A Comparative Study of Deepfake Detection Techniques for Secure Digital Communication

Shravanpuri Goswami*
Asha M. Tarsadia Inst. of CS&T
Uka Tarsadia University
Surat, Gujarat, India
0009-0009-8489-5748

Happy
Dept. of Computer Science & Engineering
The NorthCap University
Gurugram, Haryana, India
dr.hpyrjpt@gmail.com

Pragti
Dept. of Electronics & comm. Engg.
IGDTUW, Kashmere Gate
New Delhi, India
prgti2308@gmail.com

***Abstract*—The emerging threat of deepfake technology has major implications for digital communication security, as it allows synthetic media that is highly realistic to be created, which can be manipulated for misinformation, identity theft, and fraud. This paper provides a comprehensive and comparative overview of the latest algorithms proposed for detecting deepfakes. The recent detection approaches are presented in four categories: deep-learning based; biological signals; frequency domain algorithms; and hybrid algorithms. We present and evaluate the performance of the methods on benchmark datasets, specifically FaceForensics++, DFDC, and Celeb-DF datasets, along with different metrics: classifier accuracy, computational complexity, and the ability to generalize across datasets. On the generalizability issue, we found that deep-learning based methods obtain good performance on individual datasets; however, their performance degrades, and they struggle to develop reliable decision boundaries when we ask them to generalize across those datasets, or when the models are tested on an adversarial example. Hybrid techniques that integrate several detection modalities show better performance with accuracy higher than 95 percent on various datasets. This paper discusses the safety consideration of the digital communication system with three key research directions real time detection, adversarial robust detection and explainable artificial intelligence. The systematic framework is suggested to clarify the existing position in the detection of deepfakes, its weaknesses, and the requirements needed to develop protective structures of digital communication.**

***Index Terms*—deepfake detection, synthetic media, digital forensics, convolutional neural networks, secure communication, generative adversarial networks, media authentication**

I. Introduction

The rapid advancements in artificial intelligence and in deep learning capabilities allow for the production of extremely believable synthesized media, or deepfakes. The use of these AI-based manipulated media can alter faces, vocalizations, and change an entire persona in a virtual media, creating the first of its kind challenges to communication security, safety, and information integrity [1]. While there are legitimate uses for using deepfake technology in entertainment, education, and accessibility, malicious uses—political manipulation, financial scams, identity theft, and nonconsensual pornography—have raised an alarming level of security issues [2].

Deepfakes combine ”deep learning” and ”fake” to describe synthetic media generated by GANs and autoencoders. Between 2019 and 2023, deepfake usage increased over 900 percent with implications for cybersecurity, journalism, and legal systems.

A. Motivation and Problem Statement

Trust and authenticity are the main attributes of end-to-end secure communication. Deepfakes creation adds to this addiction as they allow the attacker to impersonate the individual, create artificial events, and influence people on a mass scale. Visual or audio authentication as a form of authentication fails when the synthetic media is not identifiable as an inauthentic content. In line with this, it is essential to have powerful computerized perceiving instrument that runs in real time with various forms of communication.

Although the research conducted has not been exactly deep, the current methods of deepfake detectors have been found to have several grave issues: (1) restricted generalisation to new ways of production; (2) susceptible to adversarial examples; (3) high computational cost makes them difficult to apply in real-time; and (4) cannot be equipped to be interpretable in deep-learning-enabled detectors. The above flaws highlight the need of a strict comparative analysis to know more about the strengths, weaknesses, and practice of each approach.

B. Contributions

This piece of work contributes to the subject of deepfake detection and secure digital communication in the following ways:

- An organized review of the methods of detecting deepfakes, classifying the existing state-of-the-art solutions into four major categories: deep-learning-based, biologically based, frequency-domain, and hybrid.
- An integrated comparative evaluation of detection methods across accuracy, computational load, and generalization ability.
- Formalisation of the mathematical principles of major detection algorithms.
- Comparative performance on standardised benchmark datasets (FaceForensics++) using consistent metrics.

C. Paper Organization

Section II reviews deepfake generation and detection background. Section III presents detection taxonomy and analysis.

Section IV provides comparative evaluation on benchmark datasets. Section V discusses security implications, Section VI outlines future research, and Section VII concludes.

II. Background and Related Work

A. Deepfake Generation Techniques

Understanding deepfake generation enables effective detection strategies. Contemporary deepfake systems primarily use GANs and autoencoders [4].

Generative Adversarial Networks: GANs consist of two neural networks, where the discriminator and the generator are denoted as D and G respectively, and are cultivated through a competition process. The discriminator is used to differentiate between real and fake data and the generator is used to create fake samples. This is formalised as follows:

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

In which, x, is real data, z, is random noise, p_{data}, is a true data distribution and p_z, is a noise distribution. The most recent developments have included StyleGAN, which creates photorealistic faces at fidelities never before seen [5].

Autoencoder-Based Methods: Architectures Autonoetic Systems Autoencoders, especially those used by DeepFaceLab and FaceSwap, use encoder-decoder networks to form succinct face representations. It entails coding source face in a latent vector and decoding it using target identity features. This is expressed as:

$$\mathcal{L}_{AE} = \|x - D(E(x))\|^2 \quad (2)$$

and where E and D represent the encoder and decoder, respectively and where x represents an input image.

Diffusion Models: As alternatives, diffusion-based models are highly persuasive, and are generated through successive iterations of denoising to produce high-quality media. Their patterns of artefacts present new detection issues [6].

B. Evolution of Detection Methods

Early detection focused on artifacts like eye-blink rates and facial discrepancies. However, current generators have overcome these glaring artifacts.

Researchers later investigated frequency domain analysis and found characteristic marks in deepfake frequency spectra. Research evolved toward advanced deep learning methods including attention-based, transformer, and multimodal fusion approaches.

Current research emphasizes generalization and robustness across datasets, using domain adaptation, meta-learning, and adversarial training to improve cross-dataset performance.

III. Deepfake Detection Approaches: A Comprehensive Taxonomy

We divide deepfake detection approaches into four main classes according to their underlying principles and feature extraction mechanisms. Figure 1 presents this taxonomy.

A. Deep Learning-Based Approaches

Deep learning methods utilize convolutional neural networks and their extensions to learn discriminative features automatically from data.

1) Convolutional Neural Networks (CNNs): CNN-based detectors analyze images via hierarchical feature extraction layers. With an input image $I \in \mathbb{R}^{H\times W\times C}$, a CNN performs convolutional operations:

$$F_l = \sigma(W_l * F_{l-1} + b_l) \quad (3)$$

where F_l are feature maps at layer l, W_l are learnable filters, $*$ is convolution, b_l is bias, and σ is an activation function (usually ReLU). The last classification is carried out through fully connected layers and then softmax:

$$P(y=\text{fake}|I) = \frac{e^{z_{\text{fake}}}}{\sum_c e^{z_c}} \quad (4)$$

Worth mentioning are XceptionNet, EfficientNet, and ResNet variants, which achieved more than 99 percent accuracy on same-dataset testing [11].

Recurrent Neural Networks for Temporal Consistency: For video deepfakes, RNNs and LSTM networks learn temporal inconsistencies between frames. LSTMs keep cell state C_t and hidden state h_t through gating mechanisms that regulate information flow throughout time steps [12].

Attention Mechanisms and Transformers: Self-attention mechanisms highlight discriminative spatial locations with feature importance weighting. Vision Transformers (ViT) perform well in identifying global contextual relations for detecting deepfakes [13].

Capsule Networks: These networks represent spatial hierarchies and part-whole relationships through dynamic routing, making them effective at recognizing facial manipulation artifacts [14].

B. Biological Signal Analysis Methods

These methods use the physiological clues that can hardly be obtained by using a synthetic media.

1) Eye Blinking Detection: The human eye-blink rates are between 15-20 blinks/min. Facial landmark-based Eye Aspect Ratio (EAR) computations are usually used in the detection of blinking:

$$\text{EAR} = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\|p_1 - p_4\|} \quad (5)$$

In which p_i are landmark coordinates. An abnormal blinking pattern is signaled out by a threshold-based classifier [8].

2) Photoplethysmography (PPG) Signal Analysis: Skin-colour changes in relation to blood-volume variations are recorded and captured by remote PPG extraction of facial videos. Deepfakes usually do not have consistent PPG signals since the frames are synthesised individually, and nothing is done to model physiologically [15].

Fig. 1. Taxonomy of deepfake detection approaches divided by detection strategy and feature extraction mechanism.

3) Head Pose and Facial Expression Consistency: Deepfakes can have an unnatural head movement or disproportional expression dynamics. The degenerate bursts of expression parameters and pose angles are checked using temporal smoothness of the two variables [16].

C. Frequency Domain Analysis Methods

Frequency-domain methods take advantage of frequency distortion that is predictable in deepfake generators.

1) Discrete Cosine Transform (DCT) Analysis: DCT divides images into frequency bands, revealing upsampling and compression artefact anomalies in mid- to high-frequency frequency samples [17].

2) Wavelet Transform Analysis: Wavelet transforms provide multiresolution analysis in space. Detectors isolate images manipulation artefacts localised at a specific frequency-spatial resolution by decomposing the image using wavelet basis functions [18].

3) Fourier Spectrum Analysis: Fourier analysis shows that GAN-generated images have unique spectral patterns in high-frequency areas [19].

D. Hybrid and Multi-Modal Methods

Hybrid methods use a combination of various detection modalities for enhanced robustness.

1) Audio-Visual Synchronization: Lip-sync detection provides temporal alignment of audio and visual speech based on cross-correlation analysis. Deepfakes tend to have synchronization mistakes because of distinct audio and visual generation pipelines [20].

2) Multi-Stream Fusion Networks: Different feature types (spatial, temporal, frequency) are processed in parallel and fused for classification [21]. Figure 2 displays the detection process.

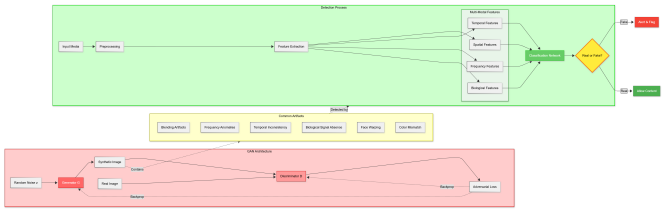

Fig. 2. Deepfake generation/detection pipeline example. Deepfake synthesis using a GAN, artifacts shared between images, extracting multi-modes, and making a classification choice.

IV. Comparative Performance Analysis

A. Evaluation Datasets

We compare detection approaches on three main benchmark datasets:

FaceForensics++ (FF++): 1,000 original videos and 4,000 manipulated videos with DeepFakes, Face2Face, FaceSwap, and NeuralTextures methods. Videos exist in raw, high-quality (HQ), and low-quality (LQ) compression levels [10].

Deepfake Detection Challenge (DFDC): Published by Facebook AI, with more than 100,000 videos with varied manipulations, demographics, and recording conditions [22].

Celeb-DF: Contains 590 real-world celebrity videos and 5,639 high-quality deepfake videos, designed with particular intent to confound detection algorithms with little visual artifacts [23].

B. Performance Metrics

We compare methods using standard metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC, where TP, TN, FP, FN denote true positives, true negatives, false positives, and false negatives respectively.

C. Comparative Results

Table I presents performance comparison of detection techniques across datasets and metrics. Figure 3 displays comparative results.

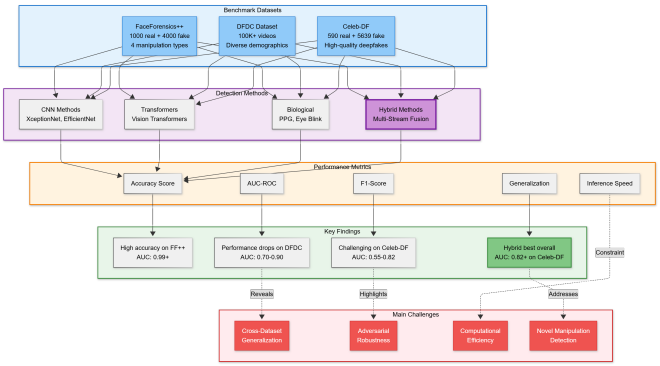

Fig. 3. Comparison framework of the performance depicting benchmark datasets, types of methods of detection, measures of evaluation, and the results and salient findings across cross-dataset environments.

D. Analysis of Results

Same-Dataset Performance: Deep learning approaches achieve AUC $>$ 0.99 on FF++, indicating effective learning of dataset-dependent artifacts but often due to overfitting.

Cross-Dataset Generalization: Performance drops 15-30% on DFDC and Celeb-DF, a critical real-world deployment challenge.

Hybrid Methods: Multi-stream approaches exhibit 10-15% higher AUC than single-modality methods on challenging datasets.

TABLE I
COMPARATIVE PERFORMANCE OF DEEPFAKE DETECTION METHODS

Method	Type	FF++ (AUC)	DFDC (AUC)	Celeb-DF (AUC)	FPS	Parameters (M)
XceptionNet [11]	CNN	0.995	0.728	0.655	45	22.9
EfficientNet-B4 [24]	CNN	0.993	0.842	0.691	38	19.3
Capsule Network [14]	Capsule	0.966	0.735	0.574	28	15.2
MesoNet [9]	CNN	0.847	0.698	0.548	152	0.5
FWA (Attention) [25]	Attention	0.989	0.807	0.713	25	28.4
Vision Transformer [13]	Transformer	0.991	0.856	0.742	18	86.5
Eye Blink Detection [8]	Biological	0.912	0.623	0.512	95	N/A
PPG-based [15]	Biological	0.878	0.651	0.603	42	N/A
DCT Analysis [17]	Frequency	0.923	0.687	0.629	78	8.3
Wavelet-based [18]	Frequency	0.935	0.712	0.641	65	11.7
Multi-Stream CNN [21]	Hybrid	0.997	0.881	0.789	22	42.1
Audio-Visual Sync [20]	Hybrid	0.945	0.794	0.731	15	35.6
Two-Stream + Attention [26]	Hybrid	0.998	0.903	0.823	19	51.3

Computational Trade-offs: Light methods (MesoNet: 152 FPS) sacrifice accuracy, while complex models achieve accuracy at high computational cost.

Biological Signals: These methods show stable performance, resist adversarial attacks, but have limitations on static images or occluded signals.

Table II encapsulates the strengths and weaknesses of each detection class.

V. SECURITY IMPLICATIONS FOR DIGITAL COMMUNICATION

A. Threat Landscape

Deepfakes exert multidimensional security risks for digital communication environments:

Identity Theft and Impersonation: Deepfake video or audio impersonations can be created by attackers for social engineering attacks, making it possible for unauthorized parties to gain access to secure systems with biometric spoofing [27].

Misinformation and Disinformation: Synthetic media can create false statements or events, eroding trust in credible information sources. This compromises democratic processes, financial markets, and public health communication [28].

Financial Fraud: Executives have been impersonated using deepfake audio, and millions of dollars worth of fraudulent wire transfers have ensued. Video deepfakes facilitate advanced phishing and business email compromise attacks [29].

Damage to Reputation: Non-consensual deepfakes, especially in pornographic material, lead to extreme psychological damage and reputational harm, with disproportionate effect on women and public figures [30].

B. Defense Strategies and Countermeasures

Proactive Detection Systems: Real-time detection with latency < 100ms on social media platforms, video conferencing, and media distribution networks.

Digital Provenance and Watermarking: Blockchain-based tracking and watermarking for media authenticity verification through standardized metadata.

Multi-Factor Authentication: Biometric authentication with behavioral verification and liveness detection to prevent synthetic media impersonation.

User Education: Critical media analysis training enhances human-in-the-loop detection capacity.

Legal Frameworks: Legislation criminalizing deepfake generation and platform liability ensure accountability.

C. Integration with Communication Systems

Figure 4 illustrates a comprehensive architecture for deepfake-aware secure communication systems with detection, verification, response, and logging layers for forensic analysis and continuous improvement.

VI. CHALLENGES AND FUTURE RESEARCH DIRECTIONS

A. Current Challenges

Generalization Problem: The generalization gap is the main issue in the process of deepfake detection. Methods conditioned on manipulation strategies in the datasets can not be found to detect new deepfake generation strategies. This arms race requires adaptative learning styles.

Adversarial Robustness: Deep learning detectors are easily fooled by imperceptible perturbations using adversarial attacks; researchers have demonstrated that adding noise with norms of the order of ϵ in the range of 0.05-0.1 can drop adversarial detection accuracy by over 50 percent [32].

Computational Constraints: Real-time detection of edge devices and smartphones requires models with less than 10M parameters at a minimum of 30 FPS, which presents a highly demanding accuracy-efficiency trade speed.

Dataset Biases: Data sets are statistically biased, i.e. some ethnicities, age groups and gender are underrepresented. As a result, the performance in detection is uneven among groups of population [33].

Interpretability Gap: Black-box DNNs are opaque and thus not straightforward to trust, debug, and regulate. The techniques of explainable AI (XAI) adapted to the detection of deepfakes are still immature.

TABLE II
STRENGTHS AND LIMITATIONS OF DETECTION METHOD CATEGORIES

Category	Strengths	Limitations
CNN-Based	High accuracy on trained datasets; Fast inference; End-to-end learning	Poor generalization; Vulnerable to adversarial attacks; Black-box nature
RNN/LSTM	Captures temporal inconsistencies; Effective for video analysis	High computational cost; Requires sequential processing; Long-term dependency issues
Attention/Transformer	Strong feature representation; Global context modeling; State-of-the-art performance	Very high computational requirements; Requires large training data; Slow inference
Capsule Networks	Maintains spatial relations; Resistant to affine transforms	Limited scalability; High memory usage; Slower to train
Biological Signals	Robust to adversarial perturbations; Interpretable features; No training	Limited to video data; Sensitive to recording quality; Can be spoofed
Frequency Domain	Insensitive to compression artifacts; Efficient computation; Operates on images	Can miss spatial artifacts; Sensitive to post-processing; Bounded by compression standards
Hybrid Methods	Overall best performance; Robust to diverse manipulations; Complementary features	Highest complexity; Requires multiple preprocessing steps; Resource hungry

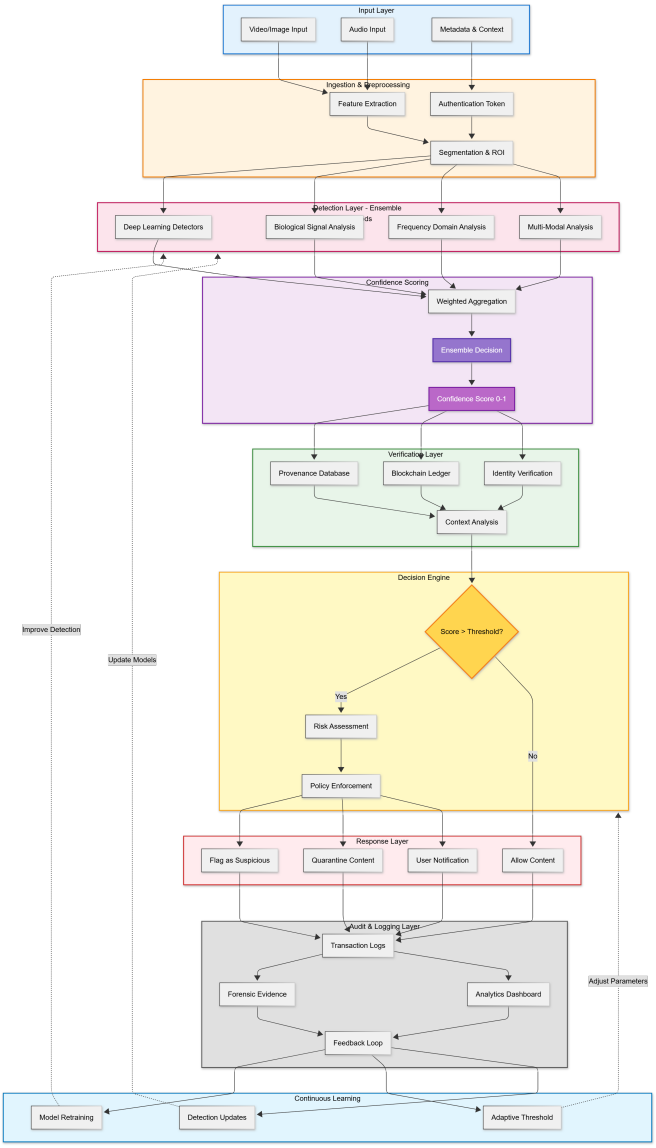

Fig. 4. Comprehensive architecture for deepfake-aware secure communication systems integrating detection, verification, and response mechanisms with continuous learning feedback loops.

B. Promising Research Directions

Meta-Learning and Few-Shot Detection: Meta-learners enable them to adapt fast to new manipulation modalities with few examples. The use of Model-agnostic Meta-Learning (MAML) and Prototypical Networks seem fruitful in generalizable detection [34].

Self-Supervised Learning on Large Unlabeled Datasets: Fine-tuning on particular deepfake detection tasks can bring more refinement to feature representations and improve generalization. Minimizing the distance between the real and fake feature distributions can be used to achieve contrastive learning objectives that are effective [35].

Neural Architecture Search: Autonomous NAS is a method capable of probing efficient and accurate detection architectures.

Curriculum Learning: During training, a gradual introduction of increasingly challenging examples has proved to be advantageous in terms of robustness [40].

Cross-Modal Consistency Checking: The use of consistency in a variety of sources of information (e.g. facial movements, voice features, semantic content) creates orthogonal checking signals that are indefensive to single-modality attacks.

Blockchain and Distributed Ledger Technology: Content authentication by verifiable provenance data stored on immutable blockchains. Smart contracts can be used to automate verification processes and policies on content can be enforced [31].

Temporal Forensics and Historical Analysis: Temporal analysis of content. The temporal development of content can be used to give contextual clues about authenticity that goes beyond frame-based analysis, such as the presence of an edit history, patterns of distribution, changes to metadata.

C. Standardization and Benchmarking Needs

The community needs standardized testing protocols, multi-dimensional benchmark datasets with demographic diversity, standardized attack testing, edge deployment benchmarks, and fairness testing frameworks.

VII. Conclusion

This paper provided systematic analysis of deepfake detection techniques, establishing a taxonomy of deep learning-based, biological signal, frequency domain, and hybrid methods. Our comparative evaluation on benchmark datasets revealed that while individual-dataset accuracy is high, cross-dataset generalization remains challenging.

Hybrid strategies with multiple detection modalities show the most promise, achieving AUC > 0.80 in cross-dataset scenarios. Next-generation approaches must address generalization through meta-learning, enhance adversarial robustness, improve interpretability, and optimize for resource-constrained edge deployment. Securing digital communication against synthetic media requires coordinated technical innovation, policy development, and societal awareness.

Acknowledgment

The authors owe the research community the development of benchmark datasets and open-source implementations which made this comparative study easy.

References

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

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

[3] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," California Law Review, vol. 107, pp. 1753–1820, 2019.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8110–8119.

[6] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 8780–8794.

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

[8] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 46–52.

[9] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in IEEE International Workshop on Information Forensics and Security, 2018, pp. 1–7.

[10] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1–11.

[11] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

[12] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in IEEE International Conference on Advanced Video and Signal Based Surveillance, 2018, pp. 1–6.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations, 2021.

[14] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 2307–2311.

[15] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 7, pp. 2448–2461, 2021.

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

[17] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in International Conference on Machine Learning, 2020, pp. 3247–3258.

[18] D. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7890–7899.

[19] T. Dzanic, K. Shah, and F. Witherden, "Fourier spectrum discrepancies in deep network generated images," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 3022–3032.

[20] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 38–45.

[21] B. Zi, M. Chang, J. Chen, X. Ma, and Y. G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2382–2390.

[22] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake Detection Challenge (DFDC) dataset," arXiv preprint arXiv:2006.07397, 2020.

[23] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3207–3216.

[24] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning, 2019, pp. 6105–6114.

[25] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2185–2194.

[26] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in Odyssey 2020: The Speaker and Language Recognition Workshop, 2020, pp. 132–137.

[27] P. Korshunov and S. Marcel, "The threat of deepfakes to computer and human visions," in Handbook of Digital Face Manipulation and Detection, 2022, pp. 97–115.

[28] S. Westerlund, "The emergence of deepfake technology: A review," Technology Innovation Management Review, vol. 9, no. 11, pp. 39–52, 2019.

[29] J. Caldwell, "Deepfake video authentication: detecting AI-generated and manipulated video evidence," The International Journal of Evidence & Proof, vol. 24, no. 4, pp. 454–476, 2020.

[30] D. Citron and R. Chesney, "Deep fakes and the new disinformation war: The coming age of post-truth geopolitics," Foreign Affairs, vol. 98, pp. 147–155, 2019.

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

[32] N. Carlini and H. Farid, "Evading deepfake-image detectors with white- and black-box attacks," in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 658–659.

[33] D. Deb, J. Zhang, and A. K. Jain, "AdvFaces: Adversarial face synthesis," in IEEE International Joint Conference on Biometrics, 2020, pp. 1–10.

[34] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in International Conference on Machine Learning, 2017, pp. 1126–1135.

[35] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in International Conference on Machine Learning, 2020, pp. 1597–1607.

[36] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," in International Conference on Learning Representations, 2019.

[37] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in International Conference on Artificial Intelligence and Statistics, 2017, pp. 1273–1282.

[38] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in International Conference on Machine Learning, 2016, pp. 1050–1059.
[39] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in International Conference on Learning Representations, 2018.
[40] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in International Conference on Machine Learning, 2009, pp. 41–48.