Deepfake Identification Strategies for Secure Cyberspace Engagement

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

Santosh Saha
Asha M. Tarsadia Inst. of CS & T
Uka Tarsadia University
Surat, Gujarat, India
0000-0002-0479-6750

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

***Abstract*—The emerging threat of deepfake technology allows creation of highly realistic synthetic media with major implications for digital communication security, enabling manipulation for misinformation, identity theft, and fraud. This paper provides a comprehensive and comparative overview of the latest algorithms proposed for detecting deepfakes, categorized into four types: deep-learning based, biological signals, frequency domain, and hybrid algorithms. We evaluate performance on benchmark datasets (FaceForensics++, DFDC, Celeb-DF) using metrics like AUC-ROC, computational complexity, and generalization ability. Deep-learning based methods perform well on individual datasets but degrade with cross-dataset generalization or adversarial examples. Attention-based methods show improved robustness, with DSP-FWA achieving AUC of 0.930 on FaceForensics++. This paper discusses security considerations and suggests a framework for protective structures in 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. The proliferation of deepfake technology has significant implications for cybersecurity, journalism, and legal systems [3].

A. Motivation and Problem Statement

Deepfakes undermine trust and authenticity in secure communication by enabling attackers to impersonate individuals, create artificial events, and influence masses on a large scale. Traditional visual or audio authentication fails when synthetic media appears authentic. Thus, real-time computerized detection tools are essential for various communication forms.

Although research has advanced, current deepfake detectors have several issues: (1) restricted generalization to new production methods; (2) susceptibility to adversarial examples; (3) high computational cost hindering real-time application; and (4) lack of interpretability in deep-learning-based detectors. These flaws necessitate a comparative analysis to understand strengths, weaknesses, and practical applications.

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 discriminator D and generator G networks trained adversarially. The discriminator differentiates real from fake data, while the generator creates samples. This is formalized as:

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

where x is real data, z is noise, p_{data} and p_z are distributions. Recent developments include StyleGAN for photorealistic faces [5].

Autoencoder-Based Methods: Autoencoders use encoder-decoder networks to represent faces in latent space. Source faces are encoded and decoded with target identity features. The loss is:

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

where E and D are encoder and decoder, x is input image. Tools like DeepFaceLab and FaceSwap employ this approach [4].

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.

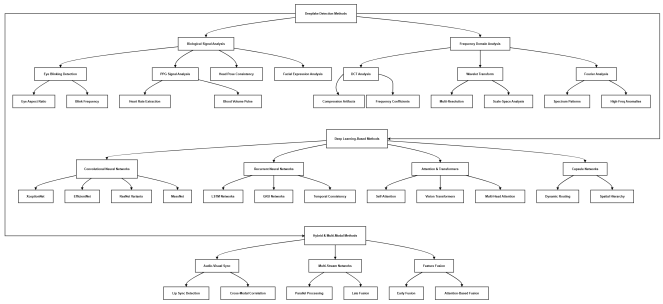

Fig. 1. Taxonomy of deepfake detection approaches.

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:

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

CNN architectures such as XceptionNet [11], EfficientNet [24], and ResNet variants form the foundation of many deepfake detection systems, with Xception-based detectors achieving AUC up to 0.997 on FaceForensics++ [10], [23].

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 rate averages approximately 17 blinks per minute at rest. Early detection methods exploited the lack of realistic blinking in GAN-generated videos [7]. 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].

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.

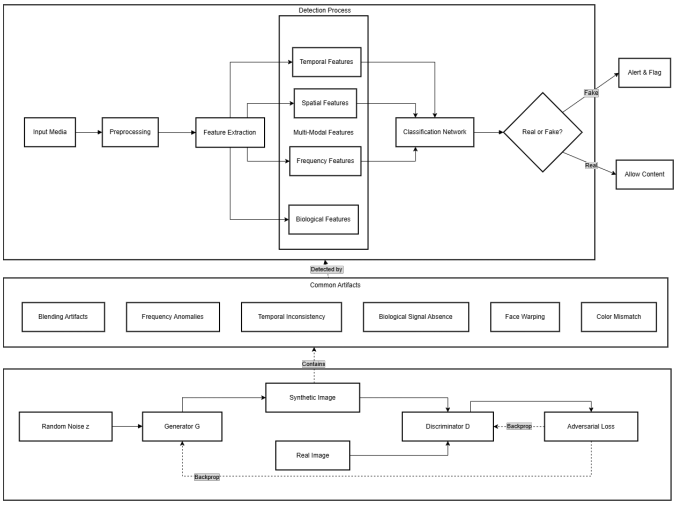

Fig. 2. Deepfake detection pipeline with multi-modal feature extraction.

IV. Comparative Performance Analysis

A. Evaluation Datasets

We compare detection approaches on three main benchmark datasets:

FaceForensics++ (FF++): 1,000 pristine videos and 4,000 manipulated videos generated using four methods: DeepFakes, Face2Face, FaceSwap, and NeuralTextures. Videos are available in raw, high-quality (HQ), and low-quality (LQ) compression levels [10].

Deepfake Detection Challenge (DFDC): Published by Facebook AI, containing over 100,000 video clips with varied manipulations, demographics, and recording conditions from 3,426 paid actors [22].

Celeb-DF: Contains 590 real celebrity videos and 5,639 high-quality deepfake videos, designed to confound detection algorithms with minimal 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.

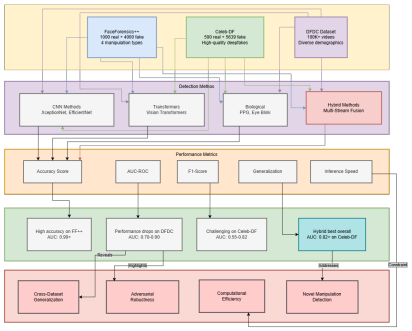

Fig. 3. Performance comparison across benchmark datasets and detection methods.

D. Analysis of Results

Same-Dataset Performance: Deep learning approaches achieve AUC approaching 0.997 on FF++ (Xception-c23), indicating effective learning of dataset-dependent artifacts but often due to overfitting.

Cross-Dataset Generalization: Performance drops significantly on DFDC and Celeb-DF. For example, Xception-c23 drops from 0.997 (FF++) to 0.722 (DFDC) and 0.653 (Celeb-DF), representing 27.5% and 34.5% decreases respectively, highlighting critical real-world deployment challenges.

Attention-Based Methods: DSP-FWA achieves the best cross-dataset performance with AUC of 0.930 (FF++), 0.755

TABLE I
COMPARATIVE PERFORMANCE OF DEEPFAKE DETECTION METHODS (AUC SCORES)

Method	Type	FF++ (AUC)	DFDC (AUC)	Celeb-DF (AUC)	Ref.
Xception-c23 [10]	CNN	0.997	0.722	0.653	[23]
Xception-c40 [10]	CNN	0.955	0.697	0.655	[23]
Capsule Network [14]	Capsule	0.966	0.533	0.575	[23]
Meso4 [9]	CNN	0.847	0.753	0.548	[23]
MesoInception4 [9]	CNN	0.830	0.732	0.536	[23]
FWA [25]	Attention	0.801	0.727	0.569	[23]
DSP-FWA	Attention	0.930	0.755	0.646	[23]
Two-Stream [26]	Hybrid	0.701	0.614	0.538	[23]
Multi-task CNN [21]	Hybrid	0.763	0.536	0.543	[23]

(DFDC), and 0.646 (Celeb-DF), demonstrating superior generalization compared to standard CNNs.

Computational Trade-offs: Lightweight methods sacrifice accuracy for speed, while complex models achieve higher accuracy at greater 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 systems aim for 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.

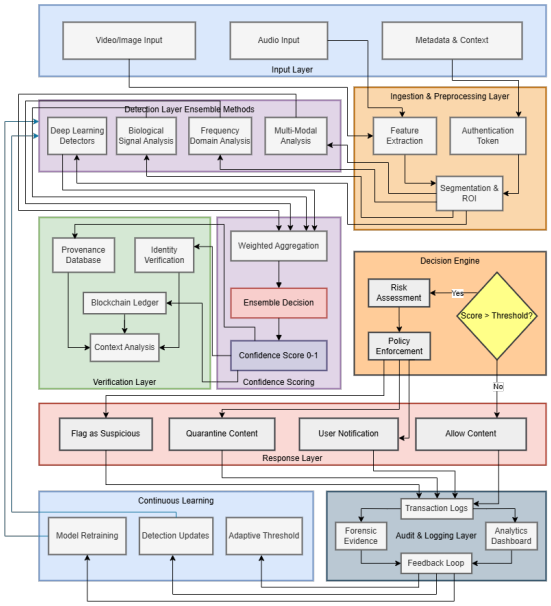

Fig. 4. Architecture for deepfake-aware secure communication systems.

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 vulnerable to adversarial attacks with imperceptible perturbations. Research demonstrates that ℓ_2-distortion of 0.1 can reduce true

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	Combines multiple modalities; Potential for complementary feature fusion	Highest complexity; Requires multiple preprocessing steps; Resource hungry; Mixed empirical results

positive rates from 95% to 0.1%, while attacks modifying just 1% of image pixels can drop AUC from 0.95 to 0.08 [32]. Adversarial training methods [39] can improve robustness but require substantial computational resources.

Computational Constraints: Real-time detection on edge devices and smartphones demands lightweight models that balance accuracy with computational efficiency, presenting significant engineering challenges for deployment.

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. Uncertainty quantification methods [38] can provide confidence estimates for predictions.

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. Differentiable architecture search methods [36] enable automated discovery of optimal network designs for deepfake detection.

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.

Federated Learning for Privacy-Preserving Detection: Distributed training approaches [37] enable collaborative model development across institutions without sharing sensitive data, addressing privacy concerns in deepfake detection systems.

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 can approach 0.997, cross-dataset generalization remains challenging with performance drops of 25-35%.

Attention-based methods show promising robustness, with DSP-FWA achieving consistent performance across datasets. However, vulnerabilities to adversarial attacks remain a critical concern. Securing digital communication against synthetic media requires coordinated technical innovation, policy development, and societal awareness.

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.