer Science from the University of Wollongong, Australia in 2012. Currently, he is a lecturer in the School of Information Technology, Deakin University, Australia. His research interests include computer vision and pattern recognition. He has published his work in highly ranked publication venues in Computer Vision and Pattern Recognition such as the Journal of Pattern Recognition, CVPR, ICCV, ECCV. He also has served a technical program committee member for many premium conferences such as CVPR, ICCV, ECCV, AAAI, ICIP, PAKDD and reviewer for the IEEE Trans. Intell. Transp. Syst., the IEEE Trans. Image Process., the IEEE Signal Processing Letters, Image and Vision Computing, Pattern Recognition, Scientific Reports.

Saeid Nahavandi received a Ph.D. from Durham University, U.K. in 1991. He is an Alfred Deakin Professor, Pro Vice-Chancellor (Defence Technologies), Chair of Engineering, and the Director for the Institute for Intelligent Systems Research and Innovation at Deakin University, Victoria, Australia. His research interests include modelling of complex systems, machine learning, deep learning, robotics and haptics. He is a Fellow of Engineers Australia (FIEAust), the Institution of Engineering and Technology (FIET) and IEEE (FIEEE).

He is the Co-Editor-in-Chief of the IEEE Systems Journal, Associate Editor of the IEEE/ASME Transactions on Mechatronics, Associate Editor of the IEEE Transactions on Systems, Man and Cybernetics: Systems, and an IEEE Access Editorial Board member.