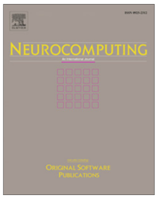
Neurocomputing 513 (2022) 351–371

Contents lists available at ScienceDirect

Neurocomputing

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



A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities

Jia Wen Seow [a,*], Mei Kuan Lim [a,*], Raphaël C.W. Phan [a], Joseph K. Liu [b]

[a] Monash University, Malaysia campus, Jalan Lagoon Selatan, 47500 Subang Jaya, Malaysia
[b] Monash University, Clayton campus, Wellington Rd, Clayton VIC 3800, Australia

ARTICLE INFO

Article history:
Received 2 April 2021
Revised 19 September 2022
Accepted 24 September 2022
Available online 28 September 2022
Communicated by Zidong Wang

Keyword:
Deepfake
Forgery generation detection
Face manipulation

ABSTRACT

When used maliciously, deepfake can pose detrimental implications to political and social forces including reducing public trust in institutions, damaging the reputation of prominent individuals, and influencing public opinions. As there is currently no specific law to address deepfakes, thus deepfake detection, which is an action to discriminate pristine media from deepfake media, plays a vital role in identifying and thwarting deepfake. This paper provides readers with a comprehensive and easy-to-understand state-of-the-art related to deepfake generation and detection. Specifically, we provide a synthesized overview and recent progress in deepfakes by categorizing our review into deepfake generation and detection. We underline publicly available deepfake generation tools and datasets for benchmarking. We also provide research insights, discuss existing gaps, and present trends for future research to facilitate the development of deepfake research.

1. Introduction

In 2017, a novel deep learning-based media forgery algorithm, better known as ‘Deepfake’ bombarded the world and wreaked havoc, threatening society’s security and privacy. It is a synthetic technique that can replace the person in an existing image or video with someone else’s characteristic or likeliness. Deepfake originated from an anonymous user under the pseudonym ’deepfake’ who uploaded numerous pornographic videos to the Reddit website by swapping the actresses’ faces with other celebrities [1].

In 2018, a Buzzfeed-published fake video of Barack Obama talking about his opinion on ’Black Panther’ and insulting the previous US President, Donald Trump [2,3], raised public awareness of Deepfake. Deepfake is more likely to target well-known people due to the rich and accessible data available on online platforms that can support deep neural network training. Imagine the consequences if the announcement video of a country leader regarding a war situation was deepfake but went viral and had deceived the public. When used maliciously, deepfake could sabotage, threaten, blackmail, inflict psychological harm, and damage reputation. It is a powerful technology that could lead to individual loss, social panic, or even threaten world peace.

In contrast, the proper use of deepfake could be beneficial to society. For instance, it can enhance the traditional pedagogical approaches to increase students’ interest in learning [4]; resurrect deceased artists for new performances [5,6], and create a memorial for those who have passed away [7]. From a medical perspective, deepfake could enable individuals who are suffering from specific forms of paralysis or physical disabilities to virtually engage with others for a better sense of participation in activities they cannot naturally take part in [4].

However, the overall negative impacts of deepfake still outweigh the positive impacts. Hence, there is a need for a robust deepfake detector to discriminate between real and fake information. In 2019, the social media giant Facebook partnered with Microsoft and various academics from different universities to create The Partnership on AI. They launched a competition entitled the ’Deepfakes Detection Challenge’ (DFDC), providing the award of up to $10 million to spur the research on deepfake detection [8]. This effort shows tremendous interest by the industry and accelerates the development of deepfake detection.

The current trend in deepfake research can be grouped into two major categories: *i. Deepfake generation*, which focused on creating, improving and stabilizing the output resolution with the least possible amount of dataset, computational power, and training time required, and *ii. Deepfake detection*, which emphasized the development of robust and generic detectors against real-time scenarios. The deepfake definition has been broadened over the years [9–

[*] Corresponding authors.
E-mail addresses: jia.seow@monash.edu (J.W. Seow), lim.meikuan@monash.edu (M.K. Lim), raphael.phan@monash.edu (R.C.W. Phan), joseph.liu@monash.edu (J.K. Liu).

https://doi.org/10.1016/j.neucom.2022.09.135
0925-2312/© 2022 Elsevier B.V. All rights reserved.

14]. In this review paper, we do not limit the understanding of deepfake to only deep learning-based face manipulation but also extend it to body motion reenactment. This review is motivated by the need to give a new impetus to deepfake research, presenting the technicality of deepfake more simply. To the best of the authors' knowledge, existing papers related to deepfake presented technical depth that may be overwhelming for new researchers in the domain.

Hence, this review paper aims to *i.* deliver a more precise overview of the different types of deepfake as well as their available generation tools and technology, *ii.* present the recent progress of deepfake detection methods with the open-source deepfake dataset, and *iii.* outline potential opportunities, research trends of deepfake in terms of application and development. Section 2 presents the standard concept of deepfake generation algorithms and the different types of deepfake and summarizes the available open-source deepfake generation tools. Then, Section 3 outlines the development of deepfake detection methodologies from the conventional handcrafted-based approaches to the recent deep learning-based techniques and listed out the publicly available deepfake training datasets. Section 4 discusses the opportunities, challenges, and future directions regarding deepfake generation and detection from both academic and industrial perspectives. Finally, a conclusion will be drawn in Section 5.

2. Deepfake Generation

Deepfake generation is a new media tampering technique that overcomes the significant flaws of traditional forgery generation approaches by reducing the manipulation traces or fingerprints that have been widely exploited for forgery detection, such as biometric or compression artifacts inconsistency [10,15]. It relies on a deep neural network that proposes learning the segmentation map or latent representation to extract input characteristics and reconstruct an entirely new fake, and yet hyperrealistic content based on the input data. With a minimal distinction between the boundary of real and fake data, detecting deepfake is more challenging than traditional manipulation media.

The three general deepfake creation models are *i.* Autoregressive model [16], *ii.* Autoencoder [17], and *iii.* Generative Adversarial Network (GAN) [18].

Autoregressive model focuses on natural image distribution instead of latent representation. It models the conditional distribution of each pixel given its previous pixels. Although it can produce high-quality images, the evaluation process is time-consuming due to its pixel-by-pixel predictions and sequential evaluation. The significant examples of this model are Pixel-RNN [19] and Pixel-CNN [20].

Autoencoder is a type of artificial neural network used for unsupervised data representation learning. It is a coupled network formed by an encoder and decoder network. The encoder converts the input to a hidden latent representation. Then, the decoder uses these representations to regenerate the data back to its original representation. The idea is to produce the output as close as possible to the input. Variational autoencoder (VAE) plays a vital role in deepfake generation.

Fig. 1 shows the basic workflow of an autoencoder. The idea of an autoencoder is to train the network to learn the input's important feature while ignoring unrelated noise. VAE is different from the conventional autoencoder. It introduces a strong assumption toward the probability distribution of latent variables, allowing it to restore complex input information better than a normal autoencoder. However, VAE is more likely to produce blurry outputs, albeit it can effortlessly generate new output after training, by sampling the distribution.

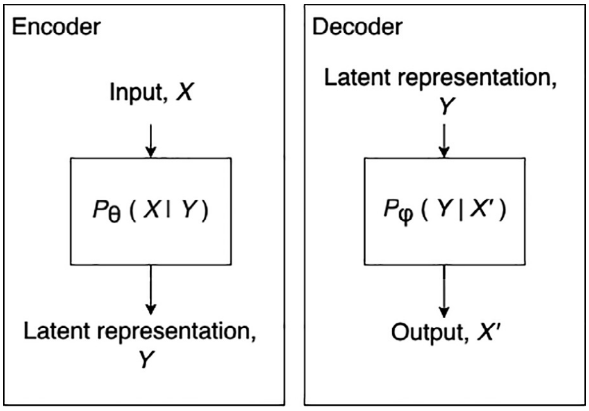


Fig. 1. The encoder, pθ converts the input data, x into a latent representation, y, and the decoder pφ reconstruct the data back as output, x'.

GAN consists of a pair of neural networks known as a generator and discriminator network. The mission of the generator network is to produce a new synthetic output based on the input's data distribution to fool the discriminator. In contrast, the discriminator network aims to differentiate precisely whether the output sample is real or fake. Both networks will continually optimize via backpropagation until they reach an equilibrium state where the fake data is indistinguishable from the real. It supports manipulation such as style transfer and image restoration that is hard to be conducted by traditional forgery generation methods [21]. Various GANs [22–29] have been published over the years to enhance the performance of software applications. For instance, ZAO [30] and FaceApp [31] show excellent performance in producing the entertainment deepfake video. The RCNN network [32] was developed in a mobile camera to enhance the camera resolution. While stacked-GAN [33] was used to solve the common low-quality deepfake synthesis issue using super-resolution to preserve more facial details in image synthesis. However, GAN requires high computational power and a vast dataset for training [22].

2.1. Types of Deepfake

In this paper, we categorized deepfake into four major groups including the entire face synthesis, reenactment (facial expression, body motion), facial attribute manipulation, and face-swapping. Fig. 2 illustrates the different deepfake types.

Entire Face Synthesis: Face synthesis results from learning the latent representation of the face dataset to generate a hyperrealistic synthetic persona. The output persona does not exist in the real world as it is produced without having a target subject. Even though this technique benefits gaming and modeling industries, the attacker could use it to fake a person's identity for illegal activities.

In 2017, Radford et al. [34] proposed a more stable generative model architecture, the DCGAN, to enhance the overall training stability. They implemented a deep convolutional concept without pooling and batch normalization to present a better image synthesis performance based on an arithmetic vector. A year later, the researchers from NVIDIA [35] introduced another network architecture named ProGAN, to further improve the output quality and stability during network training. They progressively train the input from a low-resolution and gradually improve fine details throughout the training process.

StyleGAN [36] is a network that inherits from ProGAN. Inspired by paper [37–39], the authors altered the generator architecture with an adaptive instance normalization (AdaIN) to control the generator learning at each convolutional layer. The generator tends

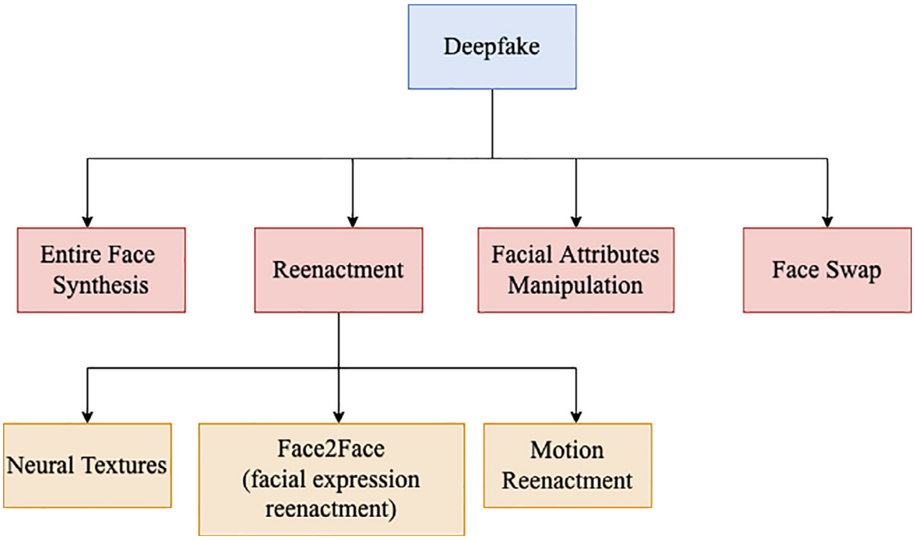


Fig. 2. The four main types of deepfake.

to synthesize a consistent style or pose based on the provided vector. They also introduced stochastic variation in controlling the placement of the synthesized person’s hairs, stubble, freckles, or skin pores. However, the researchers found that the instance normalization of StyleGAN produced significant water droplet-like artifacts in the synthetic image. This subtle trace makes it easily exposed to detection. Hence, the same research team redesigned a new normalization approach and published it as StyleGAN version 2 [40] in the same year. They successfully improved the image quality and eliminated the artifacts seamlessly. ProGAN and StyleGAN are widely used to produce synthetic face databases [41,42].

Reenactment: Unlike face synthesis, reenactment involves the transfer of facial expression or body motion between people from one to another. This technique has been popular before the appearance of deepfake. The traditional approaches leverage computer graphics to achieve reenactment results. For example, Blanz et al. reenacted the input image with a 3D morphable model extracted from the database of 3D scans [43]. Thies *et al.* promoted real-time facial expression transfer using a commodity RGB-D sensor to capture facial performance. They altered the parameters of the target to fit the source expression with a parametric model and employed a mouth retrieval synthesis to produce a high-quality outcome [44,45]. The following sections will outline the studies of several reenactment techniques, such as facial expression transfer with neural textures, the typical facial expression reenactment (Face2Face), and body motion reenactment(Puppet Master).

Neural Textures leverages feature maps, texture maps, or neural textures in the parametric vectors to synthesize photorealistic outputs. It was first introduced in [46], where the authors proposed an end-to-end deferred neural rendering network based on a convolutional encoder-decoder that integrates the knowledge of traditional computer graphics with the learnable neural textures. The authors produced a UV map that sampled the target’s neural textures and matched the source’s expression. The UV map was then fed into the neural renderer with the background image to perform synthesis and form the final reenactment result. However, the geometry proxy could easily influence the output quality. By implementing a similar concept, Fried et al. [47] focus on lip reenactment. They produced a tampered output in changing, removing, or adding dialogue of a talking-head video. The authors exploited the parametric face model to control the expression and pose alteration in different frames according to the transcript and video sequence.

Face2Face is the most common facial reenactment method in transferring expression from source to target. The attacker could control or synthesize the target to express something they would never say. A widespread incident of this reenactment method was the ’Obama deepfake video’ [2,3]. Several studies have used Barack Obama as their primary dataset for the model training. In [48], the authors performed a case study using Barack Obama’s dataset to synthesize new videos based on his voice and video stocks. They used Mel-frequency Cepstral Coefficients (MFCC) to extract the audio features from the source video, convert the target’s mouth shape into vectors, and apply Principal Component Analysis (PCA) over the frames to represent the mouth shape. Then, they trained an LSTM network to map the MFCC audio coefficients to PCA mouth shape coefficients. In the same year, Kumar et al. [49] introduced a fully trainable reenactment network to support lip synthesis based on text input using a similar dataset. The network consisted of three main modules: *i.* A text-to-speech network, CharWav, *ii.* A LSTM network for converting audio to mouth keypoints, and *iii.* A UNet-based Pix2Pix network to synthesize target video based on the mask and mouth keypoints. Similar to paper [48], Song *et al.* [50] also applied MFCC to extract audio features but they proposed a conditional recurrent network to guarantee temporal coherence and improved lip movement in an adversarial manner during training.

Numerous studies have utilized CycleGAN in their reenactment network. CycleGAN was first published in [25] to propose the translation of the image’s content from the source domain A to the target domain B without requiring paired data. The authors trained two GANs architecture with the cycle-consistency loss for learning the domain conversion between source and target in a return way to enforce the synthetic output to obtain the characteristic of the target domain while remaining its original content outframe. Nevertheless, they mentioned that the network was tailored for appearance changes and might have poor performance when the data have significant geometric or distribution differences. StarGAN [26] is a well-known CycleGAN-based translation network that focuses on multi-domain translation tasks with a single model. It eased the effort in training the transfer of multiple expressions with the support of mask vectors for different facial expression labels. In 2018, Bansal et al. [51] cooperated with Facebook Reality Lab to introduce a data-driven translation network named RecycleGAN for video retargeting. The model implemented CycleGAN with a recycle formulation, utilizing the loss function formed by recurrent loss, cycle consistency loss, recycle loss, and adversarial loss. The concept is to optimize spatio-temporal constraints to obtain better local minima during style transformation. It is worth mentioning that RecycleGAN has successfully ameliorated the mode collapse issue of conditional-GAN and achieved hyperrealism output in the facial reenactment task. To improve on mapping boundary latent space from source to target in facial reenactment, Wu *et al.* [52] leveraged CycleGAN for boundary transformation and implemented Pix2Pix encoder-decoder in reconstructing the synthetic output.

In 2020, [53] published a generic face animator named Interpretable and Controllable face reenactment network (ICface) to control pose and facial expressions. There are two major stages in the network: *i.* Facial attributes extraction of the driving image, such as interpretable head pose angles, Action Unit (AU) values, and *ii.* Integration of the captured attributes with the source image using conditional-GAN. The source image will be first converted into a neutral image with a neutralized generator before training. The result shows a notable performance in pose transformation and face reenactment in lesser semantic distortion than the baselines, yet there is still room for improvement in mitigating noticeable artifacts. In the same year, the authors of [54] proposed an Ordinal Ranking Adversarial Networks based on the concept of CycleGAN and StarGAN. The generator works with a multi-scale discriminator and one-hot label to denote the ranking of the input's age and expression intensity. This combination ensures the synthesis is correctly conducted according to the specific age groups or expression intensity. They improved the precision and performance of the condition-based synthesis.

A few studies emphasized identity invariant-based pose and expression reenactment. Shen et al. presented FaceID-GAN [55] that integrated GAN with an identity classifier to retain the identity feature during adversarial training. They used a three-dimensional morphable-based face-swapping models (3DMM) to convert input images into shape, pose, and expression parameters in synthesizing the reenact face. In [56], the authors proposed DR-GAN using an encoder-decoder structure generator to learn disentangled representation by encoding the target to a feature mapping and incorporating it with the pose code and the noise vector for face translation and transformation.

As most of the synthesis neural network training required a large dataset to learn the latent representation from different perspectives, the authors of [57] introduced a few-shot learning model. They employed a meta-learning approach in mapping face landmarks to the embedding vectors. The embedder, generator, and discriminator were trained in a K-shot learning manner. Then, the Pix2Pix generator took the output with the landmarks from a different frame to produce the synthetic target output. One of the famous Face2Face networks is FaceSwapNet [58], formed by two notable landmark handling networks. The landmark swapper module consists of two encoders and a decoder, which transfers the source's expressions to the target by computing the landmarks' loss as swapped landmarks vectors. Together with the input image, it acts as the input of a landmark-guided generator in synthesizing the final reenacted face. With this approach, the authors solved the model scalability limitation when the target is not the predefined identity and supported many-to-many face reenactment.

In [59], the authors proposed a spatial–temporal scheme. They used the factorized transposed 3D convolutional filters to enhance the optimization from both spatial and temporal aspects to produce high-quality output. The proposed model allows generating deepfake video by taking a static image with the desired motion's or expression's label and noise as the input for neural network training. However, it is challenging to transfer the expression and pose simultaneously in high resolution since it is difficult to control the probability distribution of high-resolution textures. Meanwhile, [60] suggested a stage-wise framework. The authors first introduced an encoder decoder-based semi-supervised training using the conditional vectors(pose and expression vectors) to predict the image target boundary. Then, they mapped the predicted boundary and input image as input features using two encoders, disentangled the input structure and texture in a latent space using the LightCNN network, and decoded the concatenation of the boundary and input features to perform the final target synthesis. They significantly reduced the correlation bias in conducting pose and expression transfer in high-resolution input and created a new high-resolution MVF-HQ database to support future research.

Motion Reenactment (Puppet Master) acknowledges a high level of photorealistic motion transfer. The body motion or position is retargetted from the source to the target without changing its original appearance. The common issue of this technique is pixel-to-pixel misalignment due to different sources and targets. To deal with this problem, the authors of [61] proposed an approach to utilize deformable skip-connection with the nearest neighbor loss. The idea is to decompose the source's global information into a local affine transformation set according to the target pose, then deform the source's feature map. They then applied a common skip-connection to transfer the transformed tensor and fused it with other corresponded tensors in the decoder to generate the synthetic output.

Different from feature mapping-based image reconstruction in [61], Neverova et al.[62] promoted a warping module to perform texture mapping from a source to the target, forming a high-quality texture restoration for different viewpoints and body movements. They applied a conditional generative model to target pose prediction and performed texture wrapping with Spatial Transformer Network(STN) based on each surface area's UV coordinates.

In [63], the authors introduced a modular neural network that aims to translate the changes in pose to image space. The four major modules of the network are *i.* Segmentation of the source image, to separate background and foreground, *ii.* Spatial-transformation of the segmented body parts, *iii.* Foreground synthesis of the transformed output to produce a hyperrealistic target appearance, and *iv*. Background synthesis using the foreground mask generated from the previous module to complete the body movement reenactment. This approach aims to depict the unseen poses using a generative neural network. Tulyakov et al. [64] separated the video into content and motion subspaces to further improve synthesization. A new video is synthesized by mapping the source content with the target motion vector. They named this approach MoCoGAN (Motion and Content Decomposed GAN). The author utilized Gaussian distribution to sample the content subspace and produced motion embedding. Both outputs were fed to the GRU network to create a vector set forming the motion representation. They then implemented an image generator to generate videos by sequentially mapping the frame with each motion vector and employing two discriminators to guarantee the output quality. The image discriminator aims to ensure each frame's photorealism, while the video discriminator guarantees the temporal coherence between the frames. Aberman *et al.* [65] proposed a two-branches network to also deal with the unseen poses. The first branch ensures the learning of pose-to-frame mapping, and the second one focuses on temporal coherence to convert the unseen poses into sequences that match the source video.

In [66], Kim et al. emphasized the reenactment of head pose, eye gaze, and facial expression using the novel monocular face reconstruction technique to obtain a low-dimensional parametric representation of source and target. They modified the posture, eyes, and expression parameters from source to target while retaining the source's identity and background illumination. Then, they rendered the conditioning input images according to the modified parameters. The rendering-to-translation network applied a space–time encoder and an image decoder to convert the conditioning input images to the synthetic video portrait. A similar character model-to-image translation network was published in [67]. The authors reconstructed the target from a static image to a 3D character model and trained it with the motion data to produce video-realistic output. However, they pointed out some limitations that significantly degraded the network performance, such as non-linearity of articulated motions, performance discontinuities due to self-occlusion, quality degradation due to imperfect monocular tracking, and inability to capture the challenging pose.

NVIDIA published a popular video-to-video translation network named Vid2Vid-GAN to initiate the video translation process [68]. They initiated the process of video translation by matching the source's conditional distribution to the target, synthesizing the target background and foreground using the source's segmentation mask. However, the model's performance is unreliable and inconsistent due to insufficient semantic labels in the training. To mitigate this issue, the authors proposed a few-shot Vid2Vid framework [69] by utilizing a network weight generation module with an attention mechanism. The proposed network required only several target images for the synthesis task of unseen data, but the training highly relies on the semantic estimation input, which restricted its performance.

The paper [70–72] implemented a Pix2Pix network for movement transfer. They all went through the similar preprocessing tasks, such as extracting the source's pose or motion into a set of keypoints, landmarks, or segmentation masks, then mapping it with the source foreground before feeding into Pix2Pix-GAN for synthesis. Nevertheless, the authors applied different enhancement approaches to support their solutions. In [70], the authors proposed to use FaceGAN to improve the realism of the face region. In contrast, Liu et al. [71] combined the upper body keypoints (UBKP), facial actions units, and pose (FAUP) to increase the facial detail for training. Zhou *et al.* [72] encoded position, orientation, and body parts as a Gaussian smoothed heat map to refine foreground synthesis to alleviate incoherent body artifacts.

In [73], the authors introduced an identity invariant siamese generative adversarial network (PS-GAN) to resynthesize the input according to the pose-guided image. The network consisted of two identical branches *i.* A generative network with pose-attentional transfer blocks (PATBs) that encoded the input and target pose-guided image into feature representation, and *ii.* A pair-conditional discriminator to differentiate the generated image from the real one. This study contributed to the Person Re-Identification (ReID) task by identifying the generated image's identity.

Facial Attributes Manipulation: Facial attribute manipulation involves the modification of specific facial characteristics, such as eye color, hairstyles, wrinkles, skin color, gender, and age. It can alter the appearance of a person according to the pre-set condition. StarGAN [26] is one of the representative domain-to-domain translation networks for this technique. Unlike [25,74–77] that focused on style translation between two domains, StarGAN implemented a mask vector methodology to support multi-domain training. Xiao et al. [78] proposed a similar multi-attribute CycleGAN-based translation network, ELEGANT. They emphasized the model training with adversarial loss, domain classification loss, and reconstruction loss. However, although it had provided notable style transfer results, the tampered outputs often contain unwanted artifacts, making it less desired than StarGAN.

Different from [26,78], AttGAN [79] focused more on producing high-quality facial attribute outputs. Instead of applying attribute-independent constraints in latent representation, the authors employed attribute classification constraints to ensure the preservation of attribute-excluding details during the modification. Even though they had achieved a favorable result in the current stage, it is still not feasible for large area appearance modification. On the other hand, Liu et al. introduced STGAN [80] to overcome the blurry output issue by embedding a selective transfer unit with an encoder-decoder network. The selective transfer unit algorithm was constructed based on the Gated Recurrent Units(GRU) mathematical model. The result showed that the synthetic image quality of STGAN is better than StarGAN and AttGAN. In [81], the authors implemented a URCA-GAN network based on Upsample Residual Channel-wise Attention Module(URCAM) and StarGAN. Their idea is to manipulate the specific content of the input's foreground that is different from the target image. The URCAM was used to determine the attention map and regulate the most distinctive features without affecting the spatial dimension of the transformation. They presented a higher-quality yet lesser-artifact output as compared to [25,26].

Li et al. [29] presented BeautyGAN to transfer makeup styles from one to another on an instance level while preserving the face identity. They translated the makeup style of a reference input to the target output in intra- and inter-domain perspectives to improve the translation consistency of makeup styles on different faces. The authors implemented pixel-level histogram loss as makeup loss to improve output realism in instance-level learning. They also applied a perceptual loss to maintain face identity and a cycle consistency loss to reduce artifacts. Similar to BeautyGAN, the authors in [82–84] utilized different loss functions to control attribute manipulation from different perspectives, such as age, identity, expression, and facial attributes. Nevertheless, BeautyGAN obtained the highest voting rate by 61.84% in the best makeup transfer ranking.

In [27], Jo et al. introduced SC-FEGAN, a GAN-based image editing system based on free-form masks, sketches, or colors. This approach can translate sketch input into a hyperrealistic texture form and fuse with the original image to present a high-quality yet artifact-less synthetic image. The SC-FEGAN algorithm highly relies on input processing, such as face segmentation and free-form input feature extraction. They implemented histogram equalization [85] and holistically-nested edge detection [86] to deal with the sketch data. The processing output will be fed for synthesis training with the input image. This study provided an insight into future opportunities to leverage simple drawing skills to accomplish a sophisticated image edition or restoration process. To naturally alter skin tone, Afifi *et al.* [87] proposed a color histogram-based generative model, HistoGAN, to focus on controlling the color of GAN's generated image. The model utilized StyleGAN as its backbone model but applied a color histogram instead of a fine-style vector for the last two blocks of StyleGAN.

Face Swap: Face-swapping involves switching faces from one to another while preserving the original face expression. In 2017, Korshunova et al. introduced a fast face swap methodology [88] to transform the identity of person A to person B with the condition that keeping the head position, facial expression, and lighting condition remain unchanged. Unlike the common style transfer, the style mentioned here will be the person's identity, and the rest will be the content. The authors utilized a modified multiscale convolutional neural network (CNN) with content loss, style loss, and light loss function to transform A's content into B. Using an affine transformation with 68 facial keypoints, they aligned the output face and stitched the background with a segmentation mask. As the neural network is trained to learn A's content, it can have multiple target B, which results in a one-to-many face swap algorithm. However, the training required many single-person image datasets for finetuning, which is unfavorable to the common application.

The authors of [89] presented a latent spaces-based face-swapping technique. They formulated a solution with RSGAN, promoting latent spaces' variational learning for face and hair regions.

The RSGAN consisted of two variational autoencoders (VAE) and a GAN network. The two VAEs are in charge of encoding the face and hair into latent representation, while the GAN takes care of synthesizing face-swap images according to the latent representation. The only drawback of RSGAN is that it can only perform face-swapping on a limited low resolution, 128x128 pixels. A few months later, the same research team published FSNET [90], which is an alternative method using the deep neural network (DNN) to perform face-swap synthesis with latent variables. In contrast to the previous paper [89], this network only has one VAE and GAN. The VAE encoded both face and non-face appearance into latent variables, and the GAN synthesized face-swap image using the latent variables of face *A* and non-face *B* appearance. The training used triplet loss to preserve the face identities of the face-swapped images. This approach is more stable than the traditional 3DMM.

Sun et al. proposed a face replacement method [91] for identity obfuscation. They first replaced the target face with a face from a different identity using a parametric Model-based Face Autoencoder (MoFA). Then, they trained a GAN to synthesize the swapped content based on the rendered face and obfuscated image. The GAN was also used to inpaint and blend the rendered face with the background to ensure output realism. In [28], the authors derived a recurrent neural network (RNN) approach to support face-swapping and facial reenactment. The model consists of three main components: *i.* Unet-based recurrent reenactment generator, *Gr*, *ii.* Pix2PixHD-based segmentation generator, *Gs*, and *iii.* Pix2PixHD-based inpainting generator, *Gc*. The *Gr* obtained pose and expression from the target's facial landmarks and used them to generate the source reenacted face, forming a reenacted face segmentation mask. In contrast, the *Gs* computed the hair and face segmentation mask of the target image. Then, the *Gc* inpainted the missing areas or occlusion part on the output of *Gs* to obtain a detailed, completed face-swapped result. They also implemented the Delaunay Triangulation and barycentric coordinates to ensure the facial temporal coherence to support face view interpolation. However, the output resolution tends to degrade in content with different angles.

Microsoft cooperated with the researchers from Peking University to publish a two-stage framework named FaceShifter[92] to handle occlusion during face-swapping for high fidelity output production. They performed face shifting using an Adaptive Embedding Integration Network (AEINet) and a Heuristic Error Acknowledging Network (HEARNet). The AEINet is an adaptive attentional denormalization generator that denormalized local feature integration at different levels. While the HEARNet leveraged the heuristic error between the input and reconstructed image to identify the occlusion position and further refined in a self-supervised manner. It shows a superior performance in preserving the identity and occlusive face accessories during face-swapping.

2.2. Available Deepfake Generation Tools

Many developers and researchers are enthusiastic about making their studies an open-source tool or applications that are friendly to non-technical people. This convenient resource is one factor that promotes deepfake circulation on the social media platform. Table 1 lists the deepfake generation tools and applications that are publicly available together with their features. It includes Face App [31] for facial attributes manipulation; DFaker, ZAO, Deep Face Lab, Face Swap, Deepfakes web β, Machine Tube, and Reface apps [93,30,94–98] for face swapping; Avatarify tool [99] for transferring facial expression; Impersonator++ and Jiggy tools [100,101] for movement transfer.

Table 1
Available deepfake generation tools or applications.

Ref	Deepfake Type	Tools	Feature
[31]	Facial Attributes Manipulation/ Face2Face	FaceApp	Support modification of facial expression and face attributes
[93]	Face Swap	DFaker	Support training of face swap model
[30]		ZAO	Allows face swap with celebrities from movie or TV show
[94]		DeepFaceLab	Allows face swap in videos
[95]		FaceSwap	Support face swap between peoples
[96]		Deepfakes web β	Support training of face swap model
[97]		MachineTube	Support face swap in image or video
[98]		Reface	Allows face swap with celebrities or movie character
[99]	Face2Face	Avatarify	Allows to transfer facial expression from one to the target avatar
[100]	Puppet Master	Impersonator ++	Support motion transfer using image synthesis
[101]		Jiggy	Allows to animate the person in static image to dance motion

2.3. Discussion

Face reenactment is the pioneer type of deepfake in the literature. With the exponential growth of deep learning, deepfake has covered not only the face but also body reenactment. This development brings pros and cons to society, as people could effortlessly acquire this technology with software applications. Fig. 3 shows the implementation frequency of the three major deepfake generation models over the years based on the discussed studies. Most researchers prefer to employ a stable GAN structure network in producing high-quality deepfake, although it requires a longer training time. This is because GAN can generate sharper outputs as compared to Autoencoder and Autoregressive models.

Deepfake could easily dominate falsified information to ruin a person or an organization with psychological and physical harm. Therefore, it is crucial to be aware of the different types of deepfake to obtain a clearer insight in making judgments between truth and fake media. A comprehensive understanding of various deepfake behaviors is essential for creating a reliable deepfake detection model.

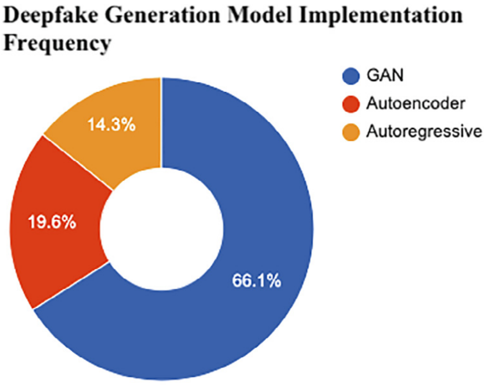


Fig. 3. Implementation frequency of the major deepfake generation models from year 2015 to 2021 based on the discussed studies.

Table 2
Summary of deepfake generation part 1.

No.	Ref.	Year	Pub.	DF Type	Technique								Output
					GAN	Au/ED	ArM	CNN	RNN	L	SC/ Att	TPF	
1	[34]	2015	ICLR	F2F	✓								Image
2	[48]	2017	ACM Trans. Graph	F2F					✓		✓	✓	Video
3	[49]	2017	NeurIPS	F2F			✓		✓			✓	Video
4	[25]	2017	ICCV	F2F	✓					✓			Image
5	[88]	2017	ICCV	FS				✓			✓		Image
6	[89]	2018	ACM SIGGRAPH	FS	✓	✓							Image
7	[90]	2018	ACCV	FS		✓							Image
8	[91]	2018	ECCV	FS	✓	✓							Image
9	[51]	2018	ECCV	F2F	✓		✓			✓			Image/ Video
10	[78]	2018	ECCV	FAM	✓								Image
11	[26]	2018	CVPR	F2F/ FAM	✓					✓			Image
12	[29]	2018	ACM-MM	FAM	✓					✓			Image
13	[22]	2018	ICLR	EFS	✓					✓			Image
14	[55]	2018	CVPR	F2F	✓			✓				✓	Image
15	[56]	2018	TPAMI	F2F	✓	✓							Image
16	[61]	2018	CVPR	MR	✓							✓	Image
17	[63]	2018	CVPR	MR	✓			✓			✓		Image
18	[62]	2018	ECCV	MR	✓						✓		Image
19	[50]	2018	IJCAI	F2F	✓				✓			✓	Video
20	[52]	2018	ECCV	F2F	✓		✓			✓			Video
21	[64]	2018	CVPR	MR	✓				✓		✓		Video
22	[66]	2018	ACM Trans. Graph	MR	✓							✓	Video
23	[68]	2018	NeurIPS	MR/ EFS	✓					✓			Video
24	[80]	2019	CVPR	FAM	✓	✓			✓				Image/ Video
25	[79]	2019	IEEE TIP	FAM	✓	✓							Image/ Video

Table 3
Summary of deepfake generation part 2.

No.	Ref.	Year	Pub.	DF Type	Technique								Output
					GAN	Au/ED	ArM	CNN	RNN	L	SC/ Att	TPF	
26	[70]	2019	ICCV	MR	✓		✓					✓	Video
27	[27]	2019	ICCV	FAM	✓						✓		Image
28	[28]	2019	ICCV	FS/ F2F			✓			✓	✓		Image/ Video
29	[36]	2019	CVPR	EFS	✓						✓		Image
30	[40]	2019	CVPR	EFS	✓								Image
31	[46]	2019	ACM Trans. Graph	NT		✓					✓	✓	Image/ Video
32	[71]	2019	ISMAR-Adjunct	MR/ F2F			✓				✓		Video
33	[47]	2019	ACM Trans. Graph	NT		✓			✓			✓	Video
34	[67]	2019	ACM Trans. Graph	MR	✓							✓	Video
35	[65]	2019	Comput Graph	MR	✓						✓		Video
36	[72]	2019	ICCV	MR			✓			✓		✓	Video
37	[57]	2019	ICCV	F2F			✓				✓		Video
38	[69]	2019	NeurIPS	MR/ F2F	✓						✓		Video
39	[58]	2019	CVPR	F2F		✓				✓		✓	Image/ Video
40	[92]	2019	CVPR	FS		✓				✓	✓		Image/ Video
41	[84]	2019	Neurocomputing	FAM	✓					✓			Image
42	[53]	2020	WACV	F2F	✓			✓					Image
43	[59]	2020	WACV	F2F	✓								Image/ Video
44	[54]	2020	IEEE TIFS	F2F	✓								Image
45	[73]	2020	Neurocomputing	MR	✓					✓			Image
46	[82]	2020	Neurocomputing	MR	✓					✓			Image
47	[83]	2020	Neurocomputing	MR	✓					✓			Image
48	[81]	2021	Neurocomputing	FAM	✓						✓		Image
49	[60]	2021	IEEE TIFS	F2F		✓		✓		✓			Image
50	[87]	2021	CVPR	FAM	✓					✓			Image

Table 2 and Table 3 summarize the key information of the discussed studies for better understanding. In each table, DF refers to Deepfake; F2F refers to Face2Face, FS refers to Face Swapping; EFS refers to Entire Face Synthesis; FA refers to Facial Attribute Manipulation; MR refers to Motion Reenactment; NT refers to Neural Texture; Au/ED refers to Autoencoder/ Encoder-Decoder; ArM refers to Autoregressive Model; L represents various loss functions; SC/ Att represents Statistical Characteristic/ Attention Mechanism; TPF represents Tertiary Preprocessing Framework, such as Char2-Wav and Morphable model. Fig. 4 presents the examples of each deepfake type.

3. Deepfake Detection

Deepfake detection can be grouped into two approaches: *i.* Conventional Methods, which rely on handcrafted features, and *ii.* Deep Learning Approaches, which emphasize on learned features. Fig. 5 outlines the taxonomy of these approaches with their subgroups. Similar to the traditional forgeries detection, such as copy-move, splicing, and inpainting, the deepfake detection pipeline also involves input preprocessing, feature extraction, and classification. However, the deep learning approaches perform feature learning between the feature extraction and classification stages.

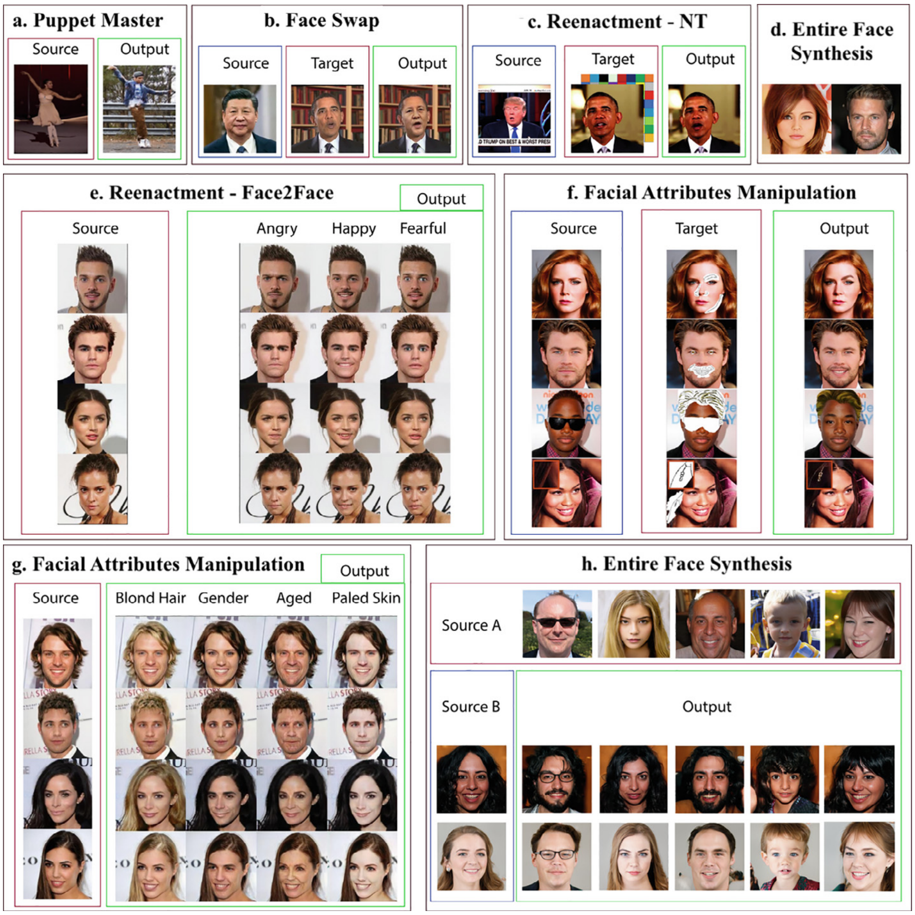


Fig. 4. Examples of deepfake. a. Puppet Master [70], b. Face Swap [28], c. Neural Texture [46], d.&h. Entire Face Synthesis [22,36], e. Face2Face [26], f.&g. Facial Attributes Manipulation [27,26].

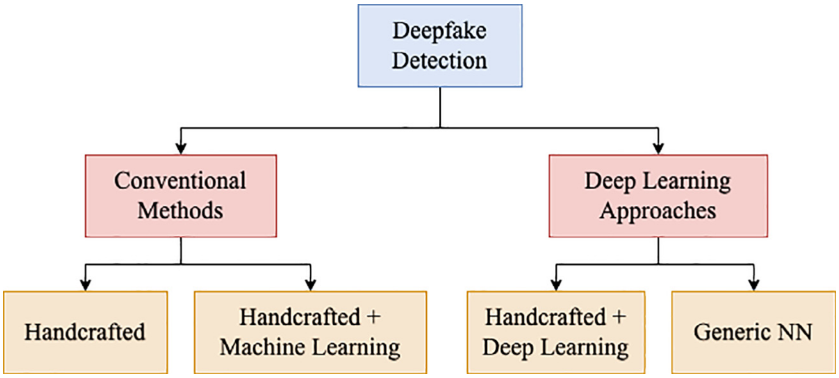


Fig. 5. Deepfake detection.

Fig. 6 presents the standard deepfake detection pipeline to illustrate the differences between the conventional and deep learning approaches in the detection process. This section reviews the studies of both conventional and deep learning-based detection methodologies.

3.1. Conventional Methods

The conventional methods involve some tedious processes for feature extraction in converting the raw data into an abstract vector representation. This is because the raw data cannot be directly input to the machine learning algorithms for classification. As a result, it needs to go through a series of sophisticated algorithms to extract meaningful information or features from the raw data before entering the classification stage. The extracted features are also known as handcrafted features. These features could be the subtle traces, pixel anomalies, edge boundary discrepancies, or the abnormal artifacts presented in the counterfeit input data. Fig. 7 shows the type of handcrafted feature.

In [102], the authors employed a Speeded Up Robust Features (SURF) algorithm with the Bag of Words model (BoW) for face swap detection. The SURF algorithm is a speeded-up version of the Scale Invariant Feature Transform (SIFT) algorithm. They used it to localize and detect features. There are two ways to select SURF

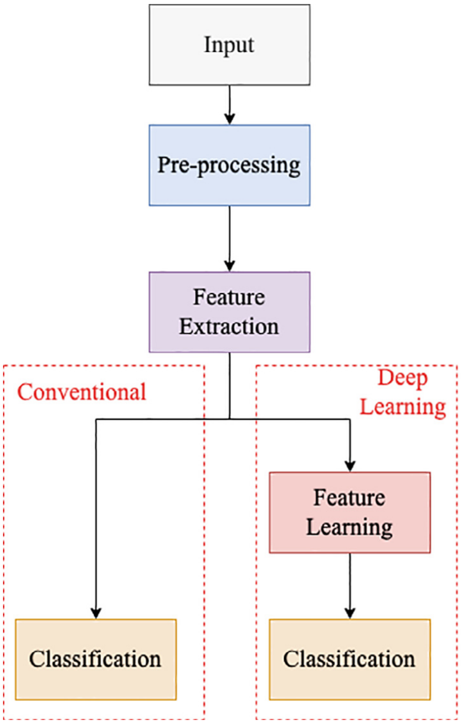


Fig. 6. Common detection pipeline.

keypoints: Grid division and Interest point detection. The selected keypoints' descriptors will be extracted, fed to the clustering system to generate features with the BoW model, and classified with Support Vector Machine (SVM). In contrast, Agarwal et al. [103] suggested feature extraction based on pixel anomalies. They applied weighted local magnitude patterns by assigning the weight inversely proportional to the absolute differences between the center and neighboring pixels. A histogram feature vector was constructed based on the output values and classified with SVM. According to the extracted features, the authors realized that the tampered image preserved high-frequency information but lost low-frequency details, exposing them to detection. For example, although the center face is well-blended, the facial keypoints of the facial features, such as the eyes, nose, and mouth, are ambiguous.

The Photo Response Non-Uniformity (PRNU) technique leverages non-uniform noise patterns for forgery detection. Koopman et al. [104] presented a pure handcrafted PRNU-based deepfake detection approach by comparing the final Welch's t-test evaluation scores for both original and deepfake videos. They first preprocessed the face region-related video frames, split them evenly over eight groups, and computed the normalized correlation scores based on each group's noise patterns. However, this method is not reliable enough as most deepfake in real-life scenarios do not have a comparison source. In [105], the authors published an unsupervised detection methodology based on a classical frequency domain analysis. This approach has no requirement with the training sample amount and mitigated the issue of paper [104]. They employed a Discrete Fourier Transform (DFT) algorithm to capture

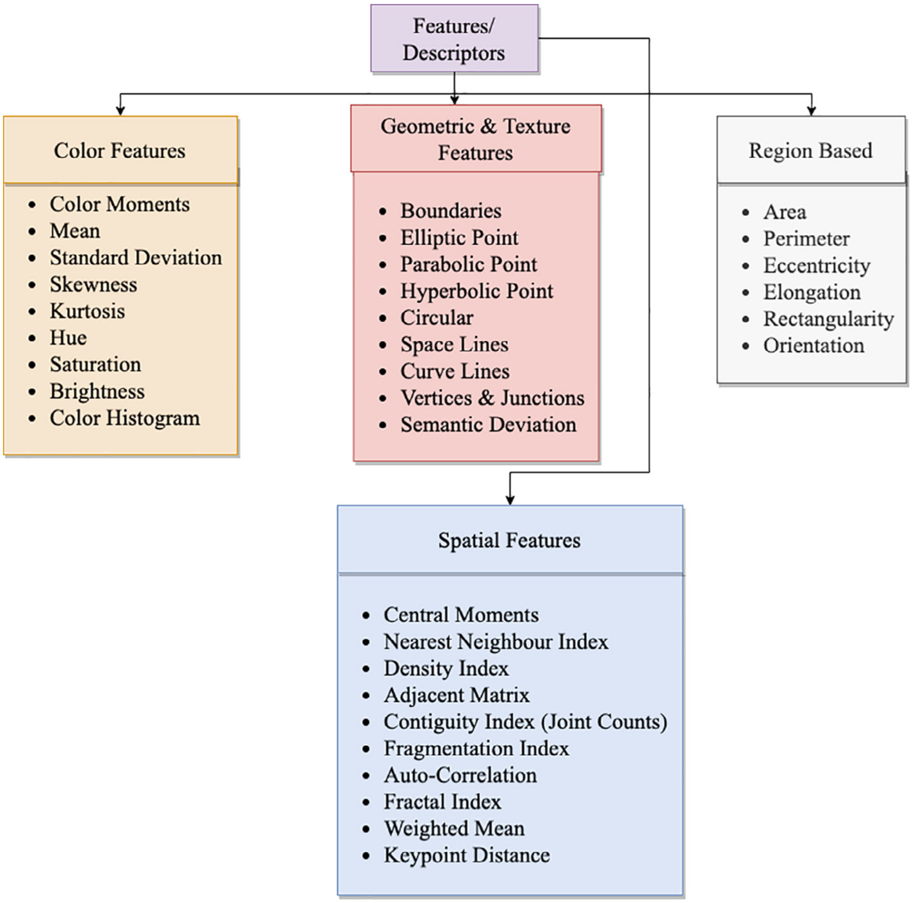


Fig. 7. Example of handcrafted features.

the deepfake image frequency by decomposing the discrete signals. Then, they applied an Azimuthal average to convert the frequency data into a 1D-representation feature vector for classification.

On the other hand, Marra et al. [106] analyzed the possibility of identifying different GANs' sources based on the fingerprints left in their deepfake output. This study shows that GANs' anomalies provided clues for deepfake detection. By working in this direction, the papers [107,108] introduced several methods to exploit the non-uniform distribution drawback of GANs during deepfake generation for deepfake detection. Both studies extracted color components from the deepfake image with different extraction approaches then converted them into feature vectors and classified them with SVM. Nevertheless, their performance could be significantly degraded with the improvement of GANs in data distribution.

Biometric artifact is another rising trend for deepfake detection. Yang et al. [109] proposed to exploit the mismatched facial landmarks of deepfake images as a clue for detection. They utilized the head orientation vector with a face detector using DLIB [110] and OpenFace2 [111]. In [112], the authors captured the unusual artifacts, such as anomalies in the reflections, eyes, or teeth details, for detection. They implemented facial geometry to train the SVM classifier with four feature vector sets: Eye, Teeth, 16-dimensional eye, teeth, and Full face crop feature vectors. These approaches exploit the drawback of deepfake in output realism.

The handcrafted feature extraction has become more challenging with the improvement of deepfake generation technologies [113]. As a result, deepfake detection research started to shift to deep learning approaches, with the aim to achieve a more flexible and reliable detection with dynamic feature learning.

3.2. Deep Learning Approaches

Learned features have proved their success in solving complex issues in various areas, such as sophisticated computer vision tasks, machine translation, face recognition, object detection, and localization [114]. Deep learning approaches employ a black box CNN feature extraction process to learn and derive features automatically from the training data using a deep neural network [10,115]. The fundamental structure of the deep learning-based methodology involves four main processes: *a.* Data Preprocessing, *b.* Feature Extraction, *c.* Feature Learning, and *d.* Classification, and it can be split into two major groups: *i.* Handcrafted feature-based feature extraction, and *ii.* Generic NN.

i. Handcrafted feature-based feature extraction

In this approach, the authors utilized specific handcrafted features or post-processing the feature extracted from NN as the input for further model training.

Biometric artifact: In [116], the authors integrated eye sequences with a long recurrent convolutional network (LRCN). They built the model using a VGG and long short-term memory (LSTM) architecture and leveraged the time-series nature of eye blinking activity to capture the temporal features for deepfake discrimination. However, the human health condition could easily influence the eye blinking frequency. Hence, this methodology might be vulnerable when dealing with people who have mental illness or nerve conduction issues. By using similar eye features, the authors of [117] presented a method that focused on evaluating the similarities between source and target eyebrow with a cosine distance metric. They hypothesized that high-resolution deepfake is more comfortable to be detected using the biometric comparison pipeline, such as eyebrow alterations. Nevertheless, this approach only applies to well-known subjects, such as politicians and celebrities, as the detection heavily relies on identity matching between the source and target and hence requires a huge amount of training samples.

Ciftci et al. [118] introduced FaceCatcher to emphasize the detection according to biological signal or Photoplethysmography (PPG). PPG is a technique that can detect subtle changes resulting in skin color due to blood pumping or peripheral circulation through the face. The variation in the PPG signal provided valuable information for deepfake detection. The authors first preprocessed the PPG signals into a feature set using the conventional signal processing methods, such as log scale, Butterworth filter, and power spectral density. Then, they applied a CNN classifier to handle the classification of the complicated feature space. By encoding the PPG feature maps with binned power spectral densities, the authors achieved an outstanding detection accuracy of 96%, but its performance might drop if the data is biased.

Different from eye sequences detection aid, Yang et al. [119] suggested using lip sequence to support deepfake discrimination based on the client talking habit. They preprocessed the raw input data by implementing a random password strategy and Dlib detector to extract the lip region, then further converted them into a lip sequence using Connectionist Temporal Classification (CTC). Next, they utilized a Dynamic Talking Habit-based Speaker Authentication network (SA-DTH-Net) to evaluate whether the extracted lip sequence conforms to the client's talking style. However, this method is not friendly to broad deepfake detection as most deepfake videos do not obtain the original actor's lip sequence for talking habit evaluation. In [120], Agarwal et al. hypothesized that the deepfake subject's mouth shape(Visemes) dynamics are sometimes inconsistent with the related spoken phoneme. They extracted the phoneme with Google's Speech-to-Text API, then manually aligned and synchronized the transcripts to the audio using P2FA [121]. They conducted experiments with different methods for visemes(classify if the mouth is open or close) measurement and realized that the CNN based-approach achieved a higher accuracy than the handcrafted-based algorithms. Nevertheless, this approach might be time-consuming as it requires many manual operations in handling phonemes and visemes alignment. Haliassos et al. [122] have published another mouth feature-based approach. Consequently, they trained the preprocessed grayscale lip-cropped frames with two lipreading pre-trained networks: a Resnet-18 model and a multi-scale temporal convolutional network (MS-TCN). The idea is to finetune the detection model with the high-level irregularities in mouth movement. They achieved a good performance in cross-dataset evaluation with an average 87.7% accuracy across five datasets. However, their model is susceptible to mouth features and failed to detect fake videos with mouth occlusion.

Spatio-Temporal Feature: In [123], the authors proposed to detect unusual motion artifacts in the deepfake video by training the network based on optical flow. Optical flow is a vector field formulated on two consecutive frames $f(t)$ and $f(t + 1)$, to extract plausible motion between the subject and the background itself. They extracted the optical flow using PWC-Net and fed it into a semi-trainable network using VGG-16 and ResNet-50 as the network backbone. While in [124], the authors developed a motion-magnified spatial–temporal representation (MMSTR) with a dual-spatial–temporal attentional network (Dual-ST AttenNet) to capture PPG variations in both spatial and temporal aspects. They preprocessed the faces to form the MMSTR maps and performed three steps in gathering the required features: *i.* Produce an adaptive spatial attention output using the MMSTR map and a spatial attention network, *ii.* Generate a block-level temporal attention using LSTM, and *iii.* Feed the motion-magnified face video to the pretrained Meso-4 network to output the frame-level temporal attention. Finally, they employed a ResNet-18 for classification by taking the features and MMSTR map as the classifier input. However, similar to [118], the PPG signals could easily be affected by other texture factors, such as skin color, sunburn, or sensitive skin.

Tariq et al. [125] applied a simple CLRNet(Convolutional LSTM Residual Network) to capture the temporal inconsistencies, such as sudden changes of brightness and facial artifacts in deepfake video. They implemented few-shot transfer learning to generalize the network by training it with several scenarios, *i.* Single-source to Single-target, *ii.* Multi-source to Single-target, and *iii.* Single-source to Multi-target.

Pixel & *Statistical Feature:* In [126], the authors adopted chrominance components for detection. They transformed the image from RGB space to a YCrCb space, extracted its edge information using a Scharr operator, and turned it into a gray level co-occurrence matrix (GLCM) for scaling. They conducted the feature extraction and classification with a depthwise separable convolution deep neural network and achieved a higher average F1 score of 0.9865 compared to other methods. Khodabakhsh et al. [127] proposed to use a ResNet-based PixelCNN++ with a universal background model (UBM). The concept is to apply a conditional probability matrix on the log-likelihood of each pixel intensity to enhance feature extraction. In [128], the authors introduced a self-supervised decoupling network(SDNN) for authenticity and compression feature learning. It exploited the compression ratio of given inputs as the self-supervised signals. The idea is to normalize the model with different compression rates so that the authenticity classifier can achieve better classification results without being affected by input compression. However, since the range of compression rates can be adjusted, the model's performance with an unseen compression rate might still be an issue.

Chen et al. [129] proposed a light-weight principal component analysis (PCA) based detection method, DefakeHop. They extracted the features from different face regions using PixelHop++ and applied subspace approximation with adjusted bias (Saab) to reduce the spatial dimension of each patch. The output was then fed to an extreme gradient boosting (XGBoost) classifier for further classification. In [130], the authors introduced a frequency-aware discriminative feature learning framework (FDFL) to solve the ambiguous feature discrimination of softmax loss and the low efficiency of handcrafted features for forgery detection. They presented a single-center loss(SCL) to bring the neutral face features to the center and push away the manipulated features. The authors found that a combined loss of SCL with softmax loss provided better results when working with the FDFL framework. However, the model has poor generalization with unseen datasets. Luo et al. [131] mentioned most CNN-detectors fail to generalize across different datasets due to overfitting in method-specific color texture. They found that image noise could efficiently remove color texture and expose forgery traces. Hence, they introduced a method using an Xception-based detector with SRM [132] high-frequency noise features. The entire model consists of three functional modules: *i.* A multi-scale high-frequency feature extraction module, *ii.* A residual guided spatial attention module, and *iii.* A dual cross-modality attention module. They adopted the suggested modules to extract more meaningful features and capture the correlation and interaction between the complementary modalities. The result shows that the model outperforms the competing approaches by more than 15% in the AUC score. In [133], Liu et al. adopted the Discrete Fourier Transform (DCT) to capture phase spectrum for deepfake detection. They hypothesized that the phase spectrum is sensitive to upsampling, which is an operation that is normally applied in deepfake generation, and assumed that the local textual information has more impact than high-level semantic information for forgery detection. However, it might be vulnerable to generation methods that do not use up-sampling.

ii. Generic NN

Generic NN is a detection methodology that comprises one or multiple neural networks conducting feature extraction and classification tasks. In contrast to other deep learning-based approaches that are integrated with handcrafted-feature, Generic NN only relies on learned-feature. *Face Recognition* & *Artifacts Discrepancies* In [134], the authors proposed a deep face recognition system named FakeSpotter. They hypothesized to monitor the neuron behavior for synthetic face detection and introduced a new neuron coverage criterion, mean neuron coverage (MNC). FakeSpotter employed a shallow layer-wise neural network architecture and specified the neuron activation using MNC. It had achieved more than 80% accuracy as compared to other detectors. Yuezu et al. [135] focused on face artifacts discrepancies. They trained several neural networks, such as VGG16, ResNet50, ResNet101, and ResNet152, to learn the discriminative features between the deepfake face areas and its neighboring regions. The idea is to exploit the artifacts introduced by the affine face warping during the deepfake generation. Similar to [135], Nirkin *et al.* [136] presented FSGAN to utilize multiple face identification networks to capture the artifact defect between the deepfake segmented face and its neighboring context. They trained two XceptionNet-based recognition systems to extract the differences between the foreground and background, then further trained them with the source embedding for further deepfake classification. Their concept is to improve the deepfake classifier's performance using recognition signals from both networks.

In [137], the authors break down the face textures into several physical decomposition groups, such as 3-dimensional shape, common texture, identity texture, ambient light, and direct light, to indicate the best combination for forgery detection. They group the ”direct light and identity texture group” as face detail and the ”3-dimensional shape, ambient light, and common texture group” as facial trend. Then, they further developed an Xception-based Forgery-Detection-with-Facial-Detail Net for deepfake detection. It is a two-stream network that combines the feature clues from both original images and facial detail. The output was fused with three approaches: *i.* Score fusion (SF), *ii.* Feature fusion (FF), and *iii.* Halfway fusion (HF). The result shows HF provided the best performance. They also integrated a supervised-detail guided attention module to enhance the forgery detection by exploring more plausible manipulated attributes.

Multi-Stream & *Multi-Stack Neural Network:* In [138], the authors suggested integrating an InceptionNet-based face classification stream with a triplet stream network for the steganalysis feature extraction. They exploited the low-level noise residual features with high-level tampering artifacts for detection enhancement. The final detection score is computed from the output scores of both streams. Another multi-stream network was presented in [139]. Kumar et al. introduced a network formed by five dedicated paralleled ResNet-18 to learn the respective face regions and capture the local facial artifacts. The combined learning of regional face areas with the full-face artifacts improved the detection performance when dealing with compressed input. Li et al. [140] implemented the face regional learning using ResNet-18 as well. They employed a Patch&Pair Convolutional Neural Networks (PPCNN) that separated the images or frames into face and non-face region patches to capture the inconsistencies between foreground and background. The embeddings of patch pairs from both branches were concatenated and passed to a classifier to compute a global decision to determine whether the input is fake or real. This study improved the neural network generalization against cross-origin deepfake.

In [141], the authors introduced a multi-stack neural network called DeepfakeStack. It consisted of two major sections: *i.* Base-Learners Creation, and *ii.* Stack Generalization. Base-Learners Creation initialized by seven deep learning models (XceptionNet, MobileNet, ResNet101, InceptionV3, DensNet121, InceptionReseNetV2, DenseNet169) that applied ImageNet weights for transfer learning. They connected them by replacing the topmost layer with two output layers and the softmax activation function. They also employed the Greedy Layer-wise Pretraining (GLP) algorithms for model training. The Stack Generalization is a meta-learner formed by a CNN classifier named DeepfakeStackClassifier(DFC). It is integrated with a larger multi-head neural network to evaluate the best detection outcome according to the predictions from each base-learner.

Shallow-CNN: In [142], Afchar et al. published two shallow network architectures to promote the learning of a simpler and localized pattern to capture the detail discriminative features with the mesoscopic features. The two proposed networks are *i.* Meso-4, formed by four convolution layers which connected to a pooling layer and followed by a fully-connected layer with a hidden layer *ii.* MesoInception-4, an enhanced version of Meso-4 with the implementation of the inception modules. It is worth mentioning that MesoInception-4 outperforms Meso-4 significantly. This study achieved a 95% detection accuracy with the FaceForensic++ dataset and was widely used as a benchmark in other deepfake detection tasks [143,144,139,145]. In [146], the authors created a shallow convolutional network(ShallowNet) that tends to detect the subtle differences between deepfake and source images with high accuracy. They realized that a shallower network with a max-pooling layer could perform better on small resolution images. However, although it can significantly reduce the training time, it failed to retain its high detection performance when dealing with high compression input [139,147].

Attention Mechanism: Fernando et al. [148] introduced a hierarchical attention memory network(HAMC), which employed an attention mechanism and a bidirectional-GRU to extract useful facial attribute features for deepfake future semantic anticipation. More precisely, the concept is to make an unseen deepfake evaluation based on the previously seen deepfake samples. They extracted the local patches of feature embeddings with a pre-trained ResNet. Then, they passed them to the bidirectional GRU and applied different weights and biases to foster the network in learning patch features at different attention levels. They adversarially trained the output encodings and ground truth with a discriminator for final deepfake classification. Similar to [148], Dang et al. [42] proposed to insert an attention map to the backbone network to enhance the feature map for deepfake classification. They applied the idea of camera model identification, using 'fingerprint' in the source image to discriminate between the deepfake and source data. The attention map consisted of various receptive fields and encoded the high-frequency fingerprint for classification. Different from [148,42], Zhao et al. [149] introduced a multiple-attentional framework that used EfficientNet-b4 as the network backbone to extract and aggregate both low-level texture and high-level semantic features of multiple attention maps for forgery detection. They applied a regional independence loss function, the bilinear attention pooling loss(BAP), and an attention-guided data augmentation mechanism to regularize each attention map in learning different semantic regions and non-overlap discriminative feature information. They assumed the subtle differences of low-level texture mostly disappear in the deeper layer and showed that enhancing the textural feature from shallow layers helps stimulate the learning of discriminative features in the forged region.

Contrastive Loss & Triplet Loss: To improve detector generalization, Hsu et al. [113] proposed a deep forgery discriminator (DeepFD) that employed contrastive loss in learning discriminative features across different GANs. The network architecture is similar to a siamese network, and the contrastive loss was computed based on pairwise learning. This study has been further enhanced with the improved algorithms in [150,151]. In [150], the authors introduced a Common Fake Feature Network (CFFN) that implemented DenseNet with the cross-layer features. It significantly improved the performance in both feature extraction and fake image recognition. They applied and trained the contrastive loss with the CFFN in a novel two-step learning policy. Similar to [150], Zhuang *et al.* [151] also adopted a two-step learning policy for model training. However, they utilized a triplet loss for optimization instead of a contrastive loss. This alteration is due to the poor performance of contrastive loss when dealing with data in the same category, such as Fake-Fake, or Real-Real. The model can prominently differentiate the positive and negative samples with triplet loss. They presented a new siamese network structure called Coupled Deep Neural Network (CDNN) to capture local and global features and had achieved a high precision of 98.6% in detection. However, the result of [113,150,151] might drop if handling the test data with distorted spatial information or different resolutions. Mittal et al. [152] applied facial and speech embedding vectors with a triplet loss function. They tend to maximize the similarity between modalities and source video and minimize the similarity between modalities and deepfake video. They integrated the memory fusion network(MSN) with CNN for deepfake prediction.

CNN Rearchitecture: Reconstruction of CNN architecture is a common technique to improve deep learning classification efficiency. It involves the alteration and enhancement in designing the kernel, filter, convolutional layer, and hyper-parameters. In [153], the authors hypothesized the residual domain could reflect the discriminative feature. They restructured the CNN by integrating a high pass filter, transformed the input image into residuals, and fed them to a three-layer convolutional network for training. They experimented with three sets of high pass filters, starting from the low to high filter dimensions. The result showed that the highest dimension high pass filter obtained the lowest detection accuracy, while the rest attained similar results. The adequate pairing of a high pass filter is required to enhance the performance of the CNN in achieving high detection accuracy.

Guo et al. [154] suggested using the adaptive convolutional layer to improve the detection accuracy. They proposed an adaptive manipulation traces extraction network (AMTEN) that extracted the feature map from the input image and used it to subtract the source image to obtain the low-level manipulation traces. The hierarchical feature extraction is formed by repeating the procedure to acquire higher-level discriminative features with the subsequent convolutional layers. In [155], Do et al. devised a VGGNet-based face detection network(VGGFace). Different to paper [153,154], it emphasized the feature extraction and hyper-parameters fine-tuning based on the face recognition.

Capsule Network: In 2017, Hinton et al. pinpointed that the lacking consideration of relative spatial and orientation relationships during network training is the major limitation of CNN. Hence, they published a more robust network architecture based on the capsule concept to mitigate this issue [156]. In [144], the authors introduced a capsule-forensic network for deepfake detection. They utilized VGG-19 for latent feature extraction and consisted of three primary and two output capsules. The authors slightly improved the algorithm published in [156] by applying Gaussian random noise to the 3D weight tensor and implementing an extra squash function before routing by iterating. The agreement between the low-level and high-level capsules will predict the probability of the input being fake or real. They then published an enhanced version [143] with a more detailed capsule architecture to yield better performance.

Recurrent Neural Network: RNN is widely used in exploring temporal discrepancies features. In [157], the authors formulated the following findings: *i.* DenseNet outperformed ResNet in feature extraction, *ii.* Face alignment in preprocessing can improve the training performance, *iii.* Evaluation on a sequence of images provided better results than a single frame input, *iv.* The bidirectional recurrent network performed better than the uni-directional recurrent network. They realized that the best approach is to preprocess face alignment using the facial landmark method rather than directly taking the raw data. They extracted the learned features using DenseNet and trained multiple RNNs at different levels of the backbone net's hierarchy. The authors of [158] also proposed to leverage temporal inconsistencies across the frames. They exploited the capsule network as a feature extractor and fed the output sequence to the LSTM to capture the temporal features. They found that the performance is better with equal interval frame selection than the single frame selection or the frame selection based on the modification level. However, it could be degraded if there are minor or consistent discrepancies across frames. Amerini et al. [159] preprocessed and transformed the input to compute a set of correlated inter-frame prediction errors. They utilized them to capture the temporal correlation among the consecutive frames via sequence learning with the CNN and LSTM. Masi et al. [160] implemented a two branches structure to encode the color domain and frequency with the Laplacian of Gaussian layers (LoG) to amplify the deepfake artifacts. They fed the combined feature maps to the bi-directional LSTM for further time-series training and classification. Similar to [160], Sun *et al.* [161] adopted a two-stream network to mine geometric features from the extracted facial landmarks. They preprocessed the video into frames and extracted the facial landmarks using Dlib. Two different feature vectors were generated from the facial landmarks and input separately to each branch of the two-stream RNN. However, although they achieved a great result on the Face Forensics++ dataset, the performance dropped when tested with the CelebDF dataset, showing their incapability in assuring model generalization. Furthermore, the complicated calibration process in mining geometric features might make it hard for duplication.

Autoencoder: Du et al. [145] introduced a locality-aware autoEncoder (LAE) to prevent overfitting in the detection model. They utilized the latent space loss and reconstruction loss to enforce the semi-supervised learning of data's intrinsic representation. In [162], the authors trained a one-class variational autoencoder (VAE) named OC-FakeDect to detect deepfake based on image reconstruction. They leveraged an anomaly score to formulate a threshold to distinguish deepfake data. The anomaly score can be obtained by computing the Root Mean Squared Error(RMSE) between the source and reconstructed images of VAE.

Multi-Person Forgery: In 2021, Zhou et al. [163] published a novel wild dataset that consists of an average of three people per scene to stimulate real-case scenarios well. They addressed a multi-temporal instance learning methodology to deal with this multi-person forgery. The approach is formed by three significant modules: *i.* A multi-temporal scale instance feature aggregation module, *ii.* An attention-based bag feature aggregation module, and *iii.* A sparse attention regulation loss, Since most approaches put effort into single face forgery detection, this study provided important insight into effectively detecting deepfake with multi-person per scene using lesser label cost.

Table 4, Table 5, Table 6, presented a comprehensive summary of the related deepfake detection approaches. There are four types of evaluation metric shown under the performances column. The presented detection result of each paper was selected following the priority of the most common metrics used, which are Accuracy (ACC), Area under the ROC Curve(AUC), Equal Error Rate (EER), and False Rejection Rate (FRR).

3.3. Available Deepfake Dataset

In the past two years, diverse datasets have been released to facilitate deepfake-related research and experiment. It is crucial to understand the dataset characteristics, including the quality, quantity, and manipulation techniques, to avoid overfitting or underfitting during practical training. Hence, this section will discuss the popular publicly available deepfake datasets. Table 7 summarized the discussed deepfake datasets.

UADFV: dataset was created by Yang et al. [116,109] for their deepfake detection experiments. They collected 49 real videos from YouTube and generated 49 fake videos using the FakeApp application. They swapped the faces in the real videos with an American actor named Nicolas Cage. Therefore, he is the only identity reflected in all fake videos. The ratio of authentic to synthetic videos is 1.0:1.0, with a resolution of 294 x 500 pixels for each video.

DeepfakeTIMIT: is a video dataset [11] modified from VidTIMIT dataset [170]. The authors applied an open-source FaceSwap-GAN approach for face-swapping. They manually selected 16 couples to comprise 620 face-swap-based deepfake videos from 32 subjects without manipulating the videos' original audio channel. The fake videos have two major quality standards: *i.* 64 × 64 pixels of low quality (LQ) images, and *ii.* 128 × 128 pixels of high quality (HQ) images.

Khodabakhsh et al. [164] proposed **Fake Faces in the Wild (FFW)** dataset as a benchmark to evaluate the generalizability of fake face detection. They gathered 150 source videos from YouTube and manipulated them utilizing deepfake and conventional tampering techniques, such as computer graphics image (CGI) and splicing.

FaceForensics++: [147] is a popular deepfake dataset that is rich in deepfake types. The authors produced 4,000 fake video sequences by tampering 1,000 Youtube source videos with four state-of-the-art automated face manipulation algorithms, including Deepfakes, Neural Texture, Face2Face, and FaceSwap. To support a more robust evaluation, they compressed the H.264 format videos with the compression rate factor of 0, 23, and 40. The dataset was then upgraded as **DeepfakeDetection dataset (DDD/ DFD)** by Google and Jigsaw [165] with another 363 original videos. They hired 28 actors to generate 3,086 high-quality deepfake videos in 16 different scenes.

Liu et al. [166] introduced a large-scale dataset called **Celeb-DF**, consisting of 590 real and 5,693 high-quality fake videos. The

Table 4
Summary of pure handcrafted feature & handcrafted-feature with machine learning based deepfake detection approaches.

No.	Ref.	Pub.	Year	Handcrafted Features			Classifer	Dataset	Performances
				Color	Texture	Spatial			
1	[102]	ICSIP	2017			✓	SVM	LFW Face Database	0.929 (ACC)
2	[103]	IJCB	2017	✓		✓	SVM	SWAPPED digital attack video face database	24.50 (EER)
3	[104]	IMVIP	2018			✓	-	Self-Created Videos	-
4	[107]	Signal Process.	2018	✓			Binary, SVM	CelebA, CelebA-HQ, Labeled Faces in the Wild	1.0 (ACC)
5	[108]	ArXiv	2018	✓		✓	SVM	LSUN	0.700 (AUC)
6	[105]	ArXiv	2019			✓	SVM	Faces-HQ, CelebA, FaceForensics++	1.0 (ACC)
7	[109]	ICASSP	2019		✓		SVM	UADFV, DARPA MediFor GAN Image,Video Challenge	0.890 (AUC)
8	[112]	WACVW	2019		✓		Logistic Regression	CelebA, ProGAN, Glow, FaceForensics	0.866 (AUC)
9	[129]	ICME	2021			✓	XGBoost	UADFV, Celeb-DF V1, Celeb-DF V2, Face Forensics ++	0.959 (AUC)

Table 5
Summary of deep Learning-based deepfake detection part 1.

No.	Ref.	Pub.	Year	Techniques							Model	Dataset	Performances
				Color	Texture	Spatial	Att.	Loss	NN Rearchitecture	Multi Stream/ Network			
1	[135]	CVPR	2018		✓						ResNet, VGG	UADFV, DeepfakeTIMIT	0.990 (AUC)
2	[116]	WIFS	2018		✓						CNN, LSTM	CEW, Eye Blinking Video (EBV)	0.990 (AUC)
3	[138]	CVPRW	2018							✓	Inception	SwapMe and FaceSwap dataset	0.928 (AUC)
4	[142]	WIFS	2018						✓		Inception	FaceForensics++	0.984 (ACC)
5	[113]	IS3C	2018					✓		✓	CNN	GAN-Generated Images Based On CelebA	0.947 (ACC)
6	[153]	IH and MMSec	2018						✓		CNN	CelebA-HQ	0.980 (ACC)
7	[155]	ISITC	2018						✓		CNN, VGG	CelebA, DC-GAN and PG-GAN generated images	0.800 (ACC)
8	[148]	IEEE TIFS	2019				✓				ResNet, GRU	FaceForensics, FaceForensics++, FakeFace in the Wild	0.999 (ACC)
9	[146]	MPS	2019						✓		CNN	CelebA, PG-GAN generated images	0.999 (ACC)
10	[123]	ICCVW	2019	✓		✓					ResNet, VGG	FaceForensics++	0.816 (ACC)
11	[144]	ICASSP	2019						✓		Capsule	FaceForensics++	0.994 (ACC)
12	[143]	Voice-Personae Project	2019						✓		Capsule	FaceForensics++	0.931 (ACC)
13	[157]	CVPR	2019						✓		DenseNet, GRU	FaceForensics++	0.969 (ACC)
14	[126]	IEEE Access	2019	✓							CNN	CASIA, GPIR, COVERAGE, BigGANs, LSUN Bedroom, PGGAN, SNGAN, StyleGAN	0.975 (ACC)
15	[151]	ICIP	2019					✓		✓	DenseNet	GAN-Generated Images Based On CelebA	0.986 (ACC)
16	[150]	Applied Sciences	2020					✓		✓	DenseNet	CelebA, ILSVRC12	0.988 (ACC)
17	[139]	WACV	2020							✓	ResNet	FaceForensics++	0.999 (ACC)
18	[125]	ArXiv	2020	✓							RNN	FaceForensics++, DDD	0.990 (ACC)
19	[124]	ACM-MM	2020			✓					ResNet, LSTM	FaceForensics++, DFDC Preview	0.997 (ACC)
20	[117]	BIOSIG	2020		✓						LightCNN, ResNet, DenseNet, SquezeNet	Celeb-DF	0.879 (AUC)
21	[158]	Master Thesis	2020			✓					Capsule, VGG, LSTM	DFDC	0.834 (ACC)
22	[127]	BIOSIG	2020	✓		✓					ResNet	FaceForensics++	0.993 (ACC)
23	[145]	CIKM	2020					✓			Autoencode, UNet	FaceForensics++	0.968 (ACC)
24	[154]	ArXiv	2020	✓		✓					CNN	CelebA, CelebA-HQ, GANs-generated dataset	0.985 (ACC)

authors synthesized the dataset using source videos collected from YouTube with the subjects of diverse ages, ethnic groups, and genders. This dataset aims to simulate real-life scenarios with various video qualities.

In late 2019, Facebook partnered with numerous technology giant companies, industry, and academic experts to organize a challenge entitled **Deepfake Detection Challenge (DFDC)**. They held a dataset collection campaign and hired numerous actors to record videos for the dataset. The host released the dataset in two phases, *i.* The preview dataset [8], and *ii.* The final version [167]. The preview dataset consists of 1,131 real videos and 4,113 fake videos. In contrast, the final version contains 19,154 real videos and 100,000 deepfake videos sourced from 3,426 paid actors. Both versions are rich in gender, skin tone, lighting conditions, head poses, age, and background.

Jiang et al. published **DeeperForensices-1.0** [168] which consisted of 50,000 real and 10,000 high-quality fake videos that originated from 100 paid actors. They employed several perturbations to guarantee the dataset diversity in simulating real-world scenarios, such as compression, blurry, and transmission errors.

He et al. [169] presented **ForgeryNet Dataset** that consisted of more than 36 mix-perturbations using 15 manipulation techniques on over 54 k subjects. It defines four significant tasks (image and video classification, spatial and temporal localization) with 9.4 M annotations. The authors selected CREMA-D [171], RAVDESS [172], VoxCeleb2 [173], and AVSpeech [174] as the source data to boost the diversity from various aspects, such as facial expression, face identity, subject angle, and case scenarios. They adopted face swapping, face reenactment, deepfake, and identity transfer as the main manipulation techniques. The whole dataset is categorized into two major groups: *i.* Image-forgery subset with 2.9 M still images, and *ii.* Video-forgery subset with more than 220 k video clips.

Zhou et al. introduced **Face Forensics in the Wild** to tackle multi-person face forgery detection [163]. The authors collected 4 k raw source videos with a minimum of 480p from YouTube and split each video into four uniform clips. Then, they randomly

Table 6
Summary of deep Learning-based deepfake detection part 2.

No.	Ref.	Pub.	Year	Techniques							Model	Dataset	Performances
				Color	Texture	Spatial	Att.	Loss	NN Rearchitecture	Multi Stream/ Network			
25	[42]	CVPR	2020			✓	✓				CNN	CelebA, Flicker-Faces-HQ (FFHQ), Face Forensics++, GANs-generated dataset	0.997 (ACC)
26	[134]	IJCAI	2020		✓		✓				CNN	CelebA, FFHQ, Face Forensics++, GANs-generated dataset, DFDC, CelebDF	0.986 (ACC)
27	[140]	TheWeb-Conf	2020							✓	ResNet	Face Forensics++, DeepfakeTIMIT, Mesonet data	0.994 (ACC)
28	[136]	ArXiv	2020		✓	✓					Inception	FaceForensics++, Celeb-DF-v2, DFDC	0.997 (AUC)
29	[160]	ECCV	2020			✓					CNN, LSTM	FaceForensics++, Celeb-DF, DFDC	0.943 (ACC)
30	[152]	ACM-MM	2020		✓			✓		✓	CNN	DeepFake-TIMIT, DFDC	0.966 (AUC)
31	[120]	CVPR	2020		✓						CNN	A2V, T2V-S, T2V-L, FakeFace in-the-wild	0.997 (ACC)
32	[159]	IH and MMSec	2020			✓					LSTM	FaceForensics++	0.943 (ACC)
33	[141]	Edge-Com	2020							✓	XceptionNet, InceptionV3, InceptionResNetV2, MobileNet, ResNet, Densenet	Self-generated dataset	0.997 (ACC)
34	[162]	CVPR	2020			✓					Autoencoder	FaceForensics++	0.982 (ACC)
35	[119]	IEEE TIFS	2021		✓						CNN	MOBIO, GRID	2.80 (FRR)
36	[128]	ICME	2021			✓					MTCNN, EfficientNet-B2	FaceForensics++	0.918 (ACC)
37	[137]	CVPR	2021	✓	✓	✓					Xception	Face Forensics ++, DFD, DFDC	0.995 (AUC)
38	[130]	CVPR	2021	✓				✓			Xception	FaceForensics++	0.967 (ACC)
39	[131]	CVPR	2021			✓					Xception	FaceForensics++, DeepfakeDetection (DFD), CelebDF, DeeperForensics-1.0	0.994 (AUC)
40	[161]	CVPR	2021		✓					✓	RNN	Face Forensics++, UADFV, CelebDF	0.999 (AUC)
41	[122]	CVPR	2021		✓						ResNet, MS-TCN	Face Forensics++, DFDC, CelebDF, DF1.0, FaceShifter	0.988 (ACC)
42	[149]	CVPR	2021		✓		✓	✓			EfficientNet-B4	Face Forensics++, DFDC, CelebDF, DF1.0	0.976 (ACC)
43	[133]	CVPR	2021			✓					Xception	Face Forensics++, DFDC, CelebDF	0.816 (ACC)
44	[163]	CVPR	2021			✓				✓	ResNet	Face Forensics++, DFDC Preview, CelebDF, FFW	0.993 (AUC)

selected a 12s sequence from each clip for forgery generation. They performed face-swapping by randomly choosing two videos as target and source from the 12 k filtered sequence collection using DeepFaceLab, FS-GAN, and a FaceSwap graphic method. The dataset has an average of three human faces in each frame and went through an automatic manipulation process with a domain-adversarial quality assessment network to save costs.

3.4. Evaluation Metric

The major evaluation metrics used in the discussed studies are Accuracy **(Acc.)**, Area Under the Curve **(AUC)**, Receiver Operating Characteristic **(ROC)**, **Recall**, **Precision**, **F1 Score**, Equal Error Rate **(ERR)**, and False Rejection Rate **(FRR)**. In most studies, the Acc., AUC, and ROC act as the primary evaluation metrics and are used as benchmark metrics to evaluate whether the model can precisely and accurately detect deepfake. The other evaluation metrics provided a better understanding to support the main detection result with a more precise analysis.

3.5. Discussion

Fig. 8 shows the implementation frequency of both deep learning and non-deep learning-based deepfake detection model according to the studies mentioned earlier. The analysis indicates that more researchers have started to employ deep learning-based approaches than traditional handcrafted features-based techniques. This trend is due to the recent deepfake quality enhancement, which leaves minimal traces and anomalies within the intrinsic feature of deepfake images or videos, which makes

Table 7
List of deepfake dataset.

Dataset	Ref.	Year	Manipulation	Ratio (Real: Fake)	Total Videos	Re-solution	Format	Source	Participant Consent
UADFV	[109]	2018	FakeApp	1.0: 1.0	49 real videos, 49 fake videos	294p x 500p	YouTube	YouTube	N
DeepfakeTIMIT	[11]	2018	FaceSwap-GAN	Only Fake	0 real videos, 620 fake videos	64p x 64p - 128p x 128p	JPG	Actor	Y
FFW	[164]	2018	Splicing, CGI, Deepfake	Only Fake	0 real videos, 150 fake videos	480p, 720p, 1080p	H.264, YouTube	YouTube	N
FaceForensics++	[147]	2019	FaceSwap, Deepfake	1.0: 4.0	1,000 real videos, 4,000 fake videos	480p, 720p, 1080p	H.264, CRF = 0, 23, 40	YouTube	N
DFD/DDD	[165]	2019	Deepfake	1.0: 8.5	363 real videos, 3,086 fake videos	1080p	H.264, CRF = 0, 23, 40	Actor	Y
Celeb-DF	[166]	2019	Deepfake	1.0: 11.3	590 real videos, 5,639 fake videos	Various	MPEG4	YouTube	N
DFDC-preview	[8]	2019	Deepfake	1.0: 3.6	1,131 real videos, 4,113 fake videos	180p - 2160p	H.264	Actor	Y
DFDC	[167]	2019	Deepfake	1.0: 5.2	19,154 real videos, 100,000 fake videos	240p - 2160p	H.264	Actor	Y
Deeper-Forensics-1.0	[168]	2020	Deepfake	5.0: 1.0	50,000 real videos, 10,000 fake videos	1080p	MP4	Actor	Y
ForgeryNet Dataset	[169]	2021	FaceSwap	1.0: 1.2	99630 real and 121617 fake videos, 1,438,201 real and 1,457,861 fake images	240p - 1080p	YouTube	Actor/ YouTube	Y/N
Face Forensics in the Wild	[163]	2021	FaceSwap	1.0: 1.0	10,000 real videos, 100,000 fake videos	>480p	YouTube	YouTube	N

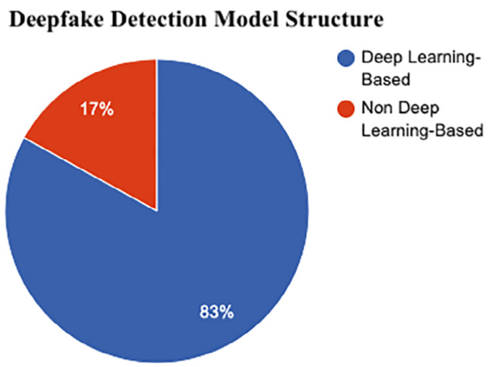


Fig. 8. The implementation frequency of deep learning & non deep learning-based deepfake detection model structure based on discussed studies.

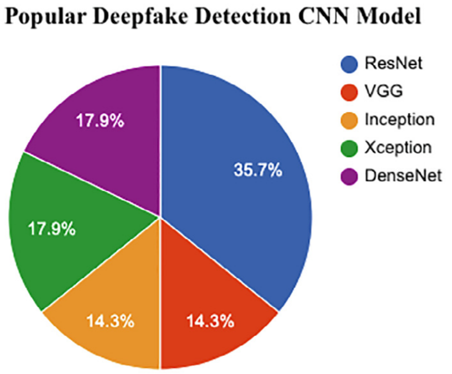


Fig. 9. The implementation frequency of the popular CNN architecture for deepfake detection model based on discussed studies.

it harder for handcrafted feature extraction. Furthermore, with the introduction of more efficient CNN architectures in the recent years, many researchers have switched their focus to learned-features extraction. Fig. 9 presented the frequency of the popular CNN used for the deepfake detector in the discussed studies.

On the other hand, instead of emphasizing a model-centric approach, recent research has shown the potential in doing data-centric methodology. In [175], the authors proposed a representative forgery mining (RFM) framework to refine the training data to improve the vanilla CNN-detector's performance. The proposed RFM framework can be used for any CNN-based detector and can provide a notable visualization result to explore the forgery region of various manipulation techniques.

4. Opportunities & Research Trends

This section discusses the potential opportunities and research direction in both deepfake generation and detection.

4.1. Deepfake Generation

The state-of-the-art deepfake creation methodologies highly rely on the GAN technology. The researchers improved the deepfake network training by integrating the tertiary concepts to achieve a more hyperrealistic and natural result with high confidence [9,176,13], such as style transfer, motion transfer, biometric artifacts, and semantic segmentation. However, the current deepfake is still imperfect and leaves room for improvement. The GAN's training is too time-consuming, resource-intensive, and easily overfitting. The output is not flawless enough to bypass the detection.

Few-shot Learning: The training dataset has been a vital issue in the deepfake area since most deepfake generation technologies require a vast amount of genuine data to support the training in generating more convincing fake content. This situation turned out to increase the overall computational resources. To address this issue, the research community emphasized training based on as little as possible amount of dataset, or more precisely, training on few-shot learning. Unlike the previous deepfake generation methodologies, a few-shot learning mechanism known as a domain field of meta-learning required only a relatively small dataset and low data labeling cost for training. One of the popular methods is *N-way K-shot* classification, which applied the ideas of transfer learning and knowledge sharing, showing the possibility of creating deepfake by matching the training likelihood distribution to the few-shot support set. In [177,57], the authors successfully implemented few-shot learning to reduce the required computational training resources. The research trend on reducing computational power and training datasets for deepfakes has consistently driven the deepfakes research development.

Deepfake Quality: Another potential trend in deepfake generation is the output quality. Due to the instability of GAN training, most deepfake outputs consist of subtle traces or fingerprints, such as unusual texture artifacts or pixel inconsistency, which make them vulnerable to the detector. Moreover, the current research community mainly focused on non-occlusive frontal face training data for deepfake generation. It might be unable to maintain its output quality when occlusions occur in the input data. Li et al. [92,178] have published a mask-guided detection approach for solving this issue. Future studies could work to ensure deepfake quality from artifact elimination, deepfake output resolution, and the ability to generate deepfake in defending attacks.

Real-time Deepfake: Putting deepfakes online to achieve a real-time face transformation effect has become a new opportunity for deepfakes. A researcher team published an open-source software [99] based on [179] to promote real-time deepfake in video conferencing. This software leverages image animation techniques based on keypoint learning and affine transformation, then directly forms the training result with the desired input image to achieve real-time reenactment. However, since it is an image animation technique, the results might consist of biometric artifacts and low fidelity for specific facial expressions and head movements. Although imperfect, it provided insights into future real-time deepfake development, which might negatively impact the medical, education, and entertainment industries.

4.2. Deepfake Detection

The competition between the deepfake generation and detection is like a cat-and-mouse game. The improvement of a generator will catalyze the enhancement of a detector. According to the study of conventional methods, designing a detector based on a specific generator's weakness(traces or anomalies) is not sustainable, reliable, and flexible. Since the generator is working towards producing artifact-less deepfake, the detector's research trend begins to shift focus on discrimination based on learned-feature rather than handcrafted-feature. However, a pre-trained CNN model cannot deal well with different deepfake scenarios and can be easily attacked by malicious users. Perhaps resolving these shortcomings will lead the performance of the deepfake detector to another level.

Adversarial Attack: The current deepfake detectors mainly emphasized on detection performance and neglected the importance of its robustness. A subtle perturbation of the input can significantly influence the performance of a trained neural network detection model to deviate from the expectation. The adversarial sample is the potent ingredient for deepfake data to evade detection. The two standard threat models of adversarial attack are the black-box, and white-box models, which are grouped according to their knowledge and access level to the target detector [180]. Wang et al. [59] proposed a stochastic-based defense mechanism that suggested switching the model's block layers to parallel channels and randomly allocating active channels in run time. As the development trend of deepfake detectors has shifted to neural network training, studying adversarial attack defense mechanisms in the deepfake field can be a new opportunity and has the potential to be explored.

One-button Solution(Generalization): Despite extensive efforts devoted to distinguishing deepfake, the detector generalization in detecting different deepfake types, data diversity, and data resolution remain a vital issue. The detector performance drastically drops when facing these scenarios, albeit it achieved an outstanding result on its original planned cases. For instance, a detector devised to tackle Face2Face might not perform well when detecting the entire face synthesis; a detector training on a particular dataset might not handle other unseen datasets, or a detector training with specific data resolution becomes vulnerable when tested with different input compression. Although few studies [127,145] have made their steps in examining generalizable detectors, none of them achieved a consistent detection accuracy with an accepTable 5% deviation when considering all these three factors in their works. Detector generalization undoubtedly is an important and rising trend of deepfake detection research.

Application & *Platform-friendly:* To protect people from being deceived by deepfake data, it is necessary to transform the deepfake detector into a reliable feature or API. Another research opportunity is developing a user-friendly tool which supports the integration with social media platforms or applications that allows people to make an accurate judgment to protect themselves from being overwhelmed by false information.

5. Conclusion

Is seeing still believing? Deepfake plays a crucial role in multimedia forgery; spotting deepfake is getting harder and more challenging. Traditional forgery countermeasure is no longer reliable for distinguishing deepfake from real videos or images. Furthermore, there is still no law enforcement to prevent this technology from being abused. Numerous giant technology companies and academics make their efforts to safeguard the media authenticity and information integrity by raising the awareness about deepfake [8]. This review presents the different deepfake types and summarized the detection methodologies in traditional and deep learning-based approaches. We presented the available resources, such as deepfake generation tools and datasets, and discussed the trends and opportunities to help accelerate research in both deepfake generation and detection.

CRediT authorship contribution statement

Jia-Wen Seow: Conceptualization, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Mei-Kuan Lim:** Conceptualization, Resources, Writing - review & editing, Supervision. **Raphaël C.-W. Phan:** Resources, Supervision. **Joseph K. Liu:** Supervision.

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

[1] M. Albahar, J. Almalki, Deepfakes: Threats and countermeasures systematic review, Journal of Theoretical and Applied Information Technology 97.

[2] C. Vaccari, A. Chadwick, Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news, Social Media + Society 6 (1) (2020) 2056305120903408. doi:10.1177/2056305120903408.

[3] J. Vincent, Watch jordan peele use ai to make barack obama deliver a psa about fake news, retrieved Oct 19, 2019 from https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed (2018).

[4] R.M. Chesney, D.K. Citron, Deep fakes: A looming challenge for privacy, democracy, and national security, California Law Review 107 (2018) 1753.

[5] D. Itzkoff, How 'rogue one' brought back familiar faces, retrieved Dec 18, 2020 from https://www.nytimes.com/2016/12/27/movies/how-rogue-one-brought-back-grand-moff-tarkin.html%20[https://perma.cc/F53C-TDYV (2016).

[6] Medium, Ai-powered digital people, retrieved Dec 18, 2020 from https://medium.com/syncedreview/ai-powered-digital-people-c0a94b7f0e8b (2020).

[7] N. Caporusso, Deepfakes for the good: A beneficial application of contentious artificial intelligence technology, in: T. Ahram (Ed.), Advances in Artificial Intelligence, Software and Systems Engineering, 2020, pp. 235–241.

[8] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, C. Canton Ferrer, The Deepfake Detection Challenge (DFDC) Preview Dataset, arXiv e-prints (2019) arXiv:1910.08854.

[9] Y. Mirsky, W. Lee, The creation and detection of deepfakes: A survey, ACM Computing Surveys (CSUR) 54 (1) (2021) 1–41.

[10] L. Verdoliva, Media forensics and deepfakes: an overview, IEEE Journal of Selected Topics in Signal Processing 14 (5) (2020) 910–932.

[11] P. Korshunov, S. Marcel, DeepFakes: a New Threat to Face Recognition? Assessment and Detection, arXiv e-prints (2018) arXiv:1812.08685.

[12] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, Deepfakes and beyond: A survey of face manipulation and fake detection, Information Fusion 64 (2020) 131–148.

[13] T. Zhang, L. Deng, L. Zhang, X. Dang, Deep learning in face synthesis: A survey on deepfakes, in: 2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET), 2020, pp. 67–70, https://doi.org/10.1109/CCET50901.2020.9213159.

[14] T.T. Nguyen, C.M. Nguyen, D. Tien Nguyen, D. Thanh Nguyen, S. Nahavandi, Deep Learning for Deepfakes Creation and Detection: A Survey, arXiv e-prints (2019) arXiv:1909.11573.

[15] N. Kanwal, A. Girdhar, L. Kaur, J.S. Bhullar, Detection of digital image forgery using fast fourier transform and local features, in: 2019 International Conference on Automation, Computational and Technology Management (ICACTM), 2019, pp. 262–267.

[16] A. van den Oord, Y. Li, O. Vinyals, Representation Learning with Contrastive Predictive Coding, arXiv e-prints (2018) arXiv:1807.03748.

[17] D.P. Kingma, M. Welling, An introduction to variational autoencoders, Foundations and Trendsö in Machine Learning 12 (4) (2019) 307–392, https://doi.org/10.1561/2200000056.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (11) (2020) 139–144.

[19] A. Van Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel recurrent neural networks (2016) 1747–1756.

[20] A. v. d. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, K. Kavukcuoglu, Conditional image generation with pixelcnn decoders, Curran Associates Inc., 2016, p. 4797–4805.

[21] Y. Chen, Y. Zhao, W. Jia, L. Cao, X. Liu, Adversarial-learning-based image-to-image transformation: A survey, Neurocomputing 411 (2020) 468–486, https://doi.org/10.1016/j.neucom.2020.06.067.

[22] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: International Conference on Learning Representations, 2018.

[23] M. Mehralian, B. Karasfi, Rdcgan: Unsupervised representation learning with regularized deep convolutional generative adversarial networks, in: 2018 9th Conference on Artificial Intelligence and Robotics and 2nd Asia-Pacific International Symposium, 2018, pp. 31–38.

[24] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, International conference on machine learning (2017) 214–223.

[25] J. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242–2251.

[26] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, IEEE (2018) 8789–8797.

[27] Y. Jo, J. Park, Sc-fegan: Face editing generative adversarial network with user's sketch and color, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1745–1753.

[28] Y. Nirkin, Y. Keller, T. Hassner, Fsgan: Subject agnostic face swapping and reenactment (2019) 7184–7193.

[29] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, L. Lin, Beautygan: Instance-level facial makeup transfer with deep generative adversarial network, in: Proceedings of the 26th ACM International Conference on Multimedia, Association for Computing Machinery, 2018, pp. 645–653, https://doi.org/10.1145/3240508.3240618.

[30] Changsha Shenguronghe Network Technology Co.,Ltd, Zao (2019). https://zaodownload.com/

[31] FaceApp Inc, Faceapp (2016). https://www.faceapp.com/.

[32] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE transactions on pattern analysis and machine intelligence 38 (2) (2015) 295–307.

[33] J. He, J. Zheng, Y. Shen, Y. Guo, H. Zhou, Facial image synthesis and super-resolution with stacked generative adversarial network, Neurocomputing 402 (2020) 359–365, https://doi.org/10.1016/j.neucom.2020.03.107.

[34] Y. Yu, Z. Gong, P. Zhong, J. Shan, Unsupervised representation learning with deep convolutional neural network for remote sensing images, in: International Conference on Image and Graphics, Springer, 2017, pp. 97–108.

[35] N. Corporation, Nvidia, retrieved Jan 2, 2021 from https://www.nvidia.com/en-us/ (2020).

[36] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4396–4405.

[37] G.-Y. Hao, H.-X. Yu, W.-S. Zheng, Mixgan: Learning concepts from different domains for mixture generation, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 2212–2219. doi:10.24963/ijcai.2018/306.

[38] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1510–1519. doi:10.1109/ICCV.2017.167.

[39] X. Huang, M.-Y. Liu, S. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, in: Computer Vision ECCV, Springer International Publishing, 2018, pp. 179–196.

[40] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan (2020) 8110–8119.

[41] J.C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, J. Fierrez, Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection, IEEE Journal of Selected Topics in Signal Processing 14 (5) (2020) 1038–1048, https://doi.org/10.1109/JSTSP.2020.3007250.

[42] H. Dang, F. Liu, J. Stehouwer, X. Liu, A.K. Jain, On the detection of digital face manipulation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5780–5789. doi:10.1109/CVPR42600.2020.00582.

[43] V. Blanz, K. Scherbaum, T. Vetter, H.-P. Seidel, Exchanging faces in images, in: Computer Graphics Forum, Vol. 23, Wiley Online Library, 2004, pp. 669–676.

[44] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, C. Theobalt, Real-time expression transfer for facial reenactment, ACM Trans. Graph. 34 (6) (2015), 183–1.

[45] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, M. Nieundefinedner, Demo of face2face: Real-time face capture and reenactment of rgb videos, in: ACM SIGGRAPH 2016 Emerging Technologies, Association for Computing Machinery, 2016, https://doi.org/10.1145/2929464.2929475.

[46] J. Thies, M. Zollhöfer, M. Nießner, Deferred neural rendering: Image synthesis using neural textures, ACM Trans. Graph. 38 (4). doi:10.1145/3306346.3323035.

[47] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D.B. Goldman, K. Genova, Z. Jin, C. Theobalt, M. Agrawala, Text-based editing of talking-head video, ACM Trans. Graph. 38 (4). doi:10.1145/3306346.3323028.

[48] S. Suwajanakorn, S.M. Seitz, I. Kemelmacher-Shlizerman, Synthesizing obama: learning lip sync from audio, ACM Trans. Graph. 36 (2017) 95:1–95:13.

[49] R. Kumar, J. Sotelo, K. Kumar, A. de Brebisson, Y. Bengio, ObamaNet: Photo-realistic lip-sync from text, arXiv e-prints (2017) arXiv:1801.01442.

[50] Y. Song, J. Zhu, D. Li, A. Wang, H. Qi, Talking face generation by conditional recurrent adversarial network, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 919–925. doi:10.24963/ijcai.2019/129.

[51] A. Bansal, S. Ma, D. Ramanan, Y. Sheikh, Recycle-gan: Unsupervised video retargeting, ECCV (2018) 119–135.

[52] W. Wu, Y. Zhang, C. Li, C. Qian, C.C. Loy, Reenactgan: Learning to reenact faces via boundary transfer, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[53] S. Tripathy, J. Kannala, E. Rahtu, Icface: Interpretable and controllable face reenactment using gans, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020.

[54] Y. Sun, J. Tang, Z. Sun, M. Tistarelli, Facial age and expression synthesis using ordinal ranking adversarial networks, IEEE Transactions on Information Forensics and Security 15 (2020) 2960–2972, https://doi.org/10.1109/TIFS.2020.2980792.

[55] Y. Shen, P. Luo, J. Yan, X. Wang, X. Tang, Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[56] L. Tran, X. Yin, X. Liu, Representation learning by rotating your faces, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (12) (2019) 3007–3021, https://doi.org/10.1109/TPAMI.2018.2868350.

[57] E. Zakharov, A. Shysheya, E. Burkov, V. Lempitsky, Few-shot adversarial learning of realistic neural talking head models, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9459–9468.

[58] J. Zhang, X. Zeng, Y. Pan, Y. Liu, Y. Ding, C. Fan, Faceswapnet: Landmark guided many-to-many face reenactment, arXiv preprint arXiv:1905.11805 2.

[59] Y. Wang, P. Bilinski, F. Bremond, A. Dantcheva, Imaginator: Conditional spatio-temporal gan for video generation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020.

[60] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, R. He, High-fidelity face manipulation with extreme poses and expressions, IEEE Transactions on Information Forensics and Security 16 (2021) 2218–2231, https://doi.org/10.1109/TIFS.2021.3050065.

[61] A. Siarohin, E. Sangineto, S. Lathuilière, N. Sebe, Deformable gans for pose-based human image generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[62] N. Neverova, R.A. Guler, I. Kokkinos, Dense pose transfer, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[63] G. Balakrishnan, A. Zhao, A.V. Dalca, F. Durand, J. Guttag, Synthesizing images of humans in unseen poses, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[64] S. Tulyakov, M.-Y. Liu, X. Yang, J. Kautz, Mocogan: Decomposing motion and content for video generation (2018) 1526–1535.

[65] K. Aberman, M. Shi, J. Liao, D. Lischinski, B. Chen, D. Cohen-Or, Deep video-based performance cloning, in: Computer Graphics Forum, Vol. 38, Wiley Online Library, 2019, pp. 219–233.

[66] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, C. Theobalt, Deep video portraits, ACM Trans. Graph. 37 (4). doi:10.1145/3197517.3201283.

[67] L. Liu, W. Xu, M. Zollhöfer, H. Kim, F. Bernard, M. Habermann, W. Wang, C. Theobalt, Neural rendering and reenactment of human actor videos, ACM Trans. Graph. 38 (5). doi:10.1145/3333002.

[68] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, B. Catanzaro, Video-to-video synthesis, Curran Associates Inc., 2018, p. 1152–1164.

[69] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, B. Catanzaro, Few-shot video-to-video synthesis, in: Advances in Neural Information Processing Systems (NeurIPS), 2019.

[70] C. Chan, S. Ginosar, T. Zhou, A.A. Efros, Everybody dance now, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5933–5942.

[71] Z. Liu, H. Hu, Z. Wang, K. Wang, J. Bai, S. Lian, Video synthesis of human upper body with realistic face, in: 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), IEEE, 2019, pp. 200–202.

[72] Y. Zhou, Z. Wang, C. Fang, T. Bui, T. Berg, Dance dance generation: Motion transfer for internet videos, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2019.

[73] Y. Chen, S. Xia, J. Zhao, M. Jian, Y. Zhou, Q. Niu, R. Yao, D. Zhu, Person image synthesis through siamese generative adversarial network, Neurocomputing 417 (2020) 490–500, https://doi.org/10.1016/j.neucom.2020.09.004.

[74] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks (2017) 1125–1134.

[75] G. Perarnau, J. van de Weijer, B. Raducanu, J.M. Álvarez, Invertible Conditional GANs for image editing, arXiv e-prints (2016) arXiv:1611.06355.

[76] M. Li, W. Zuo, D. Zhang, Deep identity-aware transfer of facial attributes, arXiv e-prints (2016) arXiv:1610.05586.

[77] W. Shen, R. Liu, Learning residual images for face attribute manipulation (2017) 4030–4038.

[78] T. Xiao, J. Hong, J. Ma, Elegant: Exchanging latent encodings with gan for transferring multiple face attributes (2018) 168–184.

[79] Z. He, W. Zuo, M. Kan, S. Shan, X. Chen, Attgan: Facial attribute editing by only changing what you want, IEEE Transactions on Image Processing 28 (11) (2019) 5464–5478, https://doi.org/10.1109/TIP.2019.2916751.

[80] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, S. Wen, Stgan: A unified selective transfer network for arbitrary image attribute editing, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3668–3677. doi:10.1109/CVPR.2019.00379.

[81] X. Nie, H. Ding, M. Qi, Y. Wang, E.K. Wong, Urca-gan: Upsample residual channel-wise attention generative adversarial network for image-to-image translation, Neurocomputing 443 (2021) 75–84, https://doi.org/10.1016/j.neucom.2021.02.054.

[82] J. Guo, Y. Liu, Attributes guided facial image completion, Neurocomputing 392 (2020) 60–69, https://doi.org/10.1016/j.neucom.2020.02.013.

[83] D. Ma, B. Liu, Z. Kang, J. Zhou, J. Zhu, Z. Xu, Two birds with one stone: Transforming and generating facial images with iterative gan, Neurocomputing 396 (2020) 278–290.

[84] J. Zeng, X. Ma, K. Zhou, Photo-realistic face age progression/regression using a single generative adversarial network, Neurocomputing 366 (2019) 295–304, https://doi.org/10.1016/j.neucom.2019.07.085.

[85] Y. Li, S. Liu, J. Yang, M. Yang, Generative face completion, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5892–5900. doi:10.1109/CVPR.2017.624.

[86] S. Xie, Z. Tu, Holistically-nested edge detection, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1395–1403. doi:10.1109/ICCV.2015.164.

[87] M. Afifi, M.A. Brubaker, M.S. Brown, Histogan: Controlling colors of gan-generated and real images via color histograms, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7941–7950.

[88] I. Korshunova, W. Shi, J. Dambre, L. Theis, Fast face-swap using convolutional neural networks, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[89] R. Natsume, T. Yatagawa, S. Morishima, Rsgan: Face swapping and editing using face and hair representation in latent spaces, in: ACM SIGGRAPH 2018 Posters, Association for Computing Machinery, 2018, https://doi.org/10.1145/3230744.3230818.

[90] R. Natsume, T. Yatagawa, S. Morishima, Fsnet: An identity-aware generative model for image-based face swapping, in: Asian Conference on Computer Vision, Springer, 2018, pp. 117–132.

[91] Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt, B. Schiele, A hybrid model for identity obfuscation by face replacement, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[92] L. Li, J. Bao, H. Yang, D. Chen, F. Wen, Advancing high fidelity identity swapping for forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[93] C. Dfaker DepFA, DFaker, https://github.com/dfaker/df (2018).

[94] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. Shift Facenheim, L. RP, J. Jiang, S. Zhang, P. Wu, B. Zhou, W. Zhang, DeepFaceLab: A simple, flexible and extensible face swapping framework, https://github.com/iperov/DeepFaceLab (2020).

[95] K. Torzdf, Andenixa, Face swap, https://github.com/deepfakes/faceswap (2020).

[96] F. Web, Deepfakes web, https://faceswapweb.com/?locale=en (2020).

[97] Mahinetube, Mahinetube, https://www.machine.tube/ (2020).

[98] NEOCORTEXT, INC., Reface app (2020). https://get.reface.app/.

[99] I. Avatarify, Avatarify: Ai face animator (2020). https://apps.apple.com/us/app/avatarify-ai-face-animator/id1512669147.

[100] W. Liu, W.L.L.M. Zhixin Piao, S. Min Jie, Gao: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis, in: The IEEE International Conference on Computer Vision (ICCV), 2019.

[101] Y. Didi, Jiggy: Magic dance gif maker (2020). https://apps.apple.com/us/app/jiggy-magic-dance-gif-maker/id1482608709

[102] Y. Zhang, L. Zheng, V.L.L. Thing, Automated face swapping and its detection, in: 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP), 2017, pp. 15–19.

[103] A. Agarwal, R. Singh, M. Vatsa, A. Noore, Swapped! digital face presentation attack detection via weighted local magnitude pattern, in: 2017 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2017, pp. 659–665.

[104] M. Koopman, A. Macarulla Rodriguez, Z. Geradts, Detection of deepfake video manipulation, 2018, pp. 133–136.

[105] R. Durall, M. Keuper, F.-J. Pfreundt, J. Keuper, Unmasking deepfakes with simple features, arXiv-eprints (2019) arXiv:1911.00686.

[106] F. Marra, D. Gragnaniello, L. Verdoliva, G. Poggi, Do gans leave artificial fingerprints? (2019) 506–511.

[107] H. Li, B. Li, S. Tan, J. Huang, Identification of deep network generated images using disparities in color components, Signal Process. 174 (2020) 107616.

[108] S. McCloskey, M. Albright, Detecting GAN-generated Imagery using Color Cues, arXiv e-prints (2018) arXiv:1812.08247.

[109] X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses, in: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 8261–8265.

[110] D.E. King, Dlib-ml: A machine learning toolkit, Journal of Machine Learning Research 10 (2009) 1755–1758.

[111] T. Baltrusaitis, A. Zadeh, Y.C. Lim, L.-P. Morency, Openface 2.0: Facial behavior analysis toolkit, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 59–66.

[112] F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes and face manipulations, in: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019, pp. 83–92.

[113] C.-C. Hsu, C.-Y. Lee, Y.-X. Zhuang, Learning to detect fake face images in the wild (2018) 388–391.

[114] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, https://doi.org/10.1038/nature14539.

[115] W. Quan, K. Wang, D. Yan, X. Zhang, Distinguishing between natural and computer-generated images using convolutional neural networks, IEEE Transactions on Information Forensics and Security 13 (11) (2018) 2772–2787, https://doi.org/10.1109/TIFS.2018.2834147.

[116] Y. Li, M. Chang, S. Lyu, In ictu oculi: Exposing ai created fake videos by detecting eye blinking, in: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7.

[117] H.M. Nguyen, R. Derakhshani, Eyebrow recognition for identifying deepfake videos, in: 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), 2020, pp. 1–5.

[118] U.A. Ciftci, I. Demir, L. Yin, Fakecatcher: Detection of synthetic portrait videos using biological signals, IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (1) (2020), https://doi.org/10.1109/TPAMI.2020.3009287, 1–1.

[119] C.Z. Yang, J. Ma, S. Wang, A.W.C. Liew, Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis, IEEE Transactions on Information Forensics and Security 16 (2021) 1841–1854, https://doi.org/10.1109/TIFS.2020.3045937.

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

[121] S. Rubin, F. Berthouzoz, G.J. Mysore, W. Li, M. Agrawala, Content-based tools for editing audio stories, in: Proceedings of the 26th annual ACM symposium on User interface software and technology, 2013, pp. 113–122.

[122] A. Haliassos, K. Vougioukas, S. Petridis, M. Pantic, Lips don't lie: A generalisable and robust approach to face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5039–5049.

[123] I. Amerini, L. Galteri, R. Caldelli, A. Del Bimbo, Deepfake video detection through optical flow based cnn, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 1205–1207.

[124] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, J. Zhao, Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4318–4327.

[125] S. Tariq, S. Lee, S.S. Woo, A Convolutional LSTM based Residual Network for Deepfake Video Detection, arXiv e-prints (2020) arXiv:2009.07480.

[126] K. Zhang, Y. Liang, J. Zhang, Z. Wang, X. Li, No one can escape: A general approach to detect tampered and generated image, IEEE Access 7 (2019) 129494–129503.
[127] A. Khodabakhsh, C. Busch, A generalizable deepfake detector based on neural conditional distribution modelling, International Conference of the Biometrics Special Interest Group (BIOSIG) 2020 (2020) 1–5.
[128] J. Zhang, J. Ni, H. Xie, Deepfake videos detection using self-supervised decoupling network, IEEE International Conference on Multimedia and Expo (ICME) 2021 (2021) 1–6, https://doi.org/10.1109/ICME51207.2021.9428368.
[129] H.-S. Chen, M. Rouhsedaghat, H. Ghani, S. Hu, S. You, C.-C. Jay Kuo, in: Defakehop: A light-weight high-performance deepfake detector, in 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021, pp. 1–6, https://doi.org/10.1109/ICME51207.2021.9428361.
[130] J. Li, H. Xie, J. Li, Z. Wang, Y. Zhang, Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6458–6467.
[131] Y. Luo, Y. Zhang, J. Yan, W. Liu, Generalizing face forgery detection with high-frequency features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16317–16326.
[132] J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images, IEEE Transactions on Information Forensics and Security 7 (3) (2012) 868–882, https://doi.org/10.1109/TIFS.2012.2190402.
[133] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, N. Yu, Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 772–781.
[134] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, Y. Liu, Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces, in: International Joint Conference on Artificial Intelligence (IJCAI), Vol. 2, 2020.
[135] Y. Li, S. Lyu, Exposing deepfake videos by detecting face warping artifacts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.
[136] Y. Nirkin, L. Wolf, Y. Keller, T. Hassner, Deepfake detection based on discrepancies between faces and their context, IEEE Transactions on Pattern Analysis & Machine Intelligence (2021), 1–1.
[137] X. Zhu, H. Wang, H. Fei, Z. Lei, S.Z. Li, Face forgery detection by 3d decomposition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2929–2939.
[138] P. Zhou, X. Han, V.I. Morariu, L.S. Davis, Two-stream neural networks for tampered face detection (2017) 1831–1839.
[139] P. Kumar, M. Vatsa, R. Singh, Detecting face2face facial reenactment in videos, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020.
[140] X. Li, K. Yu, S. Ji, Y. Wang, C. Wu, H. Xue, Fighting against deepfake: Patch&pair convolutional neural networks (ppcnn), in: Companion Proceedings of the Web Conference 2020, 2020, pp. 88–89.
[141] M.S. Rana, A.H. Sung, Deepfakestack: A deep ensemble-based learning technique for deepfake detection, in: 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), 2020, pp. 70–75.
[142] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, in: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7.
[143] H.H. Nguyen, J. Yamagishi, I. Echizen, Use of a Capsule Network to Detect Fake Images and Videos, arXiv e-prints (2019) arXiv:1910.12467.
[144] H.H. Nguyen, J. Yamagishi, I. Echizen, Capsule-forensics: Using capsule networks to detect forged images and videos, in: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) , 2019, pp. 2307–2311.
[145] M. Du, S. Pentyala, Y. Li, X. Hu, Towards generalizable deepfake detection with locality-aware autoencoder, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Association for Computing Machinery, 2020, pp. 325–334, https://doi.org/10.1145/3340531.3411892.
[146] S. Tariq, S. Lee, H. Kim, Y. Shin, S.S. Woo, Detecting both machine and human created fake face images in the wild, in: Proceedings of the 2nd international workshop on multimedia privacy and security, 2018, pp. 81–87.
[147] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images (2019) 1–11.
[148] T. Fernando, C. Fookes, S. Denman, S. Sridharan, Detection of fake and fraudulent faces via neural memory networks, IEEE Transactions on Information Forensics and Security 16 (2021) 1973–1988, https://doi.org/10.1109/TIFS.2020.3047768.
[149] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, N. Yu, Multi-attentional deepfake detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2185–2194.
[150] C.-C. Hsu, Y.-X. Zhuang, C.-Y. Lee, Deep fake image detection based on pairwise learning, Applied Sciences 10 (2020) 370, https://doi.org/10.3390/app10010370.
[151] Y. Zhuang, C. Hsu, Detecting generated image based on a coupled network with two-step pairwise learning, in: 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 3212–3216. doi:10.1109/ICIP.2019.8803464.
[152] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, Emotions don't lie: An audio-visual deepfake detection method using affective cues, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2823–2832.
[153] H. Mo, B. Chen, W. Luo, Fake faces identification via convolutional neural network, in: Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security, 2018, pp. 43–47.
[154] Z. Guo, G. Yang, J. Chen, X. Sun, Fake face detection via adaptive residuals extraction network, arXiv e-prints (2020) arXiv:2005.04945.
[155] N.-T. Do, I.-S. Na, S.-H. Kim, Forensics face detection from gans using convolutional neural network (2018).
[156] S. Sabour, N. Frosst, G.E. Hinton, Dynamic routing between capsules, in: Advances in neural information processing systems, 2017, pp. 3856–3866.
[157] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.
[158] A. Mehra, Deepfake detection using capsule networks with long short-term memory networks, retrieved Dec 18, 2020 from http://essay.utwente.nl/83028/ (August 2020).
[159] I. Amerini, R. Caldelli, 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, 2020, pp. 97–102.
[160] I. Masi, A. Killekar, R.M. Mascarenhas, S.P. Gurudatt, W. AbdAlmageed, Two-branch recurrent network for isolating deepfakes in videos, in: European Conference on Computer Vision, Springer, 2020, pp. 667–684.
[161] Z. Sun, Y. Han, Z. Hua, N. Ruan, W. Jia, Improving the efficiency and robustness of deepfakes detection through precise geometric features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3609–3618.
[162] H. Khalid, S.S. Woo, Oc-fakedect: Classifying deepfakes using one-class variational autoencoder, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 656–657.
[163] T. Zhou, W. Wang, Z. Liang, J. Shen, Face forensics in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5778–5788.
[164] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, C. Busch, Fake face detection methods: Can they be generalized?, in: 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), 2018, pp. 1–6.
[165] G.R. Nick Dufour, J. Andrew Gully, Contributing data to deepfake detection research, 2019.
[166] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision (ICCV), 2015.
[167] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. Canton Ferrer, The DeepFake Detection Challenge Dataset, arXiv e-prints (2020) arXiv:2006.07397.
[168] L. Jiang, R. Li, W. Wu, C. Qian, C.C. Loy, Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection (2020) 2889–2898.
[169] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, Z. Liu, Forgerynet: A versatile benchmark for comprehensive forgery analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4360–4369.
[170] C. Sanderson, The VidTIMIT Database, Idiap-Com Idiap-Com-06-2002, IDIAP (2002).
[171] H. Cao, D.G. Cooper, M.K. Keutmann, R.C. Gur, A. Nenkova, R. Verma, Crema-d: Crowd-sourced emotional multimodal actors dataset, IEEE transactions on affective computing 5 (4) (2014) 377–390.
[172] S.R. Livingstone, F.A. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, PloS one 13 (5) (2018) e0196391.
[173] J.S. Chung, A. Nagrani, A. Zisserman, Voxceleb2: Deep speaker recognition, arXiv e-prints (2018) arXiv:1806.05622.
[174] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W.T. Freeman, M. Rubinstein, Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation, arXiv-eprints (2018) arXiv:1804.03619.
[175] C. Wang, W. Deng, Representative forgery mining for fake face detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14923–14932.
[176] X. Tong, L. Wang, X. Pan, J. gya Wang, An overview of deepfake: The sword of damocles in ai, in: 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), IEEE, 2020, pp. 265–273.
[177] W. Liang, Z. Liu, C. Liu, DAWSON: A Domain Adaptive Few Shot Generation Framework, arXiv e-prints (2020) arXiv:2001.00576.
[178] Z. Chen, L. Xie, S. Pang, Y. He, B. Zhang, Magdr: Mask-guided detection and reconstruction for defending deepfakes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9014–9023.
[179] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, N. Sebe, First order motion model for image animation, Advances in Neural Information Processing Systems (2019) 7137–7147.
[180] K. Ren, T. Zheng, Z. Qin, X. Liu, Adversarial attacks and defenses in deep learning, Engineering 6 (3) (2020) 346–360, https://doi.org/10.1016/j.eng.2019.12.012.

Jia-Wen Seow received her dual awards in B.Eng (Hons) in Software Engineering from Taylor's University and University of West England in 2018. She joined Monash University Malaysia in 2019. Her research focuses on the design and development of a reliable detection model that is robust against most of the deepfake generation methods and adversarial attack.

Mei-Kuan Lim is a lecturer attached to the School of Information Technology at Monash University Malaysia. Her research interests include swarm intelligence, data and video analytics, computer vision and machine learning.

Raphaël C.-W. Phan is Professor at Monash University, specializing in security, cryptography and malicious AI. He has published over 90 journal papers and in excess of 120 conference papers. He was the principal investigator for a project on privacy-preserving data mining funded by the UK government & UK Ministry of Defence, and recently led projects with funding in excess of RM3 million from the Malaysian government & industry.

Joseph Liu has been developing cryptographic algorithms for secure peer-to-peer transactions since 2004. He played a role in creating the Linkable Ring Signature, which Monero, a cryptocurrency similar to Bitcoin, used to add a layer of privacy to the transaction process.