IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 August 2023, accepted 8 October 2023, date of publication 13 October 2023, date of current version 27 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3324403

SURVEY

DeepFake on Face and Expression Swap: A Review

SAIMA WASEEM[1], SYED ABDUL RAHMAN SYED ABU BAKAR[1], (Senior Member, IEEE), BILAL ASHFAQ AHMED[2,3], ZAID OMAR[1], (Senior Member, IEEE), TAISEER ABDALLA ELFADIL EISA[4], AND MHASSEN ELNOUR ELNEEL DALAM[5]

[1]Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Skudai 81310, Malaysia
[2]Department of Electrical Engineering, The University of Lahore, Lahore 55150, Pakistan
[3]Faculty of Computing, Universiti Teknologi Malaysia, Skudai 81310, Malaysia
[4]Department of Information System-Girls Section, King Khalid University, Muhayil 62529, Saudi Arabia
[5]Department of Mathematics-Girls Section, King Khalid University, Muhayil 62529, Saudi Arabia

Corresponding authors: Syed Abdul Rahman Syed Abu Bakar (syed@fke.utm.my) and Bilal Ashfaq Ahmed (bilal.ashfaq@ee.uol.edu.pk)

This work was supported by the Deanship of Scientific Research, King Khalid University, through the Large Group Research Project RGP.2/52/44.

ABSTRACT Remarkable advances have been made in deep learning, leading to the emergence of highly realistic AI-generated videos known as deepfakes. Deepfakes use generative models to manipulate facial features to create modified identities or expressions with impressive realism. These synthetic media creations can deceive, discredit, or blackmail individuals and threaten the integrity of the legal, political, and social systems. Consequently, researchers are actively developing techniques to detect deepfake content to preserve privacy and combat the dissemination of fabricated media. This article presents a comprehensive study examining existing methods of creating deepfake images and videos for face and expression replacement. In addition, it provides an overview of publicly available deepfake datasets for benchmarking, serving as important resources for training and evaluating deepfake detection systems. In addition, the study sheds light on the detection approaches used to identify deepfake face and expression swaps while discussing the challenges and issues involved. However, the focus of this study goes beyond identifying the existing barriers. It goes a step further by outlining future research directions and guiding future scientists to address the concerns that need to be addressed in deepfake detection methods. In this way, this paper aims to facilitate the development of robust and effective deepfake detection solutions for face and expression swaps, thereby contributing to ongoing efforts to protect the authenticity and trustworthiness of visual media.

INDEX TERMS Deepfake, deep learning, face manipulation, face swap, re-enactment, media forensic, generative adversarial networks.

I. INTRODUCTION

Manipulation of picture and video content isn’t new. For this purpose, many special software tools, such as Adobe Photoshop and Adobe Lightroom, have been available for decades [1]. However, the realistic modification of facial features in digital images and videos using these tools has traditionally faced limitations due to factors such as the requirement for domain expertise, complexity, and the time-consuming nature of the process. With the advent of deepfake technology, the landscape has changed dramatically over the past five years, reducing the amount of effort involved in facial manipulation [2]. The term “deepfake,” a blend of “deep learning” and “fake,” specifically refers to manipulated media content created using artificial neural networks. Deepfake techniques rely on advanced deep learning models like autoencoders and generative adversarial networks (GANs) [3], [4] to analyze a person’s facial features and behaviors, enabling the synthesis of manipulated facial images that replicate similar gestures and movements [5].

Deepfake technology raises significant global security concerns, enabling unauthorized manipulation of individual’s

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

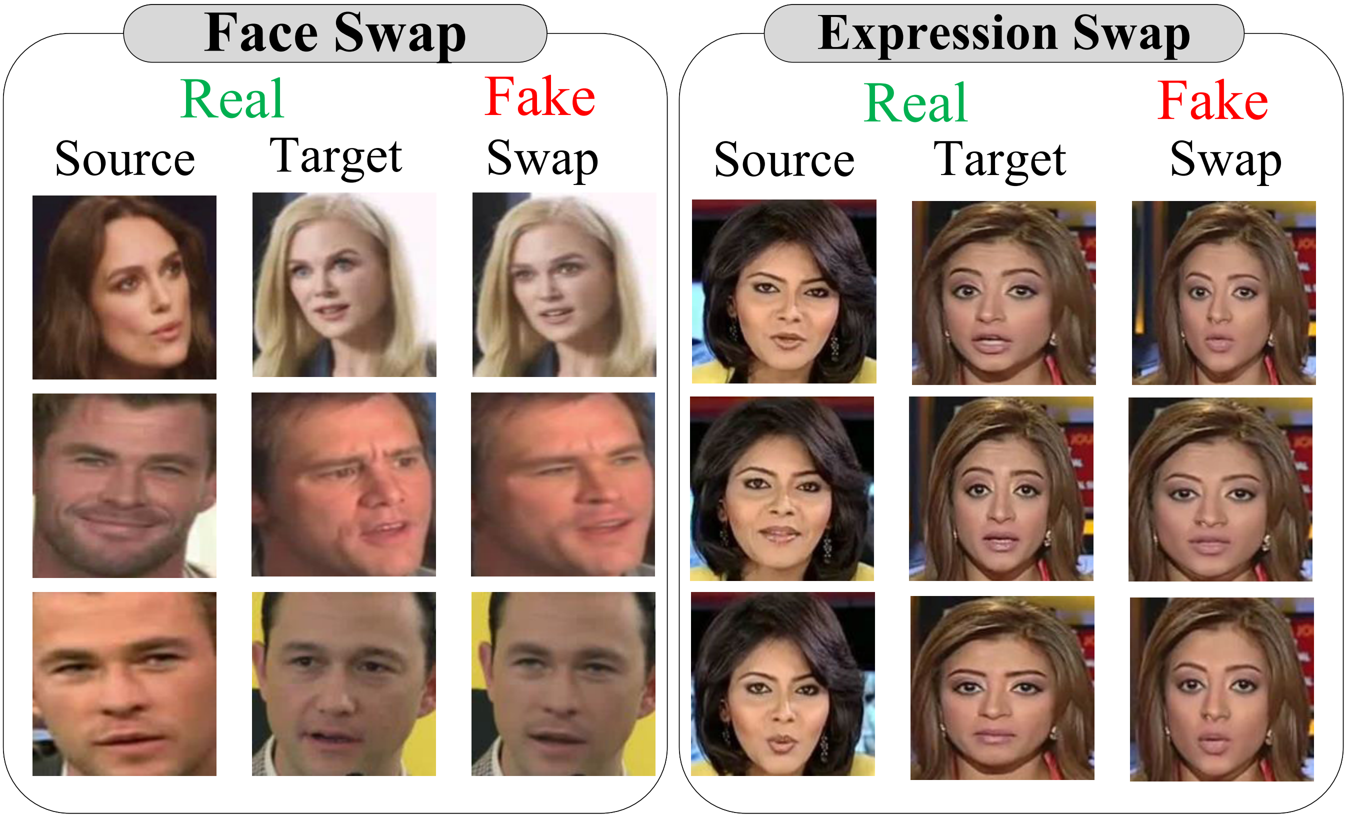

FIGURE 1. Examples of deepfake face manipulation. Face swap images obtained from the Celeb-DF dataset [6]. and expression swap images are derived from FaceForensics++ [7].

faces and expressions in political videos [8]. This technology has the potential for harmful exploitation, such as escalating tensions, spreading false information to influence elections, and misusing it on social media platforms [9]. While deepfake technology has demonstrated innovative applications, such as voice dubbing without reshooting film scenes [10], digital try-ons during shopping [11], and improved traditional teaching methods to engage students [12], the harmful uses of deepfakes outweigh these positive aspects. The accessibility and scalability of deepfake technology allow nearly anyone with device access to create compelling fake videos closely resembling authentic media [12], [13], [14]. The availability of user-friendly tools like DeepFaceLab [15], [16], smartphone apps such as Zao [17], and FakeApp [18] has simplified deepfake usage for non-professionals to swap their faces with any target person seamlessly. Deepfake face manipulations can be divided into four main groups: (1) face generation, which involves creating entirely new facial images, (2) facial attributes changes such as hair color, age, gender, glasses, etc., (3) face swap, which involves replacing the face of the original person with the face of another person, and (4) expression swap also known as re-enactment, in which the facial expression of the original person is transferred to the facial expression of the target person.

Deepfake can have different levels of risk. Out of four deepfake face manipulations, face swapping and re-enactment pose a significant threat to society [19], [20]. Figure 1 graphically summarizes these two facial changes. Face swapping is especially concerning, serving as a versatile tool for identity theft, crafting fabricated images or videos of specific individuals [21], [22], [23]. Face-swapped videos can be used to create convincing illusions of someone's presence, which can compromise biometric systems and grant unauthorized access to sensitive data [24]. In one notable case, the American AI company Kneron successfully fooled Alipay and WeChat payment processes, as well as self-service terminals in airports and train stations [25]. This shows the serious impact that deepfake face-swap technology can have on identity authentication systems that rely on biometric data. In contrast, facial re-enactment enables complete impersonation using a single image to persuade others without verbal communication. The real-time threat of facial re-enactment is evident from reported incidents in China, where stolen facial images were used to create deepfake videos [26]. They employed a smartphone with a compromised camera to trick the tax invoice system with pre-generated deepfake identities. Another concerning aspect is its potential misuse in child predator scenarios [25], where the predator hides behind a virtual avatar, needing

TABLE 1. Comparative analysis of deepfake review/survey papers.

Reference	Deepfake	Detection	Generation	Dataset	Scope
[27]	Face Manipulation	Briefly Discussed	Briefly Discussed	Limited coverage	Present deepfake face manipulation and detection techniques, no emphasis on limitations.
[28]	Face and Expression swap	Briefly Discussed	Briefly Discussed	Limited Coverage	Overview of creation/detection tools, minimal focus on limitations.
[29]	Images and Videos	Briefly Discussed	Limited Coverage	Limited Coverage	Brief overview of detection techniques, some generation aspects, not emphasizing limitations.
[30]	Images and Videos	Thoroughly Discussed	Not Discussed	Not Discussed	Focus on deepfake detection approaches and model efficacy.
[31]	Audio and Video	Briefly Discussed	Thoroughly Discussed	Limited Coverage	Provides insights on improvements, trends, limitations, and challenges in audio-visual deepfakes.
Current	Face and Expression swap	Thoroughly Discussed	Thoroughly Discussed	Thoroughly Discussed	Insights on improvements, trends, limitations, and challenges in deepfake face and expression swaps.

no resemblance to any existing individual. Reenactment is a suitable form of deepfakes to spoof face biometrics systems implementing real-time challenge-response liveness detection mechanisms.

The deepfake generation and detection field is fundamentally competitive, as both defenders (detectors) and adversaries (generators) strive to outperform each other. Significant progress has been made on both fronts in recent years. Various competitions have been launched to facilitate the development of effective deepfake detection solutions. These include the Media Forensics Challenge (MFC2018) [32] sponsored by the National Institute for Standards and Technology (NIST) [33], the Deepfake Detection Challenge (DFDC) [34], [35] initiated by Facebook in collaboration with Microsoft and academic partners from different universities, and the Deeper Forensics Challenge 2020 [36] hosted on the CodaLab platform [37] in conjunction with ECCV 2020 (The 2nd Workshop on Sensing, Understanding, and Synthesizing Humans) [38]. The research community is actively striving to enhance the detection of deepfake face and expression swap [21], [39], [40], [41], [42], [43]. While convolutional neural networks (CNNs) are commonly utilized to detect deepfake videos, their effectiveness is not absolute [44]. There are challenges and drawbacks that need to be considered. Developers of deepfake videos are becoming adept at evading detection through the use of advanced techniques such as adversarial generation algorithms, which mask distinct markers or conceal unique identifiers of deepfakes [45]. These algorithms manipulate input data to create videos that appear more authentic to detection mechanisms, making accurate detection more difficult [46]. Variations in video quality and content further complicate detection [47], as distinguishing deepfake characteristics are less prominent in certain cases.

To effectively detect deepfake face and expression swaps, it is crucial to understand the advancements and limitations of deepfake generation and detection techniques, as well as the existing detection methods and challenges to overcome for implementing effective forensic systems. Various initiatives have examined deepfake face manipulation, and some of the relevant survey and review papers are presented in Table 1. The majority of these surveys are centred on identifying deepfake content in both images and videos [29], [30]. Tolosana's survey [27] explored diverse deepfake techniques for face manipulations but lacked thorough explanations of generation methods and explicit coverage of limitations. Masood et al. [31] specifically addressed creating and detecting deepfakes in audio and video formats. Additionally, Zhang's concise survey [28] briefly covered face and expression swap detection, with limited emphasis on generation techniques. Currently, there's no comprehensive survey dedicated to exploring deepfake face and expression swap as a comprehensive topic. Such a study would provide clear insights into generating and detecting deepfake face and expression swaps, bridging gaps in previous surveys by incorporating dataset details. Additionally, such an overview would aid in identifying critical challenges and opportunities for implementing real-time forensic systems specifically designed to uncover face and expression swap deepfakes. The primary contributions of our work can be summarized as follows:

- We comprehensively examine deepfake face and expression swap generation methods, exploring their recent advancements, patterns, and the challenges they pose.
- We thoroughly analyze existing deepfake face and expression swap datasets and detection approaches, with a particular focus on their generalization,

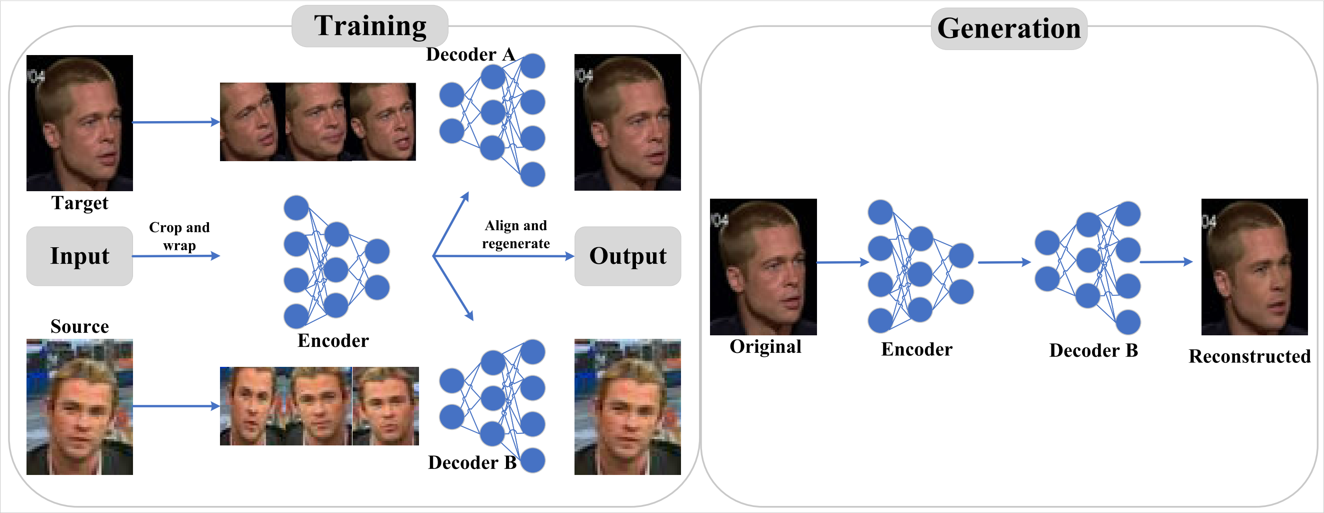

FIGURE 2. Illustration of deepfake face swap generation process.

interpretability, and robustness capabilities. These factors are critical for implementing these methods in real-time forensic systems.
- We discuss the challenges associated with deepfake detection and outline future research directions in this domain.

The paper is structured as follows: Section II discusses various algorithms proposed in recent years for generating deepfake face and expression swaps. Then, Section III presents a list of datasets used for deepfake face and expression swap detection. Next, Section IV describes different approaches to detect deepfake face and expression swaps. Afterwards, Section VI discusses the limitations of existing detection methods and future directions. Finally, Section VII concludes the paper.

II. DEEPFAKE VIDEO GENERATION

This section will discuss various methods proposed for face swapping in images and videos, using deep learning models such as GANs [48], [49] and autoencoders [50]. Table 2 and Table 3 provide a comprehensive overview of the various approaches employed for creating deepfake face and expression swaps.

A. FACE SWAP

The process involves three main steps to perform a face swap. First, the algorithm detects faces in both the source and target video. Next, the approach replaces the target face's nose, mouth, and eyes with the corresponding features of the source face. The color and lighting of the candidate's facial image are adjusted to ensure seamless integration of the two faces. Afterwards, the overlapping region undergoes match distance computation to evaluate and rank the quality of the blended candidate replacement.

In a study by Korshunova [51], a fast face swap technique was introduced using convolutional neural networks (CNNs). Training the network on a series of images allowed it to learn the target identity's appearance and generate face swap images. However, this method is limited to transforming individual images and is unsuitable for producing high-quality videos with time continuity. In 2017, a Reddit user created a deepfake video using an autoencoder [52]. The deepfake face-swap auto-encoder network utilizes a shared encoder and two decoders, while the encoder and the two decoders share parameters during the training process. A shared encoder learns to encode shared non-identities (latent vectors) underlying both source and target individuals. Two decoders reconstruct the source and target faces from their respective latent vector representations. Face swapping is accomplished by decoding the latent vector of the source face through the target face decoder. Figure 2 shows a deepfake creation process through an autoencoder. Various face-swapping applications, such as DeepFaceLab [15], [16], DFaker [53], and Deepfake tf [54] (a Deepfake framework based on TensorFlow), make use of the autoencoder technique. DeepFaceLab [15], [16] is specially designed for non-experts and includes features like residual blocks, transfer style loss and masked loss to improve face and eye consistency in the generated deepfakes. Nirkin et al. [55] presented a face swapping method that utilized face segmentation and replacement with a full convolutional network (FCN). Their technique involves identifying 2D facial landmarks in each video frame, which are used to compute the 3D posture and adjust the 3D shape for facial expressions. Using a trained FCN to predict face visibility at each pixel, the model performs face segmentation and isolates the face from the background. The source face is transformed and seamlessly integrated into the target frame using the aligned 3D face shapes as proxies. It's important

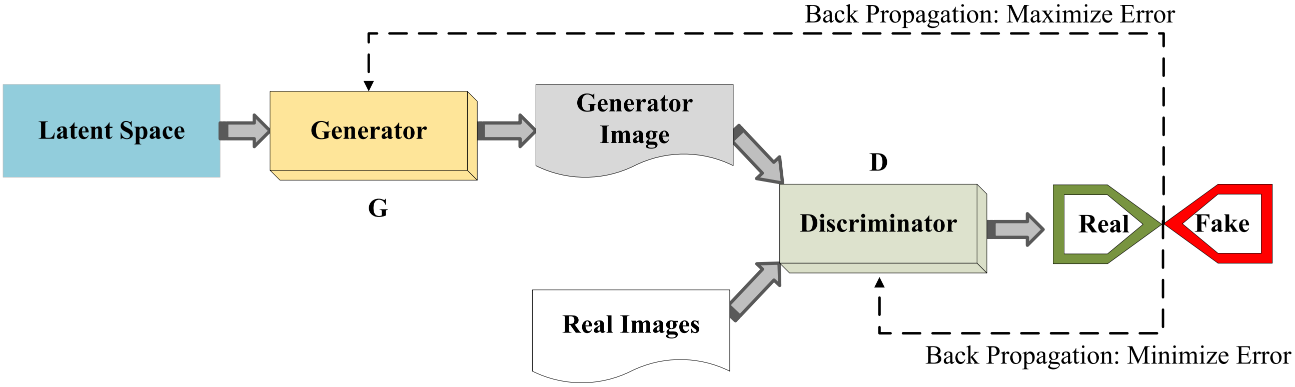

FIGURE 3. GAN network architecture.

to note that their approach was not trained end-to-end and required specific handling for occlusions in images.

GANs (Generative Adversarial Networks) represent an optimized approach for deep learning-based manipulation and effectively improve the quality of deepfakes. GANs consist of two competing neural networks: i) a generative network (G) that captures the data distribution to generate synthetic samples, and ii) a discriminative model (D) that distinguishes real examples from those generated by G. Training G is designed to increase the likelihood of D making mistakes, resulting in a two-player min-max game. The discriminator's role is to identify real and fake samples. The training process of GANs continues until an equilibrium between G and D is reached, indicating that the generator and discriminator are no longer improving. However, achieving such balance is rare, and training is usually stopped when the visual quality of the generated samples no longer shows a significant improvement [56]. Alternatively, G can be trained to generate samples that D cannot distinguish as real or fake. At this point, D is detached, and G alone can generate photorealistic fake content. The working principle of GANs is illustrated in Figure 3.

Face-swap GAN (FS-GAN) [57] employs the encoder-decoder as a generator and offers additional antagonistic and perceptive losses to the automated coding system. Added counter losses improved the efficiency of image reconstruction, and the perceptual loss helped to improve the generated face alignment with the input image. However, their approach generates over-smooth, blurred face swap results. All of the above approaches require [15], [16], [51], [52], [53], [55], [57] subject-specific training involving two video sequences to represent the source and target faces. These models have limited generalization capabilities and are primarily designed for swapping faces between specific identities. Additionally, these approaches require a significant amount of source and target facial training data.

Subject-agnostic approaches have been proposed to overcome the limitations of subject-specific or pair-specific training. These approaches aim to enhance the model's flexibility and generalization capabilities, enabling them to handle diverse face swapping scenarios without requiring extensive training data. One such approach is RSGAN [58], which uses two Variational Autoencoders (VAE) to generate separate latent vector embeddings for hair and face regions, which are then combined to create a swapped face. RSGAN requires only a single image of the source and target, reducing training requirements. However, the hair regions may not align accurately with the original image, and the technique is limited to generating face images with 128 × 128 resolution. Unlike other face swapping approaches, Natsume et al. [59] introduced FSNet, which eliminates complex pre- and post-processing stages. FSNet consists of two sub-networks: a VAE that generates the latent vector for the face area of the source image, and non-face components like hairstyles and backgrounds of the target image and a generator network that performs face swapping by merging the latent vector of the source face with the non-face component of the target image. FSNet can achieve high-quality face swapping using a single source and target image, even with diverse face orientations and illumination conditions. It also preserves the background and hairstyle from the original image but struggles with face swapping when a part of the face is occluded.

Dealing with facial occlusions in face-swapping techniques is a complicated task. Common occlusions such as hair, glasses, and hands can obscure the source's or the target's face, causing visual inconsistencies and artifacts in the face swap results. Nirkin et al. [60] proposed FSGAN, an occlusion-aware approach that uses RNN and GAN to preserve target occlusions during face swapping and re-enacting facial expressions. The FSGAN model generates a source face re-enactment based on the target's pose and expression, followed by segmenting both source and target faces. The training of the face re-enactment network involves utilizing stepwise consistency loss and Poisson blending loss in progressive stages, seamlessly blending the source face into its new environment while preserving the desired skin tone and lighting conditions. Notably, the model's subject-agnostic nature eliminates the need for fine-tuning the network for each new source. Despite these strengths, the re-enactment generator's multiple iterations can result in

TABLE 2. **Deepfake face swap generation techniques overview.**

Technique	Features	Network	Resolution	Limitations
Fast Face-swap [51]	VGGFace	CNN	256×256	Applicable to frontal face, over-smoothed swapped faces, requires large training data.
Deepfake [52]	Facial Landmarks	Encoder-Decoder	256×256	Blurry output, missing gaze direction, expressions, and illumination, needs many source and target images.
FCN [55]	Facial Landmarks	CNN	256×256	No occlusion preservation, requires extensive training data per task, color tone mismatch.
FS-GAN [57]	VGGFace	GAN	256×256	Over-smoothed faces due to missing texture features.
RSGAN [58]	Facial Landmarks	GAN	128×128	Artifacts in hair region, low output resolution, sensitive to occlusion.
FSNet [59]	Facial Landmarks	GAN	128×128	Unable to handle occluded face regions.
FSGAN [60]	Facial Landmarks	GAN+RNN	256×256	Blurry textures, limited output resolution, restricted expression range, doesn't preserve source face shape.
FaceShifter [61]	Occlusion, Style Attributes	GAN	256×256	Striped artifacts, can't preserve target's expression.
SimSwap [62]	Facial Landmarks	GAN	256×256	Obvious artifacts around face boundary, large training data needed, over-smooth faces.
MegaFS [63]	Face ROI	GAN	1024×1024	Output quality dependent on StyleGAN2 capabilities.
Hififace [64]	Facial Landmarks	GAN	512×512	Lack of realism, noticeable artifacts.
AP-GAN [65]	Facial Landmarks	GAN	256×256	Compressed representation restricts high-resolution face swaps.
High-res Face-swap [66]	Facial Landmarks	GAN	1024×1024	Output faceswap Quality depends on GAN's latent code generation.
FaceDancer [67]	Facial Landmarks	GAN	256×256	No gaze movement transfer, challenging with unusual angles or turned faces.

a blurry texture, which negatively impacts the visual quality of the generated outcomes. Additionally, FSGAN struggles to preserve target image attributes such as image resolution and the face shape of the source identity. Li et al. [61] introduced a two-stage FaceShifter approach for generating high-quality face-swapping results. In the first stage, a GAN network accurately extracts and adaptively combines the identities from the source and target images. The FaceShifter model goes beyond previous face-swapping methods by extensively utilizing target image information in creating the swapped face rather than relying on limited target image information. The second stage involves using the HEAR (Heuristic Error Acknowledging Refinement Network) to refine the occluded areas. The subject-agnostic nature of FaceShifter eliminates the need for subject-specific training, making it applicable to new face pairings. FaceShifter is subject-agnostic, produces state-of-the-art identity swap results, and handles partially occluded faces well. However, it still faces difficulties due to loose attribute constraints, leading to mismatched expressions, skin color, and poses, causing temporal discontinuity and instability in the results.

To achieve high-fidelity face swapping, SimSwap [62] introduced the ID injection module. The module separates identity information from the decoder, embedding the source face identity into the target image. This eliminates the dependency on specific decoder weights and becomes applicable to any identity. The approach incorporates a weak feature matching loss to preserve the target face features while minimizing identity modifications. Zhang et al. [65] presented a video face-swapping framework named AP-GAN to generate faces that match the target face. The framework uses a U-Net-based generator with a PE (pose and expression) block to correct pose and expression and an ID block for identity translation. The framework also uses multiscale discriminator features with perceptual loss to preserve facial characteristics, such as skin color, illumination, and occlusion. However, light, skin color, and makeup variations can still affect the temporal stability of the output video.

Xu et al. [66] introduced a framework for high-resolution (High-res) face swapping using a disentangled latent space approach with StyleGAN. The approach separates the texture and appearance attributes of facial images to preserve the desired target look and texture while transmitting source identity. When the disentangled latent code is fed into the StyleGAN generator, it produces generative features that enable high-resolution face swapping. The framework also enables the creation of face-swapped videos with consistent frames and smooth transitions by incorporating code and flow trajectory constraints. However, the accuracy of the GAN inversion method in providing precise latent codes is critical for ensuring the faithful reflection of the source image identity. Other techniques like MegaFS [63] and HifiFace [64] also use GAN-based approaches for high-resolution face swapping with different identities. MegaFS [63] utilizes pre-trained StyleGAN2 latent space to identify the corresponding latent code for generating face-swapped images, while HifiFace [64] incorporates a 3D shape-aware identity extractor to enhance metrics associated with identity transfer.

Rosberg et al. [67] proposed an approach for face swapping called FaceDancer. This method addresses challenges such as lighting, occlusions, pose variations, and semantic structure preservation. The approach incorporates the Adaptive Feature Fusion Attention (AFFA) module, which learns attention masks to merge conditioned and unconditioned features selectively. Additionally, the Interpreted Feature Similarity Regularization (IFSR) technique is utilized to enhance attribute preservation. The AFFA module lets the model decide which conditioned features to discard and which unconditioned features to keep from the target face. At the same time, the IFSR method improves facial expression, head pose, and lighting preservation while ensuring a high level of identity transfer. Furthermore, FaceDancer can scale to low-resolution images with significant distortions.

B. EXPRESSION-SWAP

Expression swap, also called deepfake re-enactment, transfers source face expressions and poses to the target while preserving the target's identity. This helps the attacker to generate uncertainty, disclose information, and manipulate facts. To illustrate this type of strategy, we use the method described in [68]. Figure 4 depicts the general facial re-enactment procedure. In the first stage, facial key points were extracted through 3D landmark detectors to render landmark images for source and target faces, and then low-dimensional parameter representation, such as pose, expression from the source, and style information from the target video, were obtained through the encoders network. Extracted source and target face features were combined to produce a mixed feature map. The decoder utilized wrapped target features and a hybrid feature map to produce a re-enacted face.

The Face2Face project [69] is a contemporary research venture that led to the development of deepfake re-enactment technology. Face2Face performs a real-time reconstruction of the face to project the source actor's facial expressions onto the target actor and then wraps it with the composite shapes derived from the source video. Dense photometric constancy measures make a face and shadow consistent in each frame. The method builds a 3D face model in the first frame and modifies each frame's expressions to re-enact the face in videos instantly. However, it should be noted that due to the utilization of coarse 3D facial reconstructions of the target face, the output video does not accurately follow the source person's head and eye movements. Additionally, the synthesized mouth movement may appear unappealing to the viewer. In Suwajanakorn's approach [70], a 3D model guided by facial features was employed to synthesize the mouth region of the face. Additionally, the interior of the face was filled using a technique similar to the Face2Face method. By leveraging audio sequence data, a recurrent neural network (RNN) was utilized to generate a sparse mouth shape for each frame in the video. The mouth textures were also synthesized and merged with the target video. However, the method could only modify the facial expression, not the 3D head's position and the constantly changing background. 3D Head Generation models excel at producing high-quality lip-sync but have a severe disadvantage in handling non-verbal cues.

Agarwal et al. [71] proposed the Audio-Visual Face Re-enactment GAN (AVFR-GAN) network. This architecture uses auditory and visual cues to generate facial re-enactments. The AVFR-GAN network starts by using pre-processed data, such as a face segmentation mask, to capture the facial structure. It also uses corresponding speech signals to improve lip synchronization. The pipeline uses an identity-aware face generator to enhance the output quality further. However, extreme facial expressions or movements not present in the training data are challenging for this approach. A technique for animating a still portrait using face landmarks and 2D wraps was introduced by Averbuch-Elor [72]. Although it can manipulate facial expressions in a frontal face and allows for moderate adjustments to head position based on a source expression, it is not suitable for generating re-enacted videos and lacks the ability to control gaze movement and preserve the target individual's identity.

These approaches are extended in Deep Video Portraits [73], which enable generative neural networks to manipulate head rotation, 3D head position, facial expression, and gaze. Deep Video Portraits use a hybrid 3D deep method to fully reanimate videos by manipulating head rotation, 3D head position, facial expression, and gaze. This involves constructing a morphable 3D facial model [74] using traditional graphics techniques and processing the rendered image with a Cycle-GAN (cGAN) [75]. Overall, the results of this approach are generally favourable, but certain challenges need to be addressed. These challenges include frame dropping due to occlusion, variations in lighting, rapid motion, face misalignment, and compression artifacts. Another approach, GANimation [76], proposed generating diverse facial expressions using a dual generator network based on emotion Action Units (AUs). Additionally,

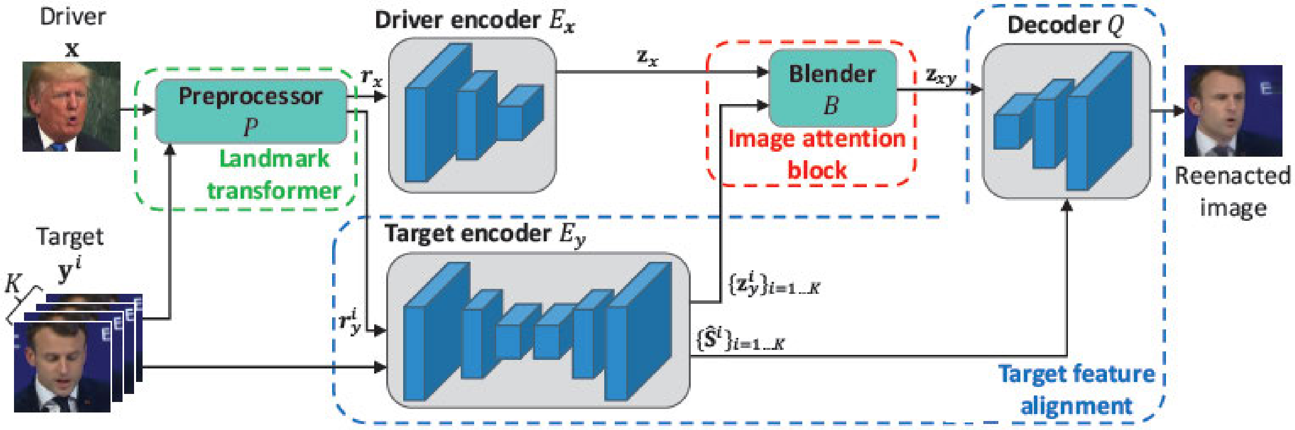

FIGURE 4. Face re-enactment process [68].

the method incorporates an attention network model that identifies and prioritizes important regions of the image for each expression. This technique enables smooth interpolation between newly created and original photos while retaining background details. However, it is worth noting that the approach has limitations in effectively handling significant variations in pose. In contrast to relying on action unit estimation, GANnotation [77] presented a deep facial re-enactment technique guided by face landmarks. This approach generates images incrementally, employing a triple loss of consistency to mitigate visual artifacts. However, it should be noted that the method primarily focuses on synthesizing images with a frontal face view for subsequent processing. Thies et al. introduced a neural texture-based pseudo-video generation process [78], intending to rectify 3D artifacts by utilizing target textures acquired from raw video data through photometric loss. The model incorporates 3D transformation and perspective effects into the learned rendering process, encoding the object's appearance within the texture space of the normalized 3D model. The technology is based on a patch-based GAN-loss similar to Pix2Pix [79]. However, it is important to note that this approach primarily focuses on re-enacting mouth movements while keeping eye movements unchanged.

The aforementioned facial re-enactment models rely on extensive datasets of source and target identities, and they are restricted to re-enacting specific individuals, making them unsuitable for generating photo-realistic re-enactments for unknown identities. A few or one-shot facial re-enactment techniques have recently been proposed to transfer facial expressions and poses using fewer or even a single target image to address this limitation. Zakharov [80] proposed a few-shot learning approach that used meta-learning to model human faces. They demonstrated how to make the Mona Lisa's face speak by using an embedded network and GAN to derive pose-invariant video information. However, their approach encountered challenges with generating accurate gaze due to landmark mismatch, resulting in a noticeable difference in the perceived personality of the re-enacted face. Zhang [81] introduced an unsupervised one-shot facial re-enactment technique, which utilizes face parsing maps and identity-specific features to guide the re-enactment process. However, it has been observed that this method produces distorted results in re-enacting a different target, indicating limitations in achieving accurate and realistic re-enactments in such scenarios. On the other hand, the First Order Motion Model (FOMM) [82] is a self-supervised network that effectively separates appearance and motion components by modeling the motion around key points using an affine transformation. FOMM can generate re-enactments with just a few training examples. Furthermore, it implements an occlusion-aware generator that calculates an occlusion mask for regions not visible in the source image. Despite this approach's significant improvements, synthesized faces still exhibit visual artifacts in continuously changing expressions. This limitation can be attributed to the dense 2D motion field estimation, which does not fully capture the intricate 3D facial motion. Hao et al. [83] proposed a one-shot FaRGAN model. This model utilizes facial landmarks to capture facial expressions and recreate poses and expressions. The generator model is composed of a transformer and an embedder, and they incorporate the spatially adaptive normalization module (SPADE) [87] to incorporate landmark information into the embedder model. However, it is important to note that the model's performance is compromised in cases where there is significant variation in appearance between the source and target faces. Furthermore, the model does not leverage gaze information in the landmark mask, limiting its ability to enhance the motion realism of the synthesized face. Ha et al. introduced the MarioNETte [68], which utilizes target feature alignment and an image attention block to incorporate target facial features into re-enacted face images. This approach allows for the recreation of the faces of unseen targets using only a few reference shots. However, it is important to note that their proposed solution does not preserve the appearance of the target face.

TABLE 3. Deepfake expression-swap techniques overview.

Technique	Features	Network	Resolution	Limitations
Face2Face [69]	Facial Landmarks	3DMM	1024×1024	Struggles with realistic mouth deformations, no head pose control, sensitive to occlusion.
Learning Lip Sync [70]	Mouth landmarks, MFCC audio	RNN	2048×1024	Requires extensive training data per expression swap, no full 3D head pose control including background.
AVFR-GAN [71]	Face Mesh, Segmentation Mask, Speech	GAN	256×256	Limited handling of varied poses.
Bringing Portraits to Life [72]	Facial Landmarks, 2D wraps	3DMM	600×800	Sensitive to large head poses, no source gaze transfer, target identity preservation.
Deep Video Portraits [73]	Parametric Face Model	CGAN	1024×1024	Sensitive to motion and occlusion.
GANimation [76]	Action Units	GAN	128×128	Lacks gaze adaptation and pose variation.
GANnotation [77]	Facial Landmarks	GAN	128×128	Lacks gaze adaptation.
Neural textures [78]	UV-map, Texture Map	3D Modeling + Patch-based GAN-loss	512×512	Unable to re-enact eye movement.
Neural Talking Head Model [80]	Facial Landmarks	GAN	256×256	Lacks gaze adaptation, noticeable target personality mismatch.
One-shot Face Reenactment [81]	Face Parsing Map, Identity Features	GAN	256×256	Distorted low-quality output.
FOMM [82]	Sparse Keypoints, Local Affine Transformation	GAN	256×256	Fails to preserve consistently changing background.
FaR-GAN [83]	Facial Landmarks	GAN	256×256	Sensitive to considerable appearance differences between source and target faces, lack of gaze adaptation.
MakeltSmile [84]	Face ROI	GAN	256×256	Limited to expressions with open face, results quality depends on training data.
FaceSwapNet [85]	Facial Landmarks	GAN	256×256	Lacks gaze adaptation.
MarioNETte [68]	Facial Landmarks	GAN	256×256	Cannot fully preserve source facial features.
PNCC GAN [86]	3D Facial ROI	GAN	256×256	Sensitive to extreme poses or expressions.

A significant challenge in facial re-enactment tools is accurately representing the opening of the mouth. Existing methods struggle with predicting the inside of the mouth and teeth, resulting in inconsistent outcomes. Previous approaches like FOMM [82] rely on sparse key points to model head pose and facial expressions, which fails to capture the rich texture details of teeth. As a result, the resulting faces may lack the necessary detail, particularly when generating smiling faces with distinct tooth structures. Fu et al. [84] introduced MakeltSmile to tackle the issue of incomplete tooth structure in face re-enactment. MakeltSmile utilizes a geometry-aware encoder to extract tooth structure information from the driving video. By incorporating tooth information from the driving face rather than the target face, this method addresses the problem of inaccurate tooth structure. MakeltSmile comprises two modules: a feature extraction module and a face generation module. The feature extraction module captures precise tooth texture from the smiling driver's face, while the face generation module faithfully reproduces faces with clear and appropriate teeth using a GAN-based technique.

Re-enactment models struggle to preserve the target identity, leading to flawed re-enactments. The FaceSwapNet model [85] addresses this challenge by enabling the transfer of facial expressions and gestures from a source to random targets using two key modules: the landmark swapper and the landmark-guided generator. The landmark swapper employs two encoders and one decoder to ensure identity consistency between the generated person and the target person by aligning the source's and target individual's landmarks. The landmark-guided generator leverages the exchanged landmark and the target person's geometry information to generate the re-enacted image. A novel triplet perceptual loss is introduced to enhance face appearance and geometry information learning. Xue et al. [86] presented an approach that leverages a GAN with a Projected Normalized Coordinate Code (PNCC) to preserve facial details and reflect the target identity accurately. The technique reconstructs the PNCC using the target identity parameters and the source pose and expression parameters estimated by 3D facial reconstruction to filter out the source identity. By adopting the reconstructed representation as the driving information, their approach achieves high-fidelity generation and ensures the preservation of target identity in face re-enactment.

C. CHALLENGES FOR DEEPFAKE CREATION

The researchers improved the deepfake generation network training by integrating the tertiary concepts to achieve a more hyperrealistic and natural result with high confidence [88], such as style transfer, motion transfer, and semantic segmentation. However, the current deepfake is still imperfect and leaves room for improvement.

- Computational Cost: The training dataset's significance in the deepfake domain is pivotal, as most deepfake technologies rely on ample genuine data for more convincing content generation. This has led to increased computational demands. To tackle this, researchers focused on minimal data use via few-shot learning [80], [81] for deepfake creation. However, these strategies encountered challenges in accurately reproducing intricate facial elements, like gaze and expressions, resulting in noticeable discrepancies in perceived identity and the overall authenticity of the generated deepfakes. The ongoing focus on minimizing computational needs and refining training datasets drives advancements in deepfake research.
- Identity leakage: When creating realistic deepfakes, preserving the target identity is difficult if there is a significant difference between the target and driving identities. This is particularly evident in face reenactment tasks where a source identity drives target expressions. During training, aspects of the facial identity data are transferred to the generated face. This event occurs when training is performed on single or multiple identities, yet data pairing is performed on the same identity. Addressing these challenges holds the key to advancing deepfake technology in the future.
- Deepfake Quality: One potential trend in deepfake generation is the improvement of output quality. Due to the instability of GAN training, most deepfake outputs contain subtle traces, such as unusual texture artifacts or pixel inconsistencies, that make them vulnerable to detection. The challenges in creating realistic deepfakes involve achieving natural emotions, seamless timing, and realistic speech rates. Additional issues include visible anomalies like frame flickering and jittering due to the lack of temporal consistency in deepfake generation frameworks that process each frame individually. Currently, deepfakes are commonly produced in controlled environments with uniform lighting and backgrounds. However, sudden changes in lighting during indoor/outdoor transitions can cause color disparities and unexpected anomalies in the output. Future research could focus on improving deepfake quality by addressing artifacts, enhancing output resolution, and defending against attacks.

III. DEEPFAKE DATASETS

Due to growing concerns about the potential risks posed by deepfake abuse, numerous groups have made valuable contributions by creating datasets to support deepfake detection. These datasets can be classified into three generations based on visual quality, quantity, and the range of deepfake techniques employed. Generally, the more recent generations of datasets offer larger volumes of data and greater diversity in synthetic methods. In this section, we present existing known deepfake datasets for face and expression swaps. Figure 5 highlights quality and variations in different generations of deepfake datasets.

A. FIRST GENERATION

- **UADFV [40]:** Yang et al. [40] created this dataset consisting of 98 videos (49 authentic YouTube videos and 49 fake videos) for their deepfake detection experiments. FakeApp [18] was employed to generate low-quality deepfake face swapping videos with a resolution of 294 × 500 pixels. Each video has an average duration of 11:14. The dataset is small enough to handle a desktop environment easily.
- **Deepfake-TIMIT (DF-TIMIT) [89], [90]:** In this particular dataset, there are a total of 960 videos. Among them, 320 videos were carefully chosen from the VidTIMIT dataset [91] as they formed 16 pairs with a striking resemblance. The faces in these selected videos were swapped utilizing the face-swap GAN, an open-source deepfake algorithm [57]. Two subsets were created: DF-TIMIT-Low-Quality (LQ) with 320 videos, each containing around 200 frames of 64 × 64 pixels, and DF-TIMIT-High-Quality (HQ) with 320 image sequences, each consisting of approximately 400 frames sized at 128 × 128 pixels.
- **Fake Faces in the Wild (FFW) [92]:** A dataset of 150 videos was presented by Khodabakhsh et al. [92]. The videos were created using both deepfake and conventional methods, such as computer graphics image (CGI) and splicing. The videos range from 60 to 2000 frames with a maximum resolution of 480 pixels.
- **FaceForensic++ (FF++) [7]:** The dataset provided contains a total of 4,000 videos, consisting of 1,000 real videos and 5,000 fake videos. Within the real video category, 1,000 videos were modified using four facial modification techniques. Two of these techniques involve face swapping deepfake methods, namely Autoencoder face-swap (DF) and Face-swap (FS), while the other two techniques involve expression swapping deepfake methods, namely Face2Face (F2F) and Neural

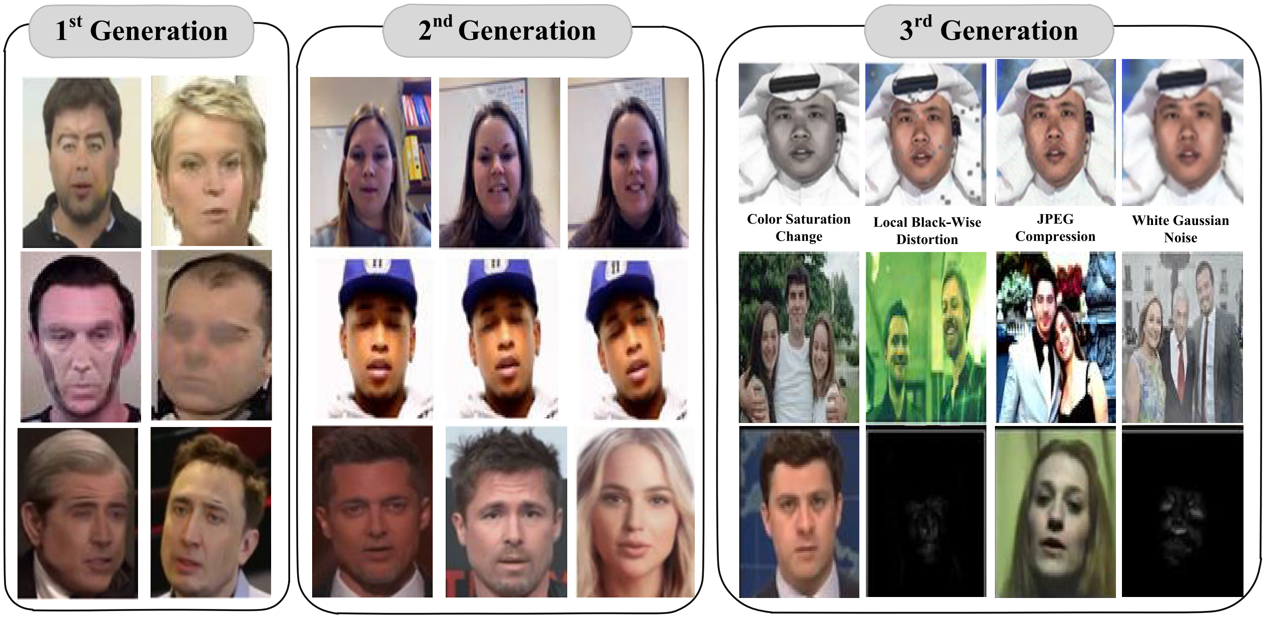

FIGURE 5. Examples of deepfake dataset evolution: Revealing weaknesses in first generation deepfake dataset, the excellent naturalness achieved in the second generation of deepfakes, advanced variations in third generation Deepfakes with perturbations, multi-face in one frame, and manipulation masks.

Texture (NT). The dataset is available in two qualities: uncompressed (Raw C0) and compressed (LQ, HQ) using the H.264 codec. The HQ videos, referred to as High-quality (C23), were compressed at a rate of 23, while the LQ videos, known as Low-quality (C40), were compressed at a rate of 40. Regarding the original video resolution, approximately 55% of the videos have a resolution of 854 × 480, which corresponds to VGA resolution. About 32.5% of the videos have a resolution of 1280×720, classified as HD. The remaining 12.5% of the videos have a resolution of 1920, commonly known as Full HD.

The quality of the images in these datasets has not explicitly been improved, causing low-resolution and unrealistic deepfake results with blurry or flawed facial features. Additionally, the datasets were created in controlled environments with specific lighting, camera angles, and backgrounds. Consequently, while these datasets are helpful for training detectors that do not require a large dataset, they are not ideal for identifying complex facial and expression swaps.

B. SECOND GENERATION

- **Deepfake Detection dataset (DFD) [93]:** The dataset, released collaboratively by Google and Jigsaw, contains 363 real videos and 3068 face-swapped fakes featuring 28 paid actors. These actors engage in actions like walking, hugging, and expressing emotions such as anger, happiness, disgust, and neutrality. The exact synthesis algorithm has not been disclosed, but it is likely a deepfake (auto-encoder) algorithm. The dataset offers videos in three compression qualities: Raw (C0), C23, and C40.
- **Deepfake Detection Challenge Preview (DFDC-P) [34]:** Facebook, Microsoft, and AWS have launched a collaborative challenge to promote the development of Deepfake detection algorithms through an organized dataset. The dataset includes 1,131 original videos featuring 66 actors and 4,113 deepfake (face-swap) videos. These videos contain three types of augmentations as perturbation techniques. However, the synthesis algorithm used in the dataset has not been disclosed.
- **Celeb-DF [6]:**This dataset includes 590 authentic YouTube videos of 59 celebrities from different backgrounds, genders, and nationalities. Additionally, it has 5,639 manipulated videos created using an advanced deepfake face swap algorithm (auto-encoder). The dataset offers a range of camera orientations, lighting conditions, and backgrounds, and the deepfakes are of high quality with no visible defects., unlike previous datasets featuring lower-quality videos with noticeable artifacts. Each video in the dataset lasts for about 13 seconds and follows a standard frame rate of 30 frames per second.
- **FaceShifter (FSh) [61]:** The dataset comprises 10000 high-quality deepfake videos generated using the FaceShifter algorithm by manipulating real videos from FF++ [7] dataset. Videos are available in three different compression qualities: Raw (C0), C23, and C40.
- **Deepfake MNIST+ [94]::** A dataset called Deepfake MNIST+ was introduced by Huang et al. [94].

TABLE 4. **“Comparative Analysis of Deepfake Datasets: Notation and Definitions. (Note: “NA” = not applicable, “–” = unknown, and “Vids” and “Imgs” represent videos and images.)”**

Dataset	Type	Real	Fake	Deepfake Generation Techniques		Perturbation Methods	Visual Quality	Year
				Expression-Swap	Face-Swap			
UADFV [40]	Vids	49	49	NA	FakeApp	0	Low	2018
DF-TIMIT [89]	Vids	320	640	NA	Faceswap-GAN	0	Low	2018
FFW [92]	Vids	-	150	NA	FakeApp, FaceSwap	0	Low	2018
FF++ [7]	Vids	1k	4k	Neural-Texture, Face2Face	Deepfake, FaceSwap	2	Low	2019
DFD [93]	Vids	363	3068	NA	Unknown Face-swapping Algorithm	2	Low	2019
DFDC-P [34]	Vids	1,131	4113	NA	Two Unknown Face-swapping Algorithms	9	Low	2019
Celeb-DF [6]	Vids	590	5639	NA	Auto-Encoder	0	High	2020
DFDC [35]	Vids	19154	100k	NTH	FSGAN, MM/NN, DeepFaceLab	19	High	2020
FSh [61]	Vids	1k	10k	NA	FaceShifter	2	High	2020
DF-1.0 [36]	Vids	50k	10k	NA	DF-VAE	7	High	2020
WDF [98]	Vids	3,805	3509	NA	Unknown Face-swapping Algorithms	-	High	2020
KoDF [99]	Vids	62,166	175,776	FOMM, ATFHP, Wav2Lip	FSGAN, DeepFaceLab, Face-Swap	0	High	2021
FFIW [100]	Vids	10k	10k	NA	DeepFaceLab, FSGAN, FaceSwap	-	High	2021
FN [101]	Imgs/ Vids	1,438,201/ 99,630	1,457,861/ 121,617	Talking Head, FOMM, ATVG Net	Deepfake, FSGAN, FaceShifter, BlendFace, MM Replace, DSS	36	High/ Low	2021
OF [102]	Imgs	16k	173k	-	GAN+Auto-Encoder	6	High/ Low	2021
MNIST+ [94]	Vids	10k	10k	FOMM	NA	2	High	2021
DF-Mobio [96]	Vids	31k	15k	NA	Faceswap-GAN	0	High	2022
FMFCC-V [103]	Vids	44,290	38,102	NA	Faceswap, Faceswap-GAN, DeepFaceLab, Recycle-GAN	12	High	2022
DF-Platter [104]	Vids	764	132,496	NA	FSGAN, FaceShifter, FaceSwap	2	High/ Low	2023

It contains 10,000 videos of human faces displaying ten different expressions and 10,000 real-face videos collected from VoxCeleb1 [95]. The facial expression swapping videos were created using the First Order Motion Model (FOMM) [82] with ten different actions, including blink, open mouth, yaw, nod, head tilted right, head tilted left, look up and smile, surprise, and embarrassment. Furthermore, challenging samples were selected using two public liveness detection APIs, where the detectors failed to detect them precisely. The videos in this dataset are compressed using the H.264 codec, with compression rates of C23 and C40.

- **DF-Mobio [96]:** This dataset contains over 46,000 videos, including 15,000 deepfakes and 31,000 authentic videos. The original videos were sourced from the Mobio dataset [97], featuring people speaking directly to the camera using a phone or laptop. The videos simulate virtual meetings on platforms like Zoom or Skype. For training the GAN model [57] to create deepfakes, 2,000 faces were captured per subject at a frame rate of 8 frames per second, with an input size of 256 × 256 pixels.

The dataset of the second generation aims to enhance the dataset's size and visual quality. Nevertheless, these datasets lack diversity in techniques employed to create manipulated data and do not possess specific task annotations, rendering them less efficient in tackling real-world challenges.

C. THIRD GENERATION

- **Deepfake Detection Challenge (DFDC) [35]:** This dataset contains approximately 100,000 manipulated videos and 19,000 original videos created for the Kaggle competition [105] as an extension of the DFDC preview dataset. The dataset includes a variety of motion and camera placement scenarios, covering different lighting conditions such as day and night. Different algorithms were used for manipulation, including DeepFaceLab [15], [16] with two resolution options: 128 × 128 (DF-128) and 256 × 256 (DF-256), an unlearned morphable mask face swap algorithm (MM/NN) [106], and Generative Adversarial Networks (GANs) techniques such as Neural Talking Heads (NTH) [80] and FSGAN [60]. To promote the development of deepfake detection models, a competition II was launched along with a hidden test set to evaluate the submitted model's performance.
- **WildDeepfake (WDF) [98]:** This dataset is unique in the third generation as it includes real and manipulated videos from the internet. The videos dataset includes 3,805 authentic videos and 3,509 deepfake face swap videos. These videos can have more than ten individuals in a scene, making it a unique and valuable resource for improving deepfake detection in real-world scenarios. The dataset includes diverse video content with events like broadcasts, films, interviews, and talks, with varying backgrounds, environments, lighting conditions, compression rates, resolutions, and video formats.
- **DeeperForensics-1.0 (DF-1.0) [36]:**The data set includes 50,000 original videos and 10,000 deepfake manipulated videos. One hundred paid actors with four different skin tones were used to create high-resolution source videos. These videos were captured with seven cameras at different locations and under nine lighting conditions, displaying eight facial expressions (happy, neutral, sad, angry, disdain, surprise, fear, and disgust). A robust face-swapping mechanism called deepfake Variational-Auto-Encoder (DF-VAE) was also developed to generate the fake videos. DeeperForensics-1.0 utilizes a range of perturbation techniques, including Gaussian blur, random flip, random brightness contrast, image compression, random cropping, color saturation, and local blockwise distortion, to simulate real-world situations and diversity in the dataset.
- **Korean Deepfake Dataset (KoDF) [99]:** This dataset focuses on Korean subjects and includes 62,166 real clips and 175,776 deepfake clips. Six different deepfake models were employed to create these fake videos. Out of these six models, three of them are face swapping models, namely FSGAN [60], DeepFaceLab [15], [16], and FaceSwap [107]. The remaining three models are re-enactment models, which include the First Order Motion Model [82], ATFHP [108], and Wav2Lip [109]. The dataset was created using a front-facing camera to guarantee minimal variation in the subject pose. The video clips are of high quality with a resolution of 1920 × 1080, and the dataset was not subjected to any perturbations such as compression, resizing, or manual editing.
- **Face Forensics in the Wild (FFIW) [100]:** The dataset consists of 4,000 videos obtained from YouTube, where each video is divided into four equal parts. A random 12-second sequence is selected from each part to generate the fake videos. Face-swapping videos were created using DeepFaceLab [15], [16], FSGAN [60], and a graphics-based FaceSwap method by randomly selecting two videos from a collection of 12,000 filtered sequences. On average, each image in the dataset includes three human faces. The dataset underwent an automated manipulation process using a domain-adversarial quality assessment network to reduce cost. The videos range from 2 to 74 seconds in length with a resolution of 854 × 480 and a frame rate of 30 frames per second, making a total of 53,000 images.
- **ForgeryNet (FN) [101]:** The dataset has been divided into two categories: a fake image set with over 2.9 million images and a fake video set with more than 220,000 videos. Both subsets also contain authentic data. To create the fake images and videos, 15 image forgery techniques and 8 video forgery techniques were used. The dataset was also mixed with 36 perturbations to make it more challenging.
- **OpenForensics (OF) [102]:** The OpenForensics dataset is designed explicitly for multi-face forgery detection and segmentation tasks, offering a large image dataset with diverse backgrounds. Additionally, the creators developed a method to generate countless fake faces without repeatedly training the auto-encoder. The dataset was created using real images obtained from Google Open Images, while fake faces were generated using a process involving GAN-based face synthesis, Poisson blending, and color-matching algorithms. The OpenForensics approach provides high-resolution face images with improved visual quality and a more natural appearance. To simulate real-world challenges, various perturbations were applied, resulting in a challenging test subset. These perturbations were categorized into color manipulation, edge manipulation, blockwise distortion, image aliasing, convolution mask transformation, and external effects, each with three levels of intensity: easy, moderate, and difficult.

- **Fake Media Forensics Challenge of China Society of Image and Graphics (FMFCC-V) [103]:** The dataset includes 38,102 deepfake videos and 44,290 authentic videos featuring 83 Asian individuals, each speaking for approximately 40 minutes. These videos showcase various head poses, facial expressions, backgrounds, resolutions, and frame rates. The videos were captured indoors and outdoors, with resolutions mainly at 480p, 720p, and 1080p and frame rates of 25fps and 30fps. The deepfake videos were generated using four different methods: Faceswap [107], Faceswap-GAN [57], DeepFaceLab [15], [16], and Recycle-GAN [110]. To simulate real-world situations, deepfake and authentic videos were subjected to twelve types of perturbations, including brightening, noise addition, blurring, darkening, contrast adjustment, and flipping. The FMFCC-V dataset offers two versions of the deepfake dataset for different applications. The long version includes approximately 16 minutes of video without any perturbations, while the short version comprises 10-second videos, with half having perturbations applied, including both authentic and deepfake videos.
- **DF-Platter [104]:** The DF-Platter dataset contains 764 YouTube videos of single and multiple subjects, showcasing a wide range of expressions, poses, backgrounds, lighting conditions, and occlusions. It employs FSGAN [60], FaceShifter [61], and FaceSwap [107] techniques to generate 132,496 high-resolution (HR) and low-resolution (LR) deepfake videos. The dataset consists of three sets: Set A for single-subject deepfakes, Sets B and C for intra-deep fakes, and multi-face deepfakes. All subjects are of Indian ethnicity and annotated with attributes such as resolution, gender, age, skin tone, and facial occlusion. The DF-Platter dataset maintains a balanced distribution of resolution and gender. The videos are available at three compression levels (c0, c23, c40), with a duration of around 20 seconds each, and provided in MPEG4.0 format, with a frame rate of 25 fps.

The third generation of the deepfake dataset demonstrates a significant improvement in sample quality compared to previous generations. It contains a wide range of video samples that showcase different subjects, backgrounds, and lighting conditions. The third generation of the dataset was generated using various deepfake models, and it incorporates different perturbations, such as compression, blurring, and noise, to imitate real-world deepfake attacks and challenges. Table 4 summarizes the various deepfake datasets available for face and expression swapping.

IV. DEEPFAKE DETECTION

Detecting deepfakes in images and videos requires analyzing the content to determine if it has been altered or remains in its original state. In the research community, detecting deepfakes is commonly approached as a binary classification problem, where videos are categorized as either genuine or fake. To accomplish this, classifiers rely on identifying features that differentiate between real and manipulated content. Researchers use various methods for extracting features, such as traditional machine learning or deep learning algorithms, to detect inconsistencies and artifacts in deepfake videos. Deep learning methods are preferred for their speed and accuracy in automatically extracting features. The process of detecting deepfakes typically involves extracting facial information from video frames, followed by using neural networks or conventional techniques to extract features. The frames are then classified based on these features. However, current deepfake detection methods face challenges regarding generalization, robustness, and interpretability, which hinder their practical applications.

- **Generalization:** It is a widely used quantitative tool to assess the performance of an algorithm on unseen datasets (not considered during training). However, many suggested deepfake detection approaches rely on supervised learning, which is susceptible to overfitting. A model trained specifically on face swapping may struggle to detect other manipulation techniques, such as expression swapping and other deepfake face swap implementations. This presents a significant challenge in real-world scenarios where new deepfake face manipulation methods are continually emerging. Retraining detectors for each new modification method becomes impractical due to the lack of adequately annotated data from new deepfake manipulation methods.
- **Robustness:** Social media platforms like Twitter, Instagram, and Facebook often compress, resize, and remove metadata from videos before they are shared on the platforms. These steps are aimed at protecting user privacy and conserving network resources. However, these practices create obstacles to deepfake detection. High compression rates, in particular, result in significant loss of image data, making deepfake detection difficult [47]. Moreover, adversarial attacks pose an additional challenge for deepfake detectors to maintain their robustness. These attacks use well-engineered perturbations such as Gaussian noise, Blur, translation, resizing, first-order gradient, etc., to increase the false positive detection rate [111]. These adversarial perturbations trick the detector [111], [112] into detecting fake content as real. Additionally, an adversary can choose perturbations that are too small to be seen by the naked eye [113], [114]. Therefore, it is essential to develop robust deepfake detection algorithms that can resist social media laundering and adversarial attacks.
- **Interpretability:** Neural network methods are used to perform batch analysis on huge sets of videos. However, the lack of interpretability in neural network-based algorithms has always been a concern. These models function as black boxes, leaving users without clear explanations for their performance. In forensic analysis,

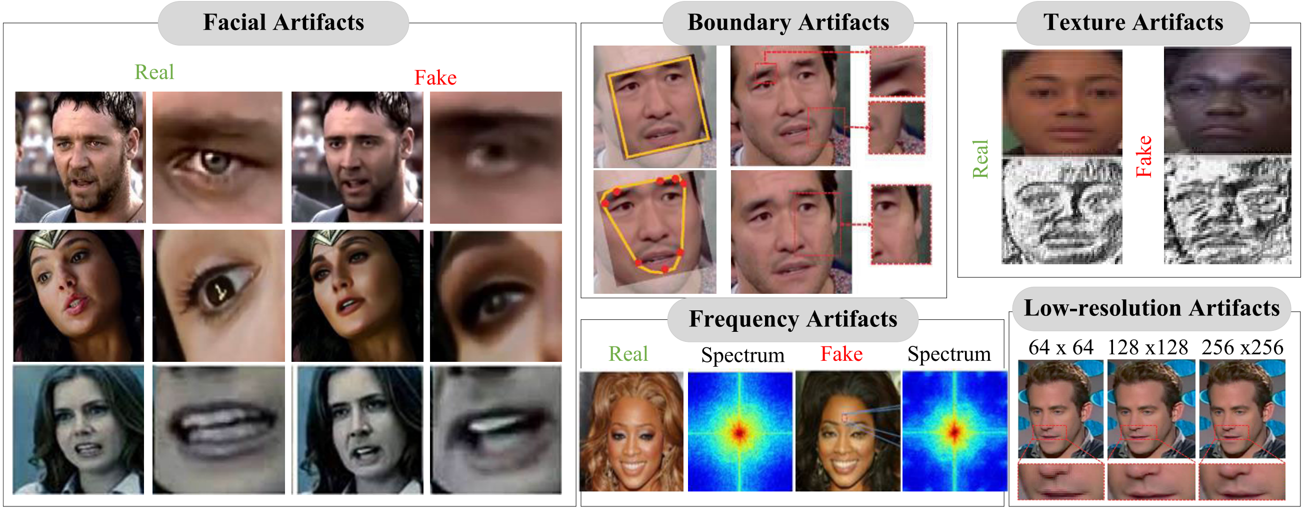

FIGURE 6. Visualization of different artifacts left by deepfake generation process.

TABLE 5. Evaluation metrics employed by state-of-the-art deepfake detection approaches.

Metric	Description
Accuracy (Acc)	The proportion of data instances accurately predicted out of the total data instances.
Precision	The ratio of correctly predicted positive data to the total number of predicted positive data.
Recall	The ratio of true positive predictions to the dataset's total number of positive instances.
F1-Score	The weighted average of precision and recall when the weights are equal.
AUC	A measure of the degree of separability between two classes of a given model.
Error Rate (EER)	The proportion of misclassified instances from the total number of instances in a dataset.

practitioners rely on these detection algorithms for a limited number of videos. Therefore, if a numerical score indicating the possibility of a video being generated through a synthesis algorithm lacks a logical foundation, it will not be useful for practitioners. In order to make the results reliable in the real context, the detection method must offer interpretability [115].

This section explores deepfake detection techniques, classifiers, notable results, evaluated datasets, and the aspects of generalization, robustness, and interpretability in detection systems. It's important to mention that researchers use different metrics to evaluate the performance of deepfake detection techniques. Table 5 provides a comprehensive list of evaluation metrics employed in state-of-the-art deepfake detection techniques.

A. ARTIFACT-BASED APPROACH

The majority of frame-based deepfake detection techniques rely on identifying the artifacts left by the Generative Adversarial Network (GAN) during the generation of the deepfake. These artifacts act as evidence of manipulation and can be extracted as features for the detection approach. The methods use hand-crafted features or neural networks to identify specific artifacts unique to deepfakes. These artifacts include a range of indicators, including biometric and biological cues, as shown in Figure 7 and visual irregularities caused by deepfake generators, as demonstrated in Figure 6.

1) VISUAL ARTIFACTS

Researchers have adopted both conventional machine learning methods and deep neural networks for deepfake detection in this approach.

The deepfake face swap algorithm generates faces with resolutions ranging from 64×64 to 256×256 [116]. However, due to its limited resolution, the algorithm faces challenges in producing small, high-quality moving facial features such as nose, eyes, hair and skin texture. Falko Matern [43] proposed a deepfake detection technique that targets the absence of details in small facial components. This technique involves extracting texture features from facial regions such as the eyes and teeth, which were used to train two different classifiers - the Multilayer Perceptron (MLP) and the logistic regression model (LogReg). The MLP-trained features derived from the eye area are effective in detecting deepfake face swaps. However, the logistic regression model incorporating eye and teeth features is more effective in detecting face re-enactment. While this method is computationally efficient, it is limited to images that meet certain conditions (e.g., open eyes, visible teeth).

Another deepfake detection approach, proposed by Yang [40], involves using the 3D head pose artifact. This method relies on the hypothesis that synthesized facial regions in deepfake videos display inconsistencies in 3D head pose estimation compared to genuine videos. To achieve this, 68 facial landmarks were utilized to estimate head

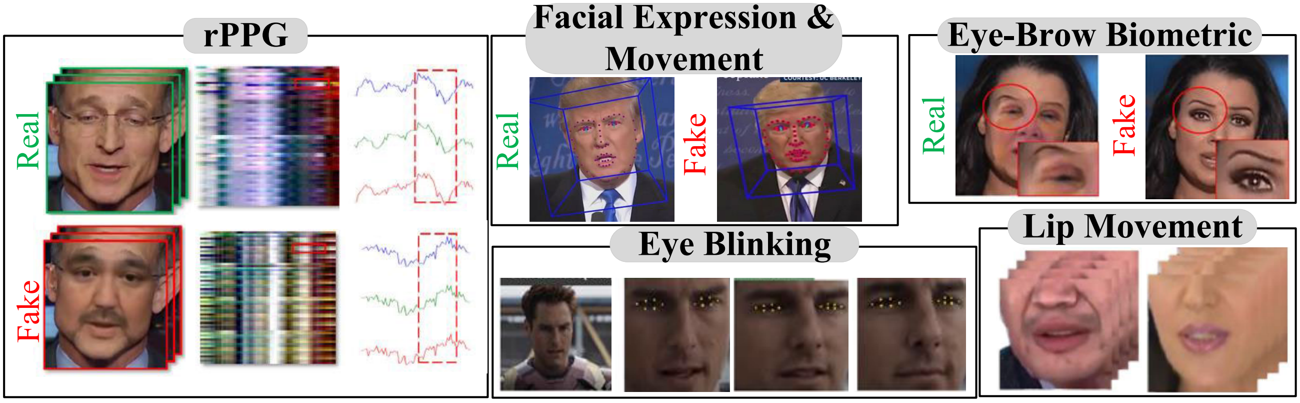

FIGURE 7. Visualization of deepfake biometric and biological artifacts for detection.

orientation and position differences. These differences were then employed to train a Support Vector Machine (SVM) classifier. The approach was effective on two small datasets, UADFV [40] and DARPA MediFor GAN Challenge [32]. However, it did not perform well on noisy and blurry images. This led to challenges in precisely predicting the positions of facial landmarks, subsequently affecting the accuracy of head pose estimations in large, visually challenging datasets [6]. In [117], Li et al. represented a method for detecting deepfake images and videos by analyzing facial symmetry. The technique identifies inconsistencies and unnatural traces in the face symmetry of the manipulated media. The dlib library [118] was used to detect faces and locate facial landmarks to estimate the 3D face pose. Face alignment was achieved by rotating around the center of the eyes. Symmetrical face patches were cropped from the images using sliding windows within the face region. Their approach utilized a Deep Residual Network (DRN) architecture and introduced a novel multi-margin angular loss function to measure the similarity of symmetry features. This loss function incorporated angular margin penalties to enhance intra-class compactness and inter-class discrepancy. Although the method exhibited robustness to data re-compression, it did not perform well on high-quality datasets [6], [35]. In light of the hypothesis that some deepfake methods are ineffective in producing realistic facial manipulations from different angles, resulting in unwanted artifacts such as blurring and irregular textures in the generated videos, Xu [119] proposed a GLCM (Gray-Level Co-occurrence Matrix) based technique to extract texture features. These features were classified using SVM to determine whether the face in the input video is original or manipulated. Both low and high-quality videos from the Deepfake-TIMIT dataset [90] and the FF++ dataset [7] were used to evaluate the performance of this approach. However, the proposed approach does not provide a generalized solution for a visually challenging dataset [6]. Another study by Kingra et al. [120] utilized Linear Binary Pattern (LBP)-encoded images to extract texture features. Unlike LBP-encoded histograms that only contain intensity information, LBP-encoded images provide both location and intensity information. The approach involves Face region detection and extraction, feature extraction by LBP, and classification using a CNN-based architecture. This approach, known as LBPNet, is able to detect deepfake images in advanced deepfake datasets [6], [34], including animations [82] and manipulated faces from mobile applications. Additionally, the model is designed to be interpretable, as the class activation maps illustrate the reasons for the predictions. However, it is important to note that LBPNet's performance is not robust for highly compressed video.

Rather than only considering spatial artifacts, researchers have investigated the flaws of deepfakes in the frequency domain. Frank et al. [121] comprehensively investigate the artifacts present in the frequency domain in various GAN architectures and datasets. They found that the upsampling techniques used in GANs cause severe artifacts. Durall [122] proposed a novel approach for analyzing artifacts in the frequency domain of deepfakes. The method involved applying the Discrete Fourier Transform (DFT) to examine the spectral distributions and calculate average amplitudes across different frequency bands. These amplitude measurements were then used as feature vectors, and fed individually into various classifiers, including SVM, K-Means, and Logistic Regression, for both supervised and unsupervised detection of deepfakes. Notably, the unsupervised classifiers demonstrated superior performance in terms of accuracy compared to the supervised classifiers. The effectiveness of this approach was demonstrated in accurately identifying medium and high-resolution deepfake images from the FF++ dataset [7]. However, detecting low-quality facial images was challenging due to the significantly narrower accessible frequency spectrum. In another approach, Kohli [123] proposed a method for detecting face deepfakes using a shallow frequency convolution neural network (fCNN). The approach involved converting facial images into the frequency domain using a 2D global Discrete Cosine Transform (DCT) and analyzing the activation map of the fCNN to derive key features for face classification.

However, the proposed method did not achieve significant generalization for the challenging Celeb-DF [6] dataset. Liu et al. [124] utilize Discrete Fourier Transform (DCT) to extract the phase spectrum. The authors emphasized the importance of local texture information in detecting forgeries, as well as the sensitivity of the phase spectrum to upsampling. Their method, Spatial Phase Shallow Learning (SPSL), combines spatial images with the phase spectrum to capture up-sampling features in facial forgery, focusing on local regions through a shallow network design. The method has demonstrated good generalization capabilities during cross-dataset evaluation. However, it is essential to note that methods relying on upsampling artifacts are ineffective for deepfake generation methods that do not incorporate upsampling.

Most deepfake generation approaches involve merging a fake face with a background image, often leading to visible irregularities around the facial area, such as uneven brightness borders. Various detectors for deepfake artifacts concentrate on these blending imperfections [125], [126], [127]. Yuezun et al. [125] proposed a method using a Convolutional Neural Network (CNN) based approach to detect distortions in facial images at the pixel level. They hypothesized that deepfake algorithms utilize affine warping to resize the images to match the target face, introducing resolution irregularities around the face boundaries. To extract facial features, the approach utilized dlib [118] and trained four CNN models: VGG16 [128], ResNet50, ResNet101, and ResNet152 [129]. Among the four, RESNET-50 was the most efficient in handling low-quality face swap videos. However, its generalization capabilities are limited for high-quality datasets [6]. Nirkin et al. [126] proposed a method that utilizes two networks, one for identifying the face region through narrow semantic segmentation and the other for recognizing the context around the face. Inconsistencies between the face and surroundings were detected by comparing these network outputs. The approach demonstrated resilience against image laundering attacks (JPEG compression, scaling) for deepfake face swap. Despite its effectiveness, this method has limitations in generalizing across high-quality deepfake datasets [6] and is ineffective against deepfake techniques that generate the entire image. The framework proposed by Kim et al. [127] utilizes a content feature extractor (CFE) and a trace feature extractor (TFE) to identify deepfake videos. The CFE focuses on extracting facial features such as wrinkles, eyes, and skin tones. At the same time, the TFE captures trace information, including subtle texture variations and facial contours, from non-facial content images. By merging features extracted from both models, the framework aims to identify deepfakes by considering a wider range of tampering indicators and not relying solely on facial information. The model demonstrates reliable performance across various compression levels of deepfake videos. However, its effectiveness on high-quality deepfake datasets has yet to be evaluated.

2) BIOLOGICAL AND BIOMETRIC ARTIFACTS-BASED APPROACH

Deepfake detection based on biological artifacts examines the subject's physiological and psychological responses to stimuli. These responses include changes in heart rate and eye movements. The premise of this approach is that these responses are difficult to control, and deepfake algorithms are unable to replicate them accurately. On the other hand, biometrics-based deepfake detection analyzes video features such as facial movements, facial expressions, voice patterns, and lip movements. The key insight behind these deepfake detectors is that each person has specific characteristics that a synthetic generator likely cannot reproduce.

Agarwal [130] proposed a biometric recognition method for distinguishing authentic faces from deepfake faces. They hypothesized that real Individual's specific facial movements and expressions during speech are absent in deepfake faces. To capture these differences, they employed the OpenFace2 [131] toolkit to extract head and face movements from a video. Facial features were derived from head rotation axes, facial action units, and 3D distances between specific mouth landmarks. Calculating the Pearson correlation between the head and facial features reduces each 10-second video clip to a 190-dimensional feature vector. SVM was utilised to predict this feature vector whether a video featured a real or fake person. However, it is important to note that the method's overall accuracy decreases significantly for detecting heavily compressed videos. Additionally, the approach is only applicable to videos that feature a frontal shot of the speaker. In another study, Agarwal et al. [132] proposed an approach for detecting deepfake face swap. They developed a Convolutional Neural Network (CNN) to identify inconsistencies in matching identities by embedding facial movements between video frames into a 256-dimensional space, capturing head poses, facial landmarks, and expressions. These features were used for spatiotemporal biometric behavior analysis. The CNN was trained to learn identity-specific mappings while VGG extracted appearance-based facial descriptors. Videos were classified as real or fake using cosine similarity based on matched identities and facial similarity. The approach requires minimal data to construct biometric reference sets and is robust toward compressed videos. However, the technique's applicability is limited to scenarios where matching biometrics for a given face is available.

Nguyen [133] developed a technique to combat deepfake attacks using a brow biometrics pipeline. Their research revealed that deepfakes often contain vulnerabilities in the eyebrow region. To test this theory, four deep-learning models were analyzed. After obtaining the feature vectors, a cosine distance metric was used to compare reference and probe eyebrows. After analyzing four CNN models [129], [134], [135], [136], ResNet and SqueezeNet models were found to be more accurate in detecting eyebrows. Despite its good detection performance, this method requires a large number of training samples due to its reliance on identity matching between the source and target. Haliassos et al. [137] proposed an approach focusing on mouth features. In their approach, they utilized pre-processed lip clip grayscale frames and trained them using two pre-trained lip reading networks: a Resnet-18 model and a multiscale temporal convolution network (MS-TCN). The goal was to fine-tune the recognition model to identify significant irregularities in mouth movement. Their approach performs well in cross-dataset analysis across five diverse datasets [6], [7], [35], [36], [61]. Furthermore, it exhibited robustness against various perturbations, such as changes in saturation, contrast, blockwise distortions, white Gaussian noise, blurring, pixelation, and video compression. However, their model could not detect fake faces with mouth occlusion and limited movement.

Liao et al. [138] developed the Facial Muscle Motion (FAMM) technique to detect compressed deepfake videos by examining facial muscle motion features from a geometric perspective. They hypothesised that the curvature of the facial skin during speech or micro-expressions causes displacement of facial feature points, resulting in unnatural muscle movements when synthesized. FAMM consists of three modules: facial landmark extraction, facial muscle motion feature construction, and prediction probability fusion. Precise landmarks are extracted through face detection, alignment, and landmark extraction. Unnatural facial motions are captured by calculating distance and angle features between adjacent frames, and time-series features are utilized to enhance unnatural muscle movements. Statistical measures such as absolute energy, absolute sum of first-order differences, time-series complexity, kurtosis, and coefficient of variation are calculated to reduce noise influence. Two classifiers, the Gate Recurrent Unit (GRU) and SVM, are trained using the difference in facial muscle motion and time-series features, and their results are combined using the Dempster-Shafer theory. The FAMM approach demonstrates robust performance for compressed and socially shared videos. However, the approach does not generalize well on unseen face manipulations.

Li et al. [21] and Jung et al. [139] proposed a deepfake detection method using eye blinks as a biological artifact. They observed that faces in deepfakes blink less often than in real videos. To determine the authenticity of a video, Li et al. [21] decomposed the video into frames and extracted eye regions based on six eye landmarks. These cropped eye regions were then processed using long-term recurrent convolutional networks (LRCN) [140] to capture temporal patterns of eye blinking. The method was evaluated on a dataset consisting of 49 interviews, presentation videos, and deepfake videos collected from the internet. However, the proposed approach did not investigate its effectiveness in detecting deepfake videos using dynamic blinking patterns. On the other hand, Jung et al. [139] introduced the DeepVision algorithm, which integrates Fast-HyperFace [146] and Eye Aspect Ratio (EAR) [147] techniques to detect and track eyes in consecutive frames. The video's authenticity was verified by estimating the number and period of blinks. The proposed technique was evaluated using a dataset incorporating a wide range of blink pattern variations, considering factors such as cognitive activity, age, gender, and time of day. However, it is important to note that blinking is commonly associated with mental illness and dopamine activation. As a result, this approach is not applicable to individuals with mental illness or abnormal neural pathways, limiting the generalizability of the proposed method.

Heart rate estimation from visual content is another biological artifact for facial video forensics. Fernandes et al. [141] introduced a method for heart rate estimation from deepfake videos in facial video forensics. They employed the Neural Ordinary Differential Equations (Neural-ODE) model to predict heart rates. They extracted features from input videos using three techniques: analyzing changes in facial skin color caused by blood flow, measuring average optical intensity in the forehead area, and magnifying and processing temporal changes in facial color using Euler's method. However, the proposed method is computationally intensive, and its robustness and generalizability need further investigation. It is essential to note that heart rate reflects hemoglobin content, which changes the skin's reflectivity over time. This natural phenomenon is often disrupted or absent in fake videos, which has prompted researchers [142], [143], [144], [145] to explore further in this area.

Recently, remote photoplethysmography (rPPG) has emerged as a promising method for detecting and analyzing changes in light absorption in facial skin tissue to identify deepfakes. Ciftci et al. [142] proposed a technique called FakeCatcher, which employs photoplethysmography (PPG) maps to capture physiological changes associated with deepfakes. By extracting rPPG signals from the face and constructing spatiotemporal representations of signal variations, PPG maps were created and used to train a CNN classifier. The approach is designed to be independent of generative models, resolutions, compressions, content, and context. The performance of the technique was evaluated on 140 online videos. The results showed good overall performance for small video segments, but its performance decreased with longer segments due to accumulated noise in biological signals.

Qi et al. [143] developed DeepRhythm, a method for deepfake detection by analyzing heartbeat rhythms in facial skin color resulting from blood flow. The approach employs dual spatiotemporal attention to adapt to changing faces and fake types. It also includes motion-magnified spatiotemporal representation (MMSTR) to capture sequential signals in facial videos, providing specific features for deepfake detection. However, it is robust against degradation methods such as JPEG compression, Gaussian blur, Gaussian noise, and temporal sampling. it still lacks satisfactory generalization performance with unseen datasets. Another study DeepfakeON-Phys [144], was designed to estimate heart rate from facial video sequences. This deep learning-based model

TABLE 6. Summary of artifacts-based deepfake detection techniques with highlighted top-performing classifiers and datasets.

Reference	Method	Classifier	Dataset	Best Performance	Deepfake Detected		Capability		
					Face-Swap	Expression-Swap	G	R	I
[125]	Warp Features	ResNet-50	**DF-TIMIT (LQ,HQ)**, UADFV	$AUC = 0.99, 0.93$	✓				✓
[21]	Blinking Pattern	LRCN	UADFV	$AUC = 0.99$	✓				
[43]	Texture Features	MLP, LogReg	Self-Built, FF++(F2F)	$AUC = 0.85, 0.86$	✓	✓			✓
[40]	3D Head Poses	SVM	**UADFV**, MediFor	$AUC = 0.89$	✓				✓
[130]	Facial Action Units	SVM	**Self-Built**, FF++	$AUC = 0.96$	✓	✓			✓
[122]	DFT Magnitude Averaging	**SVM**, K-Means, Logistic Regression	DFD	$ACC = 0.96$	✓				
[132]	Facial and Behavioral Features	ResNet101 and VGG	**Celeb-DF**, **FF++**, DFD, DFDC-P	$AUC = 0.99$	✓		✓		
[141]	Skin Color and Optical Intensity	Neural-ODE	**Self-Built**, DF-TIMIT	$Loss = 0.0215$	✓				
[139]	Blinking Pattern	EAR and HyperFace	Self-Built	$ACC = 0.87$	✓				✓
[142]	PPG Signals of Facial Regions	SVM/CNN	**UADFV**, FF++, Celeb-DF	$ACC = 0.97$	✓	✓	✓		
[143]	Skin Color	Attention CNN	**FF++**, DFDC-P	$ACC = 0.98$	✓	✓		✓	
[133]	Eyebrow Biometric	LightCNN, **ResNet**, DenseNet, **SqueezeNet**	Celeb-DF	EER=20.7% $AUC = 0.88$	✓				
[144]	Spatial and Temporal Information	Convolution Attention Network	**Celeb-DF**, DFDC-P	$AUC = 0.99$	✓		✓		
[123]	2D-GDCT	FCNN	**FF++(DF-Raw)**, FF++(Raw), FF++(C20), FF++(C40), Celeb-DF	$ACC = 0.79$	✓	✓			
[124]	DCT	XceptionNet	**FF++** Celeb-DF, DFDC	$ACC = 0.96$	✓	✓	✓		
[119]	GLCM	SVM	**DF-TIMIT (LQ,HQ)**, FF++(Raw), FF++(C20), FF++(C40), CelebDF, DFDC-P	$ACC = 0.92, 0.94$	✓			✓	
[126]	Face and Context	XceptionNet	**FF++** Celeb-DF, FSGAN, FSh	$AUC = 0.75$	✓	✓		✓	✓
[117]	Symmetrical Face Patches	DRN	**FF++(DF)**, DF-TIMIT, DFD, DFDC, Celeb-DF	$AUC = 0.99$	✓			✓	
[145]	Multiscale PPG Maps	EfficientNetV2	FF++ (C23)	$ACC = 0.90$	✓	✓			✓
[137]	Lip Clip Frames	Resnet-18 + MS-TCN	**DF1.0**, Celeb-DF, FaceShifter, FF++, DFDC	$AUC = 0.97$	✓	✓	✓	✓	

TABLE 6. ***(Continued.)*** **Summary of artifacts-based deepfake detection techniques with highlighted top-performing classifiers and datasets.**

Reference	**Method**	**Classifier**	**Dataset**	**Best Performance**	**Deepfake Detected**		**Capability**		
					Face-Swap	Expression-Swap	G	R	I
[127]	Face and Context	ResNet-18	**FF++(DF)**, **FF++(F2F)**, FF++(DF C23,C40), FF++(F2F C23,C40)	$F1 = 0.99$	✓	✓		✓	
[120]	LBP	Self-Designed Network	**FF++**, DF1.0, FOMM, Celeb-DF, DFDC-P, Mobile-Application	$ACC = 0.99$	✓	✓	✓		✓
[138]	Facial Muscle Motion	GRU+SVM	Social-website(DF, **FS**, NT, F2F, Fsh), FF++(C23), FF++(C40)	$AUC = 0.988$	✓	✓	✓	✓	

captures spatio-temporal information by analyzing color changes in faces caused by variations in oxygen concentration in the blood. It utilizes signal-processing techniques to isolate blood-related color changes from other factors like illumination and noise. The Convolutional Attention Network (CAN) has two branches: the Motion Model, which detects changes between consecutive frames to identify fakes, and the Appearance Model, which focuses on static information in frames to guide the Motion Model with relevant information. Attention masks from the Appearance Model are shared with the Motion Model at different layers, and the final output of the Motion Model serves as the output of the entire CAN. While this approach is robust to external illumination perturbations, it requires high-quality video sequences with visible faces and good lighting conditions for training. Moreover, it is ineffective against unseen datasets and computationally complex, making it unsuitable for real-time applications.

Wu et al. [145] presented multi-scale spatial-temporal photoplethysmography (PPG) maps to create an interpretable deepfake detector. Their approach assumes that different video manipulation techniques affect distinct facial regions, reflected in the PPG map of multi-scale facial regions. The method involves a two-stage network, which includes a mask-guided local attention module (MLA) that focuses on modified regions in the PPG map and a temporal transformer for capturing long-distance information between adjacent clips. The process generates a multi-scale PPG map from facial video frames and performs two-level network detection. The MLA identifies the corresponding positions in the PPG feature map for the modified facial areas. However, the method has limitations in accurately detecting deepfakes within compressed videos.

Table 6 presents the summary of Artifacts based deepfake detection algorithms. The table displays details about the author, classifier used, features employed, evaluation datasets, performance metrics, types of deepfake manipulations detected (Face-Swap and Expression-Swap), and the detection capabilities denoted by G (generalization for unseen datasets), R (robustness against perturbation attacks), and I (interpretability).

B. PIXEL AND STATISTICAL FEATURE APPROACH

In this approach, an image's pixel values are utilized as primary features, incorporating its intensity or color as input values. Statistical features capture the more complex features and relationships of an image. Table 7 provides an overview of pixel and statistical features based deepfake detection techniques. Koopman et al. [148] proposed an approach for deepfake detection based on photo response non-uniformity (PRNU) analysis. PRNU is a unique pattern noise in the camera sensor caused by manufacturing defects in silicon wafers and pixel sensitivity variations due to the physical properties of the silicon wafers. The technique suggests that examining the PRNU pattern in the facial area of video frames makes it possible to identify deepfakes where the swapped face alters the PRNU pattern. To implement their approach, videos were converted into frames and cropped to focus on the facial region. These cropped images were then divided into eight groups, and an average PRNU pattern was computed for each group. Subsequently, normalized cross-correlation values were calculated to compare the PRNU patterns among these groups. A test dataset consisting of 10 original videos and 16 deepfake face swaps using DeepFaceLab [15] was generated to evaluate their approach. The analysis shows a statistically significant difference in the mean normalized cross-correlation values between deepfakes and real fakes. These findings highlight the potential of PRNU analysis as a promising method for deepfake detection, but its effectiveness against high-quality deepfakes

TABLE 7. Summary of pixel and statistic-based deepfake detection algorithms. highly performed classifier and dataset are shown in bold.

Reference	Features	Classifier	Dataset	Best Performance	Manipulation Type Detected		Capability		
					Face-Swap	Expression-Swap	G	R	I
[148]	PRNU	-	Self-Built	Original images have higher correlation scores than deepfakes.	✓				✓
[150]	Pixel Intensity Feature Descriptor	SVM	UADFV	$ACC = 0.94$	✓				
[149]	Statistical Properties	SVM	DF-TIMIT, **FF++**, DFD	$ACC = 0.80$	✓	✓			
[151]	Statistical Properties	XceptionNET	**FF++**, Celeb-DF, DFD, DFDC, DF1.0	$AUC = 0.99$	✓	✓	✓		
[152]	Pixel Intensity	XGBoost	**UADFV**, FF++ (DF), Celeb-DF	$AUC = 1.0$	✓			✓	
[153]	Pixel Intensity	MTD-Net	**DFDC**, FF+(C23), FF++(C40), Celeb-DF, DF-1.0	$AUC = 0.99$	✓	✓	✓	✓	
[154]	Pixel Intensity Feature Descriptor	Random Forest	**DF-TIMIT(LQ)**, UADFV, FF++(DF), DFD, Celeb-DF, DFDC	$AUC = 0.99$	✓			✓	✓
[155]	Color Statistical Features	SVM	**FF++**, Celeb-DF	$AUC = 0.99$	✓				✓

is unclear. Another approach investigates the manipulation of PRNU-based content by analyzing the statistical properties of PRNUs [149]. This approach investigates the image's spectral and spatial characteristics, such as kurtosis, energy, variance range, and skewness. These features are then fed into an SVM classifier to classify frames as real or fake. Although the classifier achieved high accuracy on test datasets, it lags behind deep learning in performance. Video compression format differences can also affect classification accuracy.

Kharbat et al. [150] employed traditional edge feature detectors to extract feature points from images for deepfake video detection. These feature points were collected across video frames and used as training data for an SVM model. The underlying hypothesis was that feature points from real videos would exhibit more correlated characteristics than deepfake manipulated face videos. The results showed that Histogram of Gradient (HoG) features were more discriminative than other descriptors. However, the method has not been evaluated on high-quality compressed videos.

Chen et al. [152] introduced DefakeHop, a lightweight detection method that leverages principal component analysis (PCA) to identify deepfakes. The technique involves extracting features from different face patches using PixelHop++, a feature extraction technique. To reduce the spatial dimension of each patch, subspace approximation with an adjusted bias (saab) is used. The resulting feature representations were then fed into an extreme gradient boosting (XGBoost) classifier for binary classification. Despite its small size, DefakeHop has shown good performance across various deepfake datasets and can handle videos with different compression qualities. Yang et al. [153] presented a Multi-scale Texture Difference Network (MTD-Net) designed for deepfake detection. The model was specifically developed to extract and integrate multi-scale texture difference information, enabling robust detection of deepfake images with high compression and mixed distortion. Initially, texture difference features were extracted from cropped faces, utilizing pixel intensity and gradient information. A novel convolution operation called Central Difference Convolution (CDC) was introduced to combine intensity and gradient information to represent texture differences. Finally, the extracted textural difference features were fused at multiple scales for classification. The approach demonstrated promising performance on high-quality datasets, such as DeeperForensics-1.0, Celeb-DF, and DFDC. However, one limitation of the approach is the lack of interpretability in its classification.

Xia et al. [155] introduced an interpretable deepfake detection method. Their method looks for the differences in facial statistics between real and fake video frames in various color channels. They converted the input RGB frame to HSV and YCbCr color spaces to process the face images and used a texture map. By extracting texture difference features using first-order differential operators, the approach was able to identify disparities in the textures of real and

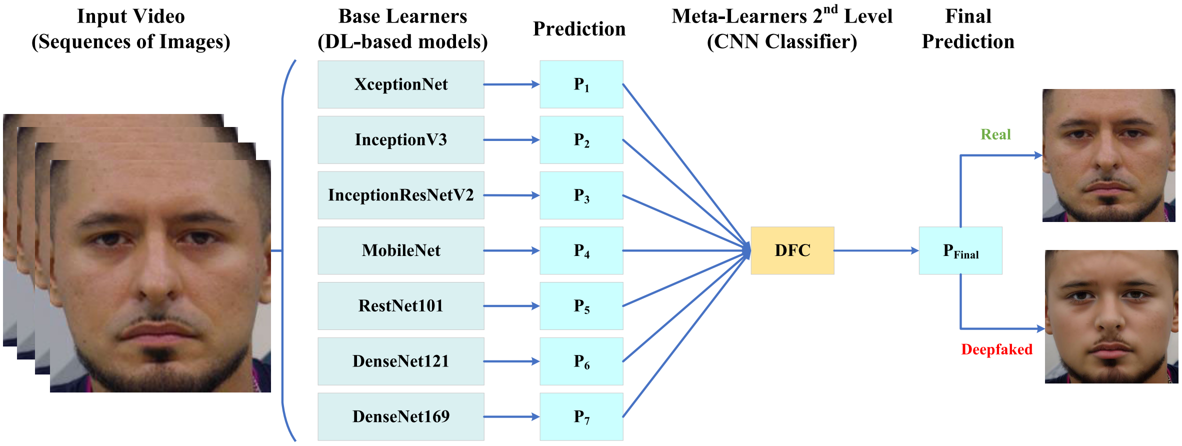

FIGURE 8. Example of multistack deepfake detection [156].

fake frames. Truncation and co-occurrence matrices were used to capture texture information while reducing data size. The combined features from different color channels were interpretable. Finally, they used SVM to classify the features. Their method has lower computational costs than deep learning-based detection, but detection performance decreases with increased compression ratios due to the loss of texture information.

According to Luo et al. [151], image noise can expose evidence of forgery by eliminating color texture. To achieve this, they developed a method using an Xception-based detector that includes high-frequency noise features. The proposed model consists of three functional modules: i. A multi-scale high-frequency feature extraction module, ii. A residual-driven spatial attention module, and iii. A cross-modality attention module. They adopted the proposed modules to extract more meaningful features and capture the correlation and interaction between the complementary modalities. The approach performs well in cross-dataset evaluation.

Wang et al. [154] proposed a computationally efficient Fused Facial Region Feature Descriptor (FFR-FD) technique to detect deepfake face swap. The method leverages the fact that deepfake face swaps display fewer feature points than real faces. By using feature points detector descriptors like SURF, SIFT, ORB, FAST&BRIEF, and A-KAZE, feature points were extracted from eight facial regions, including the entire face, mouth, inner mouth, right eyebrow, left eyebrow, right eye, left eye, and nose. These descriptors created region-specific vectors without averaging, allowing for precise information control. The vector's dimensions were reduced and connected sequentially to form FFR-FD. However, it is important to note that FFR-FD relies on visible feature points in facial images, which may be missing due to occlusions, low resolution, or other factors. In addition, the technique is less effective at detecting deepfake videos with complex backgrounds or challenging lighting conditions.

C. GENERIC NEURAL NETWORK APPROACH

The Generic Neural Network (NN) approach uses one or more neural networks to extract features and classify data. Unlike other deep learning methods that rely on manually created features, the Generic NN uses only learned features for its detections. Table 8 provides an overview of Generic Neural Network deepfake detection techniques.

Afchar [157] introduced the MesoNet CNN model for detecting deepfakes by analyzing the intrinsic features of images. To test their approach, videos were collected from the internet. However, this approach does not generalize to a high-quality Celeb-DF [6] dataset. Nguyen utilized a capsule network to detect deepfakes [158]. The network consists of three primary capsules and two output capsules for classifying real and fake images. The model extracted the feature vector through the VGG19 backbone network [128] and distributed it to the three primary capsules. The results of all three primary capsules are dynamically directed to the output classification capsules. The proposed network showed promising results on the FF++ [7] and DFD [93] datasets. However, the capsule network had lower detection rates with previously unseen deepfake videos. In [159], a self-supervised decoupling network (SDNN) for learning authenticity and compression features is proposed. As self-supervised signals, it used the compression ratio of given input images. The goal is to normalize the model with different compression rates so that the authenticity classifier can perform better classification without being influenced by the input compression. However, since the range of compression rates can be adjusted, the model's performance with an invisible compression rate can still be an issue.

Zhao et al. [160] introduced a method for detecting deepfakes that exploits the self-consistency of local source features, which are specific pieces of information within an image that remain the same regardless of the content. The hypothesis is that a manipulated image will have different source characteristics in varying locations, while an original image will have consistent source characteristics throughout.

To extract these features, a convolutional neural network (CNN) was used to create downsampled feature maps, where each vector represents the source features of a specific location in the input image. The model was trained using pair-wise self-consistency learning (PCL), which calculates the cosine similarity between all pairs of feature vectors and applies consistency loss based on whether the locations belong to the same source image. Pairs from the same source image with low similarity scores and pairs from different source images with high similarity scores are penalized. The learned feature maps were fed into a non-linear binary classifier for deepfake detection. The proposed approach was evaluated on various datasets [6], [34], [36], demonstrating its generalizability. However, the approach is ineffective on fake images that maintain consistent source characteristics throughout the entire image.

Instead of learning spatial features, Qian et al. [161] introduced a novel framework called F^3-Net, that incorporates decomposed image components with higher frequencies and local frequency statistics extracted from densely sampled spatial patches rather than learning spatial features. The fusion of these branches is achieved through a cross-attention module known as MixBlock. The F^3-Net model displays resilience in identifying deeply compressed deepfake videos. Nevertheless, the proposed technique does not generalize well across unseen deepfake datasets. Li et al. [162] presented the Frequency-Aware Discriminative Feature Learning framework (FDFL) as a solution to the issues of uncertain feature differentiation with softmax loss and the inefficiency of manually crafted features in detecting forgeries. The authors introduced a Single-Center Loss (SCL) to align neutral face features and repel manipulated ones. Combining SCL with softmax loss yielded improved results within the FDFL framework. However, it should be noted that the model showed poor generalization for expression swap datasets.

One of the biggest obstacles in training supervised classifiers to detect deepfakes is keeping them up-to-date with the latest forgery techniques. Even if they perform well on specific manipulation methods, they may struggle with new ones. However, transfer learning offers a solution by allowing knowledge gained from one task to improve the performance of related ones. Cozzolino et al. [163] developed a CNN-based method to generalize different yet related manipulations, even without specific training for each manipulation. The approach utilized an autoencoder to learn a forensic embedding capable of transferring between different manipulation domains. The network was trained to differentiate between real and fake images by activating specific regions of the latent space. By utilizing these activations, the method effectively determines the authenticity of input images. Although the proposed approach shows generalizability and works well with limited training examples for new manipulations, it has not been evaluated with compressed deepfakes to demonstrate its effectiveness for robustness. Chen et al. [164] proposed a novel two-stage deepfake detection method called FeatureTransfer that leverages transfer learning. This approach uses a CNN model, pre-trained on a dataset of deepfake images, to obtain transferrable feature vectors in both the source and target domains. These feature vectors are then fed to a domain-adversarial neural network based on back-propagation for domain adaptation (BP-DANN) training. The proposed approach exhibits improved and comparable performance compared to previous methods for cross-domain deepfake detection. However, the method is not an end-to-end detection solution and demands a large-scale deepfake dataset for pre-training the CNN, which can be time-consuming. In the field of deepfake detection, researchers have introduced multi-stream, multi-stack neural networks. Figure 8 and 9 provide a comprehensive view of these models. Multi-stream methodologies primarily rely on the fusion of multi-level or multi-domain features. Zhou et al. [165] proposed a method that combines an InceptionNet-based face classification stream with a triplet stream network to extract steganalysis features. The proposed approach enhanced detection by incorporating low-level noise residual features and high-level tampering inconsistencies. The final detection score is computed by aggregating the output scores of both streams. Nonetheless, it is worth noting that the method does not perform well on high-quality deepfake videos [6]. Another multi-stream network was presented by Kumar et al. [166]. The network consisted of five ResNet-18 models, each dedicated to learning specific facial regions and capturing local details. By combining the learning of these regional areas with the full-face, the network improved its detection performance in handling compressed input. However, the method is computationally intensive and only evaluated on expression swap deepfakes from the FF++ dataset [7] without testing its generalizability on other datasets.

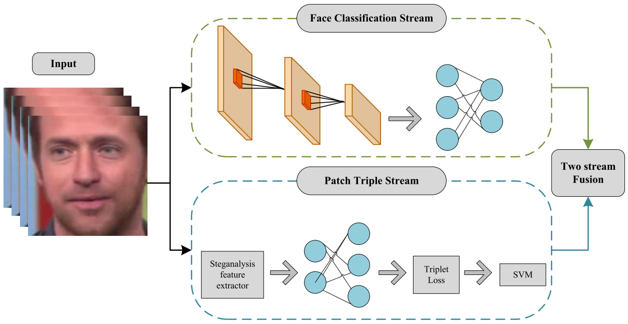

FIGURE 9. Example of multistream deepfake detection [156].

The deepfake multi-stack framework, proposed by Rana et al. [156], is an ensemble of deep learning networks for detecting manipulated videos. The model incorporates several state-of-the-art detection methods into a single classification model. It was divided into two parts: i. Base- Learners Creation, and ii. Stack Generalization. Seven deep learning models (XceptionNet, MobileNet, ResNet101, InceptionV3, DensNet121, InceptionReseNetV2, and DenseNet169) were

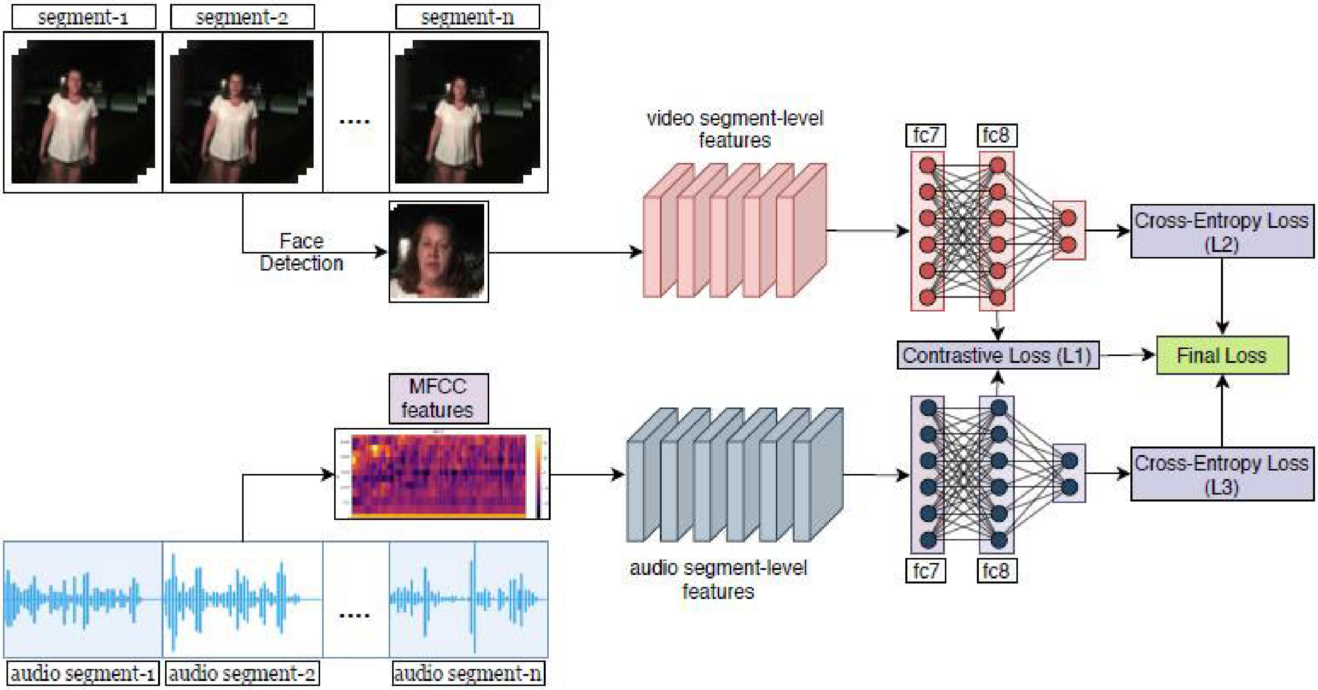

FIGURE 10. Multi-modal deepfake detection approach [168].

used with ImageNet weights to create base learners. These models were linked by replacing the top layer with two output layers and the softmax activation function. Greedy Layer-wise Pretraining (GLP) algorithms were used for model training. Stack Generalization is a CNN model called Deepfake Classifier (DFC) that learns from the predictions of the base learner using sample data on which the models were not trained. Stack Generalization combined with a larger multi-head neural network to determine the best detection result based on each base learner's predictions. While the model produced promising results, its generalization and robustness capacity is unknown. Additionally, the model size used in the proposed approach is quite large, which may result in overfitting. Another ensemble approach proposed by Bonettini et al. [167] captures high-level semantic information and improves prediction performance using different CNN models, including a modified version of EfficientNetB4. Considering hardware and time constraints, the solution is designed to be computationally efficient. Moreover, an attention mechanism has been introduced to identify the most informative parts of the input video frame for classification, thereby improving the ensembling process. Siamese training strategies have been explored with a triplet margin loss to extract additional data information and improve generalization capabilities. This training method aims to learn a feature descriptor that emphasizes the similarity between samples of the same class, prioritizing descriptive features for each class and ignoring less discriminatory features. In contrast, end-to-end training extracts features without determining their significance. The proposed solution can analyze thousands of videos in a limited time and requires less than one gigabyte of storage space. Recent research has explored attention mechanisms with CNNs for detecting deepfakes. Zhao et al. [169] proposed a multi-attention deepfake detection network that utilized EfficientNet-b4 as a backbone network to merge low-level texture features and high-level semantic features. They hypothesized that subtle low-level texture differences disappeared in the deeper layer and showed that enhancing texture features from shallow layers helps stimulate learning of discriminatory features in the forged region. They employed Bilinear Attention Pooling (BAP), regional independence loss, and an attention-guided data augmentation mechanism to regulate attention maps, capture semantic regions, and non-overlapping discriminatory feature information. The approach is not evaluated for robustness and does not generalize well on unseen datasets [6]. In another study, Ganguly et al. [170] proposed a deep learning model enhanced with a visual attention technique for distinguishing fake images from real ones. Their model incorporated a lightweight soft-attention mechanism built on top of the Xception network, focusing on identifying inconsistencies in deepfake manipulations. However, the model encountered challenges in accurately classifying expression swap face images and real images with common features, such as closed eyes. In [171], the authors introduced a Convolutional Vision-Transformer (CVT) for deepfake detection, demonstrating significant performance. However, this model still needs improvement, and further research is required to enhance its diversity, accuracy, and robustness as a solution. Zi et al. [98]

TABLE 8. Summary of generic neural network deepfake detection approach. In cases of multiple networks and datasets, highly performed models and datasets are highlighted in bold.

Reference	Model	Dataset	Best Performance	Manipulation Type Detected		Capability		
				Face-Swap	Expression-Swap	G	R	I
[165]	GoogLeNet	Self-Built	$AUC = 0.99$	✓		✓		
[157]	MesoInception-4	Self-Built	$AUC = 0.91$	✓	✓			
[158]	Capsule Network (VGG19)	FF++/DFD	$AUC = 0.96$	✓	✓			
[163]	Auto-Encoder	**FF++(F2F)**, FF++(FS)	$ACC = 0.94$	✓	✓	✓		
[166]	ResNet-18	**FF++(F2F(Raw))**, FF++(F2F(C23)), FF++(F2F(C40))	$ACC = 0.99$		✓		✓	
[162]	XceptionNet	**FF++(Raw)**, **FF++(C23)**, FF++(C40), CelebDF, DFDC	$AUC = 0.99$	✓	✓			
[161]	F^3-Net	**FF++(C23)**, FF++(C40)	$ACC = 0.98$	✓	✓		✓	
[171]	CViT	**FF++**, DFDC, UADFV	$ACC = 0.93$	✓	✓			
[174]	OpenFace + Siamese network + MFCC	**DF-TIMIT(LQ,HQ)**, DFDC	$AUC = 0.96, 0.94$	✓	✓			✓
[168]	3D-ResNet+MFCC	**DF-TIMIT(LQ,HQ)**, DFDC-P	$AUC = 0.97, 0.96$	✓				✓
[156]	XceptionNet, InceptionV3, InceptionResNetV2, MobileNet, ResNet, DenseNet	Self-Built	$ACC = 0.99$	✓	✓			
[98]	**ADDNet-2D** ADDNet-3D	**FF++(DF(C23))**, FF++(DF(C40)), DFD, WDF, DF-TIMIT(LQ, HQ),	$ACC = 0.99$	✓				
[167]	EfficientNet, XceptionNet	**FF++**, DFDC	$AUC = 0.94$	✓	✓	✓		
[164]	BP-DANN	**DFDC-P**, DF-TIMIT, FF++, DFD, Celeb-DF	$AUC = 0.98$	✓		✓		
[169]	EfficientNet-B4	**FF++(C23)**, FF++(C40), DFDC, Celeb-DF	$ACC = 0.97$	✓	✓			
[159]	EfficientNet-B2	**FF++(Raw)**, FF++(C23), FF++(C40)	$ACC = 0.99$	✓	✓		✓	
[160]	ResNet-34	**FF++**, DFDC-P, DFD, Celeb-DF, DF1.0	$AUC = 0.99$	✓	✓	✓		
[175]	R(2+1)D-18	**FF++**, DFDC	$AUC = 0.99$	✓	✓	✓		✓
[172]	Efficient-B0 and XceptionNet	FF++	$ACC = 0.83$	✓	✓			✓
[176]	ResNet-50 and Inception-v3	**DF-TIMIT(LQ,HQ)**, FF++(C23), FF++(C40), Celeb-DF	$AUC = 0.98, 0.98$	✓	✓			✓
[170]	XceptionNet	**FF++**, Celeb-DF	$ACC = 0.70$	✓	✓			
[173]	XceptionNet	**FF++**, Celeb-DF, DFDC	$AUC = 0.99$	✓	✓	✓		
[177]	Xception, ResNet18, DenseNet121	**FF++**, Celeb-DF, DFDC-P	$AUC = 0.99$	✓	✓	✓		✓

suggested ADDNets, to identify deepfakes. There are 2D and 3D variants of the model developed to use attention masks on pristine and fake faces. Compared to the state of the art, their 2D version of ADDNet (ADDNet-2D) performed better than the 3D model. However, this method searched for inconsistencies throughout the image but found no correlation between different parts of the face.

Contrastive learning techniques have gained attention in the field of deepfake detection. Contrastive learning aims to acquire features that can effectively differentiate between similar (positive) and dissimilar (negative) data points. Xu et al. [172] presented a supervised contrastive (SupCon) learning to capture the contrast between manipulated and non-manipulated images. This method involves training an encoder network with augmented data to generate normalized embeddings. A projection network then uses these embeddings to compute the supervised contrastive loss. A linear classifier is then trained on the learned representations using cross-entropy loss. Heatmaps and Uniform Manifold Approximation and Projection (UMAP) are utilized to provide an interpretable analysis. However, experimental results have shown that this approach does not generalize well on unseen manipulations. Another study by Dong et al. [173] proposed a framework that combines intra-domain and cross-domain information to improve the generalization of deepfake detection. Their approach includes two sub-networks, each processing a different view of the same input image. It integrates an SRM (Steganalysis rich model) into its data augmentation strategy to introduce frequency-aware features into the RGB feature space. The encoder network incorporates the Multi-Scale Feature Enhancement engine to detect inconsistencies in shallow feature maps and establish inter-region relationships. The model is trained using a combination of cross-entropy loss and consistency loss, minimizing invariance between different image views and supporting supervised classification. While the approach shows good generalization ability for high-quality datasets [6], [35], it performs poorly with compressed video.

In deepfake detection, multi-modal approaches, as depicted in Figure 10, demonstrate the potential of integrating different data modalities, such as video segments and their corresponding voice modalities, to enhance the robustness of inferences. Mittal et al. [174] presented a novel approach that simultaneously uses the audio (speech), and video (face) modalities and the perceived emotion features extracted from both modalities to detect any change in a video. The approach adopts a Siamese network architecture, where a real video and its deepfake counterpart were fed to obtain embedding vectors representing the modalities and perceived emotions. These vectors are then used to compute a triplet loss function that aims to minimize the similarity between the fake video modalities while maximizing the similarity between the real video modalities. The results provide interpretable insights, such as the discrepancy in perceived emotion labels between the manipulated facial modality of the fake video and the neutral speech modality of the real video. However, this technique is unable to classify cases where both speech and face modalities are manipulated. Chugh et al. [168] proposed another multi-model framework based on audio-visual dissonance for deepfake detection. Their approach involves training a bi-stream network consisting of visual and audio streams. Unlike the Mittal approach, which relies on real, fake video pairs for training, this approach uses a more traditional training protocol that is not limited to such pairs. The visual stream uses a 3D-ResNet architecture to extract features, while the audio stream uses mel-frequency cepstral coefficients (MFCC). The network is trained using a combination of contrastive loss and cross-entropy loss, ensuring that the audio and visual streams are closer to real videos and farther from fake videos. The cross-entropy loss learns discriminative features for each modality. The modality dissonance score threshold is calculated during test inference by summing the dissimilarity scores between audio and visual segments to indicate authenticity or forgery. Experimental results demonstrate that the proposed method performs well on datasets such as DFDC-P and DF-TIMIT. However, it should be noted that the approach's generalization ability on unseen datasets has not been established. Zhou et al. [175] proposed a two-plus-one stream model to distinguish video and audio deepfakes together. Unlike multimodal frameworks [168], [174] that merge inputs from different modalities, this technique models the video and audio streams separately with their labels. The proposed approach includes multiple centralized connections between the video and audio streams. This synchronization stream records the synchronization patterns between modalities. Additionally, intra and inter-attention mechanisms enhance temporal alignment between audio and video presentations. The synchronization stream is trained alongside the video and audio streams, and the final prediction is based on the output of the synchronization stream. To evaluate their approach, a new dataset was created with manipulated audio by synthesizing speech from existing video deepfake datasets. The proposed approach demonstrates the ability to generalize to previously unseen deepfake videos.

Yu et al. [176] introduced Facial Patch Mapping (FPM) as an alternative approach to training CNN instead of utilizing the entire face. FPM involves extracting smaller patches or regions from the face, which are then employed for training the CNN model. This technique utilizes a mapping engine to assign the patches to different backbone networks, reducing redundant convolution operations. The FPM method applies BM pooling module with bilinear interpolation to maintain consistent feature map sizes and minimize quantization errors. The approach trains five patch-based detectors, each focusing on specific local patches, and integrates their predictions using a local voting scheme to improve overall accuracy. While this part-based training framework provides interpretability, it struggles to generalize well on visually challenging datasets [6]. Hua et al. [177] presented an interpretable face forgery detection model by establishing patch-channel correspondence. The model includes an encoder, a feature-rearranging layer, and a binary classifier. The encoder processes a facial image and generates a range of channels, each containing information about a specific patch on the face. The feature-rearranging layer enhances interpretability by decorrelating the channels and aligning them with the corresponding patches. The model offers evidence of forgery that links to specific patches and channels, making the detection process more efficient. However, the presence of strong channel correlation and computational complexity poses limitations in terms of quantifying interpretability and optimizing patch-channel correspondence.

D. ANOMALY-BASED APPROACH

The majority of existing methods for deepfake detection focus on binary classification, assuming a clear distinction between real and fake samples. However, this binary approach requires a large number of data representing fake and authentic classes. The challenge becomes much more difficult with each new deepfake approach, as only a limited number of examples for new manipulations are accessible for training. Some techniques have approached the deepfake detection problem as a single classification task, treating real images as normal and deepfakes as anomalies. The models are solely trained on authentic images and consider fake images or videos as anomalies during evaluation. Table 9 summarizes deepfake detection methods that rely on identifying anomalous patterns.

The FakeSpotter research [178] follows an anomaly-based technique to evaluate a facial recognition network's neural activity (coverage). They employed hierarchical neuron behavior and found that their approach was highly resistant to four basic perturbation attacks: Noise, Compression, Resizing, and Blur. However, the approach is unable to generalize well on DFDC dataset [35]. Ortiz et al. [179] also followed this one-class learning paradigm to train VGGFace2 and ResNet50 models, combined with attribution-based confidence (ABC) metric to distinguish deepfake faces from real ones. This technique was not evaluated for high-quality deepfake videos [6], [35].

Khodabakhsh et al. [180] developed a technique to identify synthetic facial features. Their method involves splitting the real image and using a PixelCNN++ model to calculate the probability distribution of the pixel intensities. By considering the relationship between previous pixels, they create a probability matrix for the image that highlights the logarithmic probability of observing each pixel's intensity. This additional information provides details about the location and strength of the anomaly. A universal background model (UBM) is trained with the PixelCNN++ learned features, and a simple classifier is trained with the UBM output. The results show that the log-likelihood images are efficient in detecting synthetic facial features while reducing complexity. The model is tested on both seen and unseen manipulated data to evaluate its overall performance. However, the approach is not evaluated on challenging visual datasets [6], [35]. Another interpretable deepfake detection approach proposed by Wang [181] models the common patterns of local motion features from real videos and detects anomalies in fake videos by comparing the extracted motion patterns with the real ones. They hypothesize that co-motion patterns extracted from original videos follow the motion pattern of facial structures and are homogeneous regardless of video content variance but are less related in fake videos. To implement this technique, motion features were extracted from specific locations in the target video. These features were then categorized and converted into a correlation matrix to represent the motion patterns frame by frame. Each video's correlation matrices were grouped and weighted based on their grouping performance to create a co-motion pattern. This pattern represents the overall smoothness and correlation of video motion. The effectiveness of this approach was demonstrated on the FF++ dataset [7], which showed that it can handle high compression and random noise. However, the generalization capability of the approach to high-quality datasets has not been evaluated yet.

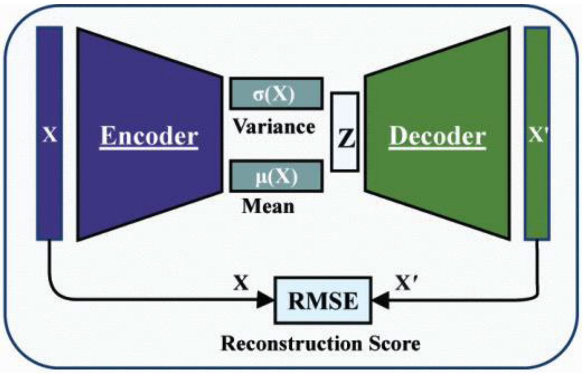

FIGURE 11. OC-Fakedect representation [182].

Khalid introduced the OC-FakeDect approach [182] for detecting DeepFake images. The approach uses a one-class Variational Auto-Encoder (VAE) to reconstruct authentic facial images. An anomaly score for the input image is calculated by comparing the mean component of the encoded image to the mean component of the reconstructed image using root mean square error (RMSE). Figure 11 illustrates the process. The strategy generalizes well to various deepfake datasets, although its robustness to perturbation attacks is unknown. Cao et al. [184] proposed the Reconstruction Classification Learning (RECCE) framework to capture and interpret the discrepancies between real and fake faces. This methodology combines reconstruction learning, multi-scale graph reasoning, and reconstruction-guided attention to learn compact representations of real faces. During training, a reconstruction network incorporating white noise is used to recreate real face images. A multi-scale graph reasoning (MGR) module is introduced in the framework, which combines latent features of encoder and decoder blocks in a bipartite graph. By considering the spatial correspondence and performing a multi-scale analysis, the MGR module improves the encoder's feature representations, thus enabling reasoning about forgery cues. The methodology uses the reconstruction difference to identify manipulated traces and employs a reconstruction-guided attention (RGA) module that focuses on probable forgery regions, computed through a difference mask. The evaluation results suggest that it is possible to distinguish fakes with unknown patterns by examining common features of real faces. Additionally, the approach demonstrates robustness against various perturbations, such as compression, blur, contrast, saturation, and pixelation.

Cozzolino [183] proposed ID-Reveal method to detect fake face videos by analyzing an individual's unique facial expressions during speech. ID-Reveal consists of three key components: a morphable 3D model to generate compressed representations of each frame, a temporal ID network to cre-

TABLE 9. Summary of anomaly-based deepfake detection approach. Highly performed classifiers and datasets are highlighted in bold.

Reference	Model	Dataset	Best Performance	Manipulation Type Detected		Capability		
				Face-Swap	Expression-Swap	G	R	I
[178]	VGG-Face3 + ResNet50	**FF++(DF)**, Celeb-DF, DFDC	$F1 = 0.98$	✓			✓	
[179]	ResNet50	DF-TIMIT	$Loss = 0.00768$	✓				
[180]	PixelCNN++	**FF++(NT)**, FF++(DF), FF++(FS), FF++(F2F)	$ACC = 0.98$	✓	✓	✓		✓
[181]	AdaBoost	**FF++(Raw)**, FF++(C23), FF++(C40), FF++(C23+noise)	$AUC = 0.98$	✓	✓		✓	✓
[182]	VAE	**DFD**, FF++	$F1 = 0.98$	✓		✓		
[183]	3DMM+ResNET	**DFD(C23,C40)**, DFDC-P, Celeb-DF	$AUC = 0.96, 0.90$	✓	✓		✓	
[184]	Encoder-Decoder	**FF++(C23)**, FF++(C40), Celeb-DF, WDF	$AUC = 0.99$	✓	✓		✓	✓
[185]	ResNet50	**DF-TIMIT**, DFDC-P, KoDF	$AUC = 0.99$	✓	✓		✓	

ate an embedded vector for facial motion and pose information, and a modified ResNet architecture with an adversarial training strategy to incorporate behavioral information. The network is trained only on real videos of different subjects. During testing, the approach requires a set of pristine videos of the target person in addition to the test video. Using these pristine examples, it calculates a distance metric to the test video using the embedding of the temporal ID network for anomaly detection. This method can detect facial reenactment even with strong video compression. However, it relies on the target person's corresponding real video to detect an anomaly in the fake video, making it unsuitable for real-world scenarios. Another anomaly-based technique proposed by Cozzolino et al. [185] focuses on exploiting the audio-visual features of the portrayed individual. The proposed technique trains its model on a large dataset of real videos consisting of over 5,000 identities with associated audio from the Voxceleb dataset [95]. The training uses a contrastive learning approach to extract different embeddings for moving faces and audio segments. In the training phase, contrast losses are used for individual modalities (audio or video) and a common contrast loss considering both modalities. The multi-way matching loss compares positive matches with all negative matches, enhancing learning stability. During the test, the method calculates POI (person of interest) similarity indices between the features of the target video and a set of reference videos of the same person of interest. These indices are normalized using the mean and standard deviation values calculated from the reference set to estimate the probability of a false alarm and to make decisions about the video's authenticity. The performance of the approach can be improved by using a diverse set of reference videos rather than increasing the quantity of data. The approach is robust to adversarial noise attacks. However, facial reenactment manipulations pose a greater challenge for the method than face-swapping, primarily due to their ability to better preserve the characteristics of the manipulated identity.

E. SPATIOTEMPORAL APPROACH

Techniques in this category focus on observing that consecutive video frames exhibit spatial and temporal consistency in pixel values. However, the deepfake generator performs frame-by-frame processing and introduces temporal discrepancies in the synthesized video. As a result, manipulated regions within a frame lack spatiotemporal consistency with neighboring frames. These inconsistencies include a sharp contrast and brightness change within small facial areas, inconsistent illumination choices, and unnatural mouth and eye movements across frames within the same video, as illustrated in Figure 12. Table 12 summarizes techniques for detecting deepfakes that involve spatiotemporal features.

Guera [39] proposed a recurrent algorithm that analyzes consistent motion patterns across adjacent frames. The algorithm utilizes a combination of convolutional neural networks (CNNs) to extract frame-specific features and long short-term memory (LSTM) networks to analyze image sequences. A fully connected network is then employed for classification using the generated sequence descriptor. The algorithm's effectiveness was evaluated on a self-built face swap video dataset comprising 600 videos and performed well even on short video sequences. However, due to the unavailability of a large dataset at the time of the research, the results were inconclusive regarding generalizability. Sabir et al. [186] developed a method to detect temporal

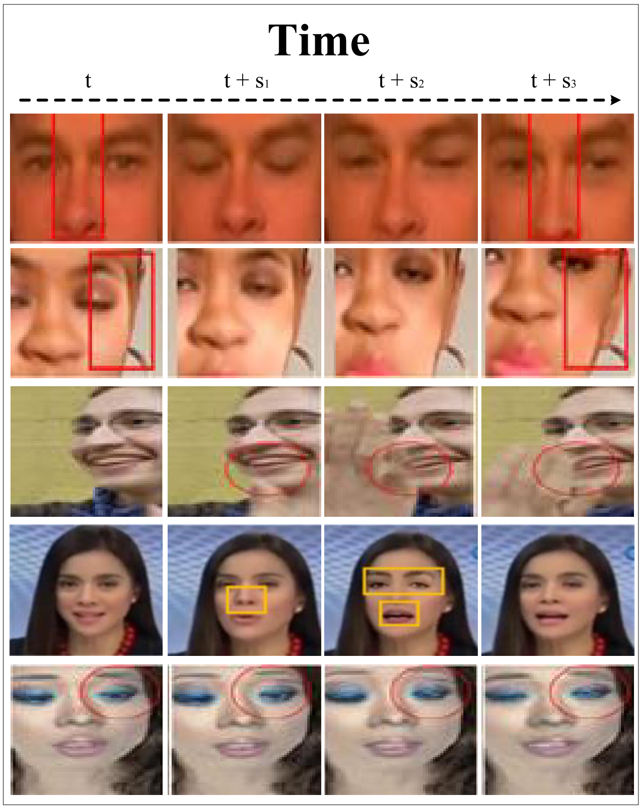

FIGURE 12. Examples of spatiotemporal inconsistencies in consecutive frames.

incoherence in deepfake videos by extracting facial features using a combination of CNN and recurrent neural network (RNN). Unlike Guera's approach [39], Sabir et al. trained their model end-to-end, combining CNN architectures such as ResNet [129] and DenseNet [136] with RNN networks. After examining different strategies for aligning and merging CNN features, they discovered that bidirectional-recurrent-DenseNet, aided by landmark-based face alignment, performed well on FF++ videos [7]. However, the method faced challenges with heavily compressed videos that disrupted frame continuity.

Using optical flow, Amerini et al. [187] proposed a technique to identify discrepancies between frames in deepfake videos. They utilized the features extracted from optical flow fields as input to a CNN model to classify deepfake and original videos. The evaluation using the FF++ dataset [7] demonstrated that this approach achieves comparable classification accuracy to state-of-the-art methods. However, it is important to note that the optical flow method has limitations in cases where assumptions of brightness constancy and small object motions are violated [188]. Consequently, the approach is unsuitable for deepfake videos involving fast-moving objects or post-processing adjustments in brightness.

Wu [189] introduced the SSTNet model, a novel approach for detecting face manipulation. This model combines spatial, steganalysis, and temporal features to achieve high accuracy, especially for compressed videos. The model first extracts faces from each frame to create a face stream. Then, the spatial and steganalysis features of the faces were extracted and combined to form the input for the module that extracts temporal features. Steganalysis suppresses image content with noisy residuals, resulting in a more concise and precise statistical explanation. The combination of residual information and temporal characteristics yields good detection accuracy, particularly for compressed videos. Montserrat et al. [190] proposed a similar method that combines RNN and CNN to extract visual and temporal features. Only frames displaying human faces were fed into the network, and a weighting method, along with a gated recurrent unit (GRU), is used to select the most informative frames for change detection automatically. This approach can process videos quickly, taking less than eight seconds on a single GPU. The weighting mechanism improves detection performance, even for high-quality deepfake videos [34]. Masi [2] also presented an approach that involves a recurrent network with two branches that remove facial content while propagating the original information. They apply Gaussian Laplacian amplification at a bottleneck layer to enhance multiband frequencies. The method analyzes aligned video face sequences, extracts discriminative features using a backbone network, and employs bi-directional LSTM for recurrent modeling supervised by a novel loss function. Experimental results on various datasets demonstrate this detection algorithm's effectiveness and generalization capabilities. Chen et al. [193] introduced the Xception-LSTM algorithm, which utilizes a novel spatiotemporal attention mechanism and Convolutional Long Short-Term Memory (ConvLSTM) to leverage intra- and inter-frame information. The spatial and temporal attention mechanism enhances spatiotemporal correlations before dimension reduction with XceptionNET. Additionally, ConvLSTM incorporates frame structure information to capture temporal dynamics. Compared to other RNN-based approaches [39], [186], this method demonstrates reduced computational requirements. However, spatiotemporal attention and increased model complexity increase the risk of overfitting, and their generalization ability is somewhat reduced.

Detecting spatiotemporal deepfakes typically involves using deep recurrent neural networks. These models learn video encodings in two stages: spatial features and sequential associations. Nguyen [192] utilizes a 3DCNN to simultaneously convolve spatial and temporal dimensions, allowing for the extraction of spatiotemporal features from just a few images. This technique also demonstrates robustness to compressed videos. However, it's challenging to determine whether spatial or temporal information contributes to the model's performance. Additionally, the performance of this method aligns with that of visual artifact based detectors, which exhibit high accuracy but poor generalization. Tariq et al. [191] proposed a deepfake video detector utilizing transfer learning. Their approach introduced the Convolutional

TABLE 10. Summary of spatiotemporal approach for deepfake. Highly performed classifier and dataset are shown in bold.

Reference	Model	Dataset	Best Performance	Manipulation Type Detected		Capability		
				Face-Swap	Expression-Swap	G	R	I
[39]	CNN + LSTM	Self-Built	$ACC = 0.97$	✓				
[186]	CNN+RNN	**FF++(DF)**, FF++(F2F), FF++(FS)	$ACC = 0.96$	✓	✓			
[187]	VGG16+ ResNet50	FF++	$ACC = 0.81$	✓	✓			
[189]	SSNET	**FF++(C23)**, FF++(C40)	$ACC = 0.98$	✓	✓		✓	
[190]	CNN+LSTM	DFDC	$ACC = 0.91$	✓	✓			
[191]	CLRNet	FF++(DF,FS,**NT**,F2F), DFD	$F1 = 0.99$	✓	✓	✓		
[2]	Bi-directional LSTM	**FF++(C23)**, FF++(C40), Celeb-DF	$AUC = 0.99$	✓	✓	✓	✓	
[41]	DPNET	**FF++(C23)**, FF++(C40), DFD, Celeb-DF, DF-1.0	EER=3.41% $AUC = 0.99$	✓	✓		✓	✓
[192]	3DCNN	**VidTimid(HQ,LQ)**, FF++(C23,C40)	$ACC = 0.99$	✓			✓	
[47]	MesoNet+ ResNet	**FF++(DF(C23,C40))**, **FF++(F2F(C23,C40))**, FF++(FS(C23,C40)), FF++(NT(C23,C40)), Celeb-DF(C23,C40))	$ACC = 0.84, 0.88$, $ACC = 0.86, 0.79$	✓	✓		✓	
[193]	Attention+ XceptionNet+ ConvLSTM	**FF++(DF)**, FF++(FS), FF++(NT), FF++(F2F), Celeb-DF, DFDC	$ACC = 1.0$	✓	✓			
[194]	Transformer	**FF++(C23)**, FF++(C40), Celeb-DF, DFDC	$ACC = 0.99$	✓	✓	✓	✓	✓
[195]	ResNet34	**DFDC-P**, FF++(C23), FF++(C40), Celeb-DF, WDF	$AUC = 0.99$	✓	✓	✓		
[196]	Graph Network	FF++(**DF**,NT,F2F,FS), Celeb-DF, DFDC-P	$AUC = 0.99$	✓	✓	✓		

LSTM-based Residual Network (CLRNet), which combines Convolutional LSTM and Residual Network architectures. CLRNet used 3D tensors to extract spatial information across consecutive frames, and then ConvLSTM cells were employed to capture temporal dependencies between frames. The core components of CLRNet consist of two types of building blocks, namely CL Block (ConvLSTM) and ID Block (Identity), resembling those found in ResNet. Through an evaluation of the CLRNet model using different transfer learning strategies, the authors demonstrated the generalizability of their approach compared to existing methods. However, it is important to note that methods based on 3D convolutions tend to have a significantly higher number of parameters than their 2D counterparts, making them computationally intensive.

To provide insight into deepfake detection predictions, a novel approach called Dynamic Prototype Network (DPNet) [41] has been proposed. DPNet integrates a prototype layer and a temporal logic verifier with a neural network architecture. This combination captures key features, including erroneous motion and temporal artifacts. By learning prototypical temporal inconsistencies within the latent space, the network organizes them into groups based on their proximity. During prediction, DPNet correlates a small set of learned dynamic prototypes with a test video, leading to a decision. Notably, the prototypes are matched to the most representative video patch in the training dataset, yielding a human-understandable representation of the trained dynamic prototypes. DPNet exhibits resilience to compressed, noisy, and distorted videos from the FF++ [7] and DF-1.0 [36] datasets. However, the approach does not generalize well on visually challenging dataset [90].

Hu et al. [47] introduced a deepfake detection approach designed explicitly for compressed videos, leveraging both temporal and frame-level properties. Their approach utilizes both frame-level and temporal properties. It combines a frame-level stream, called MesoNet stream [157], with data removal of redundant links to reduce faulty connections and minimize video compression artifacts. To reduce the size of the training dataset and decrease training time and cost, they focus on faces within frames instead of using the entire frames in their frame-level stream. Additionally, their technique incorporates a temporal stream that analyzes inconsistencies between frames by examining time-varying residual properties using ResNet [129]. The combination of these two streams determines the authenticity of the video. The proposed technique's effectiveness is evaluated through the Celeb-DF [6] and FF++ [7] datasets. The results demonstrate that the proposed method is computationally efficient and robust to different compression rates, indicating good cross-compression detection performance.

Zhao et al. [194] proposed a novel architecture of Interpretable Spatial-Temporal Video Transformer (ISTVT) for deepfake detection. This architecture includes a feature extractor that utilizes Xception blocks to extract texture features from sequences of frames. The self-attention module is divided into temporal and spatial components to address temporal inconsistencies, which is further enhanced by a self-subtraction mechanism. Moreover, the relevance propagation rules determine the relevance of the temporal and spatial self-awareness modules in each transformer block. This approach enables the creation of class-activation heatmaps for both spatial and temporal self-attention, providing interpretability. The ISTVT architecture displayed both robustness to perturbations such as JPEG compression, downscaling, and random dropout, as well as generalization ability on unseen datasets. However, the approach is computationally expensive.

A framework called Multi-Rate Excitation Network (MRE-Net) was introduced by Pang et al. [195] to detect deepfakes by capturing spatial and temporal information. To achieve this, MRE-Net uses a Bipartite Group Sampling (BGS) strategy that divides the video into multiple groups with varying sampling rates to identify inconsistencies across different distances. The framework consists of the Momentary Inconsistency Excitation (MIE) module and the Longstanding Inconsistency Excitation (LIE) module. The MIE module analyzes RGB frames and motion flows within a group, detecting short-term temporal inconsistencies. Meanwhile, the LIE module detects long-term temporal inconsistencies between neighboring groups. By integrating the outputs of both modules, MRE-Net predicts the authenticity of the video. The evaluation of MRE-Net demonstrates its performance in generalizing to unseen datasets of high quality [6], [98]. Additionally, MRE-Net exhibits robustness in handling video compression artifacts and challenging WDF dataset [98].

Shang et al. [196] developed a Spatiotemporal Graph Network (STGN) that can detect facial manipulation and forgery in videos. Their approach utilizes the Spatial Relation Graph Unit (SRGU) to model spatial relationships and capture global spatial inconsistencies through graph convolution. The Temporal Attention Graph Unit (TAGU) treats features from different frames at the same spatial location as a fully connected graph, using a cosine distance-based similarity matrix to detect temporal incoherence. The STGN architecture includes projection blocks, chart blocks, and back-projection blocks, which create a latent spatiotemporal relation space, stack SRGUs, and TAGUs to model inconsistencies, and map features back to the original space. Feature aggregation is achieved through global spatiotemporal average pooling, followed by prediction using a fully connected layer. The approach demonstrates generalization capacity for unseen manipulation. However, STGN incurs computational costs due to increased model parameters and additional graph convolution operations.

F. DEEPFAKE LOCALIZATION APPROACH

In contrast to binary classification deepfake detection models, deepfake localization approaches not only determine the authenticity of the video but also identify the exact facial areas that have been manipulated. Accurately localizing regions in facial images allows a better understanding of deepfake forgeries and the type of manipulation used, such as face or expression swaps. Table 11 summarizes the deepfake localization approaches.

Nguyen [197] developed a multi-task learning approach using autoencoders to detect and locate modified regions in facial images. Despite using a Y-shaped decoder to share information between tasks, the overall performance improvement was insignificant. Dang et al. [198] presented a CNN-based deepfake detection and localization approach incorporating an attention mechanism to optimize feature maps in the classifier model. Their proposed attention map is easy to build and can be integrated into existing backbone networks by adding a single convolutional layer that masks high-dimensional features, improving classification performance and reducing error rates. The effectiveness of their methodology was evaluated using a combination of the FF++ dataset [7] and an online video collection. Comparing this approach fairly with other techniques is difficult due to the diverse experimental protocols considered in the proposed approach. Roy-Chowdhury [199] proposed an approach to detect and localize fake facial manipulation based on a Face Expression Recognition (FER) system. The proposed approach uses a two-stream network to detect tampering. First, the FER stream extracts important information about facial expressions. Then, the second stream, responsible for manipulation detection and segmentation, uses an encoder-decoder architecture to locate manipulated regions within the facial image. While promising, this approach showed a high propensity to overfit certain datasets and is not transferrable to other unseen deepfakes.

In contrast to the encoder-decoder architecture for deepfake localization, the Face X-ray technique [42] adopted a different approach by considering noise and error levels analysis to identify blending artifacts. A fully convolutional neural network was trained to detect changes around the boundaries of deepfake generated faces with manipulated identities and expressions. The localization results of this approach are illustrated in Figure 13. While this approach aimed to generalize well on high-quality deepfake videos, its performance on low-quality data was negatively affected by compression, noise, or blur, which could remove blending boundaries and result in poor localization accuracy. In

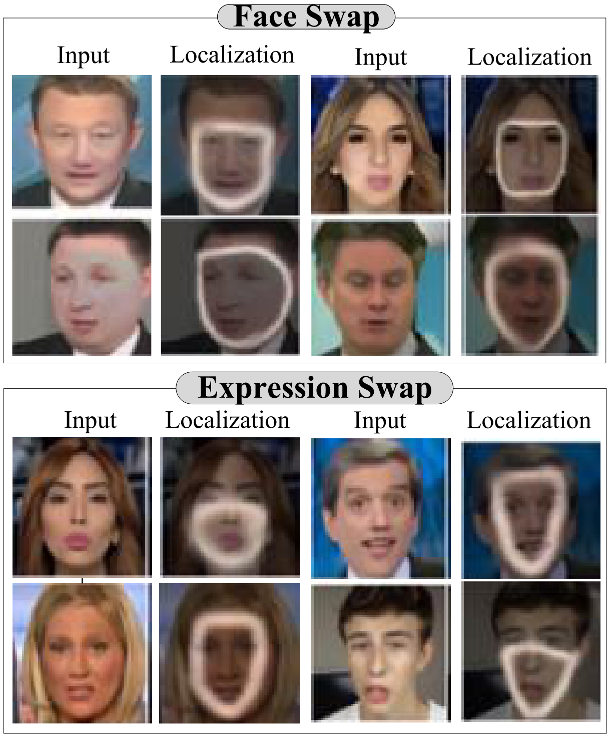

FIGURE 13. Localization of deepfake facial manipulation on FF++ dataset [42]. DF (Deepfake) and FS (Faceswap) refer to face swap manipulation, while NT (NeuralTexture) and F2F (Face-To-Face) represent expression swap manipulation.

deepfake detection and localization, traditional approaches have commonly followed a two-step pipeline involving face detection followed by face forensics. However, this approach suffers from the limitations of employing multiple models and redundant feature extraction. To address these challenges, Peng Chen et al. introduced the DLFMNet framework [200]. DLFMNet takes a different approach by integrating face detection and face forensics into one model. The framework leverages the complementary capabilities of the RGB and noise domains to capture manipulated cues. It incorporates constrained convolution layers to extract local noise features from RGB images and employs bilinear pooling to fuse multi-domain features from both streams. The Manipulation Classification Branch (MCB) at the output performs face forensics using RoIAlign features from both RGB and noise streams. The Manipulation Localization Branch (MLB) predicts pixel-level masks for the identified manipulated regions. The effectiveness and robustness of the DLFMNet framework were evaluated using the FF++ dataset. The results demonstrated that DLFMNet can detect fake faces, even in videos containing multiple faces. However, it is important to note that the performance of DLFMNet on high-quality deepfakes remains unassessed [6], [35]., leaving room for further investigation in that area.

Jian Wang et al. [202] proposed a Localization Invariance Siamese Networks (LiSiam) to ensure consistent localization across images with varying levels of quality degradation. LiSiam utilized data augmentation techniques to generate degraded images, which were processed by Siamese networks to produce segmentation maps. A novel localization invariance loss function was introduced to maintain consistent localization despite different degradation levels. Additionally, a mask-guided transformer localized suspected manipulated regions and their surroundings, while a multi-layer perceptron head incorporated co-occurrence features to make the final binary decision. LiSiam demonstrated robust and generalized localization of modified regions. However, the approach's effectiveness could be influenced by the diversity and quality of the augmented data, which should adequately represent the variations in deepfake manipulation. Waseem et al. [204] presented a novel encoder-decoder architecture based on a multi-attention mechanism designed for manipulation localization and detection. The multi-stream network concurrently captures spatial and frequency-related patterns to address deepfake detection in images with varying compression degradation. The proposed methodology uses spatial and channel attention blocks in an autoencoder stream to localize forged face regions at the pixel level. In parallel, another stream integrated spectral features with localized spatial attributes from the decoder layer through Bilinear pooling, resulting in a detection label. Although the method is resilient to compression-induced distortions, it lacks broad applicability across high-quality deepfake datasets.

M2TR, a multistream, multiscale transformer, was developed by Wang et al. [203] for deepfake detection and localization. This approach employed a two-stream architecture, with the RGB stream capturing inconsistencies between different regions within an image at multiple scales in the RGB domain, while the frequency stream used learnable frequency filters to filter out forged features in the frequency domain. A cross-modality fusion block merges information from both streams. The integrated features are then processed by fully connected layers and a decoder network for binary classification and localization of manipulated regions of the face image, respectively. Although computationally demanding, the approach has demonstrated acceptable generalization and robustness on high-quality deepfake videos.

Yu et al. [201] proposed a method for detecting and locating fake faces by identifying common traces left by different forgery methods, such as face boundary warp, PRNU noise, and biological signals. The approach involves training Specific Forgery Feature Extractors (SFFExtractors) to extract distinctive features for known forgery methods and a Common Forgery Feature Extractor (CFFExtractor) to learn shared characteristics among the SFFs and generate Common Forgery Features (CFFs). The optimization of the CFFExtractor involves multiple modules and loss functions, including feature similarity loss, domain classification loss, forgery classification loss, forgery location loss, and Automatic Weighted Loss (AWL). The Forgery Classification Module (FCM) and Forgery Location Module (FLM) perform forgery classification and localization tasks using the generated CFFs. The method demonstrates improved performance

TABLE 11. Summary of deepfake localization approach. Highly performed classifiers and datasets are highlighted in bold.

Reference	Model	Dataset	Best Performance	Manipulation Type Detected		Capability		
				Face-Swap	Expression-Swap	G	R	I
[197]	Encoder-Decoder	FF++	$AUC = 0.76$	✓	✓			
[198]	VGG16 and XceptionNet	**Self-Built(FS)**, **FF++(F2F)**, UADFV, Celeb-DF	EER=3.1 $AUC = 0.99$ EER= 3.4 $AUC = 0.99$	✓	✓			
[42]	HRNet	**FF++**, DFDC-P, DFDC, Celeb-DF	$AUC = 0.98$	✓	✓	✓		
[200]	ResNet50	**FF++(F2F(Raw,C23))**, FF++(DF(Raw,C23)), FF++(FS (Raw,C23)), FF++(NT(Raw,C23))	$AUC = 0.99$	✓	✓		✓	
[199]	Xception-Net + FER	**FF++(C23)**, FF++(C40), DFDC	$ACC = 0.99$	✓	✓			
[201]	EfficientNetB0, EfficientNetB3, Resnet50	**FF++(C23)**, FF++(C40), FF++(Raw), DFDC, Celeb-DF	$AUC = 0.92$	✓	✓	✓		
[202]	XceptionNet	Celeb-DF, **FF++(Raw), FF++(C23)**, FF++(C40)	$AUC = 0.99$	✓	✓	✓	✓	
[203]	Transformer	**FF++**, Celeb-DF, FN	$ACC = 0.99$	✓	✓	✓		
[204]	ResNet	**FF++(C23)**, **FF++(C40)**, Celeb-DF, DFDC-P, DFD	AUC=0.97	✓	✓		✓	

in deepfake manipulation detection. However, there are limitations to consider. The approach relies on the assumption that deepfake forgery techniques always leave similar traces. The proposed approach cannot detect and locate forgery if a forgery method does not share such common characteristics. Second, the diversity of original video sets can affect the system's performance.

V. DISCUSSION

The analysis shows that an increasing number of researchers are utilizing deep learning approaches instead of traditional handcrafted features-based techniques. This is because of the recent improvement in deepfake quality, which results in minimal traces and anomalies within the intrinsic feature of deepfake images or videos. It is harder to perform handcrafted feature extraction. Additionally, with the introduction of more efficient CNN architectures in recent years, many researchers have shifted their focus to learned feature extraction. Table 12 displays the experimental parameters obtained from the information presented in the referenced detection papers. Most of the detection approach utilizes Cross-entropy loss and Adam optimizer for their model training.

Different researchers use various metrics, such as Accuracy, F1-Score, AUC, EER, etc., to measure deepfake detection performance. Hence, a common benchmark is required to compare different tampering detection algorithms. The datasets used to evaluate the performance of deepfake detection algorithms must be standardized. To improve the evaluation of algorithms, the datasets must contain fake images and videos with a variety of different tampering attacks to increase the diversity of the test set.

VI. FUTURE DIRECTIONS

With the advancements in deep learning, creating deepfakes has become easier and social media platforms have made it easier for them to spread rapidly. Even if deepfakes are not shared widely, they can cause negative consequences. To address this issue, researchers are working on developing algorithms to detect deepfakes. Although significant progress has been made in this area, there are still various challenges and limitations that researchers need to overcome.

With new deepfake generation methods emerging, detecting deepfakes will become increasingly challenging, and the accuracy and efficiency of current deepfake detection approaches will decline. Most existing deepfake detection methods for face and expression swaps train and evaluate their performance on datasets that mainly include first- and second-generation deepfake datasets. These datasets, such as FaceForensic++ [7], DeepFake-TIMIT [90], and DFDC-P [34], contain low-quality controlled deepfakes with obvious artifacts, making them easy to detect. These artifacts can be removed by using post-processing techniques on the manipulated video to make them appear more natural to humans while reducing detection performance [205],

TABLE 12. Overview of experimental configuration for analyzed deepfake detection methods, with BS, LR, TE, LF, and Opt denoting Batch Size, Learning Rate, Training Epochs, Loss Function, and Optimizer respectively.

Reference	Input Size	Parameters Configuration					GPU/CPU
		BS	LR	TE	LF	Opt	
[125]	224×224	64	$1e^{-3}$	100	Cross-entropy	SGD	-
[21]	224×224	16	$1e^{-2}$	100		SGD, ADAM	-
[43]	256×256	-	-	-	-	ADAM	-
[141]	128×128	-	$1e^{-4}$	5000, 10000	-	ADAM	-
[143]	-	32	$1e^{-2}$	500	Cross-entropy	ADAM	Xeon E5-1650-v4 CPU and NVIDIA GP102L
[133]	-	-	$1e^{-3}$	200	-	-	-
[123]	256×256	32	$1e^{-3}$	40	-	SGD	Xeon W-2123 CPU
[124]	-	-	$2e^{-3}$	-	-	ADAM	-
[126]	299×299	-	-	-	Cross-entropy, logistic	-	-
[117]	64×64	600	$1e^{-4}$	60	Multi-margin angular loss	ADAM	NVDIA Tesla-P100
[145]	-	32	$1e^{-2}$	30	Cross-entropy and Attention mask	SGD	GTX-1080Ti
[137]	96×96	32	$2e^{-4}$	10	Cross-entropy	ADAM	-
[127]	256×256	32	$1e^{-4}$	100	Cross-entropy	ADAM	Intel Xeon(R) E5-2630 @2.20GHz CPU, Nvidia GeForce GTX 1080 Ti
[120]	256×256	32	$1e^{-2}$	25	-	ADAM	NVIDIA P5000
[138]	-	1024	$1e^{-3}$	-	-	ADAM	-
[150]	200×200	-	-	-	-	-	Intel(R) Core(TM) i7-8750H CPU
[151]	256×256	32	$2e^{-4}$	-	AM-Softmax	ADAM	-
[153]	224×224	64	$1e^{-2}$	50	Cross-entropy	ADAM	Intel (R) Xeon (R) CPU E5-2620 V4 and two NVIDIA GTX Titan XP GPUs
[154]	256×256	-	-	-	-	-	Nvidia GeForce RTX 2080 Ti, Intel Core i7-9700 K CPU
[165]	299×299, 128×128	32	$1e^{-1}$	16k	Triplet loss	-	-
[157]	256×256	75	$1e^{-3}$	-	-	ADAM	-
[158]	128×128	-	-	-	Cross-entropy	-	
[163]	256×256, 128×128	64	$1e^{-3}$	-	Mean squared error	ADAM	-
[166]	224×224	32	$1e^{-4}$	-	Self-Designed and Cross-entropy	ADAM	-
[162]	-	64, 128	$2e^{-3}$	36	Single-center	SGD	-
[161]	299×299	128	$2e^{-3}$	150k	Cross-entropy	SGD	-
[171]	224×224	32	$0.1e^{-3}$	50	log-loss	ADAM	-
[174]	-	128, 32	$1e^{-2}$	500, 100	Triplet	ADAM	NVIDIA GeForce GTX1080 Ti
[168]	$3 \times 25 \times 224 \times 224$	16	$0.1e^{-3}$	50	Cross-entropy	ADAM	Nvidia Titan RTX GPU 32
[98]	224×224	32	$1e^{-4}$	-	Cross-entropy	ADAM	-
[167]	224×224	32	$1e^{-5}$	20k	Triplet-margin	ADAM	Intel Xeon E5-2687W-v4 and a NVIDIA Titan V
[164]	-	128	$1e^{-4}$	10	-	ADAM	-
[169]	380×380	48	$1e^{-3}$	-	Regional-independence	ADAM	4 RTX 2080Ti GPUs
[159]	224×224	48	$1e^{-3}$	100k	Cross-entropy, Adversarial-Loss	SGD	NVIDIA GTX GeForce 1080 Ti
[160]	256×256	128	$5e^{-5}$	150	Cross-entropy	ADAM	-
[175]	-	64	$5e^{-4}$	-	Cross-entropy	ADAM	-
[172]	224×224, 299×299	-	-	30	Contrastive and Cross-entropy	Greedy	-

TABLE 12. *(Continued.)* **Overview of experimental configuration for analyzed deepfake detection methods, with BS, LR, TE, LF, and Opt denoting Batch Size, Learning Rate, Training Epochs, Loss Function, and Optimizer respectively.**

Reference	Input Size	Parameters Configuration					GPU/CPU
		BS	LR	TE	LF	Opt	
[176]	-	16	$2e^{-4}$	200	Cross-entropy	ADAM	NVIDIA TITAN V GPU
[170]	299 × 299	32	Dynamic	50	-	ADAM	Nvidia Tesla K80
[173]	-	32	$2e^{-4}$	-	Cosine-similarity, Consistency, Cross-entropy	ADAM	NIVIDA GeForce RTX 3090
[177]	256 × 256	32	$5e^{-4}$	40	Cross-entropy	SGD	-
[178]	-	-	$1e^{-4}$	-	Cross-entropy	SGD	2.20GHz Xeon CPU with 260GB RAM and two NVIDIA Tesla P40 GPUs with 24GB
[180]	64 × 64	-	$1e^{-4}$	25	-	-	-
[182]	100 × 100	128	$1e^{-3}$	300	Cross-entropy	-	
[184]	-	32	$2e^{-4}$	-	Metric-learning, Reconstruction, and Cross-entropy	ADAM	-
[185]	-	2304	$1e^{-4}$	12	Contrastive	ADAM	-
[39]	299 × 299	20,40,80	$1e^{-5}$	-	-	ADAM	-
[186]	224 × 224	-	$1e^{-4}$-	-	Cross-entropy	ADAM	-
[187]	224 × 224	256	$1e^{-4}$	-	-	ADAM	-
[189]	-	-	$1e^{-4}$	-	Cross-entropy	ADAM	-
[190]	224 × 224	2000	$1e^{-3}$	-	ArcFace, and Cross-entropy	ADAM	-
[191]	240 × 240	-	$1e^{-5}$	100	Cross-entropy	ADAM	Intel(R) Xeon(R) Silver 4114 CPU@2.20GHz with 256.0GB RAM and NVIDIA GeForce Titan RTX.
[2]	-	-	$1e^{-3}$	50	Self-Designed	ADAM	-
[41]	-	-	$2e^{-4}$, $1e^{-3}$	-	Cross-entropy, Clustering, Separation, and Diversity	-	-
[192]	128 × 128 × 16 × 3	20	$1e^{-4}$	30	-	SGD	-
[47]	256 × 256	8	$1e^{-3}$	-	Cross-entropy	ADAM	NVIDIA Titan Xp
[193]	-	1	$1e^{-5}$	-	Cross-entropy	ADAM	-
[194]	300 × 300	-	$5e^{-4}$	100	Cross-entropy	SGD	4 Tesla V100
[195]	320 × 320	3, 5	$1e^{-2}$	50	Focal	SGD	NVIDIA A100
[196]	224 × 224	4	$1e^{-4}$	90	Cross-entropy	SGD	-
[197]	-	64	$1e^{-2}$	-	Cross-entropy and Reconstruction	ADAM	-
[198]	-	16	$2e^{-4}$	-	-	ADAM	NVidia GTX 1080Ti
[42]	256 × 256	32	$2e^{-4}$	200k	Cross-entropy	ADAM	
[200]	640 × 640	16, 8	$10e^{-4}$	40k	Box-regression, Cross-entropy, Focal	ADAM	NVIDIA V100 (32GB)
[199]	-	16	$1e^{-3}$	20	Cross-entropy, Segmentation	ADAM	-
[201]	320 × 320	25	$1e^{-4}$	20	Triplet, Cross-entropy, Dice, Automatic Weighted	-	-
[202]	299 × 299	32	$2e^{-3}$	324k	Self-Designed, Cross-entropy	-	-
[203]	320 × 320	24	$1e^{-4}$	90	Cross-entropy, Segmentation, Contrastive	ADAM	-
[204]	224 × 224	16	$1e^{-4}$	-	Cross-entropy, L2	ADAM	NVIDIA GeForce RTX-3060 Ti

[206]. However, existing deepfake detectors rarely use high-quality deepfakes like OpenForensics [102], ForgeryNet [101], and DeepForensic 1.0 [36] for assessing the detection approach's robustness and generalization. To improve the effectiveness of deepfake detection models, they should be trained on datasets from various sources, including examples of both new and older deepfake datasets. In the future, researchers should focus on developing reliable, scalable, and generalizable detection techniques to address these issues.

Most deepfake detectors primarily focus on ensuring their robustness against compression, while they often overlook other types of perturbation attacks. Even minor changes in the input video or image can significantly affect the performance of detection models, causing them to deviate from expected behavior. Moreover, recent adversarial techniques aimed at misleading deep neural network-based detectors have further complicated the task of deepfake detection [111], [112], [114]. Adversarial samples are instrumental in enabling deepfake data to evade detection. Adversarial attacks are classified into two threat models: black-box and white-box, depending on the attacker's access and knowledge of the target detector. Reference [111] highlights vulnerabilities in deepfake detection by employing adversarial attacks to confuse neural network classifiers. The white-box attack reduces a specific classifier's accuracy to nearly 0% [111]. Using the Distortion-minimizing attack, only a small percentage (4%11%) of pixel modifications are needed to misclassify 89.7%100% the fake images [111]. In real-world scenarios, the black-box transfer attack is more relevant as the adversary lacks knowledge of the specific detection method. The authors assume the adversary is aware of the defense strategy. They develop an attack that achieves higher accuracy reduction and diminishes the classifier's AUC from 0.96 to 0.22 [111], rendering it unreliable. As deepfake detection relies more on neural network training, there is a valuable opportunity for research to explore robust defense mechanisms against adversarial attacks in the deepfake domain. This area holds immense potential for further investigation.

In various situations, real-time detection of fake facial content is crucial. Detection methods that are very accurate but take a long time to infer are unlikely to find widespread acceptance. As the number of social media users continues to grow and the creation and use of deepfakes become more accessible, the need for computationally efficient deepfake detection becomes even more important. Real-time functionality is essential for these approaches, as deepfake technology has the potential to cause irreversible damage. Prioritizing real-time actions is necessary because damage can be done before people realize it's fake, and the speed of social media can amplify the impact. Since smartphones play a crucial role in sharing content on social media, it is imperative to develop fast, reliable, and smartphone-compatible deepfake detection methods.

Compared to face-swap datasets, publicly available expression-swap datasets are lacking. In order to improve the accuracy of deepfake detection methods, it is crucial to train models to recognize a wider range of manipulated facial expressions, which can significantly increase the effectiveness of deepfake detection systems in identifying complex videos and images. Therefore, researchers should consider creating a more comprehensive and diverse expression exchange data set in the future.

As deepfake manipulations continue to improve, it becomes increasingly difficult to rely on a single solution to combat the multitude of deepfake threats. Addressing this problem effectively requires a combination of multiple detection systems, networks, and approaches for optimal results. In addition, research should prioritize identifying the most effective strategy to integrate all available information. Achieving this requires performing multi-asset analysis and leveraging multi-tool fusion. To counter the proliferation of deepfake, it is crucial to scrutinize all media content and supporting evidence. It is important to develop deepfake detection techniques that can recognize the input video or image based on all available supporting evidence and contextual information in which it is presented on a given platform to spread misinformation.

Current deepfake detection methods rely on machine learning techniques such as supervised classification and unsupervised clustering to detect attack patterns. However, these methods have difficulties in detecting new and unknown deepfakes. Reinforcement Learning (RL) could fundamentally change the detection of deepfakes in the future. Combining RL with game theory can help to detect and defend against anti-forensic attacks. RL can simulate an autonomous agent that can make optimal decisions even without prior knowledge of the environment. This makes it useful for tracking attacks and implementing attack-aware detectors. Deepfake detection can be modeled as a two-player zero-sum game where the sum of both players' utilities remains zero at each time step. Deep Reinforcement Learning (DRL) is particularly effective in addressing complex cyber defense challenges [207], [208], and can potentially be used to detect deepfakes and defend against anti-forensic attacks on detectors.

To counteract the far-reaching effects of deepfakes, it is crucial to integrate detection systems into social media platforms and distribution channels. These platforms can implement screening or filtering processes based on reliable detection techniques, and companies operating them can be legally obligated to remove deepfakes upon discovery [12]. Deep-fake-o-meter [209] is an example of a web-based platform that hosts over ten state-of-the-art deepfake detection models, allowing users to upload videos and receive detailed reports via email after processing by the platform's back-end algorithms. This platform serves as a valuable resource for developers and researchers to benchmark their detection algorithms against the latest methods. Another platform [210], customized for journalists, uses a video detection model and a temporal-based approach considering frame-level artifacts. The platform detects fake audio and offers an intuitive application for journalists to assess video authenticity. Collaborating these detection systems with social media platforms can further mitigate the adverse impact of deepfakes. Moreover, watermarking technologies can be integrated into content creation devices, providing a foolproof method of authenticating multimedia files and verifying when and where they were created [12]. Implementing this integration is a challenge. However, it may be possible to overcome this challenge by using blockchain technology. While blockchain technology has numerous potential applications, few current research projects focus on its use for deepfake detection. It is an excellent tool for digital provenance solutions as it can provide a sequence of unique, immutable metadata. Using blockchain technologies to solve this problem has had some success [211]. However, this field of research is still in its infancy.

VII. CONCLUSION

Advances in AI have made it easy for anyone with a smartphone to create realistic deepfake images and videos. However, the two-player nature of this research area has also led to the use of AI to detect manipulated content. Nevertheless, the quality of deepfake videos is constantly improving and posing new challenges. The two most well-known deepfake face manipulation techniques are face swapping and expression swapping can lead to harmful consequences such as revenge porn, bullying, video and news forgery, blackmail and political sabotage, making the target's life more difficult. In this article, we presented a comprehensive overview and detailed analysis of the research work on deepfake generation and detection, particularly focusing on face and expression manipulation. Furthermore, we have examined the benchmark dataset deeply for detecting these facial manipulations. The article allows a fresh perspective on the current state of deepfake research, providing valuable insights into the challenges and opportunities, as well as the trends and directions for further exploration in the field of deepfake generation and detection. We strongly aspire that this review paper can empower and accelerate the efforts of researchers and practitioners in this domain. It aims to assist them in identifying the most critical research areas while inspiring a larger community of researchers to participate in this rapidly expanding and evolving field actively. This review primarily focuses on face and expression manipulation in deepfake videos. Future studies could explore a wider range of manipulations, including voice and context alterations, to provide a more holistic understanding of the evolving deepfake landscape.

REFERENCES

[1] P. Maares, S. Banjac, and F. Hanusch, "The labour of visual authenticity on social media: Exploring producers' and audiences' perceptions on Instagram," *Poetics*, vol. 84, Feb. 2021, Art. no. 101502.

[2] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Proc. Computer Vis. (ECCV)*, A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 667–684.

[3] A. Tewari, M. Zollhöfer, F. Bernard, P. Garrido, H. Kim, P. Pérez, and C. Theobalt, "High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 357–370, Feb. 2020.

[4] J. Yi, C. Wang, J. Tao, X. Zhang, C. Yuan Zhang, and Y. Zhao, "Audio deepfake detection: A survey," 2023, *arXiv:2308.14970*.

[5] D. Pan, L. Sun, R. Wang, X. Zhang, and R. O. Sinnott, "Deepfake detection through deep learning," in *Proc. IEEE/ACM Int. Conf. Big Data Comput., Appl. Technol. (BDCAT)*, Dec. 2020, pp. 134–143.

[6] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3204–3213.

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

[8] J. Kietzmann, A. J. Mills, and K. Plangger, "Deepfakes: Perspectives on the future 'reality' of advertising and branding," *Int. J. Advertising*, vol. 40, no. 3, pp. 473–485, Apr. 2021.

[9] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, "Deepfakes: Trick or treat?" *Bus. Horizons*, vol. 63, no. 2, pp. 135–146, 2020.

[10] Z. Akhtar, "Deepfakes generation and detection: A short survey," *J. Imag.*, vol. 9, no. 1, p. 18, Jan. 2023.

[11] J. Dietmar. *Council Post: GANs and Deepfakes Could Revolutionize the Fashion Industry*. Accessed: Nov. 20, 2021. [Online]. Available: https://www.filmvideodigital.com/deepfakes/newswire/2019/5/22/forbes-gans-and-deepfakes-could-revolutionize-the-fashion-industry

[12] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *California Law Rev.*, vol. 107, p. 1753, Jul. 2019.

[13] M.-H. Maras and A. Alexandrou, "Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos," *Int. J. Evidence Proof*, vol. 23, no. 3, pp. 255–262, Jul. 2019.

[14] J. Fletcher, "Deepfakes, artificial intelligence, and some kind of dystopia: The new faces of online post-fact performance," *Theatre J.*, vol. 70, no. 4, pp. 455–471, 2018.

[15] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, R. P. Luis, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang, "DeepFaceLab: Integrated, flexible and extensible face-swapping framework," 2021, *arXiv:2005.05535*.

[16] Iperov. *Deepfacelab*. Accessed: Jul. 6, 2021. [Online]. Available: https://github.com/iperov/DeepFaceLab/blob/maste

[17] CSC Momo. *ZAO*. Accessed: Jan. 6, 2022. [Online]. Available: https://zaodownload.com

[18] FakeApp. *FakeApp 2.2.0*. Accessed: Jul. 10, 2021. [Online]. Available: https://www.malavida.com/en/soft/fakeapp/

[19] S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans, *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection*, vol. 2. Cham, Switzerland: Springer, 2019.

[20] A. Gurnani, K. Shah, V. Gajjar, V. Mavani, and Y. Khandhediya, "SAF-BAGE: Salient approach for facial soft-biometric classification–age, gender, and facial expression," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 839–847.

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

[22] J. Bateman, *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*. Washington, DC, USA: Carnegie Endowment for International Peace, 2020.

[23] N. Beuve, W. Hamidouche, and O. Deforges, "DmyT: Dummy triplet loss for deepfake detection," in *Proc. 1st Workshop Synth. Multimedia Audio-visual Deepfake Gener. Detection*. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 17–24.

[24] J. Wojewidka, "The deepfake threat to face biometrics," *Biometric Technol. Today*, vol. 2020, no. 2, pp. 5–7, Feb. 2020.

[25] A. Ali, Y. K. Jadoon, Z. Farid, M. Ahmad, N. Abidi, H. M. Alzoubi, and A. A. Alzoubi, "The threat of deep fake technology to trusted identity management," in *Proc. Int. Conf. Cyber Resilience (ICCR)*, Oct. 2022, pp. 1–5.

[26] M. Borak. (2021). *Chinese Government-Run Facial Recognition System Hacked by Tax Fraudsters: Report*. Accessed: Aug. 8, 2023. [Online]. Available: https://www.scmp.com/tech/tech-trends/article/3127645/chinese-government-run-facial-recognition-system-hacked-tax

[27] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020.

[28] T. Zhang, "Deepfake generation and detection, a survey," *Multimedia Tools Appl.*, vol. 81, no. 5, pp. 6259–6276, Feb. 2022.

[29] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake detection for human face images and videos: A survey," *IEEE Access*, vol. 10, pp. 18757–18775, 2022.

[30] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022.

[31] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Int. J. Speech Technol.*, vol. 53, no. 4, pp. 3974–4026, Feb. 2023.

[32] D. M. Turek. *Media Forensics (MediFor) challenge by Defense Advanced Research Projects Agency*. Accessed: Mar. 4, 2021. [Online]. Available: https://www.darpa.mil/program/media-forensics

[33] NIST. *Media Forensics Challenge 2018*. Accessed: Mar. 4, 2021. [Online]. Available: https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2018

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

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

[36] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2886–2895.

[37] L. Jiang. *CodaLab—Competition*. Accessed: May 4, 2020. [Online]. Available: https://competitions.codalab.org/competitions/25228

[38] ICCV. *Sensing, Understanding and Synthesizing Humans*. Accessed: Apr. 15, 2021. [Online]. Available: https://sense-human.github.io/

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

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

[41] L. Trinh, M. Tsang, S. Rambhatla, and Y. Liu, "Interpretable and trustworthy deepfake detection via dynamic prototypes," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1972–1982.

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

[43] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2019, pp. 83–92.

[44] D. A. Coccomini, R. Caldelli, F. Falchi, and C. Gennaro, "On the generalization of deep learning models in video deepfake detection," *J. Imag.*, vol. 9, no. 5, p. 89, Apr. 2023.

[45] A. Devasthale and S. Sural, "Adversarially robust deepfake video detection," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2022, pp. 396–403.

[46] S. Dhesi, L. Fontes, P. Machado, I. K. Ihianle, F. F. Tash, and D. A. Adama, "Mitigating adversarial attacks in deepfake detection: An exploration of perturbation and ai techniques," 2023, *arXiv:2302.11704*.

[47] J. Hu, X. Liao, W. Wang, and Z. Qin, "Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1089–1102, Mar. 2022.

[48] I. Goodfellow, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[49] Z. Wan, H. He, and B. Tang, "A generative model for sparse hyperparameter determination," *IEEE Trans. Big Data*, vol. 4, no. 1, pp. 2–10, Mar. 2018.

[50] A. Punnappurath and M. S. Brown, "Learning raw image reconstruction-aware deep image compressors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 1013–1019, Apr. 2020.

[51] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3697–3705.

[52] Deepfake. *Reddit*. Accessed: May 28, 2022. [Online]. Available: https://www.reddit.com/r/deepfakes

[53] Dfaker. *Dfaker/df GitHub*. Accessed: Dec. 21, 2021. [Online]. Available: https://github.com/dfaker/df

[54] StromWine. *DeepFake-Tf*. Accessed: Jan. 22, 2022. [Online]. Available: https://github.com/StromWine/DeepFake_tf

[55] Y. Nirkin, I. Masi, A. Tran Tuan, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 98–105.

[56] D. S. Trigueros, L. Meng, and M. Hartnett, "Generating photo-realistic training data to improve face recognition accuracy," *Neural Netw.*, vol. 134, pp. 86–94, Feb. 2021.

[57] S. Lu. (2022). *Faceswap-GAN*. Accessed: Mar. 22, 2023. [Online]. Available: https://github.com/shaoanlu/faceswap-GAN

[58] R. Natsume, T. Yatagawa, and S. Morishima, "RSGAN," in *Proc. ACM SIGGRAPH Posters*, Aug. 2018, pp. 1–15.

[59] R. Natsume, T. Yatagawa, and S. Morishima, "FSNet: An identity-aware generative model for image-based face swapping," in *Computer Vision–ACCV*. Cham, Switzerland: Springer, 2019, pp. 117–132.

[60] Y. Nirkin, Y. Keller, and T. Hassner, "FSGANv2: Improved subject agnostic face swapping and reenactment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 560–575, Jan. 2023.

[61] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *arXiv:1912.13457*, 2020.

[62] R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap: An efficient framework for high fidelity face swapping," in *Proc. 28th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 2003–2011.

[63] Y. Zhu, Q. Li, J. Wang, C. Xu, and Z. Sun, "One shot face swapping on megapixels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4832–4842.

[64] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, and R. Ji, "HifiFace: 3D shape and semantic prior guided high fidelity face swapping," 2021, *arXiv:2106.09965*.

[65] L. Zhang, H. Yang, T. Qiu, and L. Li, "AP-GAN: Improving attribute preservation in video face swapping," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2226–2237, Apr. 2022.

[66] Y. Xu, B. Deng, J. Wang, Y. Jing, J. Pan, and S. He, "High-resolution face swapping via latent semantics disentanglement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7632–7641.

[67] F. Rosberg, E. E. Aksoy, F. Alonso-Fernandez, and C. Englund, "FaceDancer: Pose- and occlusion-aware high fidelity face swapping," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3443–3452.

[68] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, "Marionette: Few-shot face reenactment preserving identity of unseen targets," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 10893–10900.

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

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

[71] M. Agarwal, R. Mukhopadhyay, V. Namboodiri, and C. V. Jawahar, "Audio-visual face reenactment," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5167–5176.

[72] H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen, "Bringing portraits to life," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–13, Nov. 2017.

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

[74] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schoenborn, and T. Vetter, "Morphable face models—An open framework," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 75–82.

[75] Y. Lu, Y.-W. Tai, and C.-K. Tang, ‘‘Attribute-guided face generation using conditional CycleGAN,’’ in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 1–16.

[76] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, ‘‘GANimation: Anatomically-aware facial animation from a single image,’’ in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 835–851.

[77] E. Sanchez and M. Valstar, ‘‘Triple consistency loss for pairing distributions in gan-based face synthesis,’’ 2018, *arXiv:1811.03492*.

[78] J. Thies, M. Zollhöfer, and M. Nießner, ‘‘Deferred neural rendering: Image synthesis using neural textures,’’ *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, Jul. 2019.

[79] K. Lata, M. Dave, and K. N. Nishanth, ‘‘Image-to-image translation using generative adversarial network,’’ in *Proc. 3rd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Jun. 2019, pp. 186–189.

[80] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, ‘‘Few-shot adversarial learning of realistic neural talking head models,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9458–9467.

[81] Y. Zhang, S. Zhang, Y. He, C. Li, C. C. Loy, and Z. Liu, ‘‘One-shot face reenactment,’’ 2019, *arXiv:1908.03251*.

[82] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, ‘‘First order motion model for image animation,’’ in *Advances in Neural Information Processing Systems*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019.

[83] H. Hao, S. Baireddy, A. R. Reibman, and E. J. Delp, ‘‘FaR-GAN for one-shot face reenactment,’’ 2020, *arXiv:2005.06402*.

[84] X. Fu, X. Wang, J. Liu, W. Liu, J. Dai, and J. Han, ‘‘MakeItSmile: Detail-enhanced smiling face reenactment,’’ in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8.

[85] J. Zhang, X. Zeng, Y. Pan, Y. Liu, Y. Ding, and C. Fan, ‘‘Faceswapnet: Landmark guided many-to-many face reenactment,’’ 2019, *arXiv:1905.11805*.

[86] H. Xue, J. Ling, A. Tang, L. Song, R. Xie, and W. Zhang, ‘‘High-fidelity face reenactment via identity-matched correspondence learning,’’ *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 3, pp. 1–23, Feb. 2023.

[87] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, ‘‘Semantic image synthesis with spatially-adaptive normalization,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2332–2341.

[88] X. Tong, L. Wang, X. Pan, and J. g. Wang, ‘‘An overview of deepfake: The sword of damocles in AI,’’ in *Proc. Int. Conf. Comput. Vis., Image Deep Learn. (CVIDL)*, Jul. 2020, pp. 265–273.

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

[90] P. K. Marcel. *DeepfakeTIMIT*. Accessed: Apr. 19, 2021. [Online]. Available: https://www.idiap.ch/en/dataset/deepfaketimit/index_html

[91] C. Sanderson. (2004). *The VidTIMIT Database*. Accessed: May 1, 2020. [Online]. Available: https://conradsanderson.id.au/vidtimit/

[92] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch, ‘‘Fake face detection methods: Can they be generalized?’’ in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, 2018, pp. 1–6.

[93] Google AI. *Contributing Data to Deepfake Detection Research*. Accessed: Jun. 1, 2023. [Online]. Available: http://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

[94] J. Huang, X. Wang, B. Du, P. Du, and C. Xu, ‘‘DeepFake MNIST+: A DeepFake facial animation dataset,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1973–1982.

[95] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, ‘‘Voxceleb: Large-scale speaker verification in the wild,’’ *Comput. Speech Lang.*, vol. 60, Mar. 2020, Art. no. 101027.

[96] P. Korshunov and S. Marcel, ‘‘Improving generalization of deepfake detection with data farming and few-shot learning,’’ *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 3, pp. 386–397, Jul. 2022.

[97] C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matejka, J. Cernocký, N. Poh, J. Kittler, A. Larcher, C. Lévy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes, ‘‘BI-modal person recognition on a mobile phone: Using mobile phone data,’’ in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2012, pp. 635–640.

[98] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, ‘‘WildDeepfake: A challenging real-world dataset for deepfake detection,’’ in *Proc. 28th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 2382–2390.

[99] P. Kwon, J. You, G. Nam, S. Park, and G. Chae, ‘‘KoDF: A large-scale Korean DeepFake detection dataset,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10724–10733.

[100] T. Zhou, W. Wang, Z. Liang, and J. Shen, ‘‘Face forensics in the wild,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5774–5784.

[101] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu, ‘‘ForgeryNet: A versatile benchmark for comprehensive forgery analysis,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4358–4367.

[102] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, ‘‘OpenForensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 10117–10127.

[103] G. Li, X. Zhao, Y. Cao, P. Pei, J. Li, and Z. Zhang, ‘‘FMFCC-V: An Asian large-scale challenging dataset for DeepFake detection,’’ in *Proc. ACM Workshop Inf. Hiding Multimedia Secur.* New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 7–18.

[104] K. Narayan, H. Agarwal, K. Thakral, S. Mittal, M. Vatsa, and R. Singh, ‘‘DF-platter: Multi-face heterogeneous deepfake dataset,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9739–9748.

[105] Kaggle. *Deepfake Detection Challenge*. Accessed: Feb. 1, 2023. [Online]. Available: https://www.kaggle.com/c/deepfake-detection-challenge

[106] D. Huang and F. De La Torre, ‘‘Facial action transfer with personalized bilinear regression,’’ in *Proc. Comput. Vision–ECCV 12th Eur. Conf. Comput. Vis.* Florence, Italy: Springer, Oct. 2012, pp. 144–158.

[107] DeepFakes. *Faceswap*. Accessed: May 1, 2021. [Online]. Available: https://github.com/deepfakes/faceswap

[108] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu, ‘‘Audio-driven talking face video generation with learning-based personalized head pose,’’ 2020, *arXiv:2002.10137*.

[109] J. S. Chung and A. Zisserman, ‘‘Out of time: Automated lip sync in the wild,’’ in *Proc. Comput. Vis.–ACCV Workshops, ACCV Int. Workshops*. Taipei, Taiwan: Springer, Nov. 2017, pp. 251–263.

[110] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, ‘‘Recycle-GAN: Unsupervised video retargeting,’’ in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 119–135.

[111] N. Carlini and H. Farid, ‘‘Evading deepfake-image detectors with white- and black-box attacks,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2804–2813.

[112] A. Gandhi and S. Jain, ‘‘Adversarial perturbations fool deepfake detectors,’’ in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[113] S.-Y. Lin, J.-C. Chen, and J.-C. Wang, ‘‘A comparative study of cross-model universal adversarial perturbation for face forgery,’’ in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2022, pp. 1–5.

[114] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley, ‘‘Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples,’’ in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3347–3356.

[115] S. W. Hall, A. Sakzad, and K. R. Choo, ‘‘Explainable artificial intelligence for digital forensics,’’ *WIREs Forensic Sci.*, vol. 4, no. 2, p. e1434, Mar. 2022.

[116] A. A. Maksutov, V. O. Morozov, A. A. Lavrenov, and A. S. Smirnov, ‘‘Methods of deepfake detection based on machine learning,’’ in *Proc. IEEE Conf. Russian Young Researchers Electr. Electron. Eng. (EIConRus)*, Jan. 2020, pp. 408–411.

[117] G. Li, Y. Cao, and X. Zhao, ‘‘Exploiting facial symmetry to expose deepfakes,’’ in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 3587–3591.

[118] D. E. King, ‘‘Dlib-ml: A machine learning toolkit,’’ *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.

[119] B. Xu, J. Liu, J. Liang, W. Lu, and Y. Zhang, ‘‘DeepFake videos detection based on texture features,’’ *Comput., Mater. Continua*, vol. 68, no. 1, pp. 1375–1388, 2021.

[120] S. Kingra, N. Aggarwal, and N. Kaur, ‘‘LBPNet: Exploiting texture descriptor for deepfake detection,’’ *Forensic Sci. Int., Digital Invest.*, vols. 42–43, Oct. 2022, Art. no. 301452.

[121] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, ‘‘Leveraging frequency analysis for deep fake image recognition,’’ in *Proc. 37th Int. Conf. Mach. Learn.*, H. Daumé III and A. Singh, Eds., vol. 119, Jul. 2020, pp. 3247–3258.

[122] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking DeepFakes with simple features," 2020, *arXiv:1911.00686*.

[123] A. Kohli and A. Gupta, "Detecting DeepFake, FaceSwap and Face2Face facial forgeries using frequency CNN," *Multimedia Tools Appl.*, vol. 80, no. 12, pp. 18461–18478, May 2021.

[124] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 772–781.

[125] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," 2018, *arXiv:1811.00656*.

[126] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "DeepFake detection based on discrepancies between faces and their context," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6111–6121, Oct. 2022.

[127] E. Kim and S. Cho, "Exposing fake faces through deep neural networks combining content and trace feature extractors," *IEEE Access*, vol. 9, pp. 123493–123503, 2021.

[128] B. Koonce, *Convolutional Neural Networks With Swift for Tensorflow*. Berkeley, CA, USA: Apress, 2021, pp. 63–72.

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

[130] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proc. CVPR Workshops*, vol. 1, 201, p. 389.

[131] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[132] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Detecting deep-fake videos from appearance and behavior," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2020, pp. 1–6.

[133] H. M. Nguyen and R. Derakhshani, "Eyebrow recognition for identifying deepfake videos," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2020, pp. 1–5.

[134] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[135] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5mb model size," 2016, *arXiv:1602.07360*.

[136] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

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

[138] X. Liao, Y. Wang, T. Wang, J. Hu, and X. Wu, "FAMM: Facial muscle motions for detecting compressed deepfake videos over social networks," *IEEE Trans. Circuits Syst. Video Technol.*, early access, May 22, 2023, doi: 10.1109/TCSVT.2023.3278310.

[139] T. Jung, S. Kim, and K. Kim, "DeepVision: Deepfakes detection using human eye blinking pattern," *IEEE Access*, vol. 8, pp. 83144–83154, 2020.

[140] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.

[141] S. Fernandes, S. Raj, E. Ortiz, I. Vintila, M. Salter, G. Urosevic, and S. Jha, "Predicting heart rate variations of deepfake videos using neural ODE," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1721–1729.

[142] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 15, 2020, doi: 10.1109/TPAMI.2020.3009287.

[143] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, and J. Zhao, "DeepRhythm: Exposing DeepFakes with attentional visual heartbeat rhythms," in *Proc. 28th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 4318–4327.

[144] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, and A. Morales, "DeepFakesON-Phys: Deepfakes detection based on heart rate estimation," 2020, *arXiv:2010.00400*.

[145] J. Wu et al., "Local attention and long-distance interaction of rPPG for deepfake detection," *Vis. Comput.*, 2023, doi: 10.1007/s00371-023-02833-x.

[146] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[147] T. Soukupova and J. Cech, "Eye blink detection using facial landmarks," in *Proc. 21st Comput. Vis. Winter Workshop*, Rimske Toplice, Slovenia, 2016, p. 2.

[148] M. Koopman, A. M. Rodriguez, and Z. Geradts, "Detection of deepfake video manipulation," in *Proc. 20th Irish Mach. Vis. Image Process. Conf. (IMVIP)*, 2018, pp. 133–136.

[149] F. Lugstein, S. Baier, G. Bachinger, and A. Uhl, "PRNU-based deepfake detection," in *Proc. ACM Workshop Inf. Hiding Multimedia Secur.* New York, NY, USA: Association for Computing Machinery, Jun. 2021, pp. 7–12.

[150] F. F. Kharbat, T. Elamsy, A. Mahmoud, and R. Abdullah, "Image feature detectors for deepfake video detection," in *Proc. IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2019, pp. 1–4.

[151] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16312–16321.

[152] H.-S. Chen, M. Rouhsedaghat, H. Ghani, S. Hu, S. You, and C.-C. Jay Kuo, "DefakeHop: A light-weight high-performance deepfake detector," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[153] J. Yang, A. Li, S. Xiao, W. Lu, and X. Gao, "MTD-net: Learning to detect deepfakes images by multi-scale texture difference," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4234–4245, 2021.

[154] G. Wang, Q. Jiang, X. Jin, and X. Cui, "FFR_FD: Effective and fast detection of DeepFakes via feature point defects," *Inf. Sci.*, vol. 596, pp. 472–488, Jun. 2022.

[155] Z. Xia, T. Qiao, M. Xu, N. Zheng, and S. Xie, "Towards DeepFake video forensics based on facial textural disparities in multi-color channels," *Inf. Sci.*, vol. 607, pp. 654–669, Aug. 2022.

[156] Md. S. Rana and A. H. Sung, "DeepfakeStack: A deep ensemble-based learning technique for deepfake detection," in *Proc. 7th IEEE Int. Conf. Cyber Secur. Cloud Comput. (CSCloud)/6th IEEE Int. Conf. Edge Comput. Scalable Cloud (EdgeCom)*, Aug. 2020, pp. 70–75.

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

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

[159] J. Zhang, J. Ni, and H. Xie, "DeepFake videos detection using self-supervised decoupling network," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[160] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15003–15013.

[161] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 86–103.

[162] J. Li, H. Xie, L. Yu, X. Gao, and Y. Zhang, "Discriminative feature mining based on frequency information and metric learning for face forgery detection," *IEEE Trans. Knowl. Data Eng.*, early access, Oct. 1, 2021, doi: 10.1109/TKDE.2021.3117003.

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

[164] B. Chen and S. Tan, "FeatureTransfer: Unsupervised domain adaptation for cross-domain deepfake detection," *Secur. Commun. Netw.*, vol. 2021, pp. 1–8, Jun. 2021.

[165] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1831–1839.

[166] P. Kumar, M. Vatsa, and R. Singh, "Detecting Face2Face facial reenactment in videos," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2578–2586.

[167] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 5012–5019.

[168] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other-audio-visual dissonance-based deepfake detection and localization," in *Proc. 28th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 439–447.

[169] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2185–2194.

[170] S. Ganguly, S. Mohiuddin, S. Malakar, E. Cuevas, and R. Sarkar, "Visual attention-based deepfake video forgery detection," *Pattern Anal. Appl.*, vol. 25, no. 4, pp. 981–992, Nov. 2022.

[171] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," 2021, *arXiv:2102.11126*.

[172] Y. Xu, K. Raja, and M. Pedersen, "Supervised contrastive learning for generalizable and explainable DeepFakes detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2022, pp. 379–389.

[173] F. Dong, X. Zou, J. Wang, and X. Liu, "Contrastive learning-based general deepfake detection with multi-scale RGB frequency clues," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 4, pp. 90–99, Apr. 2023.

[174] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proc. 28th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 2823–2832.

[175] Y. Zhou and S.-N. Lim, "Joint audio-visual deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14780–14789.

[176] M. Yu, S. Ju, J. Zhang, S. Li, J. Lei, and X. Li, "Patch-DFD: Patch-based end-to-end DeepFake discriminator," *Neurocomputing*, vol. 501, pp. 583–595, Aug. 2022.

[177] Y. Hua, R. Shi, P. Wang, and S. Ge, "Learning patch-channel correspondence for interpretable face forgery detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1668–1680, 2023.

[178] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu, "Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces," 2020, *arXiv:1909.06122*.

[179] S. Fernandes, S. Raj, R. Ewetz, J. S. Pannu, S. K. Jha, E. Ortiz, I. Vintila, and M. Salter, "Detecting deepfake videos using attribution-based confidence metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1250–1259.

[180] A. Khodabakhsh and C. Busch, "A generalizable deepfake detector based on neural conditional distribution modelling," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2020, pp. 1–5.

[181] G. Wang, J. Zhou, and Y. Wu, "Exposing deep-faked videos by anomalous co-motion pattern detection," 2020, *arXiv:2008.04848*.

[182] H. Khalid and S. S. Woo, "OC-FakeDect: Classifying deepfakes using one-class variational autoencoder," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2794–2803.

[183] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, "ID-reveal: Identity-aware DeepFake video detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15088–15097.

[184] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4103–4112.

[185] D. Cozzolino, A. Pianese, M. Nießner, and L. Verdoliva, "Audio-visual person-of-interest DeepFake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 943–952.

[186] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, no. 1, pp. 80–87, 2019.

[187] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1205–1207.

[188] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.

[189] X. Wu, Z. Xie, Y. Gao, and Y. Xiao, "SSTNet: Detecting manipulated faces through spatial, steganalysis and temporal features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2952–2956.

[190] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, E. Bartusiak, J. Yang, D. Güera, F. Zhu, and E. J. Delp, "Deepfakes detection with automatic face weighting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2851–2859.

[191] S. Tariq, S. Lee, and S. S. Woo, "A convolutional LSTM based residual network for deepfake video detection," 2020, *arXiv:2009.07480*.

[192] X. H. Nguyen, T. S. Tran, V. T. Le, K. D. Nguyen, and D.-T. Truong, "Learning spatio-temporal features to detect manipulated facial videos created by the deepfake techniques," *Forensic Sci. Int., Digit. Invest.*, vol. 36, Mar. 2021, Art. no. 301108.

[193] B. Chen, T. Li, and W. Ding, "Detecting deepfake videos based on spatiotemporal attention and convolutional LSTM," *Inf. Sci.*, vol. 601, pp. 58–70, Jul. 2022.

[194] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang, "ISTVT: Interpretable spatial–temporal video transformer for deepfake detection," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1335–1348, 2023.

[195] G. Pang, B. Zhang, Z. Teng, Z. Qi, and J. Fan, "MRE-net: Multi-rate excitation network for deepfake video detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3663–3676, Aug. 2023.

[196] Z. Shang, H. Xie, L. Yu, Z. Zha, and Y. Zhang, "Constructing spatio-temporal graphs for face forgery detection," *ACM Trans. Web*, vol. 17, no. 3, pp. 1–25, May 2023.

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

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

[199] G. Mazaheri and A. K. Roy-Chowdhury, "Detection and localization of facial expression manipulations," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2773–2783.

[200] P. Chen, J. Liu, T. Liang, C. Yu, S. Zou, J. Dai, and J. Han, "DLFMNet: End-to-end detection and localization of face manipulation using multi-domain features," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[201] P. Yu, J. Fei, Z. Xia, Z. Zhou, and J. Weng, "Improving generalization by commonality learning in face forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 547–558, 2022.

[202] J. Wang, Y. Sun, and J. Tang, "LiSiam: Localization invariance Siamese network for deepfake detection," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2425–2436, 2022.

[203] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, "M2TR: Multi-modal multi-scale transformers for deepfake detection," in *Proc. Int. Conf. Multimedia Retr.* New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 615–623.

[204] S. Waseem, S. A. R. S. Abu-Bakar, Z. Omar, B. A. Ahmed, S. Baloch, and A. Hafeezallah, "Multi-attention-based approach for deepfake face and expression swap detection and localization," *EURASIP J. Image Video Process.*, vol. 2023, no. 1, p. 14, Aug. 2023.

[205] T. Kim, J. Kim, J. Kim, and S. S. Woo, "A face pre-processing approach to evade deepfake detector," in *Proc. 1st Workshop Secur. Implications Deepfakes Cheapfakes*. New York, NY, USA: Association for Computing Machinery, May 2022, pp. 35–38.

[206] Y. Huang, F. Juefei-Xu, R. Wang, Q. Guo, L. Ma, X. Xie, J. Li, W. Miao, Y. Liu, and G. Pu, "FakePolisher: Making DeepFakes more detection-evasive by shallow reconstruction," in *Proc. 28th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 1217–1226.

[207] M. Alauthman, N. Aslam, M. Al-kasassbeh, S. Khan, A. Al-Qerem, and K.-K. R. Choo, "An efficient reinforcement learning-based botnet detection approach," *J. Netw. Comput. Appl.*, vol. 150, Jan. 2020, Art. no. 102479.

[208] T. T. Nguyen and V. J. Reddi, "Deep reinforcement learning for cyber security," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 3779–3795, Aug. 2023.

[209] Y. Li, C. Zhang, P. Sun, L. Ke, Y. Ju, H. Qi, and S. Lyu, "DeepFake-o-meter: An open platform for DeepFake detection," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2021, pp. 277–281.

[210] S. J. Sohrawardi, A. Chintha, B. Thai, S. Seng, A. Hickerson, R. Ptucha, and M. Wright, "Poster: Towards robust open-world detection of deepfakes," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.* New York, NY, USA: Association for Computing Machinery, Nov. 2019, pp. 2613–2615.

[211] H. R. Hasan and K. Salah, "Combating deepfake videos using blockchain and smart contracts," *IEEE Access*, vol. 7, pp. 41596–41606, 2019.

SAIMA WASEEM was born in Quetta, Pakistan. She received the B.S. degree in computer system engineering from the Balochistan University of Information Technology Engineering and Management, Quetta, in 2010, and the M.Phil. degree in computer engineering from the National University of Science and Technology, Islamabad, Pakistan, in 2014. She is currently pursuing the Ph.D. degree with the Faculty of Electrical Engineering, Universiti Teknologi Malaysia. Her research interests include computer vision, medical image processing, multimedia forensics, and machine learning.

SYED ABDUL RAHMAN SYED ABU BAKAR (Senior Member, IEEE) received the B.Sc. degree in electrical engineering degree from Clarkson University, Potsdam, New York City, USA, the M.S.E.E. degree from Georgia Tech, and the Ph.D. degree from the University of Bradford, England. He has been with the Faculty of Electrical Engineering, Universiti Teknologi Malaysia, since 1992. He is currently a Full Professor with the Electronics and Computer Engineering Department. In 2004, he established the Computer Vision, Video and Image Processing Research Laboratory and has become the Head since then. He has published more than 150 scientific papers both at national and international levels. His main research interests include computer vision and image processing with video-based security and surveillance applications, medical image processing, and biometrics. In 2019, he received the Meritorious Regional Chapter Service Award from IEEE Signal Processing Society. He was the Chair of IEEE Signal Processing Society Malaysia Chapter, from 2014 to 2018.

BILAL ASHFAQ AHMED was born in Quetta, Pakistan. He received the B.E. degree in computer system engineering from the Balochistan University of Engineering and Technology, Khuzdar, Pakistan, in 2008, and the M.Sc. degree in computer engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2013. He is currently pursuing the Ph.D. degree with the Faculty of Computing, Universiti Teknologi Malaysia, Malaysia. He is also an Assistant Professor with the Department of Electrical Engineering, The University of Lahore, Pakistan (on study leave). His research interests include SDN, the IoT, and machine learning.

ZAID OMAR (Senior Member, IEEE) received the master's degree in data communications from The University of Sheffield and the Ph.D. degree in electrical engineering from the Imperial College London, in 2012. He is currently an Associate Professor with Universiti Teknologi Malaysia. His research interests include computer vision, medical imaging, and image fusion.

TAISEER ABDALLA ELFADIL EISA received the B.Sc. degree in computer science from the Sudan University of Science and Technology, the M.Sc. degree in computer science and information from Gezira University, Sudan, and the Ph.D. degree in computer science from Universiti Teknologi Malaysia. She is currently an Assistant Professor with King Khalid University, Saudi Arabia.

MHASSEN ELNOUR ELNEEL DALAM is currently an Assistant Professor with the Saudi Arabia Department of Mathematics-Girls Section, King Khalid University, Saudi Arabia.