Computer Vision and Image Understanding 223 (2022) 103525

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Deep learning for deepfakes creation and detection: A survey

Thanh Thi Nguyen [a,*], Quoc Viet Hung Nguyen [b], Dung Tien Nguyen [a], Duc Thanh Nguyen [a], Thien Huynh-The [c], Saeid Nahavandi [d], Thanh Tam Nguyen [e], Quoc-Viet Pham [f], Cuong M. Nguyen [g]

[a] *School of Information Technology, Deakin University, Victoria, Australia*
[b] *School of Information and Communication Technology, Griffith University, Queensland, Australia*
[c] *ICT Convergence Research Center, Kumoh National Institute of Technology, Gyeongbuk, Republic of Korea*
[d] *Institute for Intelligent Systems Research and Innovation, Deakin University, Victoria, Australia*
[e] *Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh City, Vietnam*
[f] *Korean Southeast Center for the 4th Industrial Revolution Leader Education, Pusan National University, Busan, Republic of Korea*
[g] *LAMIH UMR CNRS 8201, Universite Polytechnique Hauts-de-France, Valenciennes, France*

ARTICLE INFO

Communicated by Nikos Paragios

Keywords:
Deepfakes
Face manipulation
Artificial intelligence
Deep learning
Autoencoders
GAN
Forensics
Survey

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.

1. Introduction

In a narrow definition, deepfakes (stemming from "deep learning" and "fake") are created by techniques that can superimpose face images of a target person onto a video of a source person to make a video of the target person doing or saying things the source person does. This constitutes a category of deepfakes, namely *face-swap*. In a broader definition, deepfakes are artificial intelligence-synthesized content that can also fall into two other categories, i.e., *lip-sync* and *puppet-master*. Lip-sync deepfakes refer to videos that are modified to make the mouth movements consistent with an audio recording. Puppet-master deepfakes include videos of a target person (puppet) who is animated following the facial expressions, eye and head movements of another person (master) sitting in front of a camera (Agarwal et al., 2019).

While some deepfakes can be created by traditional visual effects or computer-graphics approaches, the recent common underlying mechanism for deepfake creation is deep learning models such as autoencoders and generative adversarial networks (GANs), which have been applied widely in the computer vision domain (Vincent et al., 2008; Kingma and Welling, 2013; Goodfellow et al., 2014; Makhzani et al., 2015; Tewari et al., 2018; Lin et al., 2021; Liu et al., 2021). 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 (Lyu, 2018). 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 the 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 (Bloomberg, 2018; Chesney and Citron, 2019; Hwang, 2020). 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 (Zhou and Zafarani, 2020; Kaliyar et al., 2021; Guo et al., 2020). It can be even used to generate fake satellite images of the Earth

* Corresponding author.
E-mail address: thanh.nguyen@deakin.edu.au (T.T. Nguyen).

https://doi.org/10.1016/j.cviu.2022.103525
Received 8 February 2022; Received in revised form 18 July 2022; Accepted 27 July 2022
Available online 31 July 2022
1077-3142/

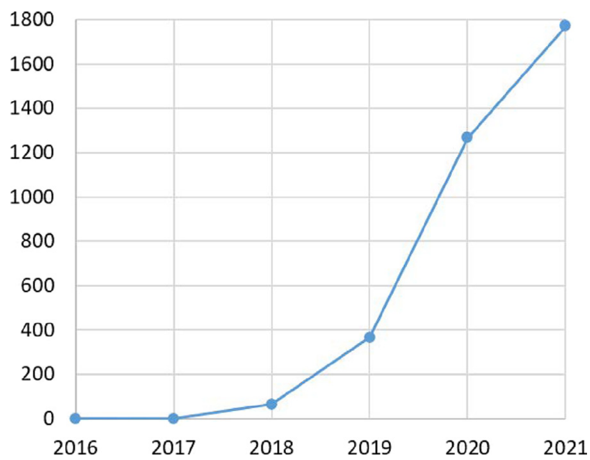

Fig. 1. Number of papers related to deepfakes in years from 2016 to 2021, obtained from https://app.dimensions.ai at the end of 2021 with the search keyword "deepfake" applied to full text of scholarly papers.

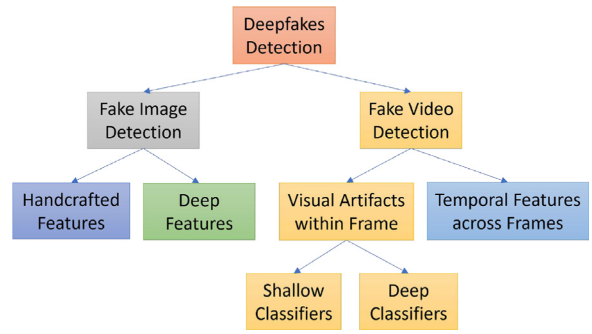

Fig. 2. Categories of reviewed papers relevant to deepfake detection methods where we divide papers into two major groups, i.e., fake image detection and face video detection.

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 (Tucker, 2019; Fish, 2019).

As the democratization of creating realistic digital humans has positive implications, there is also positive use of deepfakes such as their applications in visual effects, digital avatars, snapchat filters, creating voices of those who have lost theirs or updating episodes of movies without reshooting them (Marr, 2019). Deepfakes can have creative or productive impacts in photography, video games, virtual reality, movie productions, and entertainment, e.g., realistic video dubbing of foreign films, education through the reanimation of historical figures, virtually trying on clothes while shopping, and so on (Mirsky and Lee, 2021; Verdoliva, 2020). However, the number of malicious uses of deepfakes largely dominates that of the positive ones. The development of advanced deep neural 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 (Zakharov et al., 2019). 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 (Damiani, 2019). A recent release of a software called DeepNude shows more disturbing threats as it can transform a person to a non-consensual porn (Samuel, 2019). 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 (The Guardian, 2019). 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 (Lyu, 2020; Guarnera et al., 2020c; Jafar et al., 2020; Trinh et al., 2021; Younus and Hasan, 2020). 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 (Turek, 2019). Recently, Facebook Inc. teaming up with Microsoft Corp and the Partnership on AI coalition have launched the Deepfake Detection Challenge to catalyze more research and development in detecting and preventing deepfakes from being used to mislead viewers (Schroepfer, 2019). Data obtained from https://app.dimensions.ai at the end of 2021 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.

There have been existing survey papers about creating and detecting deepfakes, presented in Tolosana et al. (2020), Verdoliva (2020) and Mirsky and Lee (2021). For example, Mirsky and Lee (2021) focused on reenactment approaches (i.e., to change a target's expression, mouth, pose, gaze or body), and replacement approaches (i.e., to replace a target's face by swap or transfer methods). Verdoliva (2020) separated detection approaches into conventional methods (e.g., blind methods without using any external data for training, one-class sensor-based and model-based methods, and supervised methods with handcrafted features) and deep learning-based approaches (e.g., CNN models). Tolosana et al. (2020) categorized both creation and detection methods based on the way deepfakes are created, including entire face synthesis, identity swap, attribute manipulation, and expression swap. On the other hand, we carry out the survey with a different perspective and taxonomy. We categorize the deepfake detection methods based on the data type, i.e., images or videos, as presented in Fig. 2. With fake image detection methods, we focus on the features that are used, i.e., whether they are handcrafted features or deep features. With fake video detection methods, two main subcategories are identified based on whether the method uses temporal features across frames or visual artifacts within a video frame. We also discuss extensively the challenges, research trends and directions on deepfake detection and multimedia forensics problems.

2. 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 (Punnappurath and Brown, 2019; Cheng et al., 2019; Chorowski et al., 2019). The first attempt of deepfake creation was FakeApp, developed by a Reddit user using autoencoder–decoder pairing structure (Faceswap, 2022; FakeApp, 2022). In that method, the autoencoder extracts latent features of face images and the decoder is used to reconstruct the face

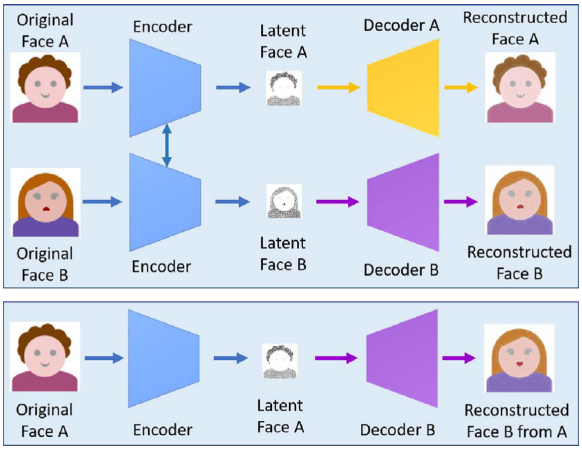

Fig. 3. A deepfake creation model using two encoder–decoder pairs. Two networks use the same encoder but different decoders for training process (top). An image of face A is encoded with the common encoder and decoded with decoder B to create a deepfake (bottom). The reconstructed image (in the bottom) is the face B with the mouth shape of face A. Face B originally has the mouth of an upside-down heart while the reconstructed face B has the mouth of a conventional heart.

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. 3 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 (DeepFaceLab, 2022b), DFaker (DFaker, 2022), DeepFake_tf (tensorflow-based deepfakes) (DeepFake_tf, 2022).

By adding adversarial loss and perceptual loss implemented in VGGFace (Keras-VGGFace, 2022) to the encoder–decoder architecture, an improved version of deepfakes based on the generative adversarial network (Goodfellow et al., 2014), i.e., faceswap-GAN, was proposed in Faceswap-GAN (2022). 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 64 × 64, 128 × 128, and 256 × 256 resolutions. In addition, the multi-task convolutional neural network (CNN) from the FaceNet implementation (FaceNet, 2022) is used to make face detection more stable and face alignment more reliable. The CycleGAN (CycleGAN, 2022) is utilized for generative network implementation in this model.

A conventional GAN model comprises two neural networks: a generator and a discriminator as depicted in Fig. 4. Given a dataset of real images x having a distribution of p_{data}, the aim of the generator G is to produce images $G(z)$ similar to real images x with z being noise signals having a distribution of p_z. The aim of the discriminator G is to correctly classify images generated by G and real images x. The discriminator D is trained to improve its classification capability, i.e., to maximize $D(x)$, which represents the probability that x is a real image rather than a fake image generated by G. On the other hand, G is trained to minimize the probability that its outputs are classified by D as synthetic images, i.e., to minimize $1 - D(G(z))$. This is a minimax game between two players D and G that can be described by the following value function (Goodfellow et al., 2014):

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

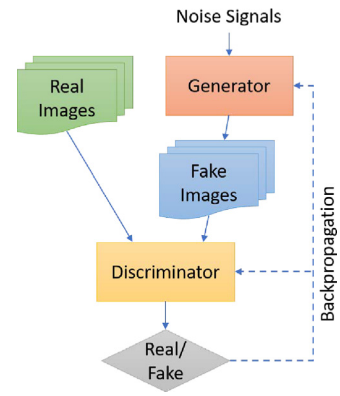

Fig. 4. The GAN architecture consisting of a generator and a discriminator, and each can be implemented by a neural network. The entire system can be trained with backpropagation that allows both networks to improve their capabilities.

After sufficient training, both networks improve their capabilities, i.e., the generator G is able to produce images that are really similar to real images while the discriminator D is highly capable of distinguishing fake images from real ones.

Table 1 presents a summary of popular deepfake tools and their typical features. Among them, a prominent method for face synthesis based on a GAN model, namely StyleGAN, was introduced in Karras et al. (2019). StyleGAN is motivated by style transfer (Huang et al., 2017) with a special generator network architecture that is able to create realistic face images. In a traditional GAN model, e.g., the progressive growing of GAN (PGGAN) (Karras et al., 2017), the signal noise (latent code) is fed to the input layer of a feedforward network that represents the generator. In StyleGAN, there are two networks constructed and linked together, a mapping network f and a synthesis network g. The latent code $z \in Z$ is first converted to $w \in W$ (where W is an intermediate latent space) through a non-linear function $f : Z \to W$, which is characterized by a neural network (i.e., the mapping network) consisting of several fully connected layers. Using an affine transformation, the intermediate representation w is specialized to styles $y = (y_s, y_b)$ that will be fed to the adaptive instance normalization (AdaIN) operations, specified as:

$$\mathrm{AdaIN}(x_i, y) = y_{s,i}\frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i} \quad (2)$$

where each feature map x_i is normalized separately. The StyleGAN generator architecture allows controlling the image synthesis by modifying the styles via different scales. In addition, instead of using one random latent code during training, this method uses two latent codes to generate a given proportion of images. More specifically, two latent codes z_1 and z_2 are fed to the mapping network to create respectively w_1 and w_2 that control the styles by applying w_1 before and w_2 after the crossover point. Fig. 5 demonstrates examples of images created by mixing two latent codes at three different scales where each subset of styles controls separate meaningful high-level attributes of the image. In other words, the generator architecture of StyleGAN is able to learn separation of high-level attributes (e.g., pose and identity when trained on human faces) and enables intuitive, scale-specific control of the face synthesis.

3. Deepfake detection

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 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

Table 1
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	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 (Liu et al., 2019) and SPADE (Park et al., 2019).
DeepFaceLab	https://github.com/iperov/DeepFaceLab	- Expand from the Faceswap method with new models, e.g. H64, H128, LIAEF128, SAE (DeepFaceLab, 2022a). - Support multiple face extraction modes, e.g. S3FD, MTCNN, dlib, or manual (DeepFaceLab, 2022a).
DFaker	https://github.com/dfaker/df	- DSSIM loss function (DSSIM, 2022) 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.
AvatarMe	https://github.com/lattas/AvatarMe	- Reconstruct 3D faces from arbitrary "in-the-wild" images. - Can reconstruct authentic 4K by 6K-resolution 3D faces from a single low-resolution image (Lattas et al., 2020).
MarioNETte	https://hyperconnect.github.io/MarioNETte	- A few-shot face reenactment framework that preserves the target identity. - No additional fine-tuning phase is needed for identity adaptation (Ha et al., 2020).
DiscoFaceGAN	https://github.com/microsoft/DiscoFaceGAN	- Generate face images of virtual people with independent latent variables of identity, expression, pose, and illumination. - Embed 3D priors into adversarial learning (Deng et al., 2020).
StyleRig	https://gvv.mpi-inf.mpg.de/projects/StyleRig	- Create portrait images of faces with a rig-like control over a pretrained and fixed StyleGAN via 3D morphable face models. - Self-supervised without manual annotations (Tewari et al., 2020).
FaceShifter	https://lingzhili.com/FaceShifterPage	- Face swapping in high-fidelity by exploiting and integrating the target attributes. - Can be applied to any new face pairs without requiring subject specific training (Li et al., 2019a).
FSGAN	https://github.com/YuvalNirkin/fsgan	- A face swapping and reenactment model that can be applied to pairs of faces without requiring training on those faces. - Adjust to both pose and expression variations (Nirkin et al., 2019).
StyleGAN	https://github.com/NVlabs/stylegan	- A new generator architecture for GANs is proposed based on style transfer literature. - The new architecture leads to automatic, unsupervised separation of high-level attributes and enables intuitive, scale-specific control of the synthesis of images (Karras et al., 2019).
Face2Face	https://justusthies.github.io/posts/face2face/	- Real-time facial reenactment of monocular target video sequence, e.g. Youtube video. - Animate the facial expressions of the target video by a source actor and re-render the manipulated output video in a photo-realistic fashion (Thies et al., 2016).
Neural Textures	https://github.com/SSRSGJYD/NeuralTexture	- Feature maps that are learned as part of the scene capture process and stored as maps on top of 3D mesh proxies. - Can coherently re-render or manipulate existing video content in both static and dynamic environments at real-time rates (Thies et al., 2019).
Transformable Bottleneck Networks	https://github.com/kyleolsz/TB-Networks	- A method for fine-grained 3D manipulation of image content. - Apply spatial transformations in CNN models using a transformable bottleneck framework (Olszewski et al., 2019).
"Do as I Do" Motion Transfer	https://github.com/carolineec/EverybodyDanceNow	- Automatically transfer the motion from a source to a target person by learning a video-to-video translation. - Can create a motion-synchronized dancing video with multiple subjects (Chan et al., 2019).
Neural Voice Puppetry	https://justusthies.github.io/posts/neural-voice-puppetry	- A method for audio-driven facial video synthesis. - Synthesize videos of a talking head from an audio sequence of another person using 3D face representation. (Thies et al., 2020).

this issue, Korshunov and Marcel (2019) produced a notable deepfake dataset consisting of 620 videos based on the GAN model using the open source code Faceswap-GAN (Faceswap-GAN, 2022). Videos from the publicly available VidTIMIT database (VidTIMIT database, 2022) 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 (Parkhi et al., 2015) and Facenet (FaceNet, 2022; Schroff et al., 2015) are unable to detect deepfakes effectively. Other methods such as lip-syncing approaches (Chung et al., 2017; Suwajanakorn et al., 2017; Korshunov and Marcel, 2018b) and image quality metrics with support vector machine (SVM) (Galbally and Marcel, 2014) produce very high error rate when applied to detect deepfake videos from this newly produced dataset. 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 (Fig. 2). The latter is distinguished

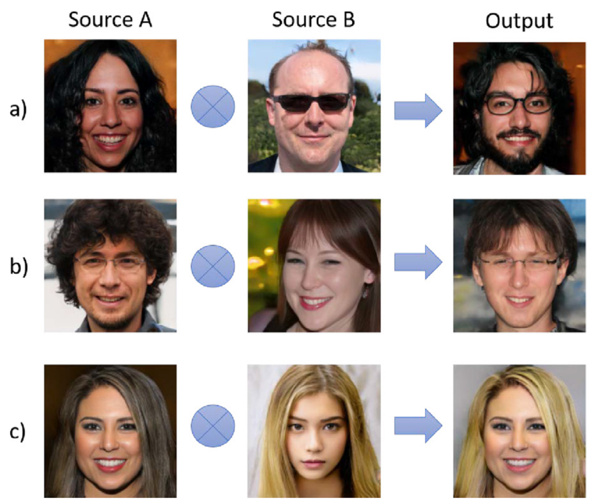

Fig. 5. Examples of mixing styles using StyleGAN: the output images are generated by copying a specified subset of styles from source B and taking the rest from source A. (a) Copying coarse styles from source B will generate images that have high-level aspects from source B and all colors and finer facial features from source A; (b) if copying the styles of middle resolutions from B, the output images will have smaller scale facial features from B and preserve the pose, general face shape, and eyeglasses from A; (c) if copying the fine styles from source B, the generated images will have the color scheme and microstructure of source B (Karras et al., 2019).

into two smaller 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.

3.1. Fake image detection

Deepfakes are increasingly detrimental to privacy, society security and democracy (Chesney and Citron, 2018a). 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 image synthesis process. Recent methods, e.g., de Lima et al. (2020) and Amerini and Caldelli (2020), have commonly applied deep learning to automatically extract salient and discriminative features to detect deepfakes.

3.1.1. Handcrafted features-based methods

Most works on detection of GAN generated images 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. (2019) 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 allows the forensic classifier to learn more intrinsic and meaningful features, which has better generalization capability than previous image forensics methods (Yang et al., 2016; Bayar and Stamm, 2016) or image steganalysis networks (Qian et al., 2015). Zhang et al. (2017) used the bag of words method to extract a set of compact features and fed it into various classifiers such as SVM (Wang et al., 2017), random forest (RF) (Bai, 2017) and multi-layer perceptrons (MLP) (Zheng et al., 2016) for discriminating swapped face images from the genuine. Among deep learning-generated images, those synthesized 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.

On the other hand, Agarwal and Varshney (2019) cast the GAN-based deepfake detection as a hypothesis testing problem where a statistical framework was introduced using the information-theoretic study of authentication (Maurer, 2000). 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 by this method.

3.1.2. Deep features-based methods

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 (Korshunova et al., 2017).

Hsu et al. (2020) 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 Chopra et al. (2005) is used. The CFFN encompasses several dense units with each unit including different numbers of dense blocks (Huang et al., 2017) to improve the representative capability for the input images. Discriminative features between the fake and real images are extracted through the CFFN learning process based on the use of pairwise information, which is the label of each pair of two input images. If the two images are of the same type, i.e., fake–fake or real–real, the pairwise label is 1. In contrast, if they are of different types, i.e., fake–real, the pairwise label is 0. The CFFN-based discriminative features are then fed to a neural network classifier 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 dataset is obtained from CelebA (Liu et al., 2015), 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 64×64, including deep convolutional GAN (DCGAN) (Radford et al., 2015), Wasserstein GAN (WGAN) (Arjovsky et al., 2017), WGAN with gradient penalty (WGAN-GP) (Gulrajani et al., 2017), least squares GAN (Mao et al., 2017), and PGGAN (Karras et al., 2017). 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 dataset is extracted from the ILSVRC12 (Russakovsky et al., 2015). The large scale GAN training model for high fidelity natural image synthesis (BIGGAN) (Brock et al., 2018), self-attention GAN (Zhang et al., 2019) and spectral normalization GAN (Miyato et al., 2018) are used to generate fake images with size of 128×128. 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 Farid (2009), Mo et al. (2018), Marra et al. (2018) and Hsu et al. (2018).

Likewise, Guo et al. (2021) proposed a CNN model, namely SCnet, to detect deepfake images, which are generated by the Glow-based facial forgery tool (Kingma and Dhariwal, 2018). The fake images synthesized by the Glow model (Kingma and Dhariwal, 2018) have the facial expression maliciously tampered. These images are hyper-realistic with perfect visual qualities, but they still have subtle or noticeable manipulation traces, which are exploited by the SCnet. The SCnet is able to automatically learn high-level forensics features of image data thanks to a hierarchical feature extraction block, which is formed by stacking four convolutional layers. Each layer learns a new

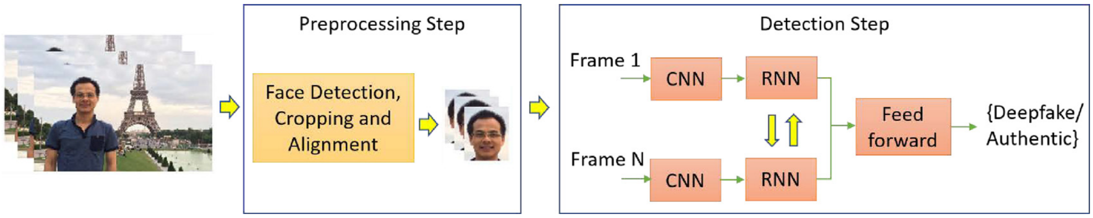

Fig. 6. A two-step process for face manipulation detection where the preprocessing step aims to detect, crop and align faces on a sequence of frames and the second step distinguishes manipulated and authentic face images by combining convolutional neural network (CNN) and recurrent neural network (RNN) (Sabir et al., 2019).

set of feature maps from the previous layer, with each convolutional operation is defined by:

$$f_j^{(n)} = \sum_{i=1}^{i} f_i^{(n-1)} * \omega_{ij}^{(n)} + b_j^{(n)} \tag{3}$$

where $f_j^{(n)}$ is the jth feature map of the nth layer, $\omega_{ij}^{(n)}$ is the weight of the ith channel of the jth convolutional kernel in the nth layer, and $b_j^{(n)}$ is the bias term of the jth convolutional kernel in the nth layer. The proposed approach is evaluated using a dataset consisting of 321,378 face images, which are created by applying the Glow model (Kingma and Dhariwal, 2018) to the CelebA face image dataset (Liu et al., 2015). Evaluation results show that the SCnet model obtains higher accuracy and better generalization than the Meso-4 model proposed in Afchar et al. (2018).

Recently, Zhao et al. (2021) proposed a method for deepfake detection using self-consistency of local source features, which are content-independent, spatially-local information of images. These features could come from either imaging pipelines, encoding methods or image synthesis approaches. The hypothesis is that a modified image would have different source features at different locations, while an original image will have the same source features across locations. These source features, represented in the form of down-sampled feature maps, are extracted by a CNN model using a special representation learning method called pairwise self-consistency learning. This learning method aims to penalize pairs of feature vectors that refer to locations from the same image for having a low cosine similarity score. At the same time, it also penalizes the pairs from different images for having a high similarity score. The learned feature maps are then fed to a classification method for deepfake detection. This proposed approach is evaluated on seven popular datasets, including FaceForensics++ (Rossler et al., 2019), DeepfakeDetection (Dufour and Gully, 2019), Celeb-DF-v1 & Celeb-DF-v2 (Li et al., 2020b), Deepfake Detection Challenge (DFDC) (Dolhansky et al., 2020), DFDC Preview (Dolhansky et al., 2019), and DeeperForensics-1.0 (Jiang et al., 2020). Experimental results demonstrate that the proposed approach is superior to state-of-the-art methods. It however may have a limitation when dealing with fake images that are generated by methods that directly output the whole images whose source features are consistent across all positions within each image.

3.2. Fake video detection

Most image detection methods cannot be used for videos because of the strong degradation of the frame data after video compression (Afchar et al., 2018). Furthermore, videos have temporal characteristics that are varied among sets of frames and they are thus challenging for methods designed to detect only still fake images. This subsection focuses on deepfake video detection methods and categorizes them into two smaller groups: methods that employ temporal features and those that explore visual artifacts within frames.

3.2.1. 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. (2019) 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 (Huang et al., 2017) and the gated recurrent unit cells (Cho et al., 2014) to exploit temporal discrepancies across frames (see Fig. 6). The proposed method is tested on the FaceForensics++ dataset, which includes 1000 videos (Rossler et al., 2019), and shows promising results.

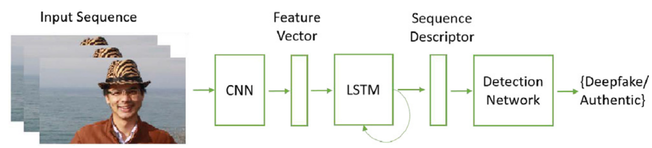

Fig. 7. A deepfake detection method using convolutional neural network (CNN) and long short term memory (LSTM) to extract temporal features of a given video sequence, which are represented via the sequence descriptor. The detection network consisting of fully-connected layers is employed to take the sequence descriptor as input and calculate probabilities of the frame sequence belonging to either authentic or deepfake class (Güera and Delp, 2018).

Likewise, Güera and Delp (2018) 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. 7. An accuracy of greater than 97% was obtained using a dataset of 600 videos, including 300 deepfake videos collected from multiple video-hosting websites and 300 pristine videos randomly selected from the Hollywood human actions dataset in Laptev et al. (2008).

On the other hand, the use of a physiological signal, eye blinking, to detect deepfakes was proposed in Li et al. (2018) 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 s, and each blink would take 0.1 and 0.4 s. 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. (2018) crop eye areas in the videos and distribute them into long-term recurrent convolutional networks (LRCN) (Donahue et al., 2015) 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.

Recently, Caldelli et al. (2021) proposed the use of optical flow to gauge the information along the temporal axis of a frame sequence for video deepfake detection. The optical flow is a vector field calculated on two temporal-distinct frames of a video that can describe the movement of objects in a scene. The optical flow fields are expected to be different between synthetically created frames and naturally generated ones (Amerini et al., 2019). Unnatural movements of lips, eyes, or of the entire faces inserted into deepfake videos would introduce distinctive motion patterns when compared with pristine ones. Based on this assumption, features consisting of optical flow fields are fed into a CNN model for discriminating between deepfakes and original videos. More specifically, the ResNet50 architecture (He et al., 2016) is implemented as a CNN model for experiments. The results obtained using the FaceForensics++ dataset (Rossler et al., 2019) show that this approach is comparable with state-of-the-art methods in terms of classification accuracy. A combination of this kind of feature with frame-based features is also experimented, which results in an improved deepfake detection performance. This demonstrates the usefulness of optical flow fields in capturing the inconsistencies on the temporal axis of video frames for deepfake detection.

3.2.2. 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.

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 (Simonyan and Zisserman, 2014), ResNet50, ResNet101 and ResNet152 (He et al., 2016). 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 Li and Lyu (2018). The proposed method is evaluated on two deepfake datasets, namely the UADFV and DeepfakeTIMIT. The UADFV dataset (Yang et al., 2019) contains 49 real videos and 49 fake videos with 32,752 frames in total. The DeepfakeTIMIT dataset (Korshunov and Marcel, 2018b) includes a set of low quality videos of 64 × 64 size and another set of high quality videos of 128 × 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 (Afchar et al., 2018), HeadPose (Yang et al., 2019), and the face tampering detection method two-stream NN (Zhou et al., 2017). 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. Nguyen et al. (2019) proposed the use of capsule networks for detecting manipulated images and videos. The capsule network was initially introduced to address limitations of CNNs when applied to inverse graphics tasks, which aim to find physical

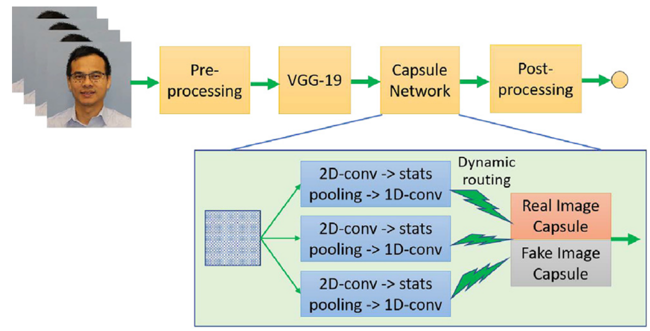

Fig. 8. Capsule network takes features obtained from the VGG-19 network (Simonyan and Zisserman, 2014) to distinguish fake images or videos from the real ones (top). The pre-processing step detects face region and scales it to the size of 128 × 128 before VGG-19 is used to extract latent features for 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 (Nguyen et al., 2019).

processes used to produce images of the world (Hinton et al., 2011). The recent development of capsule network based on dynamic routing algorithm (Sabour et al., 2017) 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 demonstrated in Fig. 8. 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 datasets covering a wide range of forged image and video attacks. They include the well-known Idiap Research Institute replay-attack dataset (Chingovska et al., 2012), the deepfake face swapping dataset created by Afchar et al. (2018), the facial reenactment FaceForensics dataset (Rössler et al., 2018), produced by the Face2Face method (Thies et al., 2016), and the fully computer-generated image dataset generated by Rahmouni et al. (2017). The proposed method yields the best performance compared to its competing methods in all of these datasets. 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.

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. (2019) 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 datasets show the great performance of the proposed approach against its competing methods. The first dataset, namely UADFV, consists of 49 deep fake videos and their respective real videos (Yang et al., 2019). The second dataset comprises 241 real images and 252 deep fake images, which is a subset of data used in the DARPA MediFor GAN Image/Video Challenge (Guan et al., 2019). Likewise, a method to exploit artifacts of deepfakes and face manipulations based on visual features of eyes, teeth and facial contours was studied in Matern et al. (2019). 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 dataset 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 Koopman et al. (2018) 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. The PRNU analysis is widely used in image forensics (Rosenfeld and Sencar, 2009; Li and Li, 2011; Lin and Li, 2016; Scherhag et al., 2019; Phan et al., 2018) and advocated to use in Koopman et al. (2018) 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. Normalized cross correlation scores are calculated for comparisons of PRNU patterns among these groups. A test dataset was created, consisting of 10 authentic videos and 16 manipulated videos, where the fake videos were produced from the genuine ones by the DeepFaceLab tool (DeepFaceLab, 2022b). The analysis shows a significant statistical difference in terms of mean normalized cross correlation scores between deepfakes and the genuine. This analysis therefore suggests that PRNU has a potential in deepfake detection although a larger dataset 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 (2019) 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, which is used to store video and its metadata in a decentralized and content-addressable manner (IPFS, 2022). 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 solution meeting security requirements. This approach is generic, and it can be extended to other types of digital content, e.g., images, audios and manuscripts.

4. Discussions and future research directions

With the support of deep learning, deepfakes can be created easier than ever before. The spread of these fake contents is also quicker thanks to the development of social media platforms (Zubiaga et al., 2018). 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 (Chesney and Citron, 2018b). 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 2. 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 (Floridi, 2018). Detection methods are still in their early stage and various methods have been proposed and evaluated but using fragmented datasets. An approach to improve performance of detection methods is to create a growing updated benchmark dataset 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 (Dolhansky et al., 2020).

Improving performance of deepfake detection methods is important, especially in cross-forgery and cross-dataset scenarios. Most detection models are designed and evaluated in the same-forgery and in-dataset experiments, which do not ensure their generalization capability. Some previous studies have addressed this issue, e.g., in Wang et al. (2020), Caldelli et al. (2021), Zhao et al. (2021), Cozzolino et al. (2018) and Marra et al. (2019), but more work needs to be done in this direction. A model trained on a specific forgery needs to be able to work against another unknown one because potential deepfake types are not normally known in the real-world scenarios. Likewise, 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 (Gandhi and Jain, 2020; Hussain et al., 2021; Carlini and Farid, 2020; Yang et al., 2021; Yeh et al., 2020). These are real challenges for detection method development and a future study 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 (Chesney and Citron, 2018b). 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 (Chesney and Citron, 2018b). 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 (Hasan and Salah, 2019) 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 (Read, 2019). 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 analyzing digital information. The development of machine 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

Table 2
Summary of prominent deepfake detection methods.

Methods	Classifiers/ Techniques	Key features	Dealing with	Datasets used
Eye blinking (Li et al., 2018)	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.
Intra-frame and temporal inconsistencies (Güera and Delp, 2018)	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 (Li and Lyu, 2018)	VGG16 (Simonyan and Zisserman, 2014), ResNet models (He et al., 2016)	Artifacts are discovered using CNN models based on resolution inconsistency between the warped face area and the surrounding context.	Videos	- UADFV (Yang et al., 2019), containing 49 real videos and 49 fake videos with 32752 frames in total. - DeepfakeTIMIT (Korshunov and Marcel, 2018b)
MesoNet (Afchar et al., 2018)	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 datasets are 98% and 95% respectively.	Videos	Two datasets: deepfake one constituted from online videos and the FaceForensics one created by the Face2Face approach (Thies et al., 2016).
Eye, teach and facial texture (Matern et al., 2019)	Logistic regression and neural network (NN)	- Exploit facial texture differences, and missing reflections and details in eye and teeth areas of deepfakes. - Logistic regression and NN are used for classifying.	Videos	A video dataset downloaded from YouTube.
Spatio-temporal features with RCN (Sabir et al., 2019)	RCN	Temporal discrepancies across frames are explored using RCN that integrates convolutional network DenseNet (Huang et al., 2017) and the gated recurrent unit cells (Cho et al., 2014)	Videos	FaceForensics++ dataset, including 1000 videos (Rossler et al., 2019).
Spatio-temporal features with LSTM (Chintha et al., 2020)	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++ (Rossler et al., 2019) and Celeb-DF (5639 deepfake videos) (Li et al., 2020b) datasets and the ASVSpoof 2019 Logical Access audio dataset (Todisco et al., 2019).
Analysis of PRNU (Koopman et al., 2018)	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 (DeepFaceLab, 2022b).
Phoneme-viseme mismatches (Agarwal et al., 2020b)	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) and others are created using synthesis techniques, i.e. Audio-to-Video (A2V) (Suwajanakorn et al., 2017) and Text-to-Video (T2V) (Fried et al., 2019).
Using attribution-based confidence (ABC) metric (Fernandes et al., 2020)	ResNet50 model (He et al., 2016), pre-trained on VGGFace2 (Cao et al., 2018)	- The ABC metric (Jha et al., 2019) 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 (Fernandes et al., 2019) and YouTube are used to generate two deepfake datasets by commercial website https://deepfakesweb.com and another deepfake dataset is DeepfakeTIMIT (Korshunov and Marcel, 2018a).
Using appearance and behavior (Agarwal et al., 2020a)	Rules based on facial and behavioral features.	Temporal, behavioral biometric based on facial expressions and head movements are learned using ResNet-101 (He et al., 2016) while static facial biometric is obtained using VGG (Parkhi et al., 2015).	Videos	The world leaders dataset (Agarwal et al., 2019), FaceForensics++ (Rossler et al., 2019), Google/Jigsaw deepfake detection dataset (Dufour and Gully, 2019), DFDC (Dolhansky et al., 2019) and Celeb-DF (Li et al., 2020b).
FakeCatcher (Ciftci et al., 2020)	CNN	Extract biological signals in portrait videos and use them as an implicit descriptor of authenticity because they are not spatially and temporally well-preserved in deepfakes.	Videos	UADFV (Yang et al., 2019), FaceForensics (Rössler et al., 2018), FaceForensics++ (Rossler et al., 2019), Celeb-DF (Li et al., 2020b), and a new dataset of 142 videos, independent of the generative model, resolution, compression, content, and context.
Emotion audio–visual affective cues (Mittal et al., 2020)	Siamese network (Chopra et al., 2005)	Modality and emotion embedding vectors for the face and speech are extracted for deepfake detection.	Videos	DeepfakeTIMIT (Korshunov and Marcel, 2018a) and DFDC (Dolhansky et al., 2019).

(*continued on next page*)

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 (Maras and Alexandrou, 2019). 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

Table 2 (*continued*).

Methods	Classifiers/ Techniques	Key features	Dealing with	Datasets used
Head poses (Yang et al., 2019)	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.
Capsule-forensics (Nguyen et al., 2019)	Capsule networks	- Latent features extracted by VGG-19 network (Simonyan and Zisserman, 2014) are fed into the capsule network for classification. - A dynamic routing algorithm (Sabour et al., 2017) 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 datasets: the Idiap Research Institute replay-attack (Chingovska et al., 2012), deepfake face swapping by Afchar et al. (2018), facial reenactment FaceForensics (Rössler et al., 2018), and fully computer-generated image set using (Rahmouni et al., 2017).
Preprocessing combined with deep network (Xuan et al., 2019)	DCGAN, WGAN-GP and PGGAN.	- Enhance generalization ability of deep learning 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 dataset: CelebA-HQ (Karras et al., 2017), including high quality face images of 1024 × 1024 resolution. - Fake datasets: generated by DCGAN (Radford et al., 2015), WGAN-GP (Gulrajani et al., 2017) and PGGAN (Karras et al., 2017).
Analyzing convolutional traces (Guarnera et al., 2020a)	KNN, SVM, and linear discriminant analysis (LDA)	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 Cho et al., 2019, StarGAN Choi et al., 2018, AttGAN (He et al., 2019), StyleGAN (Karras et al., 2019), StyleGAN2 (Karras et al., 2020)).
Bag of words and shallow classifiers (Zhang et al., 2017)	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 (Huang et al., 2007), containing 13,223 images with resolution of 250 × 250.
Pairwise learning (Hsu et al., 2020)	CNN concatenated to CFFN	Two-phase procedure: feature extraction using CFFN based on the Siamese network architecture (Chopra et al., 2005) and classification using CNN.	Images	- Face images: real ones from CelebA (Liu et al., 2015), and fake ones generated by DCGAN (Radford et al., 2015), WGAN (Arjovsky et al., 2017), WGAN-GP (Gulrajani et al., 2017), least squares GAN (Mao et al., 2017), and PGGAN (Karras et al., 2017). - General images: real ones from ILSVRC12 (Russakovsky et al., 2015), and fake ones generated by BIGGAN (Brock et al., 2018), self-attention GAN (Zhang et al., 2019) and spectral normalization GAN (Miyato et al., 2018).
Defenses against adversarial perturbations in deepfakes (Gandhi and Jain, 2020)	VGG (Parkhi et al., 2015) and ResNet (He et al., 2016)	- Introduce adversarial perturbations to enhance deepfakes and fool deepfake detectors. - Improve accuracy of deepfake detectors using Lipschitz regularization and deep image prior techniques.	Images	5000 real images from CelebA (Liu et al., 2015) and 5000 fake images created by the "Few-Shot Face Translation GAN" method (Few-Shot Face Translation GAN, 2022).
Face X-ray (Li et al., 2020a)	CNN	- Try to locate the blending boundary between the target and original faces instead of capturing the synthesized artifacts of specific manipulations. - Can be trained without fake images.	Images	FaceForensics++ (Rossler et al., 2019), DeepfakeDetection (DFD) (Dufour and Gully, 2019), DFDC (Dolhansky et al., 2019) and Celeb-DF (Li et al., 2020b).
Using common artifacts of CNN-generated images (Wang et al., 2020)	ResNet-50 (He et al., 2016) pre-trained with ImageNet (Russakovsky et al., 2015)	Train the classifier using a large number of fake images generated by a high-performing unconditional GAN model, i.e., PGGAN (Karras et al., 2017) and evaluate how well the classifier generalizes to other CNN-synthesized images.	Images	A new dataset of CNN-generated images, namely ForenSynths, consisting of synthesized images from 11 models such as StyleGAN (Karras et al., 2019), super-resolution methods (Dai et al., 2019) and FaceForensics++ (Rossler et al., 2019).
Using convolutional traces on GAN-based images (Guarnera et al., 2020b)	KNN, SVM, and LDA	Training the expectation–maximization algorithm (Moon, 1996) to detect and extract discriminative features via a fingerprint that represents the convolutional traces left by GANs during image generation.	Images	A dataset of images generated by ten GAN models, including CycleGAN (Zhu et al., 2017), StarGAN (Choi et al., 2018), AttGAN (He et al., 2019), GDWCT (Cho et al., 2019), StyleGAN (Karras et al., 2019), StyleGAN2 (Karras et al., 2020), PGGAN (Karras et al., 2017), FaceForensics++ (Rossler et al., 2019), IMLE (Li et al., 2019b), and SPADE (Park et al., 2019).
Using deep features extracted by CNN (Guo et al., 2021)	A new CNN model, namely SCnet	The CNN-based SCnet is able to automatically learn high-level forensics features of image data thanks to a hierarchical feature extraction block, which is formed by stacking four convolutional layers.	Images	A dataset of 321,378 face images, created by applying the Glow model (Kingma and Dhariwal, 2018) to the CelebA face image dataset (Liu et al., 2015).

support the determination of the authenticity of digital media and have obtained accurate and reliable results, e.g., Su et al. (2017) and Iuliani et al. (2018), 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 (Malolan et al., 2020). 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. Researchers have recently attempted to create white box and explainable detection methods. An example is the approach proposed by Giudice et al. (2021) in which they use discrete cosine transform statistics to detect so-called specific GAN frequencies to differentiate between real images and deepfakes. Through the analysis of particular frequency statistics, that method can be used to mathematically explain whether a multimedia content is a deepfake and why it is. More research must be conducted in this area and 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.

5. Conclusions

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. This survey provides a timely overview of deepfake creation and detection methods and presents a broad discussion on challenges, potential trends, and future directions in this area. This study therefore will be valuable for the artificial intelligence research community to develop effective methods for tackling deepfakes.

CRediT authorship contribution statement

Thanh Thi Nguyen: Conceptualization, Methodology, Investigation, Writing. **Quoc Viet Hung Nguyen:** Conceptualization, Writing – review & editing. **Dung Tien Nguyen:** Methodology, Writing – original draft. **Duc Thanh Nguyen:** Visualization, Writing – original draft. **Thien Huynh-The:** Validation, Writing – review & editing . **Saeid Nahavandi:** Validation, Writing – review & editing. **Thanh Tam Nguyen:** Visualization, Writing – review & editing. **Quoc-Viet Pham:** Validation, Writing – review & editing . **Cuong M. Nguyen:** Investigation, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Afchar, Darius, Nozick, Vincent, Yamagishi, Junichi, Echizen, Isao, 2018. MesoNet: A compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security, WIFS. IEEE, pp. 1–7.

Agarwal, Shruti, Farid, Hany, El-Gaaly, Tarek, Lim, Ser-Nam, 2020a. Detecting deep-fake videos from appearance and behavior. In: IEEE International Workshop on Information Forensics and Security, WIFS. IEEE, pp. 1–6.

Agarwal, Shruti, Farid, Hany, Fried, Ohad, Agrawala, Maneesh, 2020b. 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.

Agarwal, Shruti, Farid, Hany, Gu, Yuming, He, Mingming, Nagano, Koki, Li, Hao, 2019. Protecting world leaders against deep fakes. In: Computer Vision and Pattern Recognition Workshops, vol. 1, pp. 38–45.

Agarwal, Sakshi, Varshney, Lav R., 2019. Limits of deepfake detection: A robust estimation viewpoint. arXiv preprint arXiv:1905.03493.

Amerini, Irene, Caldelli, Roberto, 2020. 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.

Amerini, Irene, Galteri, Leonardo, Caldelli, Roberto, Del Bimbo, Alberto, 2019. Deepfake video detection through optical flow based CNN. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 1205–1207.

Arjovsky, Martin, Chintala, Soumith, Bottou, Léon, 2017. Wasserstein generative adversarial networks. In: International Conference on Machine Learning. PMLR, pp. 214–223.

Bai, Shuang, 2017. Growing random forest on deep convolutional neural networks for scene categorization. Expert Syst. Appl. 71, 279–287.

Bayar, Belhassen, Stamm, Matthew C., 2016. 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.

Bloomberg, 2018. How faking videos became easy and why that's so scary. https://fortune.com/2018/09/11/deep-fakes-obama-video/.

Brock, Andrew, Donahue, Jeff, Simonyan, Karen, 2018. Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096.

Caldelli, Roberto, Galteri, Leonardo, Amerini, Irene, Del Bimbo, Alberto, 2021. Optical flow based CNN for detection of unlearnt deepfake manipulations. Pattern Recognit. Lett. 146, 31–37.

Cao, Qiong, Shen, Li, Xie, Weidi, Parkhi, Omkar M, Zisserman, Andrew, 2018. VGGFace2: A dataset for recognising faces across pose and age. In: The 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018. IEEE, pp. 67–74.

Carlini, Nicholas, Farid, Hany, 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.

Chan, Caroline, Ginosar, Shiry, Zhou, Tinghui, Efros, Alexei A, 2019. Everybody dance now. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5933–5942.

Cheng, Zhengxue, Sun, Heming, Takeuchi, Masaru, Katto, Jiro, 2019. Energy compaction-based image compression using convolutional autoencoder. IEEE Trans. Multimed. 22 (4), 860–873.

Chesney, Robert, Citron, Danielle Keats, 2018a. Deep fakes: A looming challenge for privacy, democracy, and national security. Democr. National Secur. 107.

Chesney, R., Citron, D.K., 2018b. Disinformation on steroids: The threat of deep fakes. https://www.cfr.org/report/deep-fake-disinformation-steroids.

Chesney, Robert, Citron, Danielle, 2019. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. Foreign Aff. 98, 147.

Chingovska, Ivana, Anjos, André, Marcel, Sébastien, 2012. On the effectiveness of local binary patterns in face anti-spoofing. In: Proceedings of the International Conference of Biometrics Secial Interest Group, BIOSIG. IEEE, pp. 1–7.

Chintha, Akash, Thai, Bao, Sohrawardi, Saniat Javid, Bhatt, Kartavya, Hickerson, Andrea, Wright, Matthew, Ptucha, Raymond, 2020. Recurrent convolutional structures for audio spoof and video deepfake detection. IEEE J. Sel. Top. Sign. Proces. 14 (5), 1024–1037.

Cho, Wonwoong, Choi, Sungha, Park, David Keetae, Shin, Inkyu, Choo, Jaegul, 2019. Image-to-image translation via group-wise deep whitening-and-coloring transformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10639–10647.

Cho, Kyunghyun, Van Merriënboer, Bart, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, Bengio, Yoshua, 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Choi, Yunjey, Choi, Minje, Kim, Munyoung, Ha, Jung-Woo, Kim, Sunghun, Choo, Jaegul, 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.

Chopra, Sumit, Hadsell, Raia, LeCun, Yann, 2005. Learning a similarity metric discriminatively, with application to face verification. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'05, vol. 1. IEEE, pp. 539–546.

Chorowski, Jan, Weiss, Ron J, Bengio, Samy, van den Oord, Aäron, 2019. Unsupervised speech representation learning using WaveNet autoencoders. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (12), 2041–2053.

Chung, Joon Son, Senior, Andrew, Vinyals, Oriol, Zisserman, Andrew, 2017. Lip reading sentences in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR. IEEE, pp. 3444–3453.

Ciftci, Umur Aybars, Demir, Ilke, Yin, Lijun, 2020. FakeCatcher: Detection of synthetic portrait videos using biological signals. IEEE Trans. Pattern Anal. Mach. Intell..

Cozzolino, Davide, Thies, Justus, Rössler, Andreas, Riess, Christian, Nießner, Matthias, Verdoliva, Luisa, 2018. ForensicTransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510.

CycleGAN, 2022. https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.

Dai, Tao, Cai, Jianrui, Zhang, Yongbing, Xia, Shu-Tao, Zhang, Lei, 2019. Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11065–11074.

Damiani, J., 2019. A voice deepfake was used to scam a CEO out of $243,000. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/.

DeepFaceLab, 2022a. DeepFaceLab: Explained and usage tutorial. https://mrdeepfakes.com/forums/thread-deepfacelab-explained-and-usage-tutorial.

DeepFaceLab, 2022b. https://github.com/iperov/DeepFaceLab.

DeepFake_tf, 2022. DeepFake_tf: Deepfake based on tensorflow. https://github.com/StromWine/DeepFake_tf.

Deng, Yu, Yang, Jiaolong, Chen, Dong, Wen, Fang, Tong, Xin, 2020. Disentangled and controllable face image generation via 3D imitative-contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5154–5163.

DFaker, 2022. https://github.com/dfaker/df.

Dolhansky, Brian, Bitton, Joanna, Pflaum, Ben, Lu, Jikuo, Howes, Russ, Wang, Menglin, Canton Ferrer, Cristian, 2020. The deepfake detection challenge dataset. arXiv preprint arXiv:2006.07397.

Dolhansky, Brian, Howes, Russ, Pflaum, Ben, Baram, Nicole, Ferrer, Cristian Canton, 2019. The deepfake detection challenge (DFDC) preview dataset. arXiv preprint arXiv:1910.08854.

Donahue, Jeffrey, Anne Hendricks, Lisa, Guadarrama, Sergio, Rohrbach, Marcus, Venugopalan, Subhashini, Saenko, Kate, Darrell, Trevor, 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.

DSSIM, 2022. https://github.com/keras-team/keras-contrib/blob/master/keras_contrib/losses/dssim.py.

Dufour, Nick, Gully, Andrew, 2019. Contributing data to deepfake detection research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

FaceNet, 2022. https://github.com/davidsandberg/facenet.

Faceswap, 2022. Faceswap: Deepfakes software for all. https://github.com/deepfakes/faceswap.

Faceswap-GAN, 2022. https://github.com/shaoanlu/faceswap-GAN.

FakeApp, 2022. FakeApp 2.2.0. https://www.malavida.com/en/soft/fakeapp/.

Farid, Hany, 2009. Image forgery detection. IEEE Signal Process. Mag. 26 (2), 16–25.

Fernandes, Steven, Raj, Sunny, Ewetz, Rickard, Pannu, Jodh Singh, Jha, Sumit Kumar, Ortiz, Eddy, Vintila, Iustina, Salter, Margaret, 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.

Fernandes, Steven, Raj, Sunny, Ortiz, Eddy, Vintila, Iustina, Salter, Margaret, Urosevic, Gordana, Jha, Sumit, 2019. Predicting heart rate variations of deepfake videos using neural ODE. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 1721–1729.

Few-Shot Face Translation GAN, 2022. https://github.com/shaoanlu/fewshot-face-translation-GAN.

Fish, T., 2019. Deep fakes: AI-manipulated media will be 'weaponised' to trick military. https://www.express.co.uk/news/science/1109783/deep-fakes-ai-artificial-intelligence-photos-video-weaponised-china.

Floridi, Luciano, 2018. Artificial intelligence, deepfakes and a future of ectypes. Philos. Technol. 31 (3), 317–321.

Fried, Ohad, Tewari, Ayush, Zollhöfer, Michael, Finkelstein, Adam, Shechtman, Eli, Goldman, Dan B, Genova, Kyle, Jin, Zeyu, Theobalt, Christian, Agrawala, Maneesh, 2019. Text-based editing of talking-head video. ACM Trans. Graph. 38 (4), 1–14.

Galbally, Javier, Marcel, Sébastien, 2014. Face anti-spoofing based on general image quality assessment. In: The 22nd International Conference on Pattern Recognition. IEEE, pp. 1173–1178.

Gandhi, Apurva, Jain, Shomik, 2020. Adversarial perturbations fool deepfake detectors. In: IEEE International Joint Conference on Neural Networks, IJCNN. IEEE, pp. 1–8.

Giudice, Oliver, Guarnera, Luca, Battiato, Sebastiano, 2021. Fighting deepfakes by detecting GAN DCT anomalies. arXiv preprint arXiv:2101.09781.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, Bengio, Yoshua, 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27, 2672–2680.

Guan, Haiying, Kozak, Mark, Robertson, Eric, Lee, Yooyoung, Yates, Amy N, Delgado, Andrew, Zhou, Daniel, Kheyrkhah, Timothee, Smith, Jeff, Fiscus, Jonathan, 2019. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In: IEEE Winter Applications of Computer Vision Workshops, WACVW. IEEE, pp. 63–72.

Guarnera, Luca, Giudice, Oliver, Battiato, Sebastiano, 2020a. Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 666–667.

Guarnera, Luca, Giudice, Oliver, Battiato, Sebastiano, 2020b. Fighting deepfake by exposing the convolutional traces on images. IEEE Access 8, 165085–165098.

Guarnera, Luca, Giudice, Oliver, Nastasi, Cristina, Battiato, Sebastiano, 2020c. Preliminary forensics analysis of deepfake images. In: AEIT International Annual Conference, AEIT. IEEE, pp. 1–6.

Güera, David, Delp, Edward J., 2018. Deepfake video detection using recurrent neural networks. In: 15th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS. IEEE, pp. 1–6.

Gulrajani, Ishaan, Ahmed, Faruk, Arjovsky, Martin, Dumoulin, Vincent, Courville, Aaron, 2017. Improved training of Wasserstein GANs. arXiv preprint arXiv:1704.00028.

Guo, Bin, Ding, Yasan, Yao, Lina, Liang, Yunji, Yu, Zhiwen, 2020. The future of false information detection on social media: New perspectives and trends. ACM Comput. Surv. 53 (4), 1–36.

Guo, Zhiqing, Hu, Lipin, Xia, Ming, Yang, Gaobo, 2021. Blind detection of glow-based facial forgery. Multimedia Tools Appl. 80 (5), 7687–7710.

Ha, Sungjoo, Kersner, Martin, Kim, Beomsu, Seo, Seokjun, Kim, Dongyoung, 2020. Marionette: Few-shot face reenactment preserving identity of unseen targets. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, (07), pp. 10893–10900.

Hasan, Haya R., Salah, Khaled, 2019. Combating deepfake videos using blockchain and smart contracts. IEEE Access 7, 41596–41606.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian, 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

He, Zhenliang, Zuo, Wangmeng, Kan, Meina, Shan, Shiguang, Chen, Xilin, 2019. AttGAN: Facial attribute editing by only changing what you want. IEEE Trans. Image Process. 28 (11), 5464–5478.

Hinton, Geoffrey E., Krizhevsky, Alex, Wang, Sida D., 2011. Transforming auto-encoders. In: International Conference on Artificial Neural Networks. Springer, pp. 44–51.

Hsu, Chih-Chung, Lee, Chia-Yen, Zhuang, Yi-Xiu, 2018. Learning to detect fake face images in the wild. In: 2018 International Symposium on Computer, Consumer and Control, IS3C. IEEE, pp. 388–391.

Hsu, Chih-Chung, Zhuang, Yi-Xiu, Lee, Chia-Yen, 2020. Deep fake image detection based on pairwise learning. Appl. Sci. 10 (1), 370.

Huang, Gao, Liu, Zhuang, Van Der Maaten, Laurens, Weinberger, Kilian Q, 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Huang, Gary B., Ramesh, Manu, Berg, Tamara, Learned-Miller, Erik, 2007. Labeled Faces in the Wild: a Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07–49, University of Massachusetts, Amherst.

Hussain, Shehzeen, Neekhara, Paarth, Jere, Malhar, Koushanfar, Farinaz, McAuley, Julian, 2021. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3348–3357.

Hwang, T., 2020. Deepfakes: a Grounded Threat Assessment. Technical report, Centre for Security and Emerging Technologies, Georgetown University.

IPFS, 2022. IPFS powers the Distributed Web. https://ipfs.io/.

Iuliani, Massimo, Shullani, Dasara, Fontani, Marco, Meucci, Saverio, Piva, Alessandro, 2018. A video forensic framework for the unsupervised analysis of MP4-like file container. IEEE Trans. Inf. Forensics Secur. 14 (3), 635–645.

Jafar, Mousa Tayseer, Ababneh, Mohammad, Al-Zoube, Mohammad, Elhassan, Ammar, 2020. Forensics and analysis of deepfake videos. In: The 11th International Conference on Information and Communication Systems, ICICS. IEEE, pp. 053–058.

Jha, Susmit, Raj, Sunny, Fernandes, Steven, Jha, Sumit K, Jha, Somesh, Jalaian, Brian, Verma, Gunjan, Swami, Ananthram, 2019. Attribution-based confidence metric for deep neural networks. Adv. Neural Inf. Process. Syst. 32, 11826–11837.

Jiang, Liming, Li, Ren, Wu, Wayne, Qian, Chen, Loy, Chen Change, 2020. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2889–2898.

Kaliyar, Rohit Kumar, Goswami, Anurag, Narang, Pratik, 2021. DeepFakE: Improving fake news detection using tensor decomposition-based deep neural network. J. Supercomput. 77 (2), 1015–1037.

Karras, Tero, Aila, Timo, Laine, Samuli, Lehtinen, Jaakko, 2017. Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

Karras, Tero, Laine, Samuli, Aila, Timo, 2019. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410.

Karras, Tero, Laine, Samuli, Aittala, Miika, Hellsten, Janne, Lehtinen, Jaakko, Aila, Timo, 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.

Keras-VGGFace, 2022. Keras-VGGFace: VGGFace implementation with Keras framework. https://github.com/rcmalli/keras-vggface.

Kingma, Diederik P., Dhariwal, Prafulla, 2018. Glow: Generative flow with invertible 1×1 convolutions. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 10236–10245.

Kingma, Diederik P., Welling, Max, 2013. Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114.

Koopman, Marissa, Rodriguez, Andrea Macarulla, Geradts, Zeno, 2018. Detection of deepfake video manipulation. In: The 20th Irish Machine Vision and Image Processing Conference, IMVIP. pp. 133–136.

Korshunov, Pavel, Marcel, Sébastien, 2018a. Deepfakes: A new threat to face recognition? Assessment and detection. arXiv preprint arXiv:1812.08685.

Korshunov, Pavel, Marcel, Sébastien, 2018b. Speaker inconsistency detection in tampered video. In: The 26th European Signal Processing Conference, EUSIPCO. IEEE, pp. 2375–2379.

Korshunov, Pavel, Marcel, Sébastien, 2019. Vulnerability assessment and detection of deepfake videos. In: 2019 International Conference on Biometrics, ICB. IEEE, pp. 1–6.

Korshunova, Iryna, Shi, Wenzhe, Dambre, Joni, Theis, Lucas, 2017. Fast face-swap using convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3677–3685.

Laptev, Ivan, Marszalek, Marcin, Schmid, Cordelia, Rozenfeld, Benjamin, 2008. Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.

Lattas, Alexandros, Moschoglou, Stylianos, Gecer, Baris, Ploumpis, Stylianos, Triantafyllou, Vasileios, Ghosh, Abhijeet, Zafeiriou, Stefanos, 2020. AvatarMe: Realistically renderable 3D facial reconstruction "in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 760–769.

Li, Lingzhi, Bao, Jianmin, Yang, Hao, Chen, Dong, Wen, Fang, 2019a. FaceShifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457.

Li, Lingzhi, Bao, Jianmin, Zhang, Ting, Yang, Hao, Chen, Dong, Wen, Fang, Guo, Baining, 2020a. Face X-Ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5001–5010.

Li, Yuezun, Chang, Ming-Ching, Lyu, Siwei, 2018. In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In: 2018 IEEE International Workshop on Information Forensics and Security, WIFS. IEEE, pp. 1–7.

Li, Chang-Tsun, Li, Yue, 2011. Color-decoupled photo response non-uniformity for digital image forensics. IEEE Trans. Circuits Syst. Video Technol. 22 (2), 260–271.

Li, Yuezun, Lyu, Siwei, 2018. Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656.

Li, Yuezun, Yang, Xin, Sun, Pu, Qi, Honggang, Lyu, Siwei, 2020b. 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.

Li, Ke, Zhang, Tianhao, Malik, Jitendra, 2019b. Diverse image synthesis from semantic layouts via conditional IMLE. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4220–4229.

de Lima, Oscar, Franklin, Sean, Basu, Shreshtha, Karwoski, Blake, George, Annet, 2020. Deepfake detection using spatiotemporal convolutional networks. arXiv preprint arXiv:2006.14749.

Lin, Xufeng, Li, Chang-Tsun, 2016. Large-scale image clustering based on camera fingerprints. IEEE Trans. Inf. Forensics Secur. 12 (4), 793–808.

Lin, Jiacheng, Li, Yang, Yang, Guanci, 2021. FPGAN: Face de-identification method with generative adversarial networks for social robots. Neural Netw. 133, 132–147.

Liu, Ming-Yu, Huang, Xun, Mallya, Arun, Karras, Tero, Aila, Timo, Lehtinen, Jaakko, Kautz, Jan, 2019. Few-shot unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10551–10560.

Liu, Ming-Yu, Huang, Xun, Yu, Jiahui, Wang, Ting-Chun, Mallya, Arun, 2021. Generative adversarial networks for image and video synthesis: Algorithms and applications. Proc. IEEE 109 (5), 839–862.

Liu, Ziwei, Luo, Ping, Wang, Xiaogang, Tang, Xiaoou, 2015. Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3730–3738.

Lyu, Siwei, 2018. Detecting 'deepfake' videos in the blink of an eye. http://theconversation.com/detecting-deepfake-videos-in-the-blink-of-an-eye-101072.

Lyu, Siwei, 2020. Deepfake detection: Current challenges and next steps. In: IEEE International Conference on Multimedia & Expo Workshops, ICMEW. IEEE, pp. 1–6.

Makhzani, Alireza, Shlens, Jonathon, Jaitly, Navdeep, Goodfellow, Ian, Frey, Brendan, 2015. Adversarial autoencoders. arXiv preprint arXiv:1511.05644.

Malolan, Badhrinarayan, Parekh, Ankit, Kazi, Faruk, 2020. Explainable deep-fake detection using visual interpretability methods. In: The 3rd International Conference on Information and Computer Technologies, ICICT. IEEE, pp. 289–293.

Mao, Xudong, Li, Qing, Xie, Haoran, Lau, Raymond YK, Wang, Zhen, Paul Smolley, Stephen, 2017. Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2794–2802.

Maras, Marie-Helen, Alexandrou, Alex, 2019. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. Int. J. Evidence Proof 23 (3), 255–262.

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

Marra, Francesco, Gragnaniello, Diego, Cozzolino, Davide, Verdoliva, Luisa, 2018. Detection of GAN-generated fake images over social networks. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval, MIPR. IEEE, pp. 384–389.

Marra, Francesco, Saltori, Cristiano, Boato, Giulia, Verdoliva, Luisa, 2019. Incremental learning for the detection and classification of gan-generated images. In: 2019 IEEE International Workshop on Information Forensics and Security, WIFS. IEEE, pp. 1–6.

Matern, Falko, Riess, Christian, Stamminger, Marc, 2019. Exploiting visual artifacts to expose deepfakes and face manipulations. In: IEEE Winter Applications of Computer Vision Workshops, WACVW. IEEE, pp. 83–92.

Maurer, Ueli M., 2000. Authentication theory and hypothesis testing. IEEE Trans. Inform. Theory 46 (4), 1350–1356.

Mirsky, Yisroel, Lee, Wenke, 2021. The creation and detection of deepfakes: A survey. ACM Comput. Surv. 54 (1), 1–41.

Mittal, Trisha, Bhattacharya, Uttaran, Chandra, Rohan, Bera, Aniket, Manocha, Dinesh, 2020. Emotions don't lie: A deepfake detection method using audio-visual affective cues. 3, arXiv preprint arXiv:2003.06711.

Miyato, Takeru, Kataoka, Toshiki, Koyama, Masanori, Yoshida, Yuichi, 2018. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957.

Mo, Huaxiao, Chen, Bolin, Luo, Weiqi, 2018. Fake faces identification via convolutional neural network. In: Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. pp. 43–47.

Moon, Todd K., 1996. The expectation-maximization algorithm. IEEE Signal Process. Mag. 13 (6), 47–60.

Nguyen, Huy H., Yamagishi, Junichi, Echizen, Isao, 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP. IEEE, pp. 2307–2311.

Nirkin, Yuval, Keller, Yosi, Hassner, Tal, 2019. FSGAN: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7184–7193.

Olszewski, Kyle, Tulyakov, Sergey, Woodford, Oliver, Li, Hao, Luo, Linjie, 2019. Transformable bottleneck networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7648–7657.

Park, Taesung, Liu, Ming-Yu, Wang, Ting-Chun, Zhu, Jun-Yan, 2019. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2337–2346.

Parkhi, Omkar M., Vedaldi, Andrea, Zisserman, Andrew, 2015. Deep face recognition. In: Proceedings of the British Machine Vision Conference, BMVC. pp. 41.1–41.12.

Phan, Quoc-Tin, Boato, Giulia, De Natale, Francesco G.B., 2018. Accurate and scalable image clustering based on sparse representation of camera fingerprint. IEEE Trans. Inf. Forensics Secur. 14 (7), 1902–1916.

Punnappurath, Abhijith, Brown, Michael S., 2019. Learning raw image reconstruction-aware deep image compressors. IEEE Trans. Pattern Anal. Mach. Intell. 42 (4), 1013–1019.

Qian, Yinlong, Dong, Jing, Wang, Wei, Tan, Tieniu, 2015. Deep learning for steganalysis via convolutional neural networks. In: Media Watermarking, Security, and Forensics, vol. 9409, p. 94090J.

Radford, Alec, Metz, Luke, Chintala, Soumith, 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Rahmouni, Nicolas, Nozick, Vincent, Yamagishi, Junichi, Echizen, Isao, 2017. Distinguishing computer graphics from natural images using convolution neural networks. In: IEEE Workshop on Information Forensics and Security, WIFS. IEEE, pp. 1–6.

Read, M., 2019. Can you spot a deepfake? Does it matter? http://nymag.com/intelligencer/2019/06/how-do-you-spot-a-deepfake-it-might-not-matter.html.

Rosenfeld, Kurt, Sencar, Husrev Taha, 2009. A study of the robustness of PRNU-based camera identification. In: Media Forensics and Security, vol. 7254, p. 72540M.

Rössler, Andreas, Cozzolino, Davide, Verdoliva, Luisa, Riess, Christian, Thies, Justus, Nießner, Matthias, 2018. FaceForensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179.

Rossler, Andreas, Cozzolino, Davide, Verdoliva, Luisa, Riess, Christian, Thies, Justus, Nießner, Matthias, 2019. FaceForensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, et al., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Sabir, Ekraam, Cheng, Jiaxin, Jaiswal, Ayush, AbdAlmageed, Wael, Masi, Iacopo, Natarajan, Prem, 2019. Recurrent convolutional strategies for face manipulation detection in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, vol. 3, (1), pp. 80–87.

Sabour, Sara, Frosst, Nicholas, Hinton, Geoffrey E., 2017. Dynamic routing between capsules. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 3859–3869.

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

Scherhag, Ulrich, Debiasi, Luca, Rathgeb, Christian, Busch, Christoph, Uhl, Andreas, 2019. Detection of face morphing attacks based on PRNU analysis. IEEE Trans. Biom. Behav. Identity Sci. 1 (4), 302–317.

Schroepfer, M., 2019. Creating a data set and a challenge for deepfakes. https://ai.facebook.com/blog/deepfake-detection-challenge.

Schroff, Florian, Kalenichenko, Dmitry, Philbin, James, 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.

Simonyan, Karen, Zisserman, Andrew, 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Su, Lichao, Li, Cuihua, Lai, Yuecong, Yang, Jianmei, 2017. A fast forgery detection algorithm based on exponential-Fourier moments for video region duplication. IEEE Trans. Multimed. 20 (4), 825–840.

Suwajanakorn, Supasorn, Seitz, Steven M, Kemelmacher-Shlizerman, Ira, 2017. Synthesizing Obama: Learning lip sync from audio. ACM Trans. Graph. 36 (4), 1–13.

Tewari, Ayush, Elgharib, Mohamed, Bharaj, Gaurav, Bernard, Florian, Seidel, Hans-Peter, Pérez, Patrick, Zollhofer, Michael, Theobalt, Christian, 2020. StyleRig: Rigging StyleGAN for 3D control over portrait images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6142–6151.

Tewari, Ayush, Zollhoefer, Michael, Bernard, Florian, Garrido, Pablo, Kim, Hyeongwoo, Perez, Patrick, Theobalt, Christian, 2018. High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 357–370.

The Guardian, 2019. Chinese deepfake app Zao sparks privacy row after going viral. https://www.theguardian.com/technology/2019/sep/02/chinese-face-swap-app-zao-triggers-privacy-fears-viral.

Thies, Justus, Elgharib, Mohamed, Tewari, Ayush, Theobalt, Christian, Nießner, Matthias, 2020. Neural voice puppetry: Audio-driven facial reenactment. In: European Conference on Computer Vision. Springer, pp. 716–731.

Thies, Justus, Zollhöfer, Michael, Nießner, Matthias, 2019. Deferred neural rendering: Image synthesis using neural textures. ACM Trans. Graph. 38 (4), 1–12.

Thies, Justus, Zollhofer, Michael, Stamminger, Marc, Theobalt, Christian, Nießner, Matthias, 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.

Todisco, Massimiliano, Wang, Xin, Vestman, Ville, Sahidullah, Md, Delgado, Héctor, Nautsch, Andreas, Yamagishi, Junichi, Evans, Nicholas, Kinnunen, Tomi, Lee, Kong Aik, 2019. ASVspoof 2019: Future horizons in spoofed and fake audio detection. arXiv preprint arXiv:1904.05441.

Tolosana, Ruben, Vera-Rodriguez, Ruben, Fierrez, Julian, Morales, Aythami, Ortega-Garcia, Javier, 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. Inf. Fusion 64, 131–148.

Trinh, Loc, Tsang, Michael, Rambhatla, Sirisha, Liu, Yan, 2021. Interpretable and trustworthy deepfake detection via dynamic prototypes. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1973–1983.

Tucker, Patrick, 2019. The newest AI-enabled weapon: 'deep-faking' photos of the earth. https://www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/.

Turek, M., 2019. Media forensics (MediFor). https://www.darpa.mil/program/media-forensics.

Verdoliva, Luisa, 2020. Media forensics and deepfakes: An overview. IEEE J. Sel. Top. Sign. Proces. 14 (5), 910–932.

VidTIMIT database, 2022. http://conradsanderson.id.au/vidtimit/.

Vincent, Pascal, Larochelle, Hugo, Bengio, Yoshua, Manzagol, Pierre-Antoine, 2008. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning. pp. 1096–1103.

Wang, Xin, Thome, Nicolas, Cord, Matthieu, 2017. Gaze latent support vector machine for image classification improved by weakly supervised region selection. Pattern Recognit. 72, 59–71.

Wang, Sheng-Yu, Wang, Oliver, Zhang, Richard, Owens, Andrew, Efros, Alexei A, 2020. CNN-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8695–8704.

Xuan, Xinsheng, Peng, Bo, Wang, Wei, Dong, Jing, 2019. On the generalization of GAN image forensics. In: Chinese Conference on Biometric Recognition. Springer, pp. 134–141.

Yang, Chaofei, Ding, Leah, Chen, Yiran, Li, Hai, 2021. Defending against gan-based deepfake attacks via transformation-aware adversarial faces. In: IEEE International Joint Conference on Neural Networks, IJCNN. IEEE, pp. 1–8.

Yang, Xin, Li, Yuezun, Lyu, Siwei, 2019. Exposing deep fakes using inconsistent head poses. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP. IEEE, pp. 8261–8265.

Yang, Pengpeng, Ni, Rongrong, Zhao, Yao, 2016. Recapture image forensics based on Laplacian convolutional neural networks. In: International Workshop on Digital Watermarking. Springer, pp. 119–128.

Yeh, Chin-Yuan, Chen, Hsi-Wen, Tsai, Shang-Lun, Wang, Sheng-De, 2020. Disrupting image-translation-based deepfake algorithms with adversarial attacks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. pp. 53–62.

Younus, Mohammed Akram, Hasan, Taha Mohammed, 2020. Effective and fast deepfake detection method based on haar wavelet transform. In: International Conference on Computer Science and Software Engineering, CSASE. IEEE, pp. 186–190.

Zakharov, Egor, Shysheya, Aliaksandra, Burkov, Egor, Lempitsky, Victor, 2019. Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9459–9468.

Zhang, Han, Goodfellow, Ian, Metaxas, Dimitris, Odena, Augustus, 2019. Self-attention generative adversarial networks. In: International Conference on Machine Learning. PMLR, pp. 7354–7363.

Zhang, Ying, Zheng, Lilei, Thing, Vrizlynn L.L., 2017. Automated face swapping and its detection. In: The 2nd International Conference on Signal and Image Processing, ICSIP. IEEE, pp. 15–19.

Zhao, Tianchen, Xu, Xiang, Xu, Mingze, Ding, Hui, Xiong, Yuanjun, Xia, Wei, 2021. Learning self-consistency for deepfake detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15023–15033.

Zheng, Lilei, Duffner, Stefan, Idrissi, Khalid, Garcia, Christophe, Baskurt, Atilla, 2016. Siamese multi-layer perceptrons for dimensionality reduction and face identification. Multimedia Tools Appl. 75 (9), 5055–5073.

Zhou, Peng, Han, Xintong, Morariu, Vlad I., Davis, Larry S., 2017. Two-stream neural networks for tampered face detection. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW. IEEE, pp. 1831–1839.

Zhou, Xinyi, Zafarani, Reza, 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Comput. Surv. 53 (5), 1–40.

Zhu, Jun-Yan, Park, Taesung, Isola, Phillip, Efros, Alexei A, 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232.

Zubiaga, Arkaitz, Aker, Ahmet, Bontcheva, Kalina, Liakata, Maria, Procter, Rob, 2018. Detection and resolution of rumours in social media: A survey. ACM Comput. Surv. 51 (2), 1–36.