Deepfake Technologies: Generation, Detection and Emerging Challenges

Kazuhiko Sumi
Dept. of Integrated Information Technology, *Aoyama Gakuin University*
Sagamihara, Japan
ORCID: 0000-0002-9165-5912

***Abstract*—Deepfake technologies have rapidly advanced due to recent progress in deep generative models, enabling the creation of highly realistic synthetic facial images and videos. Although these techniques have demonstrated significant potential in entertainment, digital human creation, and content production, they also pose substantial risks, including privacy violations, misinformation, and social manipulation. The continuous evolution of generative models—from early autoencoder-based approaches and GAN-driven methods to the latest diffusion models—has not only improved visual fidelity but also broadened the accessibility of deepfake creation tools. In parallel, deepfake detection has emerged as a critical research area. Detection methods span a wide spectrum, including biological-signal analysis, artifact-based forensics, learning-based classification, and multi-modal fusion. Despite steady improvements, ensuring robustness against unseen manipulation techniques and next-generation generative models remains a major challenge. The growing diversity of self-supervised and unsupervised training models for detection will be promising solutions. This paper provides a comprehensive overview of the current landscape of deepfake generation and detection. It summarizes technological foundations, representative methods, and recent advances.**

***Index Terms*—Deepfake, generative models, diffusion models, deepfake detection, digital media security**

I. Introduction

Deepfake technology has existed for many years, but around 2020 it underwent a major breakthrough. Since then, the quality of synthetic media has rapidly advanced to the point where ordinary viewers can no longer reliably distinguish real content from fabricated content, and user-friendly tools have made such technology accessible to the general public. As a result, the misuse of deepfakes for misinformation and disinformation has become an increasingly serious social concern.

Deepfake technology refers to computational methods that synthesize images or videos of a person—mimicking the appearance and behavior of a real individual. The term "deepfake" has been invented, because these techniques rely on deep networks.

Deepfakes can be broadly categorized into four primary classes: (1) Face Swap, (2) Face Reenactment, (3) Entire Face Generation, and (4) Face Editing. These specific explanations and the description of deepfake generation methods are discussed in Section II.

Deepfake technology has been expected to contribute positively to society in many fields. In the arts, it enables the creation of novel visual expressions; in commercial media production, it dramatically increases efficiency—for example, in producing posters or marketing materials. In filmmaking, tasks that traditionally required extensive manual effort can now be automated. Moreover, technology enables the construction of realistic virtual avatars, potentially allowing individuals to increase their productivity in virtual environments. However, when deepfake technology is used to fool people, it poses a significant threat to social stability. Indeed, the U.S. Government Accountability Office (GAO) warned in its 2024 report that "malicious use of deepfakes could erode trust in elections, spread disinformation, undermine national security, and empower harassers." [1]

In response to these risks, deepfake detection technologies have been extensively studied. Such techniques determine whether a given image or video of a person is real (captured from an actual individual) or fake (computer-generated). Detection methods correspond to the four categories of face swap, face reenactment, entire face generation, and face editing.

When understanding detection methods, it is easier to grasp the approach by classifying them based on how they extract features from images rather than which category of deepfakes they detect. Broadly speaking, these methods fall into two categories: cue-based detection, where humans identify artificial artifacts not present in real images and attempt to detect them, and data-driven detection, which uses machine learning to find artificial artifacts across numerous examples. Data-driven methods have become the mainstream approach in recent years. This paper explains these methods in greater detail in Section III.

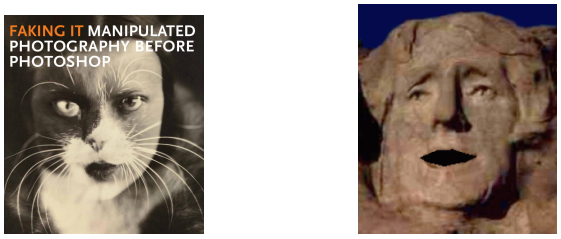
Deepfake detection and deepfake generation are evolving in a competitive race to outdo each other, showing no signs of stopping. The current challenges and future outlook, as understood from benchmarks and technological trends, are discussed in the final Section IV.

The purpose of this article is to provide researchers and practitioners in information technology fields, who may not specialize in deepfake generation or detection, with a clear overview of deepfake detection technology. Rather than presenting an exhaustive survey, this article focuses on representative deepfake generation techniques, their corresponding detection methods, and major evaluation benchmarks. We also explain the current open problems and discuss future perspectives. For more comprehensive surveys, readers are referred to recent review papers [2] [3] [4] [5] [6].

979-8-3315-7111-5/26/$31.00 ©2026 IEEE

II. Deepfake Technologies

Deepfake technology is a computational method using deep neural networks to synthesize non-existent images or videos that mimic the appearance and behavior of real individuals. However, the technology to create non-existent images and footage has existed since the invention of photography. Fig. 1a shows a fake image produced using masking and multiple exposures when transferring a photograph onto a dry plate. Only those who mastered advanced photographic techniques could create such artistic works. By the 1990s, image editing software like Photoshop emerged, but manipulating another person's face still required considerable operational knowledge. Around the same time, research was conducted on processing static images of human faces into videos where they appear to speak, synchronized with audio [7]. This was achieved using a model-based approach: extracting facial landmarks, statistically learning the movement of each landmark relative to the audio, and then deforming the face image accordingly (see Fig. **??**. At the 2007 Tsukuba World Expo, the Future Cast System [8] was demonstrated. This system 3D-scanned visitors' faces and synthesized them onto movie characters. In both cases, only researchers and engineers with advanced expertise could master these facial manipulation techniques. The invention of the Variational Autoencoder [9] in 2013 and the Generative Adversarial Network (GAN) [10] in 2014, it became possible to generate new images from existing ones using these deep neural networks [11]. Research applying this to facial images emerged, marking the beginning of modern deepfakes (See Fig. 3 [12]). Even today, many deepfakes utilize techniques based on GANs.



(a) (b)

Fig. 1: Artificial generated face image and videos. (a): Photographic fake face "Io+Gatto" (Wluz 1932) [13], (b): Talking statue of Jefferson in Voice Puppetry (Brand 1999) [7]

Subsequently, diffusion models (DM) were used in generative models to generate facial images [14] [15], achieving improved resolution and clarity, allowing the creation of more realistic facial images. Furthermore, the fusion of diffusion model-based techniques with Large Language Models led to the cloud-based release of AI tools like DALL-E, SORA, Stable Diffusion, and Midjourney. These image generation tools, which create images and videos from prompts and reference images, have made deepfakes accessible to anyone with ease.

Current deepfake technology can be classified into five types, as shown in Figure 4, based on the survey paper [6].

Fig. 2: Future Cast System: the 3D scanned audience face is imposed into the character in the movie. (2008) [8]

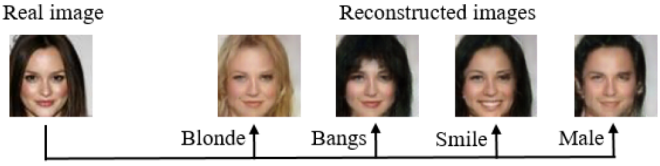


Fig. 3: The first face manipulation with GAN (2016) [12]

These are referred to as Face (identity) swap (FS), Face Reenactment (FR), Cross-Modal Manipulation (CMM), Entire Face Synthesis (EFS), and Face (attribute) Editing (FE). In the latest benchmark paper DF40 [16], deepfake are classified into four types; Cross-Modal Manipulation is included in Face Reenactment.

Face Swap (FS) is a type of deepfake that replaces the facial region of a source person with that of a target person. Early FS methods involved setting a mask over the facial region of the source person's image and replacing the image within the mask with the target person's facial image. Only the facial region was replaced; the background remained unchanged. Variational Autoencoders (VAE) and Generative Adversarial Networks (GAN) were used as models to edit the target facial image geometrically and photometrically to match the source. Deepfake datasets generated using this method include Celeb-DF [17] and DFDC [18]. More recently, methods have been used that do not employ masks. Instead, the source image is encoded into a latent space, and the encoding information of the target face is added to this latent space. The entire image is then decoded using an image generation model. A representative example is UniFace [19].

Face reenactment is a type of deepfake that replaces the expressions and movements of a source person with those of a target person. To achieve this, the source person's facial image is segmented into feature points and a mesh with those points as vertices. The source facial image is then modified to align with the feature point coordinates generated from the target person. The model used for such transformations is the 3DMM [20]. Recently, NeRF [21] and 3D Gaussian Splatting [22] are frequently employed as 3D facial models, while some examples still utilize DM.

Cross-Modal Face Manipulation (CFM) is a type of Face

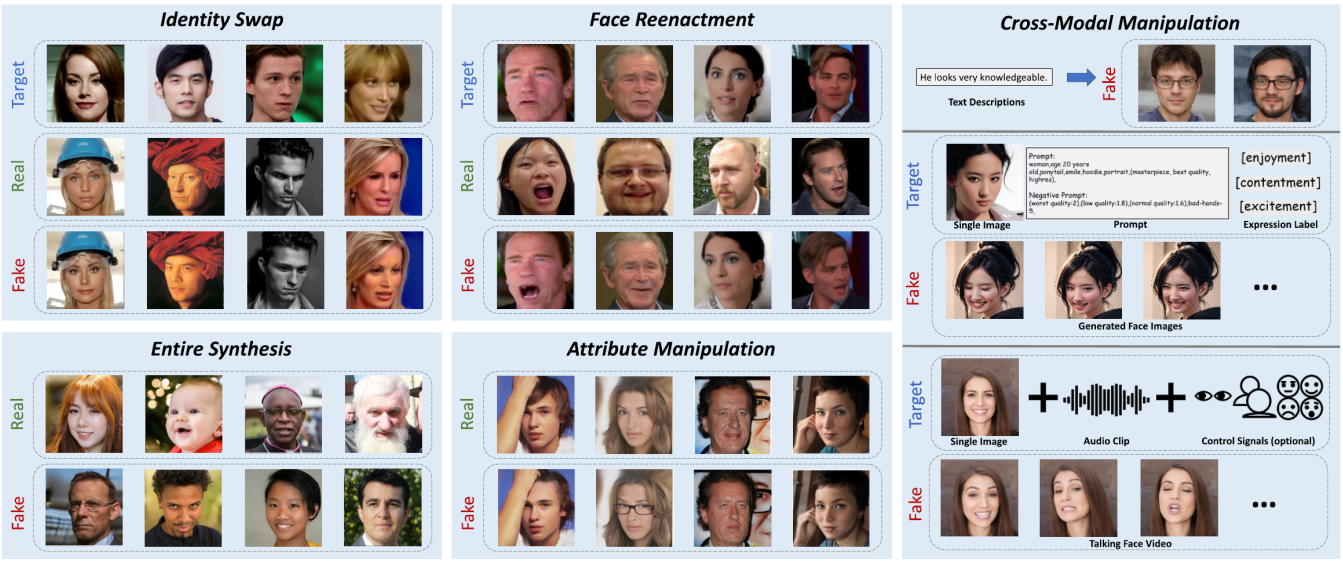


Fig. 4: Illustration of different DeepFake face types: Identity (Face) Swap, Face Reenactment, Entire Synthesis, Attribute Manipulation, and Cross-Modal Manipulation. This figure is quated from Xie 2026. [6]

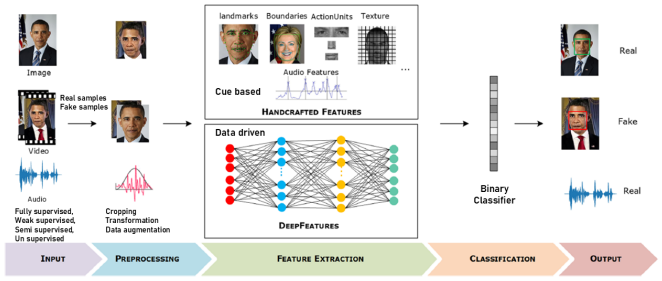


Fig. 5: General deepfake detection pipeline. This figure is quoted from Masood 2023 [25]

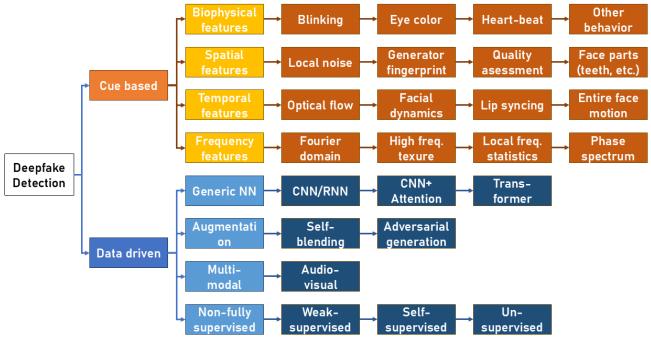


Fig. 6: Taxonomy of deepfake detection method

III. Deepfake Detection Technologies

Reenactment (FR) and is sometimes called Talking Head Generation. Instead of providing expressions or facial poses as parameters, it uses input from different modalities like speech audio or text. In this case, the model learns the temporal relationship between the audio/text and facial deformations, then uses the same techniques as FR to generate images of the face speaking.

Entire Face Synthesis, also called Face AIGC (AI Generated Contents), generates facial images without mimicking specific individuals using GAN-based methods like StarGAN [23] or DM-based methods like Stable Diffusion [24]. This is currently the most actively developed area, showing significant improvements in resolution and fidelity.

Face Manipulation involves altering attributes such as age or gender in facial images, or adding elements like makeup, glasses, or beards. Techniques used include GANs and DM.

Deepfake detection is a binary classification problem, where a test image or video is input, and it is determined whether it is real or generated (fake). When the subject is a face, as shown in Figure 5, pre-processing involves cropping the face region, extracting features from it, and then using a classifier to distinguish real/fake. The methods for extracting these features differ. As shown in Figure 6, they can be broadly classified into cue-based methods and data-driven methods. Cue-based methods extract features by identifying human-defined cues specific to fakes. These can be categorized into four approaches: Biophysical cue, Spatial domain cue, Temporal domain cue, and Frequency domain cue. In contrast, data-driven methods derive these features through machine learning from training samples, rather than having them designed by humans. Data-driven methods can be categorized into Generic networks, Data augmentation-based methods, Multi-modal based methods, and Multi-supervision-based methods. Each is explained in

more detail below.

Biophysical-based cues are human-designed feature extraction methods for identifying biological characteristics of human faces that deepfakes cannot fully replicate. Examples include blink frequency and speed, iris color differences between eyes, facial color changes due to pulse, and other behavioral characteristics. Spatial domain cues focus on extracting features where deepfake generators exhibit artificial traces (artifacts) in the spatial characteristics of images, differing from real images. Particularly in face swaps (FS), since the central face image is swapped while the periphery remains from the source image, many methods detect differences between the center and periphery. Examples include local noise, unique patterns (fingerprints) inherent to the generator, differences in image quality, and the resolution or fidelity of facial components (eyes, lips, teeth). These too are becoming increasingly difficult to detect due to advances in DM generators.

Temporal domain cues perform fake detection based on consistency and motion characteristics between video frames. Examples include consistency of motion vectors (optical flow), naturalness of dynamic facial changes, consistency between speech and lip movements, and continuity of overall facial motion.

Frequency domain cues perform spatial and temporal domain cues in the frequency domain. They map images or videos to the frequency domain using FFT and utilize frequency domain features. Examples include comparing Fourier spectra (especially differences between central and peripheral regions in the FS), high-frequency texture components, statistics of local features, and phase spectra.

These cue-based methods align with human intuition, providing good interpretability for classification results. Furthermore, since they rely on features independent of the deepfake generation technique, they remain effective against unknown generation methods different from those used during training. On the other hand, many deepfake generators, from VAE to GAN, were unable to accurately reproduce the features emphasized by cue-based methods. However, recent DMs have significantly improved in resolution and fidelity, making it increasingly difficult to detect deepfakes using such human-designed features. Consequently, the latest state-of-the-art deepfake detection has shifted from cue-based methods to the data-driven methods described next.

Next, we discuss data-driven approaches. Generic networks employ general deep learning networks (RNN, CNN, Transformer) that take images as input and output real/fake classification results. For still images, CNNs are typically used, while for video, features are extracted frame-by-frame and processed by an RNN. Recently, influenced by Vision Transformers, examples using Transformers have become increasingly common. A representative example of such methods is XceptionNet which remains a baseline reference in this field. The structure of XceptionNet is similar to Fig. 5, without audio input/output and with the Xception model as its backbone. Generic networks fundamentally require a large number of training samples annotated as real/fake, and the effort to secure the data is excessive. Therefore, the data augmentation and non-fully supervised methods described next have proven effective. Data augmentation generates similar samples from a small number of samples through augmentation, instead of preparing numerous fake samples. Among these, the self-blending method SBI [26] generates pseudo-fake samples by applying variations likely to occur in fake generation to real samples. It demonstrates high detection capability against unknown deepfake generators.

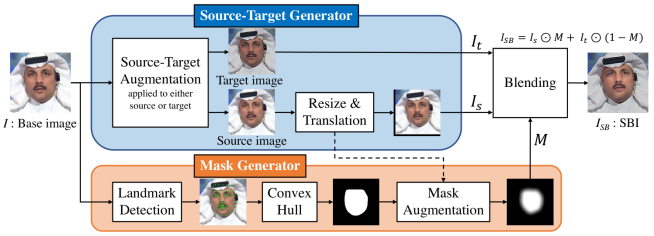


Fig. 7: An example of data augmentation (Shiohara 2022) [26]

Multi-modal approaches for audio-video face reenactment combine audio and video features for detection, rather than detecting fakes separately for each modality. Non-fully supervised approaches replace supervised learning with weak supervised, self-supervised, and unsupervised learning for generic networks. For example, when multiple people appear in an image but only some are fake, or when only a portion of a video is fake, accurately annotating exactly what is fake is extremely difficult. Therefore, only the information “a fake exists within this” is annotated. This is called weak supervised learning, and it allows for the preparation of a much larger amount of data. On the other hand, SBI, which modifies real samples rather than generating fake samples using actual deepfake tools, can be considered a self-supervised approach. One self-supervised method involves training reconstruction using only real samples through encode-decode with VAE or DM, then exploiting the large reconstruction error in fake samples. Representative methods include OC-FakeDect using One Class VAE [27] and DIRE using DM [28]. Fig. 8 illustrates how the vectors representing real and fake input images in DIRE are mapped through reconstruction. For real images, the DM process causes significant movement, whereas fake images, having already undergone the DM process, exhibit minimal movement. Thus, DIRE demonstrats strong performance in detecting DM-generated deepfakes.

Meanwhile, methods like [29] [30] have been proposed that start with pseudo real/fake labels and gradually update them to correct labels without any real/fake annotations. These methods have garnered attention for their high detection performance against the latest DM-based deepfakes and unknown deepfakes.

According to the latest benchmarking [16], cue-based methods and methods using Generic Networks demonstrate sufficiently high detection performance of 0.99 or higher against VAE- and GAN-based deepfake detection methods published

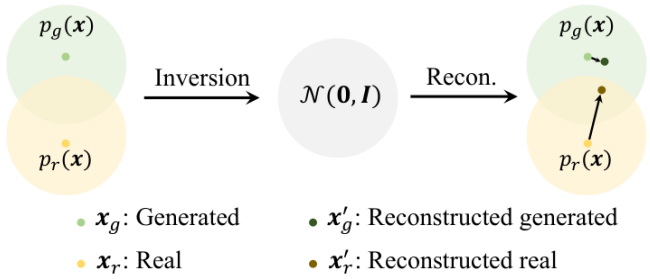


Fig. 8: Illustration of the difference between a real sample and a generated sample from DIRE perspective (Wang 2023) [28]

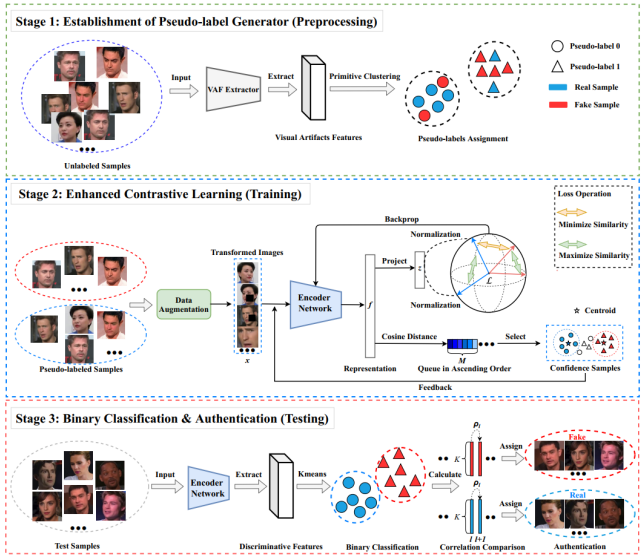


Fig. 2: Illustration of the proposed fully unsupervised framework: on the stage 1, the proposed pseudo-label generator is established, where the primitive clustering is completed; on the stage 2, the proposed enhanced contrastive learning is conducted, where the discriminative features are extracted; on the stage 3, based on the inter-frame correlation, binary classification and authentication are realized in order to distinguish between real and fake video.

Fig. 9: An example of unsupervised learning based deepfake detection (Qiao 2024) [29]

around 2022. These methods achieve this performance using AUC [1] values above 0.99 for VAE- and GAN-based deepfakes published before 2022. However, their detection performance for newly released DM-based deepfakes peaks at 0.7, and in some cases even below 0.4. In contrast, self-supervised and unsupervised data-driven methods achieve performance above 0.7 even against DM-based deepfakes, suggesting promising future progress.

IV. Open Problems and Future Directions

Even in the latest deepfake detection, the following issues remain unresolved. First, improving detection accuracy against DM-based deepfakes. Detection performance against VAE- and GAN-based deepfakes has already reached a sufficient level. However, the realism and fidelity of DM-based image generation to real images are steadily improving, making detection performance enhancement an urgent task. Furthermore, detection performance against unknown generation methods different from those used in training—i.e., improving generality—and enhancing explainability (specifying which features at which locations led to the fake determination) are also necessary. Additionally, since deepfakes can spread via social media, robust detection against post-processing like image compression and resolution conversion used on social media is required.

As shown in Figure 10, from 2000 to 2025, Large AI Models trained via self-supervised learning on large datasets using Transformers emerged one after another, with image and video generation among their applications. Correspondingly, numerous detection models have also been proposed. Specifically, for improving detection accuracy: Progress in learning models through self-supervised/unsupervised learning and the scaling up of datasets have been pursued. Following the development of large language models through self-supervised learning on vast amounts of text, fake detection is likely to advance by improving accuracy through learning on large volumes of images without annotations. Unfortunately, increasing the scale of training data leads to higher training costs. Solutions to this issue will likely require waiting for general technological advancements, such as reducing the training costs of large language models. Furthermore, it is also expected that explanations for fake determinations can be provided, such as identifying which image patch generated the signal indicating a fake or comparing predicted fake images with the target image. As discussed, the emergence of DM image generation has dramatically increased the difficulty of deepfake detection. However, new technologies are emerging, and the co-evolution of generation and detection will likely continue.

[1]AUC is the area under the Receiver Operator Curve (ROC), calculated by varying the detection threshold as a parameter, where the ROC is defined as the sum of the False Negative Identification Rate and the False Positive Identification Rate. The closer this value is to 1.0, the higher the detection performance.

Acknowledgment

In preparing this review paper, I would like to express my gratitude for the opportunity to reference numerous survey papers and original research papers, as well as to cite their figures and tables. A part of this research was conducted as a project of the Center for Advanced Information Technology Research (CAIR), Aoyama Gakuin University.

References

[1] U.S. Government Accountability Office, "Science & tech spotlight: Combating deepfakes," U.S. Government Accountability Office, Washington, DC, Tech. Rep. GAO-24-107292, Mar. 2024.
[2] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," vol. 54, no. 1, Jan. 2021. [Online]. Available: https://doi.org/10.1145/3425780
[3] L. Guarnera, O. Giudice, and S. Battiato, "Mastering deepfake detection: A cutting-edge approach to distinguish gan and diffusion-model images," vol. 20, no. 11, 2024.
[4] G. Pei, J. Zhang, M. Hu, Z. Zhang, C. Wang, Y. Wu, G. Zhai, J. Yang, C. Shen, and D. Tao, "Deepfake generation and detection: A benchmark and survey," 2024.
[5] L. Lin, N. Gupta, Y. Zhang, H. Ren, C.-H. Liu, F. Ding, X. Wang, X. Li, L. Verdoliva, and S. Hu, "Detecting multimedia generated by large ai models: A survey," 2025.

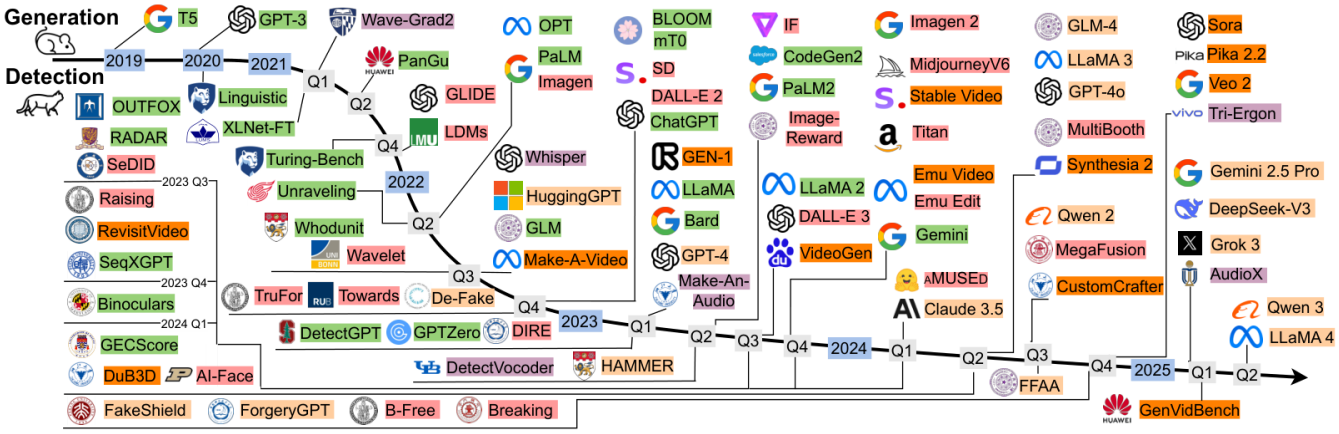


Fig. 10: Rapid evolution of Large AI Models both of generation and detection. This figure is quated from Lin 2026. [5]

[6] S. Xie, T. Qiao, S. Li, X. Zhang, J. Zhou, and G. Feng, "Deepfake detection in the aigc era: A survey, benchmarks, and future perspectives," *Information Fusion*, vol. 127, p. 103740, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253525008024

[7] M. Brand, "Voice puppetry," in *SIGGRAPH 1999: 26th Annual Conference on Computer Graphics and Interactive Techniques*, 1999, p. 21–28.

[8] A. MAEJIMA, S. WEMLER, T. MACHIDA, M. TAKEBAYASHI, and S. MORISHIMA, "Instant casting movie theater: The future cast system," *IEICE Transactions on Information and Systems*, vol. E91.D, no. 4, pp. 1135–1148, 2008.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[10] 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 (NeurIPS)*, vol. 27, 2014, p. 139–144.

[11] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[12] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Alvarez, "Invertible conditional gans for image editing," in *NIPS 2016 Workshop on Adversarial Training*, 2016.

[13] M. Fineman, *Faking It: Manipulated Photography Before Photoshop*. New York / New Haven: The Metropolitan Museum of Art / Yale University Press, 2012, exhibition catalogue of "Faking It: Manipulated Photography Before Photoshop", Oct.11 2012–Jan.27 2013, The Met.

[14] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[16] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, J. Zhu, D. Luo, C. Wang, S. Ding, Y. Wu, and L. Yuan, "Df40: Toward next-generation deepfake detection," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2024.

[17] Y. Li, P. Sun, H. Qi, Q. Li, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3207–3216.

[18] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[19] C. Xu, J. Zhang, Y. Han, G. Tian, X. Zeng, Y. Tai, Y. Wang, C. Wang, and Y. Liu, "Designing one unified framework fornbsp;high-fidelity face reenactment andnbsp;swapping." Berlin, Heidelberg: Springer-Verlag, 2022.

[20] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '99. USA: ACM Press/Addison-Wesley Publishing Co., 1999, p. 187–194. [Online]. Available: https://doi.org/10.1145/311535.311556

[21] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 104–121.

[22] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)*, vol. 42, no. 4, July 2023.

[23] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 8789–8797.

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.

[25] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023.

[26] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 18 720–18 729.

[27] H. Khalid and S. S. Woo, "Oc-fakedect: Classifying deepfakes using one-class variational autoencoder," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[28] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "Dire for diffusion-generated image detection," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22 388–22 398.

[29] T. Qiao, S. Xie, Y. Chen, F. Retraint, and X. Luo, "Fully unsupervised deepfake video detection via enhanced contrastive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[30] L. Zhang, T. Qiao, M. Xu, N. Zheng, and S. Xie, "Unsupervised learning-based framework for deepfake video detection," *IEEE Transactions on Multimedia*, vol. 25, pp. 4785–4799, 2023.