A Study on Deepfake Detection with Continual Learning

Nekkanti Mownika[1], N. Swapna Goud[1], and Y. V. R. Naga Pawan[2]

[1] School of Engineering, Anurag University, Hyderabad, India

[2] Department of CSE, Anurag Engineering College, Kodad, India

24eg305a24@anurag.edu.in[1], swapnagoudcse@anurag.edu.in[1], ynpawan@gmail.com[2]*

Abstract:

The rapid advancement of deepfake technology has made it increasingly challenging to distinguish between real and manipulated digital content, raising significant concerns for public trust, privacy, and security. While detection methods have progressed to identify forgeries generated by GANs and diffusion models, most approaches rely on static and homogeneous training data, which does not reflect the evolving landscape of deepfake generation. Although several review studies exist, there is no comprehensive survey providing a systematic overview with unified evaluation metrics. Continual deepfake detection, which aims to adapt to new manipulations without forgetting previously learned knowledge, has received limited attention and faces challenges such as taskidentification dependencies and computational overhead. This survey consolidates existing research on detection methods, continual learning strategies, datasets, and associated challenges. It provides a taxonomy of approaches and highlights open problems to guide the development of robust, adaptable, and future-proof deepfake detection systems.

Keywords: Continual Learning, Catastrophic Forgetting, Deepfake Detection, Generative Models, Multi-modal Analysis, Prompt-based Optimization.

1 Introduction

Our digital lives have come a long way, from just capturing special moments to creating realistically rendered synthetic ones. One of the most interesting and divisive outcomes of this evolution is the emergence of *deepfakes*. The term combines "deep learning" and "fake", referring to AI-generated content such as images, videos, text, and audio that convincingly blur the boundary between reality and fabrication. and offers promising applications in fields like accessibility, entertainment, and education. Although deepfake offers promising applications in fields like accessibility, entertainment, and education. However, their misuse has raised severe concerns. These

D. R. Reddy et al. (eds.), *Proceedings of the First International Conference on Advances in Forensics and Cyber Technologies (ICFACT 2025)*, Advances in Computer Science Research 127,
https://doi.org/10.2991/978-94-6239-610-4_31

include the spread of misinformation, identity theft, fraud, cyberbullying, and political propaganda. These threats not only endanger individual privacy and security but also undermine public trust and the integrity of digital media. Given these emerging threats, the development of robust deepfake detection methodologies constitutes a research priority with significant practical implications. Such detection frameworks serve not only as technical countermeasures but also as essential safeguards for individual privacy rights, institutional integrity, and broader social cohesion.

Deepfakes are generated or manipulated using advanced deep learning techniques, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), variable autoencoders (VAEs), and diffusion models (DMs). Detecting deepfake content remains a research challenge due to the high realism of the generated outputs and the continuous evolution of manipulation techniques, including face-swapping, lipsyncing, puppeteering, voice conversion, and natural language processing (NLP) [1]. The widespread availability of user-friendly deepfake tools such as DeepfakesWeb, DeepFaceLab, FaceApp, ChatGPT, DALL-E2, and Midjourney has enabled even non-expert users to produce highly convincing synthetic content[2]. Although this democratization of generative technology has opened up creative avenues, it also poses significant challenges for security, privacy, and public trust. In recent years, deepfake content has increased at an alarming rate. According to CyberNews[1], the number of fake videos online doubled every six months between 2018 and 2020, growing from approximately 8,342 in mid-2018 to over a 1 million in April 2021. More recent estimates indicate that by 2023, nearly 500,000 fake videos were circulating on social networks. As illustrated in Fig. 1, this trend is projected to escalate dramatically, with the volume of deepfake videos expected to reach nearly a 8 million by 2025, a trajectory that underscores the continued exponential growth of this phenomenon.

Deepfake detection methods have been the subject of extensive investigation. The majority of current research creates deepfake detectors in static environments using uniformly huge datasets. However, as deepfakes continue to develop in increasingly complex and varied forms, new techniques must be developed on a continuous basis. Deepfakes actually constantly appear in practice using a variety of hidden designs, making detection techniques a dynamic problem. The issue of *catastrophic forgetting* must be mitigated by detectors that can adjust to new forgeries while maintaining the knowledge of earlier ones. This problem has not been thoroughly examined in the literature, despite its importance, which is why this survey is necessary. We specifically analyse the state-of-the-art detection techniques, investigate how constant learning can help with catastrophic forgetting, and point out unresolved issues to direct future studies.

[1] https://cybernews.com/cyber-pros-spooked-by-deepfake-statistics/

1.1 Motivation

The dynamic nature of deepfakes necessitates detection techniques that transcend static training assumptions. Models may gradually learn from novel and invisible modifications with continuous deepfake detection, all without sacrificing performance on more established forgeries. There is still a dearth of research

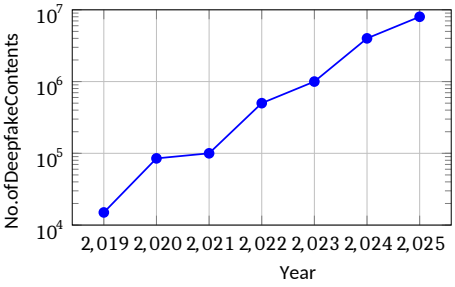

Fig.1. Estimated Growth of Deepfake Content (2019–2025).

in this field, and current methods, particularly those that use vision-language models, frequently run into problems like task-identification dependencies and high computing costs. By offering a thorough analysis of detection techniques, looking at continual learning approaches, evaluating datasets, and pointing out unresolved issues, this survey seeks to close this gap. We hope to direct the creation of reliable, flexible, and future-proof deepfake detection systems by compiling current knowledge and offering a structured taxonomy.

1.2 Scope of the Survey

This subsection reviews state-of-the-art (SOTA) surveys on deepfakes. Table. 1 provides a comparative analysis between our survey and existing surveys. This survey provides a comprehensive review of deepfake generation and detection research in text, image, video, and audio modalities, addressing notable gaps in previous surveys. Prior work has largely focused on facial manipulations and video deepfakes while underrepresenting detailed examination of emerging challenges such as catastrophic forgetting and continual learning, which are critical for developing adaptive and robust detection systems in dynamic real-world scenarios.

1.3 Survey Contributions

The key contributions of this survey are as follows: (1) A Unified Generation Models: GANs, CNNs, and Diffusion Models. (2) A Comparative Analysis of Generation &

Detection Methods - A comprehensive review and comparative analysis of state-of-the-art detection methods based on their precision, efficiency, and scalability. (3) Addressing Catastrophic Forgetting – An analysis of catastrophic forgetting in detection models and strategies for continual learning. (4) Taxonomy of Challenges and Open Problems – A new taxonomy of key challenges, including dataset gaps, adversarial robustness, and generalization of cross-modalities, to guide future research.

Table 1. Comparative Analysis of Existing Deepfake Surveys

Author	Year	Strengths	Limitations
Yadav et al. [80]	2019	Discussed benefits and drawbacks of deepfakes and their generation methods.	Limited coverage on detection; no taxonomy, audio detection, or dataset listing.
Ramadhani et al. [81]	2020	Presented a taxonomy focused on video deepfake detection; briefly mentioned datasets.	Audio detection not addressed.
Katarya et al. [82]	2020	Covered image, video, and audio deepfake detection concisely.	Lacked detailed explanations and dataset list.
Tolosana et al. [18]	2020	Detailed overview of facial manipulations and detection methods.	Limited to facial content; audio and datasets not covered.
Yu et al. [83]	2020	Examined deepfake detection techniques in depth.	Brief coverage of datasets; no audio detection.
Mirsky et al. [84]	2020	Focused on deepfake generation techniques.	Detection, taxonomy, datasets, and audio not fully discussed.
Nguyen et al. [85]	2021	Discussed generation tools, detection taxonomy, and video deepfakes.	Audio detection and datasets missing.
Abdulreda et al.	2022	Extensive focus on	Limited coverage of

[86]		facial manipulation techniques.	detection, audio, and datasets.
Masood et al. [87]	2022	Emphasized ML-based detection for audio and video.	Did not cover adversarial training foundations.
Tao Zhang [88]	2022	Analyzed datasets and benchmarking challenges.	Lacked GAN-based adversarial training details.
Patil et al. [89]	2023	Highlighted biological classifiers and distance metrics in detection.	No taxonomy of classifiers.
Dhesi et al. [90]	2023	Explored adversarial learning approaches.	Limited explanation of perturbation generation.
Yogesh et al. [91]	2023	Comprehensive taxonomy for audio and video detection, datasets, and features.	Did not cover signal-level features and transfer learning.
Gan Pei et al. [1]	2024	Up-to-date survey with structured benchmarking across applications.	Limited detail on ethics, challenges, and dataset diversity.
This survey	2025	Covers single- and multimodal detection, including continual learning and catastrophic forgetting.	-

1.4 Organization of the Survey

This survey is organized as follows. In Section 2, we present a detailed taxonomy of deepfake contents, covering different types of manipulations including image, video, audio, and multimodal forgeries. Section 3 reviews the deepfake detection methods, categorizing them based on features, modalities, and learning approaches, from classical techniques to modern deep learning solutions. In Section 4, we discuss catastrophic forgetting and continual learning, highlighting challenges in maintaining detector performance across evolving deepfake techniques. Section 5 provides a comprehensive overview of datasets, detailing their modalities, sizes, and characteristics that are essential for training and evaluation. Section 6 offers a

discussion on results and current insights and future scopes, addressing limitations and open research challenges in deepfake detection. Finally, Section 7 concludes the survey with a summary of findings and reflections on the current state and future prospects of deepfake research.

2 Taxonomy of Deepfake Contents

Manipulated content comes in various forms, each posing unique detection challenges and potential risks to individuals. Deepfakes are broadly categorised into image, video, and audio types. Video deepfakes often consist of manipulated frames, essentially images. Some deepfakes combine audio and video to create lip-sync videos. Fig. 2 summarises the taxonomy of deepfake content.

Image and video deepfakes primarily rely on manipulated techniques, which can be classified into four main types: entire face synthesis, identity swap, attribute manipulation, and expression swap. Entire face synthesis uses GANs such as StyleGAN to generate non-existent synthesis faces, often for fake profiles [3]. The identity swap replaces one person's face with another, as seen in the ZAO app, enabling the spread of fake news or inappropriate content. Attribution alters features such as skin tone, hair colour, or gender, while expression swap changes facial expressions, demonstrated by tools like Face2Face and NeuralTextures. Full-body puppetry, another form of manipulation, transfers a person's body movements onto another's body. Audio manipulations include voice swapping, which changes voice to mimic another, and text-to-speech, which generates audio from text. Voice conversion techniques are further divided into parallel (altering voice characteristics while keeping content intact) and non-parallel (changing both voice and content), both can be exploited for misinformation or scams.

All aforementioned Deepfake manipulations pose distinct challenges, making it difficult to design a universal system capable of accurately identifying all variants across diverse generation techniques. This remains an active research area with ongoing efforts toward developing generalized detection frameworks.

2.1 Deepfake Generation

Autoencoders were among the earliest models used for deepfake generation, with early tools like *FakeApp* enabling face-swapping. They learn compact latent

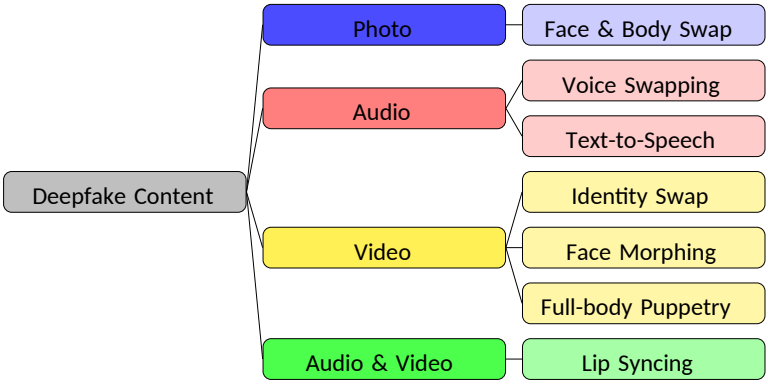

Fig.2. Taxonomy of Deepfake Content

representations of images while minimizing reconstruction loss, preserving key visual features such as skin tone, texture, and facial structure. VAEs extend autoencoders by modelling the latent space probabilistically, allowing smooth interpolation and realistic content generation [93]. Formally, encoder $E : R^m \rightarrow R^l$ and decoder $D : R^l \rightarrow R^m$ are trained to minimize the reconstruction loss expressed as argmin $_{E,D}\, E\left[e(x,(D \circ E)(x))\right]$, where e is the reconstruction loss function and $D \circ E$ denotes the composition of encoder and decoder. Where e is the reconstruction, loss and $D \circ E$ denotes the composition of encoder and decoder. VAEs consist of an encoder, a probabilistic latent space, and a decoder, which together enable high-quality controllable image synthesis for deepfakes.

Generative Adversarial Networks(GANs), introduced by Goodfellow et al.[82] comprise a generator that synthesizes data and a discriminator that distinguishes real from synthetic inputs. Training is formulated as a min-max optimization problem, where the generator seeks to minimize and the discriminator to maximize the following objective $\min_G \max_D V(D,G) = \mathbb{E}_{x\sim\rho(x)}\left[\log D(x)\right] + \mathbb{E}_{z\sim\rho(z)}\left[\log\left(1 - D(G(z))\right)\right]$.

Through iterative training, the generator learns to produce realistic outputs capable of fooling the discriminator. GANs have been applied in diverse generative tasks, including image-to-image transition, image completion, and text-to-image synthesis. Key variants include Progressive, Growing GAN(ProGAN) for high-resolution generation, VAE-GAN[7] combining variational autoencoding with adversarial learning, CycleGAN[4] for domain translation, FSGAN[11] for face swapping, StarGAN[8] v2 for multi-domain style transfer, STGAN[10] for attribute editing, MD-GAN[9] for multi-modal synthesis.

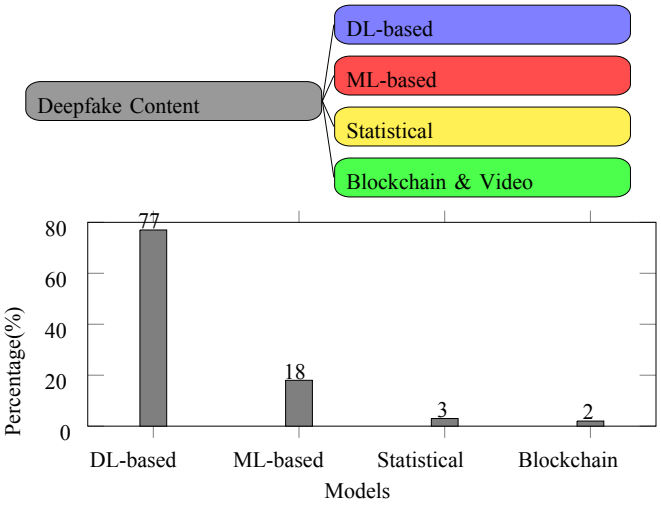

Fig.3. Taxonomy of popular deepfake detection techniques and classification distribution.

Diffusion Models[1] generate data through a progressive denoising (diffusion) process. The denoising diffusion probabilistic model (DDPM)[12] has gained significant attention for its exceptional generative performance, particularly in handling large-scale, high-resolution images. Latent Diffusion Models (LDMs)[13] extend this capability by operating in a compressed latent space, offering greater flexibility and efficiency in modeling complex data distributions. In video generation, diffusion models have demonstrated strong potential[14,15]. For example, Stable Video Diffusion (SVD)[16] fine-tunes a base text-to-video model using an image-to-video conversion task. Similarly, AnimateDiff [17] introduces a motion modeling module to a frozen text-to-image model, which is then trained on video clips to refine motion priors, enabling realistic and coherent motion generation.

Popular platforms and tools for deepfake generation are compiled in Table 2. It emphasises their main uses, supporting technologies, and working environments, which range from online and mobile applications to deep learning frameworks focused on research. Current deepfake techniques, such as face swapping, picture/video synthesis, multi-domain translation, and high-quality image production, are well summarised in the table.

3 Deepfake Detection Methods: A categorical Overview

Deepfake technology relies primarily on deep learning (DL) to manipulate images and videos. Although DL-based techniques are the most prevalent, other methods are also employed to detect these forgeries. This section categorizes and describes the most popular deepfake detection techniques. We classify the research into the following categories based on the techniques applied: DL-based, ML-based, statistical, and blockchain. A breakdown of the distribution of these methods is illustrated in Figure 3.

Table 2. Popular Deepfake Generation Tools and Platforms

Tool	Description	URL	Platform
ZAO	Allows users to superimpose faces from a single image onto videos and movies.	https://apps.apple.com/cn/app/id1465199127	Application
AutoFaceSwap	Enables face-swapping via webcam or drag-and-drop for live videos.	https://www.microsoft.com/en-us/p/auto-face-swap/9nblggh3m5nq	Application
FaceApp	Provides image and video editing to change facial appearance, age, or gender.	https://apps.apple.com/gb/app/faceapp-ai-face-editor/id1180884341	Application
FaceSwap	Deep learning-based method for swapping faces in images and videos.	https://github.com/deepfakes/faceswap	TensorFlow
FSGAN	RNN-based face-swapping and face reenactment for single images or videos.	https://github.com/YuvalNirkin/fsgan	PyTorch
FaceSwap-GAN	GAN-based face-swapping with adversarial and perceptual loss for images and videos.	https://github.com/shaoanlu/faceswap-GAN	TensorFlow

FewShot	GAN-based model for face translation, including crossethnicity transformations.	https://github.com/shaoanlu/ fewshot-face-translation	TensorFlow -GAN
StyleGAN / StyleGAN2	Style-based GANs improving image quality, weight modulation, and progressive growth removal.	https://github.com/NVlabs/stylegan https://github.com/NVlabs/stylegan2	TensorFlow
StyleGAN2-ADA	Adaptive discriminator augmentation to stabilize training with limited data.	https://github.com/NVlabs/stylegan2-ada	TensorFlow
DFaker	Generates reconstructed images (64x64 input to 128x128 output) using DSSIM loss.	https://github.com/dfaker/df	TensorFlow
DeepFake_tf	TensorFlow implementation similar to DFaker for face reconstruction.	https://github.com/StromWine/DeepFake_tf	TensorFlow
Deepfakes Web	Web-based face-swapping and video creation tool using deep learning.	https://deepfakesweb.com/	Web
StarGAN-V2	GANs for multi-domain image-to-image translation across multiple domains.	https://github.com/yunjey/stargan https://github.com/clovaai/stargan-v2	PyTorch
DeepFaceLab	Generates high-quality faceswapping videos.	https://github.com/iperov/DeepFaceLab	TensorFlow
DiscoFace	Disentangled and controllable face generation using 3D contrastive learning.	https://github.com/microsoft/DiscoFaceGAN	TensorFlow

Machine Learning methods, especially tree-based models, offer interpretable deepfake detection and easy tuning. GANs generate realistic fake faces,and ML techniques detect artifacts or inconsistencies in facial features, landmarks, and head poses. Lightweight models like MPLs can efficiently identify deepfakes. The accuracy can reach up to 98%, but it depends on the type of dataset, the selection of feature.

Deep learning-based methods dominates deepfake detection due to their ability to automatically learn discriminative features. CNNs (e.g., Meso-4) were early models for video and image analysis, while deeper networks, attention mechanisms, and capsule networks improved performance. RNNs and temporal modelling capture facial improvements frame-by-frame in videos or images , and autoencoders or optical-flow methods reduce overfitting. Feature extraction includes spatiotemporal cues, facial landmarks, micro-expressions, and patchbased analysis. Multi-modal approaches combine audio-visual cues, and ensemble learning or adversarial training further boosts accuracy, often exceeding 99%. State-of-the-art deepfake detection primarily relies on deep learning, treating the task as binary classification. CNN-based models like XceptionNet, EfficientNet, MesoNet, and ResNet excel at learning discriminative spatial features, while Vision Transformers and Capsule Networks capture spatial-temporal inconsistencies. Hybrid approaches combining CNNs with RNNs or optical flow exploit temporal cues, and transfer learning with SVMs improves performance on limited data. Recent diffusion-based detectors reveal vulnerabilities to small perturbations, highlighting the need for more robust detection systems.

Statistical measurement-based methods detect deepfakes by analyzing statistical differences between original and manipulated content. Techniques include examining photo-response non-uniform (PRNU), a unique sensor noise pattern in images and computing normalized cross-correlation scores across video frames. Hypothesis testing and statistical frameworks, such as ExpectationMaximization and distance-based measures, help quantify discrepancies between real and GAN-generated images. The effectiveness often depends on the quality of the GAN output, with higher-resolution and more accurate GANs being harder to detect.

Blockchain-based methods can verify the authenticity and provenance of digital content in a secure, decentralized manner. Public blockchains allow tracing videos or images back to their source, which make them useful for deepfake detection. Frameworks have been proposed to track the origin of suspect content, integrate decentralized storage (e.g., IPFC), and maintain tamper-proof records. Some approaches use deep learning models (e.g., LSTM, CNN) to extract and hash high-dimensional features, which are then stored on permissioned blockchains, giving content owners control and ensuring verifiable authenticity.

Understanding the Dominance of Deep Learning:

Every class has advantages and disadvantages; an integrative approach may provide a more robust solution. However, DL-based approaches are widely used

because they are highly effective at extracting and selecting features, making them particularly adept at detecting fake media content. Deepfake generation involves the use of advanced generative models to imitate genuine content. This poses difficulties for standard methods that may have trouble adjusting to the complex patterns present in synthetic media [94]. Deep learning architectures, such as Convolutional Neural Networks (CNNs) and GANs, can find complex and subtle features in deepfake content because they are deep and do not work in a straight line [32].

4 Catastrophic Forgetting & Continual Learning

Forgetting refers to the phenomenon where previously acquired information in a machine learning system degrades over time. Forgetting was not a significant concern since the models were trained and evaluated on fixed datasets. The concept of *catastrophic forgetting* was first formally introduced by McCloskey and Cohen [32]. They demonstrated that neural networks, when trained sequentially on different tasks, tend to forget previously learned tasks when new tasks are learned. To address this issue, incremental learning arrived. Although literature is rich in continual learning, recently, wang et.al [32] discussed the two shades of forgetting emphasizing its benefits and harms along with challenges in addressing forgetting. In some situations, there are many cases where forget becomes necessary. Firstly, forgetting can mitigate overfitting, as it allows the model to forget irrelevant details and focus on the most pertinent patterns in the training data. Additionally, by discarding unnecessary information, forgetting facilitates the learning of new knowledge, as the model can make better use of its capacity to acquire and adapt to novel information. Lastly, forgetting helps protect privacy by discarding sensitive user information, ensuring that such data is not retained in the model's memory. Fig.4 presents catastrophic challenges.

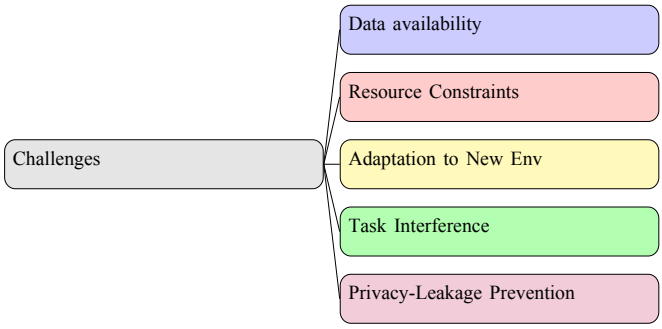

Fig.4. Catastrophic forgetting challenges continual learning.

4.1 Continual Learning

Artificial neural networks have been reported to exhibit, and in some cases surpass, human-level performance on individual rigid tasks. However, these networks remain static entities of knowledge for those specific tasks, which can lead to catastrophic forgetting (i.e., forgetting old tasks) when attempting to learn new tasks. The main objective of Incremental Learning (IL) is to address this issue. As presented in Fig. 5, early continual learning approaches addressed catastrophic forgetting by introducing regularization terms that constrain network parameters, preventing them from forgetting previously learned knowledge when updated with new tasks [34–37]. While these methods mitigate forgetting to some extent, performance often deteriorates after multiple sequential updates. Rehearsal-based methods, which maintain a memory buffer of past data, have demonstrated improved performance over memory-free approaches [38– 40]. However, storing sensitive data for future rehearsal raises privacy concerns. To address this, parameter-isolation techniques leverage pre-trained models and fine-tune only a small subset of parameters for each new task, selected at inference via query-key mechanisms [41–44]. Notably, S-Prompts [45] tackles domainincremental learning (DIL) by tuning CLIP [46] models on domain-independent sets of vision and language prompts, using query-key matching to select the most suitable prompt set under high domain shifts. MoP-CLIP [47] further improves out-of-distribution performance by employing a mixture of prompt-tuned CLIP models.

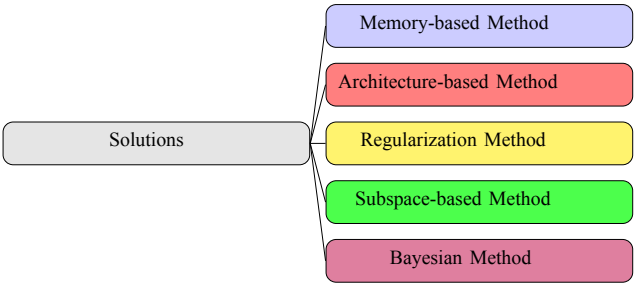

Fig.5. Existing solutions for catastrophic forgetting in continual learning.

5 Datasets

Several datasets have been widely used for deepfake detection research. The UFDAV dataset consists of 49 original videos and an equal number of deepfake versions [53].

FaceForensics contains 1,004 authentic videos and 2,008 manipulated videos, created using both source-to-video and self-reenactment methods [55]. FaceForensics++ (FF-DF) includes 1,000 original videos and 1,000 face-swapped videos [32]. Celeb-DF and Celeb-DF v2 offer larger collections of deepfake content generated from celebrity videos [52]. DF-TIMIT provides low and high-quality deepfake videos generated via faceswap-GAN [56]. Other notable datasets include the Google/Jigsaw Deepfake Detection dataset [57], DFDC [58,59], DeeperForensics 1.0 [60], KoDF [62], Deepfake MNIST+ [63], Wavefake for audio manipulations [64], ForgeryNet [65], and DF-W, which contains deepfakes collected from real-world online sources [66].

6 Results and Discussion

To provide a comparative understanding of continual learning approaches in deepfake detection, we summarize the results reported in the Prompt2Guard framework [26]. Table 3 presents a comparison of state-of-the-art continual learning methods in terms of *Average Accuracy (AA)* and *Accuracy Forgetting (AF)*.

Table 3. Performance comparison of continual deepfake detection methods under varying buffer sizes.

Method	Prompts	Buffer Size	AA ↑	AF ↑
LRCIL [27]	×	100 samples/class	76.39	-4.39
iCaRL [39]	×	100 samples/class	79.76	-8.73
LUCIR [67]	×	100 samples/class	82.53	-5.34
LRCIL [27]	×	50 samples/class	74.01	-8.62
iCaRL [39]	×	50 samples/class	73.98	-14.50
LUCIR [67]	×	50 samples/class	80.77	-7.85
DyTox [68]	✓	50 samples/class	86.21	-1.55
EWC [35]	×	0 samples/class	50.59	-42.62
LwF [34]	×	0 samples/class	60.94	-13.53
DyTox [68]	✓	0 samples/class	51.27	-45.85
L2P [41]	✓	0 samples/class	61.28	-9.23
S-iPrompts [45]	✓	0 samples/class	74.51	-1.30
MoP-CLIP [47]	✓	0 samples/class	88.54	-0.79
S-liPrompts [45]	✓	0 samples/class	88.65	-0.69
Prompt2Guard[26]	✓	0 samples/class	90.28	-0.71

6.1 Observations

The comparison shows that prompt-based continual learning strategies consistently outperform classical rehearsal- and regularization-based methods. Among these,

Prompt2Guard achieves the highest Average Accuracy (90.28%) with minimal forgetting (-0.71), reflecting superior adaptability to sequentially evolving deepfake manipulations.

Methods such as *L2P* and *S-iPrompts* effectively mitigate forgetting but show lower overall accuracy compared to *MoP-CLIP* and *Prompt2Guard*. This indicates that conditioned prompt optimization enhances both generalization and stability across different manipulation types.

6.2 Insights for Multimodal Deepfake Detection

While most continual learning frameworks focus primarily on visual features, integrating additional modalities such as video, audio and text can further enhance robustness against sophisticated forgeries. Recent works suggest that multimodal fusion improves cross-domain consistency and provides complementary cues for identifying subtle inconsistencies in generated media.

6.3 Discussion

Numerous significant patterns and difficulties are revealed by the literature on deepfake detection. First, audio and multimodal detection techniques are still understudied, and the majority of current algorithms concentrate on single modalities, especially photos and videos. Though they have demonstrated impressive performance on static datasets, methods based on convolutional neural networks (CNNs), transformers, and GAN-based discriminators sometimes find it difficult to generalise to hidden manipulations or changing deepfake generating approaches.

A possible paradigm for addressing catastrophic forgetting is continuous deepfake detection, which allows models to adjust to novel perturbations without requiring retraining. Existing methods, however, have drawbacks such reliance on task identification, significant computing overhead, and a lack of evaluation measures for equitable cross-study comparison. Furthermore, the majority of datasets utilised in continuous learning situations are either tiny or undiversified, which restricts their usefulness in the actual world.

The lack of standardised evaluation processes, the scarcity of multimodal datasets, and the paucity of research on resource-efficient and privacy-preserving detection techniques are some of the limitations that comparative analysis reveals. Filling in these gaps could enhance the model's deployment in real-world situations as well as its resilience and adaptability.

Future studies should concentrate on creating multimodal continual learning frameworks, utilising self-supervised and prompt-based approaches, and investigating strategies for effective and private implementation. Building robust, future-proof

deepfake detection systems that can manage quickly changing threats will require such work.

7 Conclusion

The rapid evolution of deepfake generation techniques presents serious obstacles to public trust and detection. This survey provided a thorough analysis of datasets, continual learning techniques, and detection methods, emphasising their shortcomings in terms of responding to novel perturbations. Even though continuous deepfake detection shows promise in reducing catastrophic forgetting, issues including task identification, computational expense, and the absence of standardised evaluation still exist. To create deepfake detection systems that are reliable, flexible, and ready for the real world, future research should concentrate on multimodal detection, continuous learning frameworks, and practical deployment.

References

1. Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., Zhai, G., Yang, J., Shen, C., Tao, D.: Deepfake generation and detection: A benchmark and survey. arXiv preprint arXiv:2403.17881 (2024)
2. Rana, M.S., Nobi, M.N., Murali, B., Sung, A.H.: Deepfake detection: A systematic literature review. IEEE Access 10, 25494–25513 (2022)
3. Heidari, A., Navimipour, N.J., Dag, H., Unal, M.: Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 14(2), e1520 (2024)
4. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. ICCV, pp. 2223–2232 (2017) 5. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2Face: Real-time face capture and reenactment of RGB videos. In: Proc. CVPR, pp. 2387– 2395 (2016)
6. Liu, K., Perov, I., Gao, D., Chervoniy, N., Zhou, W., Zhang, W.: DeepFaceLab: Integrated, flexible and extensible face-swapping framework. Pattern Recognition 141, 109628 (2023)
7. Ibrahim, B.I., Nicolae, D.C., Khan, A., Ali, S.I., Khattak, A.: VAE-GAN based zero-shot outlier detection. In: Proc. ISCSIC, pp. 1–5 (2020)
8. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proc. CVPR, pp. 8789–8797 (2018)
9. Chan, C.C.K., Kumar, V., Delaney, S., Gochoo, M.: Combating deepfakes: MultiLSTM and blockchain as proof of authenticity for digital media. In: Proc. AI4G, pp. 55–62 (2020)
10. Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., Wen, S.: STGAN: A unified selective transfer network for arbitrary image attribute editing. In: Proc. CVPR, pp. 3673–3682 (2019)

11. Nirkin, Y., Keller, Y., Hassner, T.: FSGAN: Subject agnostic face swapping and reenactment. In: Proc. ICCV, pp. 7184–7193 (2019)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 6840–6851 (2020)
13. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. CVPR, pp. 10684–10695 (2022)
14. Gupta, A., Yu, L., Sohn, K., Gu, X., Hahn, M., Li, F.-F., Essa, I., Jiang, L., Lezama, J.: Photorealistic video generation with diffusion models. In: Proc. ECCV, pp. 393–411 (2024)
15. Liu, J., Wang, Q., Fan, H., Wang, Y., Tang, Y., Qu, L.: Residual denoising diffusion models. In: Proc. CVPR, pp. 2773–2783 (2024)
16. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
17. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
18. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion 64, 131–148 (2020)
19. Banerjee, S., Yadav, S.K., Dhara, A., Ajij, M.: A survey: Deepfake and current technologies for solutions. Journal of AI Research 58(4), 123–145 (2025)
20. Whyte, C.: Deepfake news: AI-enabled disinformation as a multi-level public policy challenge. Journal of Cyber Policy 5(2), 199–217 (2020)
21. Singh, P., Dhiman, B.: Exploding AI-generated deepfakes and misinformation: A threat to global concern in the 21st century. Available at SSRN 4651093 (2023)
22. Matli, W.: Extending the theory of information poverty to deepfake technology. International Journal of Information Management Data Insights 4(2), 100286 (2024)
23. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proc. CVPR, pp. 4401–4410 (2019)
24. Moore, R., Lopes, J.: Paper templates. In: Proc. TEMPLATE'06, 1st International Conference on Template Production. SCITEPRESS (1999)
25. Smith, J.: The Book, 2nd edn. The Publishing Company, London (1998)
26. Laiti, F., Liberatori, B., De Min, T., Ricci, E.: Conditioned prompt-optimization for continual deepfake detection. In: International Conference on Pattern Recognition, pp. 64–79. Springer (2024)
27. Popescu, A.C., Farid, H.: Exposing digital forgeries by detecting traces of resampling. IEEE Transactions on Signal Processing 53(2), 758–767 (2005)
28. Agarwal, S., Farid, H.: Photo forensics from JPEG dimples. In: Proc. WIFS, pp. 1–6 (2017)
29. Huh, M., Liu, A., Owens, A., Efros, A.A.: Fighting fake news: Image splice detection via learned self-consistency. In: Proc. ECCV, pp. 101–117 (2018)
30. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proc. CVPR, pp. 1251–1258 (2017)

31. Chai, L., Bau, D., Lim, S.N., Isola, P.: What makes fake images detectable? Understanding properties that generalize. In: Proc. ECCV, pp. 103–120 (2020)
32. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to detect manipulated facial images. In: Proc. ICCV, pp. 1–11 (2019)
33. Wang, S.Y., Wang, O., Owens, A., Zhang, R., Efros, A.A.: Detecting photoshopped faces by scripting Photoshop. In: Proc. ICCV, pp. 10072–10081 (2019)
34. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(12), 2935–2947 (2017)
35. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences 114(13), 3521–3526 (2017)
36. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.: Memory aware synapses: Learning what (not) to forget. In: Proc. ECCV, pp. 139–154 (2018)
37. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.S.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: Proc. ECCV, pp. 532–547 (2018)
38. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: A strong, simple baseline. Advances in Neural Information Processing Systems 33, 15920–15930 (2020)
39. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: iCaRL: Incremental classifier and representation learning. In: Proc. CVPR, pp. 2001–2010 (2017)
40. Wu, Y., Chen, Y., Wang, L., et al.: Large scale incremental learning. In: Proc. CVPR, pp. 374–382 (2019)
41. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proc. CVPR, pp. 139–149 (2022)
42. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: DualPrompt: Complementary prompting for rehearsal-free continual learning. In: Proc. ECCV, pp. 631–648. Springer (2022)
43. Smith, J.S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., Kira, Z.: Coda-Prompt: Continual decomposed attentionbased prompting for rehearsal-free continual learning. In: Proc. CVPR, pp. 11909– 11919 (2023)
44. De Min, T., Mancini, M., Alahari, K., Alameda-Pineda, X., Ricci, E.: On the effectiveness of LayerNorm tuning for continual learning in vision transformers. In: Proc. ICCVW, pp. 3577–3586 (2023)
45. Wang, Y., Huang, Z., Hong, X.: S-prompts learning with pre-trained transformers: An Occam's razor for domain incremental learning. Advances in Neural Information Processing Systems 35, 5682–5695 (2022)
46. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proc. ICML, pp. 8748–8763 (2021)
47. Nicolas, J., Chiaroni, F., Ziko, I., Ahmad, O., Desrosiers, C., Dolz, J.: MOP-CLIP: A mixture of prompt-tuned CLIP models for domain incremental learning. arXiv preprint arXiv:2307.05707 (2023)

48. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proc. ICML, pp. 4904–4916 (2021)
49. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: CoCa: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
50. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for visionlanguage models. In: Proc. CVPR, pp. 16816–16825 (2022)
51. Lee, D., Song, S., Suh, J., Choi, J., Lee, S., Kim, H.J.: Read-only prompt optimization for vision-language few-shot learning. In: Proc. ICCV, pp. 1401–1411 (2023)
52. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: A large-scale challenging dataset for deepfake forensics. In: Proc. CVPR, pp. 3207–3216 (2020)
53. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: Proc. ICASSP, pp. 8261–8265 (2019)
54. Alzantot, M., Wang, Z., Srivastava, M.B.: Deep residual neural networks for audio spoofing detection. arXiv preprint arXiv:1907.00501 (2019)
55. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179 (2018)
56. Korshunov, P., Marcel, S.: Deepfakes: A new threat to face recognition? Assessment and detection. arXiv preprint arXiv:1812.08685 (2018)
57. Google AI: Deep Fake Detection Dataset. https://ai.googleblog.com/2019/09/contributingdata-to-deepfake-detection.html (2019)
58. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C.: The deepfake detection challenge (DFDC) preview dataset. arXiv preprint arXiv:1910.08854 (2019)
59. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge (DFDC) dataset. arXiv preprint arXiv:2006.07397 (2020)
60. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In: Proc. CVPR, pp. 2889–2898 (2020)
61. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: WildDeepfake: A challenging realworld dataset for deepfake detection. In: Proc. ACM Multimedia, pp. 2382–2390 (2020)
62. Kwon, P., You, J., Nam, G., Park, S., Chae, G.: KoDF: A large-scale Korean deepfake detection dataset. In: Proc. ICCV, pp. 10744–10753 (2021)
63. Huang, J., Wang, X., Du, B., Du, P., Xu, C.: Deepfake MNIST+: A deepfake facial animation dataset. In: Proc. ICCV, pp. 1973–1982 (2021)
64. Frank, J., Schönherr, L.: WaveFake: A data set to facilitate audio deepfake detection. arXiv preprint arXiv:2111.02813 (2021)
65. He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., Liu, Z.: ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In: Proc. CVPR, pp. 4360–4369 (2021)
66. Pu, J., Mangaokar, N., Kelly, L., Bhattacharya, P., Sundaram, K., Javed, M., Wang, B., Viswanath, B.: Deepfake videos in the wild: Analysis and detection. In: Proc. WebConf, pp. 981–992 (2021)

67. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: Proc. CVPR, pp. 831–839 (2019)
68. Douillard, A., Ramé, A., Couairon, G., Cord, M.: DyTox: Transformers for continual learning with dynamic token expansion. In: Proc. CVPR, pp. 9285–9295 (2022)
69. Wang, Z., Liu, Z., Wang, Y., et al.: Learning to prompt for continual learning. In: Proc. CVPR, pp. 139–149 (2022)
70. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision 130(9), 2337–2348 (2022)
71. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: Proc. ICML, pp. 8748–8763 (2021)
72. Jia, M., Tang, L., Chen, B.C., et al.: Visual prompt tuning. In: Proc. ECCV, pp. 709–727 (2022)
73. Wang, Z., Zhang, Z., Lee, C.Y., et al.: DualPrompt: Complementary prompting for rehearsal-free continual learning. In: Proc. ECCV, pp. 631–648 (2023)
74. Gao, Y., Xie, S., et al.: Continual learning with foundation models: A survey. arXiv preprint arXiv:2304.01235 (2023)
75. Liu, Y., et al.: Deepfake detection: A comprehensive survey. ACM Computing Surveys 56(3), 1–42 (2023)
76. Verdoliva, L.: Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing 14(5), 910–932 (2020)
77. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Face X-ray for more general face forgery detection. In: Proc. CVPR, pp. 5001–5010 (2020)
78. Cozzolino, D., Poggi, G., Verdoliva, L.: Forensic analysis of GAN-generated images. In: Proc. IEEE International Conference on Image Processing (ICIP), pp. 4322– 4326 (2018)
79. Zhang, X., Yi, J., Tao, J.: Towards robust audio deepfake detection: A continual learning benchmark (EVDA). In: Proc. ICASSP, pp. 1–5 (2024)
80. Yadav, D., Salmani, S.: Deepfake: A survey on facial forgery technique using generative adversarial network. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 852–857. IEEE (2019)
81. Ramadhani, K.N., Munir, R.: A comparative study of deepfake video detection method. In: 2020 3rd International Conference on Information and Communications Technology (ICOIACT), pp. 394–399. IEEE (2020)
82. Katarya, R., Lal, A.: A study on combating emerging threat of deepfake weaponization. In: 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 485–490. IEEE (2020)
83. Yu, P., Xia, Z., Fei, J., Lu, Y.: A survey on deepfake video detection. IET Biometrics 10(6), 607–624 (2021)
84. Mirsky, Y., Lee, W.: The creation and detection of deepfakes: A survey. ACM Computing Surveys 54(1), 1–41 (2021)
85. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–8. IEEE (2019)

86. Abdulreda, A.S., Obaid, A.J.: A landscape view of deepfake techniques and detection methods. International Journal of Nonlinear Analysis and Applications 13(1), 745–755 (2022)
87. Masood, M., Nawaz, M., Malik, K.M., Javed, A., Irtaza, A., Malik, H.: Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence 53(4), 3974–4026 (2023)
88. Zhang, T.: Deepfake generation and detection: A survey. Multimedia Tools and Applications 81(5), 6259–6276 (2022)
89. Patil, K., Kale, S., Dhokey, J., Gulhane, A.: Deepfake detection using biological features: A survey. arXiv preprint arXiv:2301.05819 (2023)
90. Dhesi, S., Fontes, L., Machado, P., Ihianle, I.K., Tash, F.F., Adama, D.A.: Mitigating adversarial attacks in deepfake detection: An exploration of perturbation and AI techniques. arXiv preprint arXiv:2302.11704 (2023)
91. Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I.E., Nyameko, R., Aluvala, S., Vimal, V.: Deepfake generation and detection: Case study and challenges. IEEE Access 11, 143296–143323 (2023)
92. Yan, Z., Zhang, Y., Yuan, X., Lyu, S., Wu, B.: DeepfakeBench: A comprehensive benchmark of deepfake detection. arXiv preprint arXiv:2307.01426 (2023)
93. Kietzmann, J., Lee, L.W., McCarthy, I.P., Kietzmann, T.C.: Deepfakes: Trick or treat? Business Horizons 63(2), 135–146 (2020)
94. Naitali, A., Ridouani, M., Salahdine, F., Kaabouch, N.: Deepfake attacks: Generation, detection, datasets, challenges, and research directions. Computers 12(10), 216 (2023)