Wiley
International Journal of Intelligent Systems
Volume 2025, Article ID 9987535, 32 pages
https://doi.org/10.1155/int/9987535

WILEY

Review Article

A Review of Deepfake and Its Detection: From Generative Adversarial Networks to Diffusion Models

Baoping Liu,[1] **Bo Liu**,[1] **Tianqing Zhu**,[2] **and Ming Ding**[3]

[1]*School of Computer Science, University of Technology Sydney, Sydney, New South Wales 2007, Australia*
[2]*Faculty of Data Science, City University of Macau, Macau, China*
[3]*Data61, CSIRO, Sydney, New South Wales 2015, Australia*

Correspondence should be addressed to Bo Liu; bo.liu@uts.edu.au

Received 5 November 2024; Accepted 29 March 2025

Academic Editor: Yu-an Tan

Deepfake technology, leveraging advanced artificial intelligence (AI) algorithms, has emerged as a powerful tool for generating hyper-realistic synthetic human faces, presenting both innovative opportunities and significant challenges. Meanwhile, the development of Deepfake detectors represents another branch of models striving to recognize AI-generated fake faces and protect people from the misinformation of Deepfake. This ongoing cat-and-mouse game between generation and detection has spurred a dynamic evolution in the landscape of Deepfake. This survey comprehensively studies recent advancements in Deepfake generation and detection techniques, focusing particularly on the utilization of generative adversarial networks (GANs) and diffusion models (DMs). For both GAN-based and DM-based Deepfake generators, we categorize them based on whether they synthesize new content or manipulate existing content. Correspondingly, we examine various strategies employed to identify synthetic and manipulated Deepfake, respectively. Finally, we summarize our findings by discussing the unique capabilities and limitations of GANs and DM in the context of Deepfake. We also identify promising future directions for research, including the development of hybrid approaches that leverage the strengths of both GANs and DM, the exploration of novel detection strategies utilizing advanced AI techniques, and the ethical considerations surrounding the development of Deepfake. This survey paper serves as a valuable resource for researchers, practitioners, and policymakers seeking to understand the state-of-the-art in Deepfake technology, its implications, and potential avenues for future research and development.

Keywords: Deepfake; Deepfake detection; diffusion models; generative adversarial networks

1. Introduction

The development of deep learning (DL) and artificial intelligence (AI) has dramatically accelerated the synthesis of media content, such as images and videos. Within AI, a specific area of focus lies in the generation of synthetic and manipulated human faces, commonly referred to as Deepfake techniques. Initially gaining public attention through amusing events where individuals could swap celebrities' faces or alter politicians' speeches via lip-syncing, Deepfake images and videos have become increasingly sophisticated. Unlike manual manipulation or creation using graphics applications, well-trained Deepfake generation models excel in both performance and speed of generation. Variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models (DMs) are among the Deepfake generation models capable of synthesizing nonexistent faces by learning the data distribution of real-world human faces. This significantly lowers the barrier to artistic creation. At the same time, these models can also be employed to manipulate existing media, allowing for the creation of Deepfake content with alarming realism and ease. When pioneering works such as GANs and DMs are introduced, subsequent efforts focus on enhancing their structure and improving attributes such as training stability, convergence speed, and generation diversity. These improved

generative models are then applied to various tasks, including new content synthesis and media manipulation. To comprehensively evaluate the generation and manipulation performance of Deepfake models, some metrics are proposed. Inception score (IS) measures the quality of generated images based on the classification confidence of a pretrained inception model. It computes the KL divergence between the predicted class probabilities for each generated sample and the marginal class distribution. A higher IS indicates both sharpness and diversity in the generated images. The Frechet Inception Distance (FID) measures the similarity between the feature distributions of real and generated images by comparing their means and covariances in the Inception network's feature space. Lower FID values indicate higher quality and similarity between generated and real data. Peak signal-to-noise ratio (PSNR) measures the quality of generated images by comparing the pixel-wise error between the generated image and the ground truth. Higher PSNR values indicate better image quality. Structural similarity index (SSIM) evaluates the similarity between the generated and real images by considering luminance, contrast, and structure. A value closer to 1 indicates higher similarity. These metrics evaluate the performance of Deepfake generators in terms of generation quality (FID, SSIM, PSNR, and IS) and diversity (IS and FID).

While Deepfakes showcase AI's creative and manipulative potential, concerns about their potential misuse in fraud, disinformation, and privacy violations have also arisen. The rapid progression from early, rudimentary Deepfakes to current seamless fabrications underscores the need for robust forensic detection methods. Despite proactive approaches like digital watermarking, most Deepfake detectors typically frame detection as binary classification tasks. These detectors leverage labeled fake and real images to identify various artifacts such as biometrics, texture distortion, blending edges, temporal inconsistencies, and multimodal mismatches as distinctive features. Evaluating the performance of Deepfake detectors is crucial to ensure their effectiveness. Two commonly used metrics in Deepfake detection are the area under the ROC curve (AUC) and accuracy (ACC). Accuracy measures the proportion of correct predictions (both true positives and true negatives) made by the model out of all predictions. AUC is a performance metric that evaluates the ability of a model to distinguish between classes (real vs. Deepfake). It calculates the AUC, which plots the true positive rate (TPR) against the false positive rate (FPR) at various thresholds. Compared with ACC, AUC is more robust in cases of imbalanced data since AUC provides a comprehensive measure of model performance across all classification thresholds, making it more robust than accuracy in cases of imbalanced data. However, generative methods continually evolve to evade the latest forensic techniques, leading to a cat-and-mouse game between Deepfake generation and detection. Both fields evolve through innovations in neural structure, objective functions, and data modeling, ensuring a dynamic landscape of technological advancement and countermeasure development.

Compared with VAEs, which typically generate low-resolution and blurred images, GANs and DMs have attracted much public attention due to their visually striking outputs, intuitive training procedures, and wide range of applications. Several survey works have studied GANs [1–5] and DMs [6–9]. However, as far as we are concerned, none of them have comprehensively reviewed the evolution of Deepfake from the GAN to the DM era. To fill the gap, our paper focuses on Deepfake in both the GAN and DM eras, allowing for a more thorough examination of the evolution of Deepfake and forensic techniques. Compared with survey works that generally review generative tasks [2, 7], this paper focuses on Deepfake human faces due to their deceptive nature. Within each era, we survey representative generative models capable of producing Deepfake human faces, along with Deepfake detection models. By encompassing both Deepfake generation and detection, we provide deeper insight into the ongoing cat-and-mouse game between creators and detectors than previous surveys focusing solely on generation or detection [10–12]. Unlike existing survey works that also cover both generation and detection [5, 13, 14], we categorize Deepfake generators into generative and manipulative models based on their intended task. This finer categorization enhances understanding of the application of generative models in Deepfake creation. Correspondingly, Deepfake detectors are categorized based on their ability to detect either generated or manipulated content. We compare GANs and DMs in terms of generation and detection, highlighting their respective strengths and weaknesses. Through comprehensive analysis of generation and detection techniques in the GAN and DM eras, we propose promising avenues for future research. In summary, the key contributions of this survey are as follows:

- Comprehensive examination of Deepfake generation and detection in both GAN and DM eras, marking the first comprehensive analysis of GAN-based and DM-based Deepfake to our knowledge.
- Categorization of Deepfake generation and detection based on tasks (i.e., manipulation and synthesizing), providing a more intuitive understanding of Deepfake and its detection.
- Comparative analysis of GANs and DMs in various aspects, alongside proposed research directions and protective advice based on their development.

The organization of the subsequent sections is outlined in Figure 1: GAN and DM are introduced in Section 2 and Section 3, respectively. For each section, we first introduce the generation pipelines to illustrate the source of generative capacity (2.1 and 3.2). Subsequently, we introduce some representative generation and detection works in two subsections (2.2 and 2.3 for GANs and 3.3 and 3.4 for DMs). Notably, in Section 4, we compare GANs and DMs with regard to generation and detection, and propose potential future research directions.

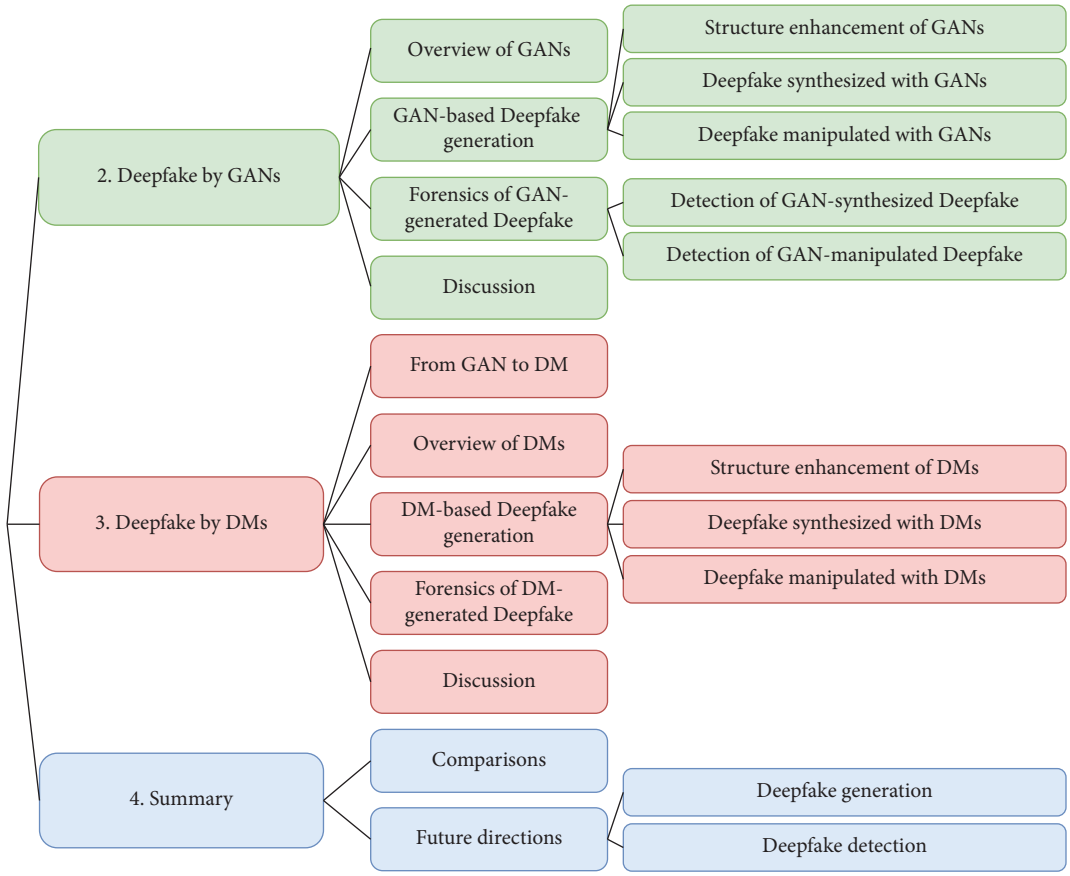

Figure 1: The structure of this paper. We mark the generation-related parts in red and the detection-related in green.

2. Deepfake by GANs

2.1. Overview of GANs. GANs were initially proposed in [15]. GANs consist of two neural networks, the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$ that are trained simultaneously through adversarial processes (shown in Figure 2). The generator network $\mathcal{G}$ takes random noise $z \sim p_z(z)$ as input, where $p_z(z)$ is a distribution over the noise vector z (i.e., Gaussian or Uniform), and produces a synthetic data point $\mathcal{G}(z)$. The goal of the generator is to make $\mathcal{G}(z)$ resemble real data as closely as possible. The discriminator network $\mathcal{D}(x)$ outputs a probability that the input data x is real (from the real data distribution $p_{\text{data}}(x)$) or fake (from the generated data distribution $p_g(x)$). The GAN's training process is framed as a minimax optimization problem, where the generator tries to minimize a loss function, and the discriminator tries to maximize its ability to correctly classify real versus fake samples. The objective function is as follows:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \left[\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log \mathcal{D}(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-\mathcal{D}(\mathcal{G}(z)))]\right]. \tag{1}$$

The generator tries to minimize this function while the discriminator tries to maximize it, and the adversarial training objective enables the generator to progressively improve its outputs based on the feedback from the discriminator. The introduction of GANs has paved the way for numerous applications, ranging from image synthesis to data augmentation.

2.2. GAN-Based Deepfake Generation. After GAN was proposed, a large number of papers have proposed various GAN-based generation methods. Some works focus on improving GAN performance by improving attributes of GANs such as training stability, training speed, and generation diversity. More works focus on adopting GANs in various generation tasks, including media synthesis and manipulation. Therefore, in this subsection, we examine GANs focusing on structural enhancement to optimize their performance. Secondly, we will investigate the synthesis of entirely new, non-existing content through GANs. Lastly, we will delve into the generation of fake content by manipulating existing media. This segment will cover a range of techniques, such as object replacement, style transfer, and detail reenactment within images. Works of each category

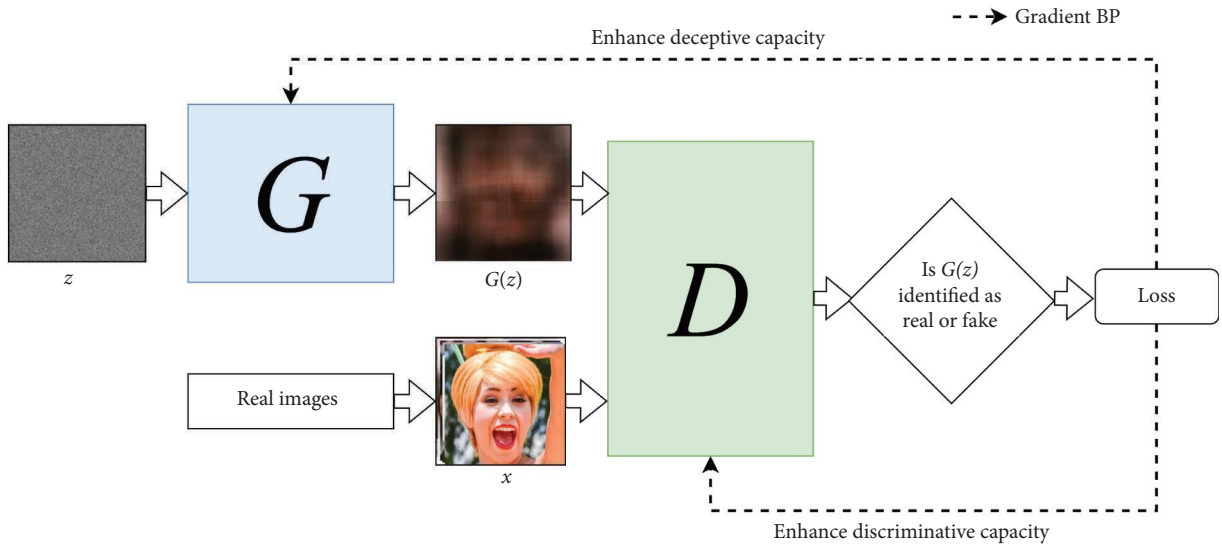

FIGURE 2: The structure of generative adversarial networks. The $\mathscr{G}$ represents the generator and the $\mathscr{D}$ represents the discriminator. The generative capacity of GANs is from the zero-sum non-cooperative game between the generator and the discriminator, which are optimized to a Nash equilibrium during training.

are listed in corresponding tables (Tables 1, 2, and 3) in terms of the proposed model (Structure), generation condition (condition, guidance to classifiers from users, can be descriptive text, class labels, etc.), specific task (Task), and modality (Modal).

2.2.1. Structure Enhancement of GANs. Deep Regret Analytic GAN [16] (DRAGAN) proposes viewing GAN training as a regret minimization process rather than divergence minimization between real and fake distributions. Through theoretical analysis of GAN convergence under this regret minimization process, the authors connect common training instabilities like mode collapse to undesirable local equilibria. They identify sharp discriminator gradients around some real data points as a characteristic of these poor local equilibria. To address this, the authors propose DRAGAN, a gradient penalty scheme that avoids degenerating local equilibrium and enables faster, more stable training with fewer mode collapses.

Wasserstein GAN [17] (WGAN) introduces a key change in the training of a discriminator. Instead of classifying inputs as real or fake, WGAN changes the discriminator to a critic that outputs real-valued estimates. This is achieved using the Wasserstein distance, a measure of distance in probability space, which provides a smoother and more meaningful gradient to the generator. Although WGAN shows improved training instability, WGAN uses weight clipping to enforce Lipschitz's constraint on the critic, which can still lead to poor performance and convergence failures. WGAN with gradient penalty [18] (WGAN-GP) proposes to penalize the gradient norm of the critic's input rather than clipping weights, which not only enhances the stability of WGANs but also supports stable training across various GAN structures with minimal hyperparameter adjustments.

CVAE-GAN [52] merges a VAE with a GAN to synthesize images in specific, fine-grained categories. It represents images as a combination of label and latent attributes within a probabilistic model, allowing for the generation of category-specific images by altering the category label and varying values in a latent attribute vector.

Progressive growing GAN [20] (PGGAN) was proposed to synthesize high-resolution human faces. The key idea of PGGAN is to improve the generator and the discriminator progressively. In detail, PGGAN starts generation from a model structure that can only generate low-resolution images. Then, PGGAN adds new layers that model increasingly fine details as training progresses, and the details of generated images, as training progresses, are also becoming more and more abundant. By adding details progressively, PGGAN can generate high-resolution images. Compared with low-resolution images, high-resolution images can better persuade recipients of content authenticity and better evade visual detection.

Least square GAN [21] (LSGAN) focuses on the vanishing gradient issues during training. This is because regular GANs model the discriminator as a classifier with sigmoid cross entropy loss. LSGAN proposed a least square loss for the discriminator. The least square loss is applied in an a-b coding scheme for the discriminator, where "a" and "b" represent the labels for fake and real data. LSGAN shows its effectiveness in improving both training stability and generated image quality.

BigGAN [22] added the orthogonal regularization and truncation trick in the large-scale training of GAN to achieve both high-fidelity images and a wide variety of generated samples.

Spectral normalization GAN [23] (SNGAN) stabilizes GAN training by normalizing discriminator weights in each iteration. Spectral normalization constrains the Lipschitz

Table 1: Representative GANs that enhance the GAN structures, where — in the condition column represents unconditional.

Ref	Structure	Condition	Task	Modal
[15]	GAN	—	Generation	Image
[16]	DRAGAN	—	Generation	Image
[17]	WGAN	—	Generation	Image
[18]	WGAN-GP	—/class	Generation	Image
[19]	CVAE-GAN	class	Generation	Image
[20]	PGGAN	—	Generation	Image
[21]	LSGAN	—/class	Generation	Image
[22]	BigGAN	—	Generation	Image
[23]	SNGAN	—/class	Generation	Image
[24]	AC-GAN	class	Generation	Image
[25]	AdaGAN	—	Generation	Image
[26]	UGAN	—	Generation	Image
[27]	U-Net GAN	—/class	Generation	Image
[28]	FastGAN	—	Generation	Image

Table 2: Representative GANs that generate Deepfake by synthesizing new content.

Ref	Structure	Condition	Task	Modal
[29]	DCGAN	—	Generation	Image
[30]	HistoGAN	—	Generation	Image
[19]	CVAE-GAN	—	Generation	face
[31]	StyleGAN	—	Generation	Image
[32]	StyleGAN2	—	Generation	Image
[33]	StyleGAN3	—	Generation	Image
[34]	StyleGAN-V	—	Generation	Video
[35]	StyleGAN-T	Text	Generation	Image
[36]	DAE-GAN	Text	Generation	Image
[37]	3D-GAN	—	Generation	3D object
[38]	VQGAN	—/class	Generation	Video
[39]	SSA-GAN	Text	Generation	Video
[40]	GT-GAN	—	Generation	Time series

Table 3: Representative GANs that generate Deepfake by manipulating existing content, where I2I represents image-to-image translation.

Ref	Structure	Condition	Task	Modal
[41]	StarGAN	Image	I2I	Image
[42]	StarGANv2	Image	I2I	Image
[43]	FSGAN	—	Faceswap/editing	Image
[44]	FSGANv2	Image	Faceswap/editing	Image
[45]	AttGAN	Image	Editing	Image
[46]	CAGAN	Text	Editing	Image
[47]	CycleGAN	Image	I2I	Image
[48]	SPMPGAN	Image	Editing	Image
[49]	Anyconst GAN	—/image	Editing	Image
[50]	T2CI-GAN	Text	Generation	Image
[51]	StyleCLIP	Text	Reenactment	Image

constant and limits the capacity of the discriminator. This prevents unstable oscillatory dynamics between the generator and discriminator during adversarial learning.

Auxiliary classifier GAN [24] (AC-GAN) improved training techniques for GANs to synthesize high-resolution 128×128 images with global coherence. The proposed label-conditioned GAN structure generates samples with clearer object details compared to lower-resolution images. Besides, the progressive growing technique also speeds up the training process and enhances its stability.

Adaptive GAN [25] (AdaGAN) aimed at addressing the problem of missing modes in GANs, which refers to the inability of a model to produce examples in certain regions of the data space. AdaGAN is an iterative procedure that enhances the training of GANs by adding a new component to a mixture model at each step. This is achieved by running a GAN algorithm on a reweighted sample, drawing inspiration from boosting algorithms.

Unrolled GAN [26] (UGAN) proposed a novel method to stabilize the training of GANs by redefining the

generator's objective in relation to an unrolled optimization of the discriminator. UGAN addresses a critical challenge in GAN training: balancing between using the optimal discriminator in the generator's objective, which is theoretically ideal but practically infeasible, and relying on the current value of the discriminator, which often leads to instability and suboptimal solutions.

U-Net GAN [27] proposes a U-Net-based discriminator structure for GANs to improve global and local image coherence, which provides detailed per-pixel feedback in addition to global image assessment. This guides the generator in synthesizing shapes/textures that are indistinguishable from real images. A per-pixel consistency regularization is also introduced using CutMix augmentation on the U-Net input. This focuses the discriminator on semantic/structural changes rather than perceptual differences. The proposed U-Net discriminator and consistency regularization improve GAN performance, enabling the generation of images with shapes, textures, and details highly faithful to the real data distribution.

FastGAN [28] studies few-shot high-resolution image synthesis with GANs using minimal computation and limited a small number of training samples. The proposed FastGAN is a lightweight GAN structure that is capable of producing superior image quality at a resolution of 1024×1024, with the training process requiring only a few hours on a single GPU.

2.2.2. Deepfake Synthesized With GANs. Deep convolutional GAN [29] (DCGAN) was proposed to synthesize new faces with a class of CNNs. By applying a set of skills to the structure of networks, such as replacing pooling layers with convolutions, adopting BatchNormalization (BN), and adopting a proper activation function, DCGAN stabilizes the training of the network. Though only able to generate low-resolution human faces, DCGAN is considered an early and classic work for fake face generation.

HistoGAN [30] was proposed to deal with the color inconsistency of GAN-generated images. Inspired by the idea of "style mixing" in StyleGAN [31, 32], HistoGAN first obtains histogram features by converting images into the log-chroma space. Then, a GAN network is established by replacing the last two blocks of StyleGAN a with the color histogram feature adopted to learn a lower dimensional representation. With the output of the network and a histogram-aware color-matching loss, the generated images better match the color histogram of the target images.

CVAE-GAN [19] combines a VAE with a GAN to synthesize images in fine-grained categories, such as faces of a specific person or objects in a category. CVAE-GAN models an image as a composition of labels and latent attributes in a probabilistic model. By varying the fine-grained category label fed into the resulting generative model, CVAE-GAN generates images in a specific category with randomly drawn values on a latent attribute vector.

The StyleGAN series comprises a progression of Deepfake generation models featuring progressively enhanced generative capabilities. StyleGAN [31] was proposed by making improvements to PGGAN. StyleGAN does not feed the learned latent code to the generator. Instead, latent code is applied for style mixing before arriving at the generator. Style mixing aims to discover the position of latent code that controls different styles by a mapping network. The latent code is mapped to style code through the network, and the style code is mixed and then fed to the generator. Style mixing allows StyleGAN to control style at each layer in the generator. Besides, StyleGAN introduced noise at each layer in the generator to improve generation diversity. By style mixing and introduction of noise in the generator, StyleGAN synthesizes high-quality images of high resolution and better diversity. StyleGAN2 [32] focuses on the water droplet disadvantage of StyleGAN. It is pointed out that the water droplet appears due to the AdaIN operation, which normalizes each feature map separately and destroys the information between feature maps. StyleGAN2 adopted weight demodulation to remove the water droplet while retaining full controllability. Noticing that the details of generated images are unnaturally fixed to pixel coordinates rather than object surfaces due to aliasing in the generator, StyleGAN3 [33] analyzed that the root cause of this is traced to careless signal processing. Interpreting all signals as continuous, StyleGAN3 derives small architectural changes to prevent unwanted information from leaking into the hierarchical synthesis. StyleGAN-V [34] focuses on video generation by treating videos as continuous signals. StyleGAN-V is a continuous-time video generator that leverages continuous motion representations through positional embeddings. StyleGAN-V also adopted a holistic discriminator that aggregates temporal information more efficiently by concatenating frame features. By training the model on high-resolution images, StyleGAN-V generates high-resolution long videos at high frame rates. StyleGAN-T [35] is a model designed to meet the specific demands of large-scale text-to-image synthesis. Compared with existing generators that require iterative evaluation to generate a single image, StyleGAN-T offers a much faster alternative as they can generate images in a single forward pass. StyleGAN-T is also designed to meet the specific demands of large-scale text-to-image synthesis, including large capacity, stable training on diverse datasets, strong alignment with text inputs, and a balance between controllable variation and text alignment.

Dynamic aspect-aware GAN [36] (DAE-GAN) is a text-to-image synthesizer representing text at multiple levels such as sentence and word aspect. DAE-GAN mimics human iterative drawing by globally enhancing the image and then locally adjusting details based on text. In detail, DAE-GAN generates an initial image from the sentence embedding and then refines it using a novel aspect-aware dynamic re-drawer (ADR). ADR alternates between global refinement with word vectors (attended global refinement) and local detail refinement with aspect vectors (aspect-aware local refinement).

3D-GAN [37] is capable of generating 3D objects by leveraging volumetric convolutional networks and GANs to create 3D objects from a probabilistic space. 3D-GAN features an adversarial criterion, which enables the generator

to implicitly understand and replicate the structure of 3D objects, resulting in high-quality outputs. 3D-GAN can generate 3D objects without reference images or CAD models, exploring the manifold of 3D objects.

VQGAN [38] aims to generate long videos by combining 3D-VQGAN with transformers to generate videos comprising thousands of frames. Trained on various datasets, VQGAN is able to generate high-quality long videos of impressive diversity and coherency. By incorporating temporal information with text and audio, VQGAN can generate meaningful long videos that are contextually rich and aligned with the given inputs.

Semantic-spatial aware GAN [39] (SSA-GAN) is designed to generate photorealistic images that are not only semantically consistent with the overall text description but also align with specific textual details. In detail, SSA-GAN enhances the fusion of text and image features and introduces a semantic mask learned in a weakly supervised manner to guide spatial transformation in the image.

GT-GAN [40] generates time series data by handling both regular and irregular time series data without requiring modifications to the model. GT-GAN integrates a range of related techniques, from neural ordinary/controlled differential equations to continuous time-flow processes, into a cohesive system. This integration is a critical factor in the GT-GAN's success.

2.2.3. Deepfake Manipulated With GANs. StarGAN [41] enables image-to-image translations across multiple domains with a single model. StarGAN's unified structure allows for simultaneous training on multiple datasets with different domains within one network. This leads to superior quality in translated images compared to existing models and provides the flexibility to translate any input image into any desired target domain. StarGAN can be used on facial attribute transfer and expression synthesis. StarGANv2 [42] presents a single, scalable framework for diverse high-quality image-to-image translation across multiple domains, unlike existing approaches with limited diversity or separate models per domain. Through experiments on CelebA-HQ and a new animal faces dataset AFHQ, StarGAN v2 demonstrates superior visual quality, diversity, and scalability compared to baselines. The release of AFHQ provides a valuable benchmark for assessing image translation models in terms of large inter- and intradomain variability. Overall, StarGAN v2 tackles the key criteria of diversity and scalability for multidomain image mapping within a single unified generator model.

FSGAN [43] enables face swapping and reenactment between pairs of faces without requiring model training on those faces specifically. Key technical contributions include an RNN-based approach to adjust for varying poses and expressions, continuous view interpolation for video, completion of occluded regions, and a blending network for seamless skin and lighting preservation. A novel Poisson blending loss integrates Poisson optimization and perceptual loss. FSGANv2 [44] is unique in that it is subject agnostic, enabling face swapping on any pair of faces without specific training on those faces. It features an iterative DL-based approach for face reenactment, capable of handling significant pose and expression variations in both single images and video sequences. For videos, it offers continuous interpolation of face views using reenactment, Delaunay triangulation, and barycentric coordinates. A face completion network manages occluded face regions, and a face blending network ensures seamless blending while preserving skin color and lighting conditions, utilizing a novel Poisson blending loss.

AttGAN [45] was proposed to make more accurate facial attribute manipulation. The proposed method is based on the observation that attribute-independent latent representation applied in some attribute manipulation methods restricts the capacity of the latent representation. Thus, the generated human faces are oversmooth and distorted. To make an improvement, AttGAN proposed reconstruction learning and adversarial learning to make the generated images visually more realistic. In detail, the proposed reconstruction learning preserves attribute-excluding details by making attribute editing, and the reconstruction task shares the entire encoder–decoder network. The proposed adversarial learning provides the whole network with the capacity for visually realistic editing. Thus, AttGAN can achieve high-quality facial attribute manipulation by making faces more natural.

CAGAN [46] tackles automatic global image editing from linguistic requests to significantly reduce labor for novices. An editing description network (EDNet) is proposed to predict embeddings for image generator cycles. Despite data imbalance, several augmentation strategies combined with an Image-Request Attention module enable spatially adaptive edits. A new semantic evaluation metric is also introduced as superior to pixel losses. Extensive experiments on two benchmarks demonstrate state-of-the-art performance, with effectiveness stemming from explicitly handling limited language-image examples and selective region-based editing guided by linguistic cues.

CycleGAN [47] is an image-to-image translation GAN, which has been used for face-swapping and expression reenactment manipulation. The basic idea under CycleGAN design is interesting. It assumes that for two domains, the source domain and the target domain, the mapped target domain representation of an input, when mapped in the opposite direction, the backward-mapped result should be close to the input, which also explains its name. Based on this idea, the cycle loss is proposed to optimize the whole network.

SPMPGAN [48] achieves semantic image editing by generating desired content in specified regions based on local semantic maps. SPMPGAN introduces a style-preserved modulation with two steps: fusing contextual style and layout to generate modulation parameters and then employing those to modulate feature maps. This injects a semantic layout while maintaining context style. A progressive structure further enables coarse-to-fine-edited content generation. By preserving style, more consistent results are obtained with less noticeable boundaries between edited regions and known pixels. Overall, explicit style modeling significantly improves semantic editing quality and coherence.

Anyconst GAN [49] enables interactive image editing with GANs by allowing fast, perceptually similar previews at reduced computational costs. Inspired by quick preview features in creative software, Anyconst GAN supports variable resolutions and channels for generation at adjustable speeds. Subset generators serve as proxies for full results, trained using sampling, adaptation, and conditional discrimination to improve visual quality over separately tuned models. Novel projection techniques also encourage output consistency across subsets. Anyconst GAN provides up to 10x cost reduction and 6–12x preview speedups on CPUs and edge devices without perceivable quality loss, crucially enabling interactivity.

T2CI-GAN [50] addresses the challenge of generating visual data from textual descriptions, a task that combines natural language processing (NLP) and computer vision techniques. While current methods use GANs to create uncompressed images from text, T2CI-GAN focuses on generating images directly in a compressed format for enhanced storage and computational efficiency. T2CI-GAN employs DCGAN and proposes two models: One trained with JPEG-compressed DCT images to generate compressed images from text and another trained with RGB images to produce JPEG-compressed DCT representations from text.

StyleCLIP [51] manipulates images using the StyleGAN framework, inspired by its ability to generate highly realistic images. The focus is on utilizing the latent spaces of StyleGAN for image manipulation, which traditionally requires extensive human effort or a large annotated image dataset for each manipulation. StyleCLIP introduces the innovative use of contrastive language-image pretraining (CLIP) models to develop a text-based interface for StyleGAN image manipulation, eliminating the need for manual labor. The approach includes an optimization scheme that modifies an input latent vector based on a text prompt using a CLIP-based loss. A latent mapper is also developed to infer text-guided latent manipulation steps for input images, allowing quicker and more stable text-based manipulation. The paper also presents a method to map text prompts to input-agnostic directions in StyleGAN's style space, facilitating interactive text-driven image manipulation.

2.3. Forensics of GAN-Generated Deepfake. Detecting Deepfakes involves distinguishing between two primary categories: synthesized Deepfake and manipulated Deepfake. Synthesized Deepfake detection focuses on identifying media generated entirely by AI algorithms, while detecting manipulated Deepfake involves recognizing authentic media that has been altered or manipulated using AI-based tools. The investigated works of synthetic Deepfake and manipulated Deepfake are listed in Tables 4 and 5, where we show the adopted or proposed classifier (Classifier), adopted datasets (Datasets) and the modality (Modal).

2.3.1. Detection of GAN-Synthesized Deepfake

2.3.1.1. Detectors Based on Spatial Artifacts. Spatial artifacts are intuitive and are adopted widely to detect GAN-generated Deepfake.

Xception was first adopted in [53] to detect Deepfake. The detector is tested on the proposed benchmark dataset, FaceForensics++. Thorough analysis shows that while manipulations can fool humans, data-driven detectors leveraging domain knowledge can achieve high accuracy even with strong compression.

DeepFD [54] adopted contrastive loss while seeking typical features of the synthesized images generated by different GANs.

Nguyen et al. [56] tackle open-world detection through incremental learning. Experiments on a dataset of various GAN-generated images demonstrate that the proposed approach can correctly discriminate when new GAN structures are presented, enabling continuous evolution with emerging data types.

Bonettini et al. [57] ensemble CNN models for video face forgery detection targeting modern techniques. An EfficientNetB4 base network is augmented with attention layers and siamese training for diversity. Combining these complementary models provides promising detection results on two datasets with over 119,000 videos. The study demonstrates integrating multiple specialized networks through attention and metric learning improves generalization to evolving forgery methods compared to individual models.

ForgeryNet [58] is introduced as an extensive benchmark dataset to advance digital forgery analysis, addressing limitations of diversity and task coverage in prior face forgery corpora. It contains 2.9M images and 221K videos with unified annotations for four key tasks: image classification (binary, 3-class, 15-class), spatial localization, video classification with random frame manipulation, and temporal localization. By providing massive scale, variety, and multitask annotations, ForgeryNet aims to promote research and innovation in facial forgery detection, spatial localization, temporal localization, and related directions.

Pairwise self-consistency learning (PCL) [65] detects Deepfake through inconsistent source features preserved despite manipulation. A novel PCL method trains models to extract these indicative cues. An inconsistency image generator (I2G) synthesizes training data with source mismatches. By exposing remnants of original input traces, the learned representations reliably detect Deepfake generated even by advanced methods. The strong cross-dataset performance demonstrates real-world viability.

FD^2 Net [67] processed biological information in a more complex way, and they proposed that a face image is a production from the intervention of the underlying 3D geometry and the lighting environment. The lighting environment can further be decomposed into direct light, common texture, and identity texture. Noticing that direct light and identity texture contain critical clues, these two components are merged and named facial detail images. The proposed Forgery-Detection-with-Facial-Detail (FD^2 Net) is mainly based on facial images and facial detail images. The two types of images are passed through a two-stream network, and outputs of the two streams are then fused and passed to a classifier.

Vision transformer (ViT) uses self-attention mechanism to process all input tokens simultaneously, allowing for

Table 4: Deepfake detectors for synthetic Deepfake.

Ref	Classifier	Datasets	Modal
[53]	Xception	FF++	Image
[54]	DeepFD	Self-built (StarGAN, DCGAN, WGAN, WGAN-GP, LSGAN, PGGAN)	Image
[55]	CNNs	FF++, self-built (ProGAN, Glow)	Image
[56]	Capsule-Forensics	FF++, self-built (ProGAN, Glow)	Image/video
[57]	CNNs	FF++, DFDC	Video
[58]	Various	ForgeryNet	Image/video
[59]	MT-SC	Self-built (CycleGAN, ProGAN, Glow, StarGAN)	Image
[60]	SVM	FF++	Image
[61]	F^3-Net	FF++	Image
[62]	CNNs, LSTM	FF++, Celeb-DF, DFDC preview	Video
[63]	MesoInception, ResNet, Xception	FF++, self-built (ProGAN, StyleGAN, StyleGAN2)	Image
[64]	FakeLocator	FF++, self-built (STGAN, AttGAN, StarGAN, IcGAN, StyleGAN, PGGAN, StyleGAN2)	Image
[65]	PCL + I2G	FF++, DFD	Image
[66]	Xception	FF++, DFD, DFDC, Celeb-DF, DF1.0	Image
[67]	FD^2 Net	FF++, DFD, DFDC	Image
[68]	Transformer	FF++, DFD, DFDC	Video
[69]	Transformer	FF++, FaceShifter, DeeperForensics, DFDC	Video
[70]	CNNs	FF++, DFDC, Celeb-DF	Image
[71]	Transformer	FF++, Celeb-DF, self-built: SR-DF (FSGAN, FaceShifter, first-order motion, IcFace)	Image
[72]	ICT	MS-Celeb-1M, FF++, DFD, CelebDFv1, CelebDFv2, DeeperForensics	Image
[73]	HCIL	FF++, Celeb-DF, DFDC preview, WildDeepfake	Video
[74]	FST-matching	FF++	Image
[75]	CNNs	FF++, Celeb-DF	Image
[76]	UIA-ViT	FF++, Celeb-DF, DFD, DFDC	Image
[77]	HFI-Net	FF++, Celeb-DF, DFDC, TIMIT	Image
[78]	CNNs	FF++, DFDC, Celeb-DF	Image
[79]	EfficientNet, transformers	FF++, DFDC	Video
[80]	CNNs	Self-built (ProGAN, SN-DCGAN, CramerGAN, MMDGAN)	Video
[81]	LRNet	FF++, Celeb-DF	Video
[82]	Transformer	FF++, FaceShifter, DeeperForensics, Celeb-DF, DFDC	Video
[63]	CNNs	Self-built (PGGAN, StyleGAN, StyleGAN2)	Image
[83]	HAMMER	DGM^4	Image
[84]	MRE-Net	FF++, Celeb-DF, DFDC, Wild Deepfake	Video
[85]	CNNs	DF_TIMIT, FF++, DFD, DFDC, Celeb-DF	Image
[86]	TI^2 Net	FF++, Celeb-DF, DFD, DFDC	Video
[87]	IIL	FF++, DFDC, Celeb-DF	Image
[88]	ISTVT	FF++, FaceShifter, DeeperForensics, Celeb-DF, and DFDC datasets	Image
[89]	AVoiD-DF	DefakeAVMiT, FakeAVCeleb, DFDC	Video

Table 5: Deepfake detectors for manipulated Deepfake.

Ref	Classifier	Datasets	Modal
[54]	DeepFD	Self-built (StarGAN, DCGAN, WGAN, WGAN-GP, LSGAN, PGGAN)	Image
[90]	RNN/VAE	Self-built (FaceApp)	Video
[91]	Face X-ray	FF++, DFD, DFDC, Celeb-DF	Image
[55]	CNNs	FF++, self-built (ProGAN, Glow)	Image
[60]	SVM	FF++	Image
[92]	ResNet-50	Self-built (ProGAN, StyleGAN, BigGAN, CycleGAN, StartGAN, GauGAN, etc.)	Image
[93]	XceptionNet, VGG16	FF++, self-built (FaceApp, StarGAN, PGGAN, StyleGAN)	Image
[63]	MesoInception4, ResNet, Xception	FF++, self-built (ProGAN, StyleGAN, StyleGAN2	Image
[64]	FakeLocator	FF++, self-built (STGAN, AttGAN, StarGAN, IcGAN, StyleGAN, PGGAN, StyleGAN2)	Image
[71]	Transformer	FF++, Celeb-DF, self-built: SR-DF (FSGAN, FaceShifter, first-order-motion, IcFace)	Image

better parallelization and capturing long-range dependencies more effectively. Khan and Dai [68] tackle Deepfake video detection through a novel incremental learning video ViT. 3D face reconstruction provides aligned UV textures and poses. Dual image-texture inputs extract complementary cues. Incremental fine-tuning further improves performance with limited additional data. Comprehensive experiments on public datasets demonstrate state-of-the-art Deepfake video detection. Sequenced feature learning from image pairs and incremental optimization enables robust identification of forged footage spread online.

Nirkin et al. [70] detect identity manipulations in faces by exposing discrepancies between the face region and surrounding context. While methods like Deepfakes aim to adjust the face to match the context, inconsistencies that reveal manipulation arise. A two-network approach is proposed, with one network focused on the segmented face and the other recognizing the face context. By comparing their recognition outputs, discrepancies indicative of manipulation are identified, complementing conventional real versus fake classifiers. Crucially, the approach generalizes to unseen manipulation methods by capitalizing on the common mask-context mismatch.

FST-matching (fake, source, target images matching) [74] provides insights into how Deepfake detection models learn artifact cues when trained on binary labels. Three hypotheses based on image matching are proposed and verified: (1) Models rely on visual concepts unrelated to the source or target as artifact-relevant. (2) FST-matching between fake, source, and target images enables implicit artifact learning. (3) Concepts learned from uncompressed data fail on compression. An FST-matching detection model is proposed that matches triplets to make compression-robust concepts explicit.

Liang et al. [75] address overfitting in CNN Deepfake detection caused by reliance on dataset content biases over manipulation artifacts. A disentanglement framework is designed to remove content interference. Content consistency and global representation contrastive constraints further enhance independence. The proposed framework significantly improves generalization by guiding focus toward artifacts rather than contextual cues specific to a dataset. Visualizations and results demonstrate state-of-the-art detection through explicit content disentanglement and constraint-driven artifact emphasis.

Yu et al. [78] improve Deepfake video detection generalization through commonality learning across manipulation methods. Specific forgery feature extractors (SFFExtractors) are first trained separately on each method using losses to ensure detection ability. A common forgery feature extractor (CFFExtractor) is then trained under SFFExtractor supervision to discover shared artifacts. The proposed strategy provides an effective way to learn broadly indicative forged video traits that transcend specific manipulation types and improve out-of-distribution detection.

Chai et al. [63] analyze spatial patterns revealing fakes using a patch-based classifier to identify more easily detectable regions. A technique is proposed to exaggerate these detectable properties. Experiments show even an adversarially trained generator leaves localized flaws that can be exploited. The analysis and visualization of spatial detection cues provide insights into artifacts persistent across datasets, structures, and training variations. Localizing and amplifying subtle clues improves understanding of what makes current forged multimedia detectable even as human perception fails.

HAMMER [83] was proposed to tackle the newly proposed detecting and grounding multimodal media manipulation (DGM4), which exposes subtle forgery across images and text. Beyond binary classification, DGM4 grounds specific manipulated regions in both modalities through deeper reasoning. A large dataset with diverse manipulation types and annotations is constructed to enable investigation. The proposed HAMMER captures fine-grained cross-modal interactions via manipulation-aware contrastive learning across encoders and modality-aware aggregation. Dedicated heads integrate uni- and multimodal reasoning for detection and spatial grounding. DGM4 defines a new problem and benchmark for tackling realistic multimodal misinformation by moving beyond binary single-modality classification to spatial localization and reasoning.

Li et al. [85] propose forensic symmetry to detect Deepfake by comparing natural features between symmetrical facial patches. A multistream model extracts symmetry features from frontal images and similarity features from profiles. Collectively representing natural traits, discrepancies in feature space indicate manipulation likelihood. Face patch pairs are mapped to an angular hyperspace to quantify natural feature differences and tampering probability. A heuristic algorithm aggregates frame-level predictions for video-level judgment. The proposed forensic technique reliably detects subtle asymmetric perturbations indicative of manipulation by explicitly encoding and contrasting innate facial symmetries. The robustness to compression and generalization across datasets highlights applicability to real-world Deepfake detection.

Dong et al. [87] analyze generalization issues in Deepfake detection models stemming from implicit identity representations. Binary classifiers unexpectedly learn dataset-specific identity traits rather than generalizable manipulation cues. This implicit identity leakage severely harms out-of-distribution detection. To mitigate, an ID-unaware model is proposed that reduces identity influence by obscuring hair/background. By identifying and addressing identity overfitting, this exploration provides critical insights into improving Deepfake detection reliability when distributions shift. The proposed techniques move toward adaptable and widely applicable models.

2.3.1.2. Detectors Based on Frequency Artifacts. Detectors are also looking for more subtle artifacts in the frequency domain to recognize Deepfake.

Frank et al. [80] proposed the first work introducing frequency information to Deepfake detection. They observed that although some GAN-generated images can be visually undetectable, obvious artifacts can be easily identified in the frequency domain. In detail, spatial images are first transformed to the frequency domain by DCT. With frequency

features, even a simple classifier can achieve a classification accuracy of over 80%. A CNN-based classifier easily achieved a classification accuracy of over 99% on both LSUN [94] and CelebA [95] datasets. However, such a simple utilization of frequency features performs poorly on more challenging datasets.

F^3-Net [61] captures both global and local frequency features of images. F^3-Net captures global frequency features with a frequency-aware decomposition (FAD) module. Together with simple fixed filters, FAD applies a set of learnable filters to decompose images into different frequency component images. The decomposed image components are stacked and fed into a CNN to generate deeper features. Local frequency statistics (LFS) is the second frequency-aware module of F^3-Net to capture local frequency information. LFS first collects frequency responses in small patches of an image with sliding window discrete Fourier transform (SWDCT). Then, LFS calculates the mean frequency responses at a series of learnable frequency bands. The features learned by FAD and LFS are then fused by a cross-attention for later prediction.

Luo et al. [66] relied on high-frequency features to make their detector more generalizable. They proposed a two-stream framework. One stream is the RGB stream, and the other one is the high-frequency stream. The RGB stream utilizes raw images, while the high-frequency stream utilizes residual images generated by SRM filters. The raw image features and residual features then go through three flows: entry flow, middle flow, and exit flow. Entry flow extracts multiscale high-frequency features by the combination of down-samplings and multiscale SRMs. The output of entry flow is then passed to middle flow, where dual cross-modality attention is applied to model the correlations between the two modalities at different scales. The exit flow contains the fusion model and the final classifier. The proposed method achieved over 99% AUC if trained and tested on images manipulated by the same techniques.

HFI-Net [77] tackles face forgery detection through hierarchical frequency modeling to improve generalization. A dual CNN-transformer network captures multiscale cues. A novel frequency-based feature refinement (FFR) module applies frequency attention to emphasize artifacts while suppressing pristine semantics. FFR enables cosharing global–local interactions to fuse complementary branches using frequency guidance. Stacking at multiple levels exploits hierarchical frequency artifacts. By thoroughly exposing frequency-domain discrepancies at multiple semantic levels, the hierarchical framework significantly advances cross-dataset and robust Deepfake detection.

2.3.1.3. Detectors Based on Temporal Artifacts. Temporal artifacts are also explored to analyze sequential information within image patches and video frames.

Videos naturally contain dynamic information, which can be used to detect their authenticity. Masi et al. [62] adopted a recurrent neural network (RNN) to better capture time serial features. It proposed a two-branch framework to detect Deepfake videos by color domain information and frequency-domain clues. The frequency branch is based on the Laplacian of Gaussian. In detail, the input is processed with two sets of filters. The first set is fixed and nonlearnable, which works as a 2D Gaussian kernel. The second set, a dimensionality reduction filter, maps back the dimensionality to the input expected by the next layer. The separated two branches are then fused by the fusion layer, and the outputs of the fusion layer then go through the backbone network and a bidirectional LSTM module for temporal feature analysis.

LRNet [81] aims at detecting Deepfake videos via temporal modeling of precise geometric features rather than vulnerable appearance cues. A calibration module enhances the precision of extracted geometry. A two-stream RNN structure sufficiently exploits temporal patterns. Compared to prior methods, LRNet has lower complexity, easier training, and increased robustness to compression and noise.

Zheng et al. [69] leverages temporal coherence for video face forgery detection through a two-stage framework. First, a fully temporal convolution network with 1×1 spatial kernels extracts short-term patterns while improving generalization. Second, a temporal transformer captures long-term dependencies. Trained from scratch without pretraining, the approach achieves state-of-the-art performance and transfers effectively to new forgery types—highlighting robust sequence modeling. By thoroughly exploiting motion cues absent in images, the work advances temporal machine learning for identifying fake video faces.

Gu et al. [73] detect Deepfake videos through hierarchical contrastive learning of temporal inconsistencies in facial movements. Unlike binary supervision methods, a two-level contrastive paradigm is proposed for local and global inconsistency modeling. Multisnippet inputs enable contrasting real and fake videos from both intra- and intersnippet perspectives. Region-adaptive and cross-snippet fusion modules further enhance discrepancy learning. The hierarchical framework advances Deepfake video detection through comprehensive modeling of subtle spatiotemporal discrepancies in facial dynamics.

TI2Net [86] detects Deepfake video by exposing temporal identity inconsistencies. It captures discrepancies in face identity across frames of the same person. Encoding identity vectors and modeling their evolution as a temporal embedding enables reference-agnostic fake detection without dataset-specific cues. The identity inconsistency representation is used for authenticity classification. Triplet loss improves embedding discriminability. By focusing on semantic identity integrity rather than superficial artifacts, TI2 Net advances reliable Deepfake video detection through a principled temporal discrepancy modeling approach.

Dynamic prototype network (DPNet) [82] uses dynamic prototypes to expose temporal inconsistencies. Most methods process videos frame-by-frame, lacking temporal reasoning, despite artifacts across frames being essential to human judgment. DPNet represents videos through evolving prototypes that capture inconsistencies and serve as visual explanations. Extensive experiments show competitive performance even on unseen datasets while providing intuitive dynamic prototype explanations. Further temporal

logic specifications verify model compliance to desired behaviors, establishing trust.

Multirate excitation network (MRE-Net) [84] leverages dynamic spatial–temporal inconsistencies for Deepfake video detection. MRE-Net uses bipartite group sampling at diverse rates to cover varying facial motion. Early stages excite momentary inconsistencies via spatial and short-term cues. Later stages capture long-term intergroup dynamics. This holistic modeling of transient, localized flaws and long-range coherence surpasses methods relying on isolated spatial or temporal cues.

Coccomini et al. [79] detect video Deepfake of faces as advances in VAEs and GANs enable highly realistic forged footage. Combining EfficientNet feature extraction with ViTs provides comparable performance to the state-of-the-art without requiring ensembles or distillation. A simple voting scheme handles multiple faces. The best model achieves 0.951 AUC on DFDC, nearing the state-of-the-art for detecting sophisticated forgeries. Unlike complex prior art, transformers and streamlined inference provide efficient and accurate Deepfake detection ready for real-world deployment. By showing strong results are possible without weighting or model combinations, this exploration demonstrates the power of transformers and presents a straightforward framework for exposing falsified video faces.

While transformers can be adopted to process video to explore temporal patterns among video frames, they can also process an image as a series of patch embeddings. Therefore, transformer-based detectors first split images into patches and represent the patches in latent space as patch embeddings. Then, self-attention is adopted to explore the sequential inconsistencies between patches to recognize Deepfake. The process considers both patch content and the correlation between patches, allowing for better exploration of structural information for Deepfake detection. Identity consistency transformer (ICT) [72] detects face forgery by exposing identity mismatches between inner and outer facial regions. It incorporates a consistency loss to determine the identity (in)congruence of face parts. Experiments demonstrate superior generalization across datasets and real-world image degradations like Deepfake videos. Leveraging semantics outperforms low-level artifacts prone to obfuscation. ICT also readily integrates external identity cues, providing particular utility for celebrity Deepfake. By honing in on high-level facial identity integrity, ICT advances robust and semantically grounded fake face detection. UIA-ViT [76] enables inconsistency-aware Deepfake detection without pixel-level supervision. Vision transformers are proposed to capture patch consistency relations via self-attention implicitly. Two key components are introduced, unsupervised patch consistency learning (UPCL) with pseudolabels for consistency representations and progressive consistency weighted assemble (PCWA) to enhance classification. Without location annotations, UPCL and PCWA guide transformers to focus on discrepancies indicative of manipulation. By capitalizing on self-attention localization and consistency-driven learning, the approach advances robust generalization of Deepfake detection without expensive pixel-level labels. Interpretable spatial–temporal video transformer (ISTVT) [88] improves the exploitation of joint spatial–temporal cues for robust Deepfake detection. A decomposed spatial–temporal self-attention captures localized artifacts and temporal dynamics. Visualization via relevance propagation explains model reasoning on both dimensions. Strong detection performance is demonstrated on FaceForensics++ and other datasets, with superior cross-dataset generalization. The transformer structure enhanced by explicit spatial–temporal modeling outperforms recurrent and 3D convolutional approaches. Interpretability insights are provided into spatiotemporal patterns indicative of manipulation learned by the model. By advancing joint spatial–temporal reasoning while remaining interpretable, ISTVT moves toward truly understanding and exposing the underlying footprint of forged multimedia.

The mismatched temporal patterns between audio and video can also help detect AI-generated videos. AVoiD-DF [89] tackles Deepfake detection by modeling audio-visual inconsistencies in multimodal media. AVoiD-DF encodes spatiotemporal cues and then fuses modalities through joint decoding and cross-modal classification. The framework exploits intermodal and intramodal discrepancies indicative of manipulation. A new benchmark dataset, DefakeAVMiT, contains videos with either visual or audio tampering via diverse methods, enabling multimodal research.

2.3.2. Detection of GAN-Manipulated Deepfake. Neural ODE [90] was proposed and trained on original videos' heart rates. Testing on commercial and database Deepfake, the model successfully predicts fake heart rates. By evidencing detectable physiological signals in synthesized video, this exploration of Neural-ODE-based heart rate prediction represents a novel approach to automatically discerning real from forged footage.

Face X-ray [91] reveals blended boundaries in forged faces without relying on manipulation technique specifics. Observing a common blending step across methods, face x-ray exposes composites via blending edges in forged images and absence in real ones. Requiring only blending logic at training, face X-ray generalizes to unseen state-of-the-art techniques, unlike existing detectors reliant on method artifacts.

Matern et al. [55] investigated detecting artifacts in video face editing to expose manipulated footage. Many algorithms exhibit classical vision issues from tracking/editing pipelines. Analyzing current facial editing approaches identifies characteristic flaws like Deepfakes and Face2Face. Simple visual feature techniques can effectively expose such manipulations, using explainable signals accessible even to non-experts. These methods easily adapt to new techniques with little data yet achieve high AUC despite simplicity.

Durall et al. [60] made a theoretical analysis and suggested that upsampling (up-convolutional) operation in generative models leaves frequency artifacts in generated images, which may lead to GAN fingerprints. They further noticed that frequency artifacts exist, especially in the high-

frequency spectrum, which is consistent with the observation in [96] that high-frequency components carry more GAN fingerprint information. Thus, the GAN fingerprint can be considered a consequence of frequency distortion in the generative model caused by upsampling operations. The observation of fingerprints makes sense, since it is closely related to frequency distortions and inspired many works to look for Deepfake clues in the frequency domain.

Wang et al. [92] explore a "universal" detector to distinguish real images from various CNN-generated images regardless of structure or dataset. With preprocessing, augmentation, and a classifier trained only on ProGAN, surprising cross-structure and dataset generalization are achieved, even with methods like StyleGAN2. The findings suggest modern CNN-generated images share systematic flaws preventing realistic synthesis, evidenced by the success of the simple universal fake detector.

Dang et al. [93] tackle detecting manipulated faces and localizing edited regions, which is crucial as synthesis methods produce highly realistic forgeries. Simply using multitask learning for classification and segmentation prediction provides limited performance. Instead, an attention mechanism is proposed to process and improve classification-focused embeddings by highlighting informative areas. This simultaneously benefits binary real versus fake classification and visualizes manipulated areas for localization. A large-scale fake face dataset enables thorough analysis, showing attention-enhanced models outperform baselines in detecting facial forgeries and revealing the edits through interpretable attention maps. Overall, selective feature improvement and visualization via attention mechanisms provide superior hybrid detection and localization of modern forged face imagery.

Chai et al. [63] proposed to recognize Deepfake by the analysis of image patches, where the final classification was not made globally on the image. Instead, a shallow network with local receptive fields was applied to first predict whether patches are real or fake. The final decision of image authenticity is made by aggravating the classification results of all patches. In this way, the general classifier achieves both detection and localization tasks.

FakeLocator [64] tackles detecting and locating GAN-based fake faces by tracing upsampling artifacts within generation pipelines. Observing such imperfections provides valuable clues. A technique called FakeLocator is proposed to produce high-accuracy localization maps at full resolution. An attention mechanism guides training for universality across facial attributes. Data augmentation and sample clustering further improve generalization over various Deepfake methods. By capitalizing on tell-tale upsampling artifacts, FakeLocator advances multimedia forensics by exposing manipulation boundaries in addition to classification.

M2TR [71] detects Deepfakes through multiscale transformers capturing subtle manipulations. Most methods map images to binary predictions, missing consistency patterns indicative of forgery. Multi-modal Multi-scale TRansformer (M2TR) applies transformers across patch scales, exposing local inconsistencies. Frequency information further improves detection and compression robustness via cross-modality fusion. A new high-quality Deepfake dataset, SR-DF, addresses the limitations of diversity and artifacts in existing benchmarks.

2.3.3. Detection Performance Comparisons. We also compare the detection performance of some state-of-the-art methods in Table 6. We compared the performance of the detector on the test subset of the training dataset (in-set) and its generalization performance on an unseen dataset (cross-set). It can be observed that the most Deepfake detectors can perform well and achieve over 90% classification accuracy and AUC, some even achieve over 99% prediction accuracy (F^3-Net and FakeLocator). However, the generalization ability of these detectors remains a significant challenge, as their performance drops considerably when evaluated on the cross-set. The generalization ability is heavily influenced by the data distribution differences between the training and testing datasets. Detectors generally perform better on datasets with similar data distributions to the training set. For instance, UIA-ViT and Face X-ray demonstrate strong generalization when tested on datasets with similar context, quality, and identities to the training set, achieving over 90% AUC and accuracy, respectively. These generalization challenges highlight the primary difficulty in deploying Deepfake detectors in real-world applications, where the diversity of data (contexts, qualities, and identities) can vary greatly.

2.4. Discussion of GAN-Based Deepfake. When new Deepfake techniques appear and bring threats, new detectors with corresponding detective capacity will be proposed to protect people from misinformation about Deepfake. The generation development enables Deepfake to survive existing detectors with more realistic outputs. The generation and detection race is naturally a cat-and-mouse race. We summarize the adversarial development of Deepfake generation and detection in Figure 3.

When Deepfake was only able to generate low-quality outputs [16, 17, 52], detectors adopt intuitive features such as abnormal color [70, 91], textures [53, 54, 56, 65], and biological information [67, 72, 86, 90].

Enhanced GANs then try to improve the generative quality in terms of realism and resolution [31–34, 36], and the enhanced image quality can evade the detection of the above naive detectors. Therefore, the detectors look into more subtle artifacts such as frequency [61, 66, 77, 80] and dynamic artifacts [73, 84, 88].

Both generation and detection are becoming increasingly powerful and advancing rapidly through mutual competition.

Therefore, when more powerful detectors have been developed to recognize GAN-generated Deepfake, the research on generation has naturally focused on improving GANs to produce more realistic Deepfake content that can evade detection. Despite numerous improvements and variations to enhance generation quality, the researchers encountered several inherent limitations when refining

Table 6: Comparison of detection performance for GAN-generated Deepfakes. Metrics for in-set and cross-set performance are accuracy/area under ROC curve (AUC).

Classifier	Artifact type	In-set	Cross-set
Xception	Spatial	99.26/—	—/—
UIA-ViT	Spatial	—/99.33	—/94.68
Capsule-Forensics	Spatial	93.11/-	—/—
F^3-Net	Frequency	99.99/99.9	—/—
Face X-ray	Frequency	—/99.53	—/95.40
LRNet	Temporal	—/99.90	—/57.4
FakeLocator	Temporal	100/	93.04/—

GAN models. Our analysis in Figure 4 illustrates the distribution of research efforts targeting structure enhancement, media manipulation, and synthesizing. We note that 36.8% of the surveyed GANs focus on overcoming known limitations of GANs, such as training instability, mode collapse, and limited diversity, highlighting the fragile nature of the GAN structure.

In detail, the adversarial design between the generator and the discriminator brings three main limitations:

- The adversarial design results in instability in GAN training. The adversarial nature of GANs necessitates a delicate balance between the generator and discriminator, imposing strict requirements on the convergence path during gradient optimization. Oscillations or divergence may appear in the training process when GAN improves in one network but leads to deterioration in the other.
- The adversarial design leads to mode collapse if the training process fails to reach a Nash equilibrium. Neither the generator nor the discriminator can unilaterally improve when mode collapse appears, and the GAN fails to converge to the optimal results.
- The adversarial design leads to insufficient diversity of generations because the generator focuses on fooling the discriminator rather than capturing the full diversity of the target distribution. Besides, the adversarial design of GANs also limits attribution in other aspects, such as interpretability and scalability.

Therefore, researchers are eager to find new generator models that are easier to train and can produce more realistic outputs. The models are also expected to have better interpretability, allowing researchers to better control the training process. Better scalability may also significantly extend generators' ability to cross-modal generation. We summarize the adversarial development of Deepfake generation and detection (from GAN to DM) in Figure 3.

3. Deepfake by DMs

3.1. From GAN to DM. Researchers are exploring alternatives to GANs due to their inherent limitations in training stability and generative performance. In recent years, DMs have emerged as a promising alternative, gaining significant traction in various generation tasks, including Deepfake generation. To illustrate the inherent connection between GAN and DM and why DM has gained more popularity in Deepfake generation, we investigate deeper into the image generation task.

Figure 5 shows the prototype of image generation and generation pipelines of three generators (i.e., GAN, VAE, and DM). Essentially, a Deepfake generator is expected to produce an image with a given latent variable x (i.e., a latent code) as input. Different generators adopt different strategies to enable the pipelines to learn from real data and empower the generative capacity. In GAN, the discriminator compares the generated image with real samples to update the generator and improve the generation performance. However, the adversarial design brings the limitations mentioned above. Instead of an adversarial setting, VAE adopted a reconstruction strategy to learn from real data, which provides another possible solution for training a Deepfake generator. However, VAE first compresses the original image to a much smaller dimension with an encoder and then uses a low-dimension latent code for generation. The low-dimension latent code greatly limits its representational capacity and the diversity of generated images.

By replacing the encoder of VAE with progressive adding noise, the latent code (the noise-perturbed image) remains its original size and enhances the representational capacity of the generator. By removing the adversarial design of GAN and adopting the progressive adding noise steps as an encoder in VAE for reconstruction, DMs achieve more stable training and better generative performance.

3.2. Overview of DMs. As shown in Figure 6, DMs model the generative task as a Markov chain by gradually adding and removing noise to an image until it reaches a prior distribution. The forward process of DMs step-by-step adds Gaussian noise to the clean image (x_0) until the image is degraded to a noise image (x_T). In detail, at each step t, DM corrupts the data x_0 (the original data) by adding noise, producing an intermediate noisy state x_t. Let $q(x_t|x_{t-1})$ be the distribution of the data at step t conditioned on the previous step x_{t-1}. Typically, the noise is Gaussian, that is,

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\varepsilon_t, \tag{2}$$

where $\varepsilon_t \sim \mathcal{N}(0, I)$ is Gaussian noise and α_t controls the variance of the noise added at step t. As $t \longrightarrow T$, the data become pure noise.

Then, the forward process is reversed by a denoising process repeatedly removing noise conditioned on the previous iterations, obtaining the reconstructed image (x_0'). In detail, the reverse process is typically modeled as a parameterized neural network $\varepsilon_\theta(x_t, t)$ that predicts the noise added at each step, and from this prediction, the model can reconstruct the data:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_{t-1}}}\varepsilon_\theta(x_t, t)\right), \tag{3}$$

The backward denoising process is optimized to generate Deepfake from noise images, serving as the source of generative capacities of DMs.

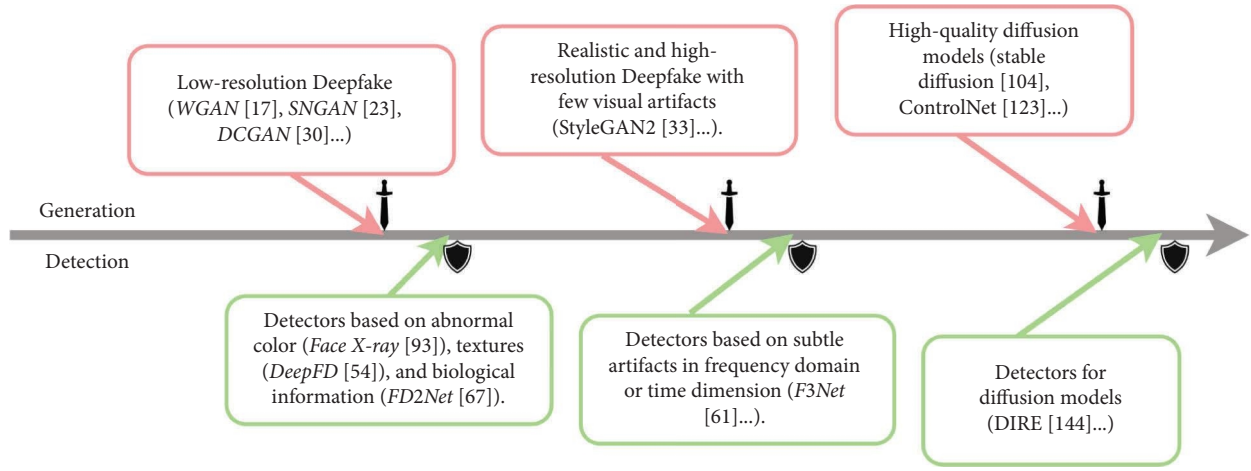

Figure 3: The cat-and-mouse race between Deepfake generators and detectors, and the evolution from GAN to diffusion model.

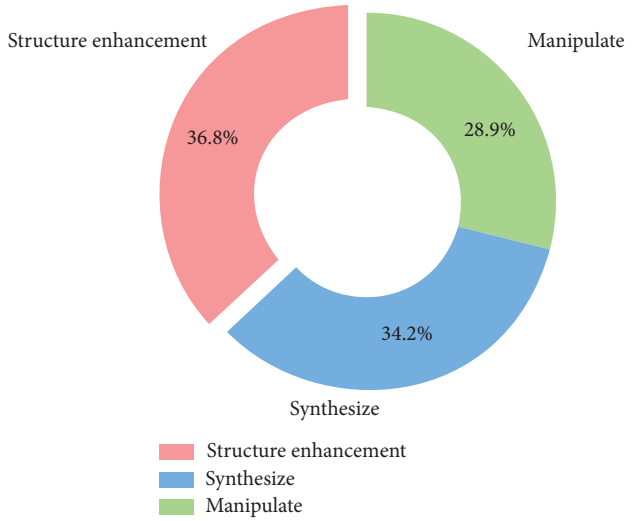

Figure 4: Distribution of GAN-based Deepfake generators.

The loss function for training a DM is typically the mean squared error (MSE) between the predicted noise $\varepsilon_\theta(x_t, t)$ and the actual noise ε_t added during the forward diffusion process. The goal is to minimize the difference between the predicted and actual noise over all time steps. The objective function for training the model is as follows:

$$\mathscr{L} = \mathbb{E}x_0, \varepsilon_t, t\left[\left\|\varepsilon_t - \varepsilon_\theta(x_t, t)\right\|^2\right], \tag{4}$$

where ε_t is the noise added to the data during the forward diffusion process. $\varepsilon_\theta(x_t, t)$ is the noise predicted by the neural network at time step t. The expectation is taken over the data x_0, the noise ε_t, and the time steps t. This loss function encourages the model to predict the noise added during the forward process and thus effectively learns how to reverse the diffusion process.

DMs can capture multimodal data density functions by training networks to optimally denoise. The iterative guidance also creates tractable sampling, allowing DMs to trade off diversity and quality without network changes. By modeling reverse dynamical systems, DMs have rapidly advanced across domains from images to video, audio, and 3D shapes, achieving best-in-class results with stability and flexibility advantages over GANs and VAEs.

3.3. DM-Based Deepfake Generation. Similar to the development of GAN, works on DMs first focus on enhancing the attributes of DM structures. Then, DM-based generators are widely studied in Deepfake generation tasks, including synthetic and manipulated Deepfake. Therefore, we categorize DM works according to their structure enhancement, synthesizing, and manipulating tasks and list them in Tables 7, 8, and 9, respectively.

3.3.1. Structure Enhancement of DMs. The trailblazing work of DMs is the denoising diffusion probabilistic model (DDPM) [97]. DDPM is a class of latent variable models drawing connections to nonequilibrium thermodynamics. Based on a link between DMs and denoising score matching with Langevin dynamics, a weighted variational training approach is proposed in DDPM. This also enables a progressive lossy decoding scheme generalizing autoregressive generation. By formalizing connections to thermodynamic diffusion and denoising objectives, this work advances DMs to match or surpass GAN performance on image datasets. Theoretical grounding and interpretable decoding are valuable properties alongside the sample quality achievements.

DDIM [98] accelerates sampling for DDPM. While DDPMs generate high-quality images without adversarial training, sampling is slow as it requires simulating a Markov chain. DDIMs instead utilize non-Markovian diffusion processes corresponding to faster deterministic generative models. The same training procedure as DDPMs is used. Experiments show DDIMs produce equal quality samples up to 50x faster than DDPMs in terms of wall-clock time. DDIMs also enable computation-quality trade-offs, semantically meaningful latent space interpolation, and low-error observation reconstruction.

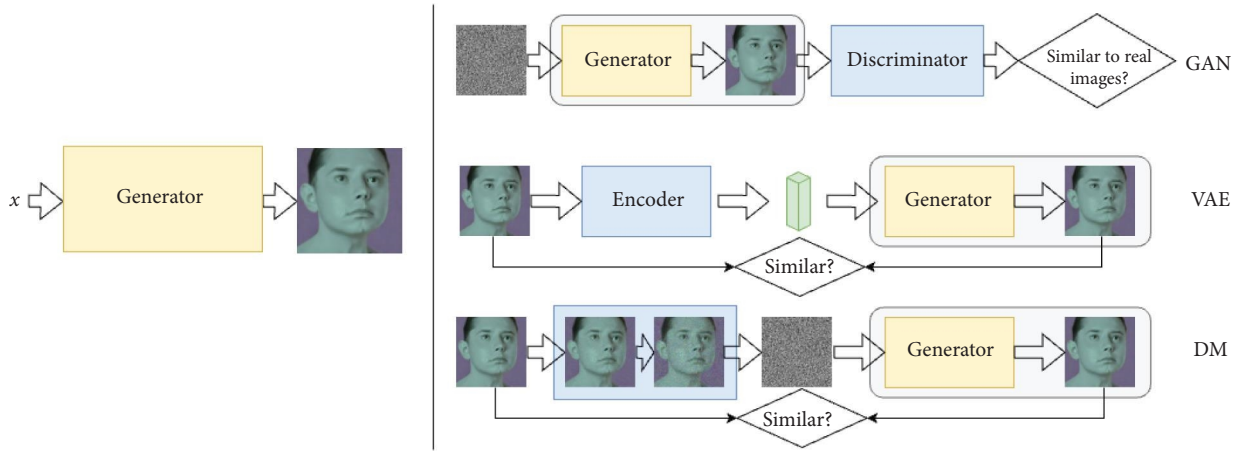

FIGURE 5: Prototype of image generator (left) and comparisons of GAN, VAE, and DM (right). We mark the generator prototype module in purple frames in all three generators.

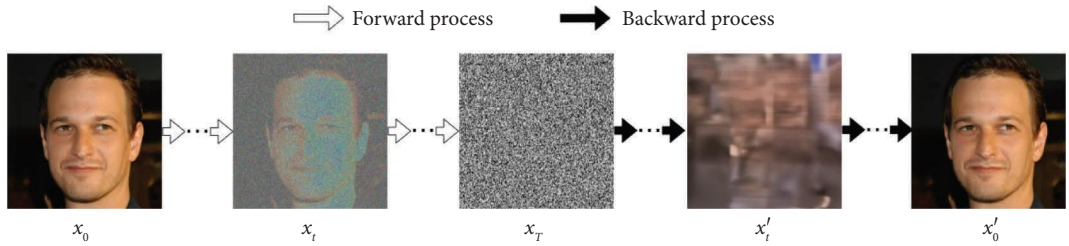

FIGURE 6: The structure of the diffusion model. The forward process gradually adds noise to an image over multiple steps, slowly transforming it from a clear image to pure noise. At each step, a small amount of Gaussian noise is added to the image. The backward process, also called the reverse process or denoising process, attempts to remove noise from a noisy image step by step, gradually reconstructing the original image. This is typically done using a neural network trained to predict and remove the noise at each step.

Song et al. [99] present a framework unifying score-based generative modeling and diffusion probabilistic models under continuous stochastic processes. A forward SDE injects noise to perturb data into a known prior, while a reverse-time SDE generates back to the data distribution by removing noise. The key insight is the reverse SDE that only requires estimating the score (gradient) of the perturbed data. Neural networks estimate the scores for accurate generative modeling. Encapsulating prior methods, new sampling procedures and capabilities are introduced—including a predictor–corrector to correct errors and an equivalent neural ODE for sampling and likelihood evaluation. The unified perspective also enables high-fidelity 1024×1024 image generation for the first time among score-based models. By bridging score modeling and diffusion through continuous processes, the work significantly improved both sampling techniques and sample quality.

Improved DDPM [100] proposed that simple modifications to the standard DDPM framework allow for high-fidelity generation and density estimation. Learning variances of the reverse diffusion further reduces sampling cost by 10x with negligible quality loss, which is important for practical deployment. Precision–recall metrics show that DDPMs cover the target distribution better than GANs. Crucially, DDPM performance smoothly scales with model capacity and compute, enabling straightforward scalability. By enhancing DMs to match or exceed GANs in sample quality, likelihood evaluation, and efficiency, this work greatly expands their capabilities and viability as generative models for images. The demonstrations of precise control over quality-compute trade-offs and scalability further attest to their value for real-world usage.

Iterative latent variable refinement (ILVR) [101] addresses controllable image generation by guiding the stochastic diffusion process in DDPMs. A method called ILVR leverages the generative reversibility of DDPMs to synthesize high-quality images matching a reference image. By iterative refinement, a single unconditional DDPM can adaptively sample diverse sets specified by references without retraining. Applications in multiscale image generation, multidomain translation, paint-to-image synthesis, and scribble-based editing demonstrate controllable generation using a single model. Enhanced control addresses a key limitation of DDPMs while retaining benefits like efficient sampling and good image fidelity. By steering random diffusion trajectories toward desired targets, ILVR significantly advances both the flexibility and practicality of DMs for computer graphics tasks demanding guided synthesis.

ADM [102] improved DMs for unconditional and conditional image generations through structure enhancements and classifier guidance, a method that effectively trades off diversity for fidelity using gradients from

Table 7: Representative DMs that enhance the DM structures, where I2I represents image2image translation and SR represents image super-resolution.

Ref	Structure	Condition	Task	Modal
[97]	DDPM	—	Generation	Image
[98]	DDIM	—	Generation	Image
[99]	DDPM-SDE	—/class	Generation	Image
[100]	Improved DDPM	—	Generation	Image
[101]	DDPM	—	I2I/editing	Image
[102]	ADM	—/class	Generation	Image
[103]	ADM	—/class	Generation	Image
[104]	VQ-DDM	—	SR	Image
[105]	CycleDiffusion	Text	I2I	Image
[106]	DPM-Solver++	—	Generation	Image

Table 8: Representative DMs that generate Deepfake by synthesizing new content.

Ref	Structure	Condition	Task	Modal
[107]	ADM	Text	Generation, inpainting	Image
[108]	DALL-E	Text	Generation	Image
[109]	VQ-DDM	—	Generation	Image
[110]	LVDM	—/Text	Generation	Video
[111]	Imagen	Text	Generation	Image
[112]	GLIDE	Text	Generation	Image
[113]	Continual diffusion	Text	Generation	Image
[114]	LFDM	Image	Generation	Video
[115]	DreamBooth	Text	Generation	Image
[116]	DDPM, VQ-DDM	Text	Generation	Image
[117]	VQ-DDM	Text	Generation	Image
[118]	DCFace	Identity/style	Generation	Image
[119]	VQ-DDM	Text	Generation	Image
[120]	Imagen Video	Text	Generation	Video
[121]	Diffused Heads	Audio	Generation	Video
[122]	DDPM	—/Image	Generation	Video
[123]	ControlNet	Various	Generation	Images
[124]	Uni-ControlNet	Various	Generation	Images
[125]	GaussianDreamer	Text	Generation	3D assets
[126]	VideoCrafter1	Image/text	Generation	Video
[127]	VideoCrafter2	Image/text	Generation	Video

Table 9: Representative DMs that generate Deepfake by manipulating the existing content.

Ref	Structure	Condition	Task	Modal
[128]	SDEdit	—	Generation, editing	Image
[129]	DiffFace	—	Generation	face
[130]	RePaint	—	Inpainting	Image
[131]	DiffEdit	class	Editing	Image
[132]	UniTune	Text	Editing	Image
[133]	DDIM	Text	Editing	Image
[134]	DiffusionCLIP	Text	Editing	Image
[135]	Blended diffusion	Text	Editing	Image
[135]	Palette	Image	I2I	Image
[136]	Imagic	Text	Editing	Image
[137]	DragDiffusion	Motion	Editing	Image

a classifier. ADM reports remarkable results on various benchmarks, achieving lower FID scores on datasets like ImageNet, which indicates higher quality and diversity in generated images. Furthermore, the paper discusses the societal impacts and limitations of DMs, highlighting their potential for wide-ranging applications while also acknowledging the challenges in sample speed and the absence of explicit latent representations.

Ho and Salimans [103] introduce classifier-free guidance for conditional DMs to improve sample quality without requiring a separate classifier. Classifier guidance leverages classifier gradients to trade off diversity and fidelity, needing an additionally trained classifier. Instead, a conditional and unconditional DMs are jointly trained. Combining the conditional and unconditional score estimates creates a similar fidelity–diversity spectrum as classifier guidance. This classifier-free technique avoids the cost and complexity of a second model. By showing the generative model alone that contains all necessary signals for controllable generation post-training, this exploration makes guidance simpler and more efficient for denoising diffusion.

Stable diffusion (SD) [104] formulates diffusion generative modeling in latent spaces of pretrained autoencoders. While pixel-based models achieve a promising synthesis, their optimization is expensive and sampling sequentially.

By balancing complexity reduction and detail preservation in latent spaces, SD reaches near-optimal trade-offs, enabling high-fidelity generation with greatly reduced computing. The latent formulation also allows convolutional high-resolution synthesis. With cross-attention layers, SD conditions are flexible on inputs like text and bounding boxes. Experiments across unconditional generation, class conditioning, inpainting, text-to-image synthesis, and super-resolution achieve new state-of-the-art results while significantly lowering cost compared to pixel models. By shifting the modeling domain from pixel to latent space, SD unlocks the practical deployment of DMs through substantial computational savings and conditional generation capabilities.

CycleDiffusion [105] formulates an alternative Gaussian latent space for DMs along with a reconstructing encoder mapping images into this space. While DMs typically utilize a denoising sequence space, unlike GANs and VAEs, CycleDiffusion explicitly models a continuous latent code and confers benefits. One key finding is emergent common latent spaces across DMs trained on related domains. This enables applications like unpaired image translation and text-to-image editing with a single encoder–decoder cycle. Additionally, the latent formulation unlocks unified guidance of DMs and GANs via energy-based control signals. Experiments on conditional image generation guided by CLIP demonstrate DMs better capture distribution modes and outliers than GANs. By exposing and harnessing continuous latent structures amid the sequential sampling process, Gaussian embedding confers new applications to leverage DMs intuitively as latent variable models.

DPM-Solver++ [106] proposed a single approach for both continuous and discrete-time DMs. By introducing more accurate numerical methods for solving the diffusion ODEs, DPM-Solver++ achieves high-quality sampling in significantly fewer steps than previous methods.

3.3.2. Deepfake Synthesized With DMs. GLIDE [107] explores DMs for text-to-image synthesis by comparing classifier-free and CLIP guidance strategies to trade off diversity and fidelity. The research demonstrates that samples from a 3.5 billion parameter text-conditional DM utilizing classifier-free guidance are more favored by humans than those from DALL-E, even when DALL-E employs CLIP reranking, a more resource-intensive process. The paper also reveals that these models can be fine-tuned for image inpainting tasks, enabling advanced text-driven image editing capabilities. This signifies an important stride in the field of AI-driven image generation and editing, offering more realistic and accurate outputs with more straightforward guidance techniques.

DALL-E [108] leverages CLIP image representations for controllable text-to-image generation. A two-stage model generates a CLIP embedding from text and then decodes this into an image. Explicitly modeling the intermediate representation improves sample diversity with minimal fidelity loss compared to end-to-end synthesis. The approach also enables intuitive image variations and edits by manipulating the compact encoding. DMs prove most effective for both stages, capturing semantics in the latent space and high-quality decoding. Compared to end-to-end conditional GANs and autoregressive models, explicit modeling and inversion of the robust CLIP space confer superior generation control and efficiency. By factorizing generation through a recognizable visual denoising manifold, the work significantly advances the state-of-the-art in flexible text-driven image synthesis.

Vector-quantized discrete diffusion model (VQ-DDM) [109] integrates vector-quantized variational autoencoders (VQ-VAEs) with DMs for discrete sequence generation. While VQ-VAEs provide rich codebooks and autoregressive decoding offers high fidelity, sequential pixel synthesis lacks global context. Denoising diffusion captures global dependencies but is expensive for images. The proposed VQ-DDM combines the strengths of both VQ-VAEs and denoising diffusions. Unlike autoregressive VQ approaches, the framework generates images with global coherence from compact codes without strict pixel order. VQ-DDM matches state-of-the-art image quality with greater efficiency than pixel diffusion while surpassing vector-quantized autoregressive models on tasks like inpainting needing nonlocal patterns. By synergizing discrete sequential synthesis with continuous diffusion mechanisms, the hybrid technique rectifies key limitations of both families for controllable image generation.

Imagen [111] is a text-to-image DM achieving new levels of realism and language understanding. By scaling up transformer structures for both text encoding and image decoding, Imagen reaches high FID on COCO without domain-specific fine-tuning. Human evaluations also show samples comparable to real data in image-text alignment. To comprehensively assess synthesis quality, the DrawBench benchmark is proposed, comparing Imagen to leading approaches on fidelity and alignment. Raters prefer Imagen over alternatives like VQ-GAN + CLIP and DALL-E 2 in side-by-side comparisons. By synergizing large language and image transformers in a diffusion framework, Imagen represents a major advance in programmatic image generation through intuitive text-based control.

Liu et al. [112] propose structured image generation using compositional DMs to address composition failures in large text-to-image models. By interpreting the diffusion process as energy-based models with combinable data distributions, separate component models specializing in visual primitives like objects and relations can be integrated at test time to render complex scenes. This compositional approach generalizes to intricate configurations unseen during training, accurately binding detailed language descriptions to image elements. Experiments demonstrate photorealistic rendering of sentence relations, facial attributes, and rare interobject arrangements—challenging cases for leading models like DALL-E 2. The framework advances structured conceptualization and binding through modular component assembly rather than using emergent reasoning in end-to-end models. By scaling synthesis in a human-inspired combinatorial manner, the technique makes significant progress on controllable image generation from intricate language specifications.

Continual diffusion [113] addresses catastrophic forgetting when sequentially adapting text-to-image models to new visual concepts with few examples. Previous fine-tuning suffers performance degradation on older concepts as new ones are added. A continually self-regularized adaptation approach called C-LoRA is proposed for stable DMs to enable continual adaptation without storing past data. Using fixed random prompts further improves generalization. Experiments on continual classification and the proposed continual diffusion setting demonstrate C-LoRA matches or outperforms rehearsal-based techniques without the storage overhead. By circumventing forgetting during compositional concept acquisition, C-LoRA advances practical applications for on-device personalization and incremental learning where storing growing user data is infeasible. Its effectiveness underscores the importance of continual adaptation techniques for the real-world deployment of customizable generative models.

Latent flow diffusion model (LFDM) [114] presents conditional image-to-video (I2V) synthesis. Unlike direct generation approaches, LFDM warps a given image over a predicted optical flow sequence in latent space to simultaneously model spatial appearance and temporal dynamics. It comprises an unsupervised latent flow autoencoder for content modeling and a DM that generates compact latent flows rather than high-dimensional videos. By factorizing spatial details from motion, LFDM efficiently models dynamic patterns while preserving the fidelity of provided imagery. Experiments across datasets demonstrate LFDM outperforming prior work in photorealism and motion coherence. Simple decoder fine-tuning further enables domain adaptation leveraging pretrained components. With both solid empirical performance and interpretability, LFDM provides an advance in controllable video generation by separating what changes from what persists.

DreamBooth [115] enables personalized image generation by fine-tuning text-to-image models to embed user-provided subjects using a few example images. A unique identifier bind provides individuals with novel contextual synthesis, preserving their traits. Built on pretrained semantic knowledge, an autogenous class-specific loss allows extrapolation to varied scenes and views beyond the reference set. Applications in recontextualization, text-guided view manipulation, and stylization showcase controllable generation while keeping identity intact. A new dataset and protocol rigorously benchmark such conditional synthesis quality. By synergizing external personal datasets with the rich priors of large generative models, the technique significantly advances the control, customizability, and specificity of text-driven image production. Enabling lay users to adapt such models to familiar domains with minimal data also underscores a move toward accessible and trustworthy AI utilization.

Meng et al. [116] distill classifier-free guided DMs to accelerate inference while retaining sample quality. Despite effectiveness for high-fidelity generation, evaluating conditional and unconditional models is computationally expensive. A distillation pipeline first matches the combined output and then progressively transfers sampling to more efficient diffusion. For pixel models, comparable scores to the original are achieved using as few as four steps on CIFAR-10 and ImageNet 64×64, accelerating sampling up to 256×. For latent DMs like SD, 10× speedups are attained with 1–4 steps and no visual decline. The approach also enables efficient guided editing and inpainting, generating quality results in just 2–4 denoising iterations. By drastically reducing sampling costs, the distillation framework unlocks the deployment of large conditional DMs under practical time constraints, advancing their use for interactive applications.

Bansal et al. [117] introduce a universal guidance technique enabling conditional image generation from DMs using arbitrary modalities without retraining. Most conditioned DMs only accept specific forms of input like text. The proposed algorithm leverages gradient signals from any external guidance source to steer image sampling, segmentation maps, face recognition, object detectors, or classifiers. Demonstrations with diverse side information validate the capability to specialize generic generative priors for new domains purely through gradient-based conditioning—no network changes needed. By flexibly binding external recognition networks, the approach significantly advances DM adaptability and specificity without costly dataset-driven fine-tuning.

Dual condition face generator (DCFace) [118] tackles synthetic face dataset generation for recognition via DMs conditioned on subject identities and stylistic factors. Controlling interclass similarity and intraclass variability poses challenges for GANs and 3D approaches. DCFace disentangles identity vectors and patch-based style codes to enable consistent identity synthesis under customizable pose, lighting, and other transformations. An identity loss through sampling stabilizes recognition cues. The interpretable and reliable factor manipulation provides a significant advance in configurable synthetic faces spanning intrapersonal imaging dynamics at scale.

Papa et al. [119] provide critical analysis on generating and detecting realistic fake faces using DMs to inform ethical deployment. By identifying text prompts for photorealistic human synthesis with SD, malicious usage risks are exposed should access go unchecked. The study sounds caution about the advancing creative abilities of generative models and the urgency for tamper detection advances to catch up. By demonstrating the ease of synthesis paired with solutions for exposing fakes, the exploration attempts to steer DM progress responsibly toward transparency and monitoring against potential harms.

Imagen Video [120] is a cascaded text-to-video (T2V) DM generating high-quality HD footage. A base generator coupled with interleaved spatial and temporal super-resolution networks enables controllable video synthesis from language descriptions. Techniques transferred from image diffusions combined with progressive distillation produce high-fidelity and efficient sampling. Experiments demonstrate diverse video generation spanning artistic styles, 3D environments, and complex multistage actions described in input prompts. The model exhibits strong world knowledge and understanding of spatiotemporal dynamics

in rendered scenes. By advancing multiscale adversarial learning to sequential data, Imagen Video significantly pushes the state-of-the-art in programmatic video synthesis through intuitive text interfaces.

Diffused Heads [121] tackles talking face video generation from a single image and audio using DMs. A proposed autoregressive diffusion approach synthesizes natural head movements and facial expressions like blinks from only a static face and speech clip. Backgrounds are also preserved within the generated talking head video. Evaluation of two datasets shows the method achieves great realism of motion and expressions. By advancing one-shot controlled generation, the technique reduces data needs for personalized video synthesis down to a selfie-style photo. The solid empirical results validate DMs as a promising paradigm for highly conditional sequential data generation grounded in limited supervision.

Harvey et al. [122] tackle long-duration video modeling using diffusion probabilistic models. The proposed framework can conditionally complete arbitrary frame subsets to enable efficient iterative optimization for coherent event sequencing over 25+ minutes. New datasets and metrics based on autonomous driving scenarios provide semantic grounding. By scaling iterative reversible generation with flexible sparse conditioning, the approach surpasses prior video diffusion techniques on closed-loop coherence over extreme time horizons. The capability to extrapolate realistic event chains from sparse keyframes underscores DMs as uniquely suited for long-range visual prediction. The advance demonstrates video understanding through physics-inspired temporal diffusion rather than data-driven discriminative cues alone.

ControlNet [123] proposed to add spatial conditioning controls to large, pretrained text-to-image DMs. ControlNet enables DM generation under various conditions, such as sketches, depth maps, and human poses. ControlNet greatly facilitates the real-world applications of DM-based image generation. Uni-ControlNet [124] enables the simultaneous use of multiple control modes beyond text, incorporating local controls like edge maps, depth maps, and segmentation masks, as well as global controls such as CLIP image embeddings. Uni-ControlNet utilizes only two additional adapters on top of pretrained text-to-image models, eliminating the need for training from scratch and maintaining constant adapters regardless of the number of controls used. Compared with ControlNet, Uni-ControlNet significantly reduces fine-tuning costs and model size, making it more practical for real-world deployment.

VideoCrafter1 [126] presents high-quality video generation via T2V and I2V generation. The T2V model can generate high-quality video, and the I2V model produces videos that strictly adhere to the condition image. VideoCrafter2 [127] focuses on creating high-quality video models without access to premium video datasets. Instead, Videocrafter2 combines low-quality videos with synthesized high-quality images to train an effective model. Based on the analysis of the interaction between spatial and temporal modules in video models extended from SD, the authors noticed that training all modules together creates a stronger coupling between spatial and temporal components. Leveraging this, they fine-tuned spatial modules with high-quality images, improving overall video quality without degrading motion.

There are also some DMs adopted in 3D objects' generation:

Latent video diffusion model (LVDM) [110] is a lightweight video DM with a low-dimensional 3D latent space. The long video generation ability of LVDM is based on the proposed hierarchical diffusion structure. Besides, LVDM proposed conditional latent perturbation and unconditional guidance to mitigate performance degradation in long video generation. Extensive experiments demonstrate LVDM's superiority in generating realistic and longer videos. Additionally, LVDM is also extended to large-scale T2V generation tasks.

GaussianDreamer [125] is a framework generating 3D assets from text prompts. In detail, GaussianDreamer leverages 3D Gaussian splatting representation to bridge the gap between these two types of models. It uses 3D DMs to provide initial priors and then enhances geometry and appearance using 2D DMs. The process incorporates noisy point growing and color perturbation techniques to improve the initialized Gaussians.

3.3.3. Deepfake Manipulated With DMs. Stochastic differential editing (SDEdit) [128] enables easy image synthesis and editing by users through stochastic diffusion processing. Balancing realism and faithfulness to inputs like strokes poses challenges for current GAN approaches. SDEdit leverages DM priors without task-specific training or inversions. Given an input image and edits, SDEdit adds and then denoises noise through the SDE to increase realism while preserving user edits. Human studies show SDEdit outperforms GAN baselines substantially in realism and overall satisfaction on stroke-based synthesis/editing and compositing. By building on diffusion generative modeling, SDEdit advances widely applicable guided synthesis without model retraining or input inversion. The demonstrations also attest to the suitability of iterative denoising for user creativity, interacting with an editor rather than a one-shot synthesis.

DiffFace [129] is the first DM for high-fidelity facial identity swapping. DiffFace adopts an identity-conditional model to handle desired face generation. Flexible guidance during sampling then composites target attributes while translating identity via facial recognition networks. A target-preserving blend further retains background details. Benefits of DiffFace over GANs include training stability, quality, and control. User studies validate that DiffFace matches or exceeds state-of-the-art face-swapping techniques in realism and identity preservation. By harnessing iterative denoising for facial generation and composition, DiffFace sets a new bar for Deepfakes that are perceptually indistinguishable from reality while preventing unwanted use through responsible disclosure and monitoring.

RePaint [130] tackles free-form image inpainting using denoising DMs to enable extreme mask generalization.

RePaint leverages an unmodified pretrained unconditional DDPM as a generative prior. Conditioning is achieved by only sampling unmasked pixels in reverse diffusion, allowing high-quality and diverse completions for any inpainting mask. Evaluations on faces and generic images with both standard and extreme hole shapes show RePaint outperforms autoregressive and GAN methods on nearly all mask distributions. By simply harnessing pretrained generative reversibility, RePaint pushes DMs toward unconstrained inpainting without assumptions about missing regions. The extreme generalization demonstrates the suitability of iterative denoising for open-ended conditional image synthesis.

DiffEdit [131] leverages text-conditional DMs for semantic image editing without masks. Previous techniques treat editing as conditional inpainting, needing user input to specify regions to change. Instead, DiffEdit automatically identifies areas to edit by contrasting model predictions under different text prompts. Latent inference preserves unchanged content. Without masks, DiffEdit achieves state-of-the-art generative editing results on ImageNet and robustness on more challenging datasets like COCO and text-to-image samples. The capability to reconcile language-driven enhancements with input visuals moves beyond editing as inpainting toward true spatial reasoning over linguistic concepts and visual regions. By contrasting sampler behavior rather than solely using discriminators, DMs prove uniquely positioned for programmatic yet minimally invasive image manipulation.

UniTune [132] enables text-driven image editing by fine-tuning generative models on individual images. While text-to-image generation has seen great progress, editing capabilities lag behind, typically needing additional signals like masks to specify regions of change. Instead, the proposed UniTune method adapts any baseline generator using just example imagery and text descriptions. Initializing sampling from a base image and interpolating details further improves edit quality. Without masks or other guidance, UniTune performs a wide range of edits while maintaining fidelity, even for significant visual changes previously infeasible. Experiments across use cases demonstrate versatile text-controlled image manipulation via simple domain transfer learning rather than architecting separate editors. By harnessing DM reversibility and adaptability, the technique delivers on the promise of intuitive human–AI creativity through natural language alone.

Hertz et al. [133] tackle intuitive text-driven image editing without masks by analyzing text-image attention correlations in generative models. While text-to-image synthesis has seen great progress, editing capabilities lag behind, even small text changes often lead to completely different results. Existing techniques rely on masks or sketches to spatially constrain edits. Instead, cross-attention layers are identified as critical for binding words to spatial layouts. By monitoring and modifying attention distributions via textual edits alone, targeted yet minimally invasive manipulations are enabled. Demonstrations include localized revisions by replacing words, global style adjustments through additions, and granular attenuation of cue magnitudes without specifying regions. Evaluations across diverse images and prompts validate prompt-to-prompt editing quality and fidelity. By exposing and harnessing an interpretable interface for spatial-semantic alignment, the technique delivers truly programmatic image manipulation directly through natural language.

DiffusionCLIP [134] is a technique for text-guided image manipulation using DMs to address limited GAN inversion capability and artifacts. While GAN inversion with CLIP conditioning enables zero-shot editing through language, reconstructing images with novel compositions or attributes often fails. Instead, leveraging DMs' full inversion potential and high-fidelity synthesis empowers robust real image editing between unseen domains, which is validated on diverse ImageNet samples. A novel noise combination approach further enables straightforward multi-attribute editing. Human evaluations confirm the superior manipulation fidelity over GAN-based methods, better generalizing to complex real-world distributions. By building on iterative denoising rather than adversarial discrimination, DiffusionCLIP makes significant progress toward broadly applicable programmatic image editing through intuitive textual interfaces.

Blended diffusion [135] enables intuitive local text-driven image editing using CLIP guidance and diffusion blending. While global language-conditioned image generation has seen great progress, region-level editing capabilities lag behind. The proposed approach takes an ROI mask and textual description to steer a denoising DM toward desired edits spatially. Fusing the rewritten region with unaltered areas is achieved by progressively blending latent codes and input image copies under increasing noise. Qualitative and quantitative evaluations validate background preservation and text alignment exceeding related techniques and baselines. Demonstrated applications include object addition/removal/replacement, background substitution, and content extrapolation—all through descriptive language edits. By synergizing language-vision guidance with differentiable generative reconstruction, the technique significantly advances the state-of-the-art in programmatic visual editing toward flexible region-based modifications from natural textual requests.

Palette [135] establishes conditional DMs as a unified framework for advancing image-to-image translation across challenging restoration and synthesis tasks. A simple diffusion implementation outperforms specialized GAN and regression techniques on colorization, inpainting, upsampling, and JPEG artifact removal without structure customization. Studies reveal the impact of loss functions and self-attention on sample quality. A standardized benchmark and protocol leveraging ImageNet, human judgment and automated metrics provide rigorous assessment showing strong generalization of a single model trained over multiple domains. By demonstrating both superior task performance and flexibility from a general-purpose diffusion translator, this exploration signals denoising-based generation as a prime candidate for further developing this practical graphics application space.

Imagic [136] achieves complex single-image manipulation with only text queries. Prior text-based editing techniques remain limited to specific domains like overlays or style transfer and often need multiple inputs. Instead, the proposed Imagic approach performs nonrigid semantic edits like adjusting object pose and scene composition within one natural photo using pretrained generative guidance. For instance, making a dog sit or a bird spread its wings in provided shots. Unlike previous work, this text-controlled editing framework requires no additional masks, segmentation, or alternate views. A new benchmark assesses editing quality, with human studies preferring Imagic over prior techniques. Demonstrations across domains exhibit significant advances in open-ended spatial rewriting grounded in language alone. By harnessing inversion and adaptability unique to DMs, the technique makes strides toward flexible image reconfiguration through intuitive textual description.

DragDiffusion [137] extends interactive image editing techniques based on generative guidance to DMs for improved real-world applicability. While most GAN-based editors constrain capacity, editing via latent code optimization unlocks precise spatial control leveraging versatile diffusion priors. Although DMs generate images iteratively, coherent results are achieved by only optimizing a single step's latent, which enables efficient high-quality manipulation. Experiments across challenging cases with diverse objects, styles, and configurations validate generalizability exceeding GAN counterparts. By porting emerging neural editing paradigms to powerful denoising generative models, DragDiffusion makes precise semantic image manipulation more tractable for practical workflows. The technique underscores the suitability of decoder-based reversible synthesis for performer-centric creativity, interacting directly in output space rather than constrained latent configurations.

3.4. Forensics of DM-Generated Deepfake. The realistic outputs of DMs raise an urgent need for more discriminative detectors capable of detecting DM-generated Deepfake (Table 10).

Boháček and Farid [138] provide rigorous quantification of 3D geometric and photo-metric realism in AI-generated faces. Unlike graphics rendering techniques grounded in explicit scene modeling, learning-based image synthesis like GANs and DMs captures scene statistics without underlying 3D representations. Estimating facial geometry and lighting from real and synthesized data reveals limitations in capturing innate structure. Analyzing tricky details like eyes and mouth exposes flaws in adhering to natural constraints. While progress in 2D realism has accelerated through data-driven generation, experiments reveal room for improvement in adhering to the depth, physics, and semantics constraining real imagery. By moving beyond blind metrics to structured analyses, the study calls for encoding stronger inductive biases toward 3D world consistency in synthesis models. Broader findings also contribute to forensic detection capabilities by surfacing latent geometric anomalies in fake imagery.

Corvi et al. [139] investigate detecting DM-based image forgeries as an emerging challenge given their photorealism risks malicious use. While GAN-based image forensic methods have progressed, understanding traces in new diffusion synthesis and detector efficacy requires study as adoption accelerates across art and media. Initial experiments expose identifiable artifacts and limitations of existing detectors specialized for GAN imagery when applied to diffusion outputs. Assessing performance under challenging compression and resizing representative of social media distribution channels reveals significant gaps. By establishing diffusion media forensics as a crucial new domain and benchmarking state-of-the-art generalization, this study calls attention to the data and capability deficiencies that must be addressed for reliable detection, urging the community toward proactive solutions as adoption spreads.

Ricker et al. [140] evaluate GAN-targeted classifiers, which reveal limited cross-generalization and require adapter networks to reliably expose diffusion forgeries sharing few spectral artifacts with adversarial methods. Frequency analysis provides empirical insights into diffusion image properties and generator fingerprinting for improved deception. Ultimately, the dramatic detector performance gap between prevailing generative paradigms signals the need for specialized data and networks, rather than generalist solutions. By benchmarking status quo capability and characterizing frequency-domain differentiation, this work lays the crucial foundation for understanding and advancing defense against this new generation of AI media synthesis.

Sha et al. [141] detect and attribute text-to-image generated imagery to address synthesis model misuse risks. Fake images from systems like DALL-E 2 and SD are shown to share common artifacts detectable from real data across generators. Furthermore, discriminative model fingerprints enable reliable attribution to original structures. Studying detection difficulty over prompts that vary in topic and length reveals easier deception for person descriptions in the 25–75 token range. Overall, experiments provide early warning against threat models where anonymized distribution could preclude responsibility.

Lorenz et al. [142] propose an efficient detector using multi–local intrinsic dimensionality (LID) analysis. Originating from adversarial example detection, multi-LID exposes DM traces missed by prevalent GAN-focused classifiers. Near-perfect detection and generator attribution are demonstrated in realistic uncompressed settings, with graceful performance decline under compression. Rigorous benchmarking against diverse model structures and resolutions validates superiority over common approaches reliant on dataset-specific cues rather than intrinsic signals. The work establishes a solid and widely generalizable dynamical systems footprint for reliable provenance tracking by introducing powerful invariant metric failures. Findings also underscore the importance of adversarial threat modeling and standardized evaluation in driving real-world viability beyond COCO and bedrooms.

Diffusion reconstruction error (DIRE) [143] is a forgery detection technique exploiting reconstruction error signals. Existing methods struggle even when augmented with

Table 10: Detectors of DM-generated Deepfake.

Ref	Target	Classifier	Datasets	Modal
[138]	Synthetic	LR	Self-built (stable diffusion)	Image
[139]	Synthetic	Various	Self-built (DALL-E Mini, DALL-E, GLIDE, latent diffusion, stable diffusion, ADM)	Image
[140]	Synthetic	Various	Self-built (DDPM, IDDPM, ADM, PNDM, LDM, midjourney)	Image
[141]	Synthetic	CNNs	Self-built (LD, GLIDE, DELL-E2)	Image
[142]	Synthetic	Multi-LID	CiFake, ArtiFact, DiffusionDB, LAION-5B, SAC, stable Diffusion-v2.1, LSUN-bedroom	Image
[143]	Synthetic	DIRE	DiffusionForensics	Image
[144]	Synthetic	SeDID	Self-built (DDPM)	Image

diffusion data, motivating analysis into architectural biases. A proposed DIRE representation measures input-reconstruction deviations using pretrained generators. Observing real images cannot be approximated well, DIRE provides a robust bridge distinguishing real and generated data. Extensive experiments under compression and perturbations validate DIRE generalizes broadly to expose unseen DMs. A collected benchmark with diverse architectural distributions further verifies superiority over previous error metrics and leakage-hunting classifiers tuned to specific generative loopholes rather than fundamental dynamics. By exposing reconstruction tractability as the "no free lunch" cue missing in current detectors, this work provides practical and widely applicable counterfeit detection for an emerging generative paradigm.

Stepwise error for diffusion-generated image detection [144] (SeDID) uses reverse process determinism and denoising deviations to expose forgeries. A statistical variant and learned network blueprint leverage unique attributes missing from alternative generative paradigms. Evaluations against DDPMs and text-to-image systems demonstrate significant advances in exposing fabricated diffusion imagery over status quo methods tuned for GANs. By pioneering detector structures specialized to an ascendant generative class through empirical insights, SeDID marks seminal progress against AI-synthesis threats. The techniques outline a paradigm shift from adversarial discrimination and dataset overfitting to formal dynamical analysis for reliable and generalizable fake provenance.

3.5. Discussion of DM-Based Deepfake. Compared with GAN-based Deepfake generation, the training of DM-based Deepfake generators is more stable. Therefore, 23.8% of the surveyed papers focus on enhancing the structure of DMs (Figure 7), which is significantly lower than the ratio of GAN-based generators.

The larger latent space of DM significantly improves the generation diversity and quality so that fewer artifacts are left in the output images, bringing challenges to existing GAN-based detectors. New detectors are then proposed [142, 143] to make sure the detection side stays caught up in the cat-and-mouse race between Deepfake generation and detection. Considering the widespread application of text-to-image generation, corresponding detectors are also proposed to protect people from misinformation [141, 144]. However, we also notice that with the explosive growth of DM-based Deepfake generation, more detectors are expected to keep the detection side up, especially in the consideration of the cross-modal realistic Deepfake.

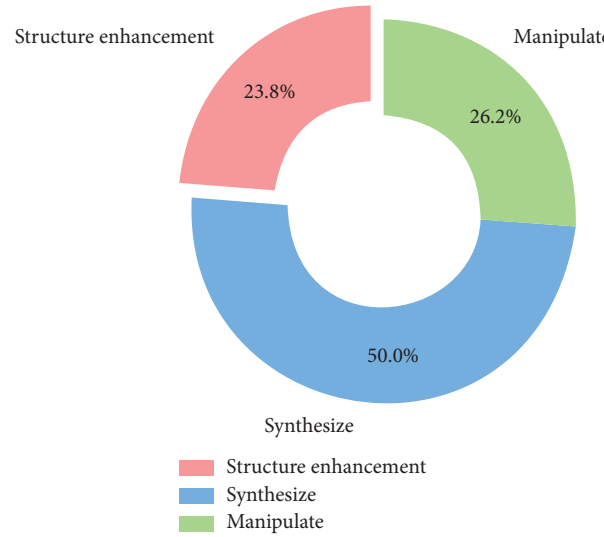

Figure 7: Distribution of DM-based Deepfake generators.

4. Summary

4.1. Comparisons. We comprehensively compare the difference between GAN and DM from the perspective of generation and detection (Table 11). The comparisons facilitate our understanding of the evolution of Deepfake.

4.1.1. GAN-Based and DM-Based Deepfake Generation

4.1.1.1. Generative Capacity. GANs and DMs obtain generative capacity by modeling data with different network designs. A GAN employs an adversarial training process between a generator and a discriminator. A GAN synthesizes artificial samples from random noise by simulating the data distribution of real images, while the discriminator is trained to distinguish between real and fake samples. The generator tries to maximize the probability of fooling the discriminator, while the discriminator aims to minimize being fooled. This adversarial contest enables the generator to continuously improve at synthesizing realistic samples until it can reliably generate data, i.e., indistinguishable from real data. DMs start with a real data sample and add noise gradually over multiple time steps. This diffuses and pushes the data away from its original form into a latent space. The

Table 11: Comparison of GANs and DMs.

Aspects	GANs	DMs
Source of generative capacity	Adversarial training of the generator and discriminator	Learning of denoise patterns
Structure	Generator–discriminator	Diffusion-denoising model, much larger size than GANs
Training	Adversarial training with potential for mode collapse	Avoid mode collapse but more time-consuming
Inference	Faster inference	Slower inference due to forward and backward process
Latent space	Compressed latent space	Full-size latent space
Image quality	High-quality, diverse images	Realistic details, especially in specified tasks like medical image generation
Controllability	Limited control	Greater controllability, suitable for tasks like text-to-image generation
Interpretability	Limited interpretability due to the adversarial nature of the generator and the discriminator	Good interpretability: The backward process closely follows a probabilistic, step-by-step generation procedure.
Cross-modal generation	Limited ability	Effective in handling complex and long-range dependencies between text and images
Detection	Ineffective on DM-generated media, subject to generalization challenge	Able to generalize to GAN-generated media, subject to generalization challenge

key is that this diffusion process can be reversed by removing the noise, which allows the model to go from latent space back to the original data sample. DMs are trained to learn how much noise needs to be removed at each step to reverse the diffusion most accurately. Once trained, DMs can generate new samples by starting with pure noise and removing the learned amounts of noise over multiple time steps to gradually reveal a new sample.

4.1.1.2. Structure. Different network structures are designed to carry on the data modeling in GANs and DMs. GAN generators adopt a series of convolutional and upsampling layers to transform the noise into an image, where skills such as batch normalization and dropout are also adopted for regularization. The discriminators in GANs are usually CNN-based classifiers that differentiate real and fake images. Compared with GANs, DMs are much larger in terms of parameters and model size. At each step of the DM backward process, the denoising model (e.g., U-Net) consists of an encoder and a decoder neural network. The encoder takes in the noisy data and encodes it into a latent representation. The decoder takes this latent code and decodes it, reconstructing the data by removing some noise. The widely adopted deep U-Net has a large number of channels and layers, requiring many parameters and leading to larger DM models.

4.1.1.3. Training and Inference. The training and inference of GANs and DMs are also different. GANs are subject to mode collapse due to the adversarial training objectives. At the initial steps of GAN training, the generator generates diverse data to deceive the discriminator. However, the generator may converge to a local optimal solution and generate data in a single mode. Not trained to a globally optimal solution, GANs fail to generate samples of high diversity. Based on simple yet reliable denoising steps, DMs avoid mode collapse during training and feature more effective training. However, DMs usually require more training iterations because of the larger parameter sizes. During inference, a well-trained GAN can usually generate a sample instantly in a very short time. Differently, the inference steps of DMs are more time-consuming than GANs. While the larger model size naturally brings more computation, DM inference must go through the forward and backward process fully, significantly extending the inference duration.

4.1.1.4. Latent Space. In GANs, the generator model maps from a compressed latent code to a higher dimensional output, such as an image. This latent code is deliberately constrained to a much smaller dimension than the output data. This latent bottleneck forces the generator to learn a compressed representation of the training data distribution. Contrastingly, DMs do not enforce a compressed latent space. The encoder takes the noisy data and outputs a latent representation at the same dimensionality as the original data. This allows the full information content to be retained at each time step. The decoder then reconstructs the data from this full-size latent code by removing the corresponding noise. This is advantageous as there is no information loss due to compression. The model can represent all details in the high-dimensional latent space. Images generated from bigger latent codes are potentially of higher quality and more realistic, although they require more computation.

4.1.1.5. Generation performance. Although it is usually considered that DM is better at generating realistic images due to the noncompressing design, GANs such as StyleGAN3 and BigGAN also generate high-resolution and high-quality images. Besides, GANs also generate images of better diversity because of the upsampling layers in the generators. At the same time, DMs have impressed people with their realistic details in specified tasks (e.g., medical image generation). Compared with GANs, DMs have shown more success in cross-modal generation, especially text-to-image generation. Such an advantage is mainly owed to the extraordinary controllability of DM. DMs allow users to explicitly add conditions and specify more fine-grained control over the generated images. Users can specify detailed prompts or conditions, such as specific objects, attributes, or scenes they want to appear in the generated image. This level of control is beneficial for text-to-image generation tasks where users have specific requirements or preferences for the generated images. Furthermore, the multistage generation process of DMs is better at handling complex and long-range dependencies between text and images, facilitating the modeling of complicated cross-modal relationships.

4.1.2. Detection of GAN-Generated and DM-Generated Deepfake. We notice that most Deepfake detectors are designed and trained to recognize GAN-generated Deepfake. However, compared with GAN-generated Deepfake, DM-generated Deepfake contains more realistic textures and fewer visual flaws, making DM-generated Deepfake challenging for GAN-target detectors, especially those designed to capture visual artifacts such as abnormal biometrics, unnatural textures, and blending edges.

In fact, Deepfake detectors encounter generalization challenges even when trained and tested on different datasets generated by the same technique (GANs or DMs). The generalization challenge is essential because fake content generated by different models and datasets naturally has different data distributions. Detectors are trained to capture the potential data distributions of the training dataset and become less effective on unseen testing datasets. While a lot of works have proposed methods such as enlarging training datasets, capturing generalizable semantic representations, or adopting external knowledge (multimodal synchronization) to improve the generalization ability, few are practically usable to detect diverse real-world Deepfake, not to mention the generalization to unseen new-generation techniques. From the perspective of attack and defense, the generalization challenge is tough because it is like in an attack-defend competition, an attacker wins by just successfully cracking one vulnerability, while a defender wins only when they can defend all attacks. When new and more powerful generation models and techniques keep emerging and developing, it is difficult for passive Deepfake detectors (defenders) to cope.

4.2. Future Directions

4.2.1. Deepfake Generation. The advent of DMs has ushered in a new era for the Deepfake generation, significantly blurring the line between AI-generated content and reality. These powerful algorithms have dramatically enhanced the quality and realism of synthetic media, pushing the boundaries of what's possible in artificial content creation. To further advance the capabilities of Deepfake generators, we propose two main directions of development: video generation and multimodal content generation. Video generation aims to create increasingly realistic and dynamic synthetic footage, potentially revolutionizing fields from entertainment to education. On the other hand, multimodal content generation seeks to integrate various forms of media—such as video, audio, and text—to create more comprehensive and immersive synthetic experiences. At the same time, as Deepfake technology grows more powerful, we must ensure that it remains secure and ethically guided. Therefore, we expand the discussion of future Deepfake generation in these three directions:

1. Improvement of Realism and Detail in Video Generation
   - Hyper-realistic Textures: Developing techniques to improve texture generation in videos to achieve hyper-realism [145]. This includes better rendering of fine details like skin texture, hair movement, and environmental interactions.
   - Dynamic Lighting and Shadow Generation: Enhancing models to better understand and generate dynamic lighting conditions and shadows that match real-world physics, thereby increasing the realism of generated videos.
   - Enhanced Long-term Dependency Modeling: Improving temporal coherence over a longer duration. Although we have seen DM-powered T2V generators such as OpenAI Sora[1] [146, 147], video generation is still challenging due to the challenging long-term dependency modeling. Therefore, more efforts are expected to avoid jitters and inconsistencies in generated video sequences.
   - Contextual Awareness: Developing models that understand and maintain the context over extended sequences, allowing for more coherent and logical story progression in generated videos. The key to this is enhancing the model's ability to understand cause-and-effect relationships and the logical flow of events over time.
2. Multimodal Generation:
   - Synchronization and Contextual Harmony: Improving the synchronization between different modalities (e.g., lip-syncing in videos) and ensuring that all generated content is contextually harmonious. Existing attempts such as [148] only generate naive and low-quality audio–video content, calling for more research attention on high-quality and long-duration multimodal generation. In addition, the large-scale deployment of large language models will have a fundamental impact on multimodal data generation [149].
   - Fine-Grained Control and Manipulation: Researching methods for providing fine-grained control over the attributes of multimodal Deepfake content [137], enabling users to manipulate specific aspects of generated media while preserving realism.
3. Evaluation metrics:
   - Context-Aware Generation Metrics: Evaluating whether a generated image or video is realistic within a given scenario and whether the overall narrative of generated content aligns with intended contexts to overcome.
   - Human-Centric Evaluation Metrics: Incorporating human perception into generation metrics. For instance, collecting subjective human evaluations of realism, believability, and creativity for generated images or videos bridges the gap between objective quality evaluation and subjective human perception.
4. Ethical and Secure Generation:
   - Privacy Preservation and Consent Mechanisms: Exploring techniques to preserve individuals' privacy and ensuring that individuals depicted in Deepfake content have given explicit consent for their likeness to be used, including robust verification methods and digital consent frameworks [150, 151].
   - Ethical Frameworks and Guidelines: Establishing clear ethical guidelines for developing and using Deepfake codes/frameworks/technologies to prevent misuse and protect individuals' rights [152, 153]. This may involve the detection of misused Deepfake and the mitigation of harmful consequences of Deepfake.

4.2.2. Deepfake Detection. As Deepfake technology continues to advance, the field of Deepfake detection must evolve correspondingly. Through the above discussion of Deepfake detection, we notice that Deepfake detectors mainly focus on spotting generation artifacts like facial inconsistencies or lighting errors. Looking ahead, future methods might include embedding watermarks in AI-generated content to improve traceability and verification. Additionally, advanced detectors could analyze the AI generation pipeline itself to better identify and understand content origins. As Deepfake technology evolves to include multimodal and video content, detection strategies will need to integrate methods for handling diverse media types. Addressing these challenges will also require ethical

considerations and collaborative efforts among researchers, technologists, and policymakers. Therefore, we discuss future directions of Deepfake detection in these critical areas:

1. Advanced Proactive Detection Techniques:
   - Imperceptible Watermarks: Creating more sophisticated watermarking techniques that are virtually undetectable to the human eye but easily identifiable by detection algorithms [154]. While emerging new Deepfake generators bring challenges to Deepfake detectors, proactive Deepfake detection is hopeful to protect people from the abuse of Deepfake.
   - Dynamic Watermarking for Video Content: Developing watermarks that change dynamically over time or in response to attempts to tamper with or remove them, thereby enhancing traceability and security.
2. Detection of Generation Advanced Artifacts:
   - Pipeline Signature Analysis: Investigating the unique signatures left by different Deepfake generation pipelines, such as specific noise patterns, artifact distributions, or model biases. Apart from the existing works that capture the raw artifacts [72, 86] and reconstruction artifacts [143], resolution and scaling artifacts may also be explored in Deepfake detectors.
3. Temporal and Cross-Modal Analysis:
   - Temporal Coherence Analysis: Developing techniques to assess the temporal coherence of video content, detecting inconsistencies in motion, lighting, or scene transitions that suggest manipulation.
   - Cross-Modal Discrepancy Detection: Leveraging discrepancies between audio, video, and metadata to identify content that has been manipulated. For instance, mismatches between lip movement and spoken words, or inconsistencies in background noise and visual setting.
4. Evaluation Metrics:
   - Threshold-independent evaluation metrics: Introducing adaptive metrics that dynamically adjust based on the deployment context, where the threshold for decision-making could be adapted in real-time to account for factors like risk tolerance or user preferences (e.g., stricter detection thresholds for sensitive content, more relaxed thresholds for noncritical use cases).
   - Evaluating Temporal Consistency in Video Deepfakes: Creating metrics that assess the coherence of facial movements, eye blinking, and other dynamic features over time and evaluate how well-detection models identify temporal artifacts, such as unnatural motion patterns, inconsistent lighting, or abrupt changes in the background that are typical in video Deepfakes.
5. Community and Framework Development:
   - Open Datasets and Benchmarks: Creating and sharing extensive datasets of Deepfake content, along with benchmarks for evaluating the performance of detection models, to facilitate collaborative research.
   - Legal and Ethical Frameworks: Establishing legal and ethical frameworks that guide the development and deployment of Deepfake detection technologies, ensuring they are used responsibly and do not infringe on privacy rights.

By addressing these areas, the field can move toward more proactive, robust, and adaptable methods for detecting Deepfake, safeguarding individuals and societies against the potential misuse of this powerful technology. The evolution of Deepfake detection strategies will likely be an ongoing process, requiring continuous innovation and collaboration among researchers, technologists, and policymakers.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

This work is supported by ARC Projects (LP220200808, DP230100246, and DP250100463) from the Australian Research Council, Australia.

Endnotes

[1]https://openai.com/sora.

References

[1] N. Aldausari, A. Sowmya, N. Marcus, and G. Mohammadi, "Video Generative Adversarial Networks: A Review," *ACM Computing Surveys* 55, no. 2 (2022): 1–25, https://doi.org/10.1145/3487891.

[2] M. Farajzadeh-Zanjani, R. Razavi-Far, M. Saif, and V. Palade, *Generative Adversarial Networks: A Survey on Training, Variants, and Applications* (2022).

[3] N. Gao, H. Xue, W. Shao, et al., "Generative Adversarial Networks for Spatio-Temporal Data: A Survey," *ACM Transactions on Intelligent Systems and Technology (TIST)* 13, no. 2 (2022): 1–25, https://doi.org/10.1145/3474838.

[4] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications," *IEEE Transactions on Knowledge and Data Engineering* 35, no. 4 (2023): 3313–3332, https://doi.org/10.1109/tkde.2021.3130191.

[5] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering Malicious Deepfakes: Survey, Battleground, and Horizon," *International Journal of Computer Vision* 130,

no. 7 (2022): 1678–1734, https://doi.org/10.1007/s11263-022-01606-8.

[6] H. Cao, C. Tan, Z. Gao, et al., "A Survey on Generative Diffusion Models," *IEEE Transactions on Knowledge and Data Engineering* 36, no. 7 (2024): 2814–2830, https://doi.org/10.1109/tkde.2024.3361474.

[7] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion Models in Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 9 (2023): 10850–10869, https://doi.org/10.1109/tpami.2023.3261988.

[8] L. Lin, Z. Li, R. Li, X. Li, and J. Gao, "Diffusion Models for Time-Series Applications: A Survey," *Frontiers of Information Technology & Electronic Engineering* (2023): 1–23.

[9] L. Yang, Z. Zhang, Y. Song, et al., "Diffusion Models: A Comprehensive Survey of Methods and Applications," *ACM Computing Surveys* 56, no. 4 (2023): 1–39, https://doi.org/10.1145/3626235.

[10] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," *IEEE Access* 10 (2022): 25494–25513, https://doi.org/10.1109/access.2022.3154404.

[11] D. Yadav and S. Salmani, "Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (IEEE, 2019), 852–857.

[12] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A Survey on Deepfake Video Detection," *IET Biometrics* 10, no. 6 (2021): 607–624, https://doi.org/10.1049/bme2.12031.

[13] J. W. Seow, M. K. Lim, R. C. Phan, and J. K. Liu, "A Comprehensive Overview of Deepfake: Generation, Detection, Datasets, and Opportunities," *Neurocomputing* 513 (2022): 351–371, https://doi.org/10.1016/j.neucom.2022.09.135.

[14] P. Swathi and S. Sk, "Deepfake Creation and Detection: A Survey," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)* (IEEE, 2021), 584–588.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Networks," *Communications of the ACM* 63, no. 11 (2020): 139–144, https://doi.org/10.1145/3422622.

[16] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, *On Convergence and Stability of Gans* (2017).

[17] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in *International Conference on Machine Learning* (PMLR, 2017), 214–223.

[18] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved Training of Wasserstein Gans," *Advances in Neural Information Processing Systems* 30 (2017).

[19] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "CVAE-GAN: Fine-Grained Image Generation Through Asymmetric Training," in *Proceedings of the IEEE International Conference on Computer Vision* (2017), 2745–2754.

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen, in *Progressive Growing of GANs for Improved Quality, Stability, and Variation* (2017).

[21] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least Squares Generative Adversarial Networks," in *Proceedings of the IEEE International Conference on Computer Vision* (2017), 2794–2802.

[22] A. Brock, J. Donahue, and K. Simonyan, in *Large Scale GAN Training for High Fidelity Natural Image Synthesis* (2018).

[23] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, *Spectral Normalization for Generative Adversarial Networks* (2018).

[24] A. Odena, C. Olah, and J. Shlens, "Conditional Image Synthesis With Auxiliary Classifier GANs," in *International Conference on Machine Learning* (PMLR, 2017), 2642–2651.

[25] I. O. Tolstikhin, S. Gelly, O. Bousquet, C.-J. Simon-Gabriel, and B. Schölkopf, "ADAGAN: Boosting Generative Models," *Advances in Neural Information Processing Systems* 30 (2017).

[26] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, *Unrolled Generative Adversarial Networks* (2016).

[27] E. Schonfeld, B. Schiele, and A. Khoreva, "A U-Net Based Discriminator for Generative Adversarial Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 8207–8216.

[28] B. Liu, Y. Zhu, K. Song, and A. Elgammal, "Towards Faster and Stabilized GAN Training for High-Fidelity Few-Shot Image Synthesis," in *International Conference on Learning Representations* (2020).

[29] A. Radford, L. Metz, and S. Chintala, *Unsupervised Representation Learning With Deep Convolutional Generative Adversarial Networks* (2015).

[30] M. Afifi, M. A. Brubaker, and M. S. Brown, "HistoGAN: Controlling Colors of GAN-generated and Real Images via Color Histograms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 7941–7950.

[31] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), 4401–4410.

[32] 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), 8110–8119.

[33] T. Karras, M. Aittala, S. Laine, et al., "Alias-Free Generative Adversarial Networks," *Advances in Neural Information Processing Systems* 34 (2021): 852–863.

[34] I. Skorokhodov, S. Tulyakov, and M. Elhoseiny, "StyleGAN-V: A Continuous Video Generator With the Price, Image Quality and Perks of StyleGAN2," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 3626–3636.

[35] A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila, *StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-To-Image Synthesis* (2023).

[36] S. Ruan, Y. Zhang, K. Zhang, et al., "Dae-GAN: Dynamic Aspect-Aware GAN for Text-To-Image Synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 13960–13969.

[37] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling," *Advances in Neural Information Processing Systems* 29 (2016).

[38] S. Ge, T. Hayes, H. Yang, et al., "Long Video Generation With Time-Agnostic VqGAN and Time-Sensitive Transformer," in *European Conference on Computer Vision* (Springer, 2022), 102–118.

[39] W. Liao, K. Hu, M. Y. Yang, and B. Rosenhahn, "Text to Image Generation With Semantic-Spatial Aware GAN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 18187–18196.

[40] J. Jeon, J. Kim, H. Song, S. Cho, and N. Park, "Gt-GAN: General Purpose Time Series Synthesis With Generative Adversarial Networks," *Advances in Neural Information Processing Systems* 35 (2022): 36999–37010.

[41] 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* (2018), 8789–8797.

[42] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN V2: Diverse Image Synthesis for Multiple Domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 8188–8197.

[43] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject Agnostic Face Swapping and Reenactment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), 7184–7193.

[44] Y. Nirkin, Y. Keller, and T. Hassner, "FSGANV2: Improved Subject Agnostic Face Swapping and Reenactment," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 1 (2023): 560–575, https://doi.org/10.1109/tpami.2022.3155571.

[45] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "ATTGAN: Facial Attribute Editing by Only Changing What You Want," *IEEE Transactions on Image Processing* 28, no. 11 (2019): 5464–5478, https://doi.org/10.1109/tip.2019.2916751.

[46] W. Jiang, N. Xu, J. Wang, et al., "Language-Guided Global Image Editing via Cross-Modal Cyclic Mechanism," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 2115–2124.

[47] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proceedings of the IEEE International Conference on Computer Vision* (2017), 2223–2232.

[48] W. Luo, S. Yang, H. Wang, B. Long, and W. Zhang, "Context-Consistent Semantic Image Editing With Style-Preserved Modulation," in *European Conference on Computer Vision* (Springer, 2022), 561–578.

[49] J. Lin, R. Zhang, F. Ganz, S. Han, and J.-Y. Zhu, "Anycost GANs for Interactive Image Synthesis and Editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 14986–14996.

[50] B. Rajesh, N. Dusa, M. Javed, S. R. Dubey, and P. Nagabhushan, "T2CI-GAN: Text to Compressed Image Generation Using Generative Adversarial Network," in *International Conference on Computer Vision and Image Processing* (Springer, 2022), 292–307.

[51] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-Driven Manipulation of StyleGAN Imagery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 2085–2094.

[52] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Towards Open-Set Identity Preserving Face Synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), 6713–6722, https://doi.org/10.1109/cvpr.2018.00702.

[53] A. Rossler, 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), 1–11, https://doi.org/10.1109/iccv.2019.00009.

[54] C.-C. Hsu, C.-Y. Lee, and Y.-X. Zhuang, "Learning to Detect Fake Face Images in the Wild," in *2018 International Symposium on Computer, Consumer and Control (IS3C)* (IEEE, 2018), 388–391.

[55] F. Matern, C. Riess, and M. Stamminger, "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)* (IEEE, 2019), 83–92.

[56] H. Nguyen, J. Yamagishi, and I. Echizen, *Use of a Capsule Network to Detect Fake Images and Videos* (2019).

[57] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video Face Manipulation Detection Through Ensemble of CNNs," in *2020 25th International Conference on Pattern Recognition (ICPR)* (IEEE, 2021), 5012–5019.

[58] Y. He, B. Gan, S. Chen, et al., "ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 4360–4369.

[59] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, "Incremental Learning for the Detection and Classification of GAN-Generated Images," in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)* (IEEE, 2019), 1–6.

[60] R. Durall, M. Keuper, and J. Keuper, "Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks are Failing to Reproduce Spectral Distributions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 7890–7899.

[61] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues," in *European Conference on Computer Vision* (Springer, 2020), 86–103.

[62] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-Branch Recurrent Network for Isolating Deepfakes in Videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16* (Springer, 2020), 667–684.

[63] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What Makes Fake Images Detectable? Understanding Properties that Generalize," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16* (Springer, 2020), 103–120.

[64] Y. Huang, F. Juefei-Xu, Q. Guo, Y. Liu, and G. Pu, "Fakelocator: Robust Localization of GAN-based Face Manipulations," *IEEE Transactions on Information Forensics and Security* 17 (2022): 2657–2672, https://doi.org/10.1109/tifs.2022.3141262.

[65] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning Self-Consistency for Deepfake Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 15023–15033.

[66] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing Face Forgery Detection With High-Frequency Features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 16317–16326.

[67] X. Zhu, H. Wang, H. Fei, Z. Lei, and S. Z. Li, "Face Forgery Detection by 3D Decomposition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 2929–2939.

[68] S. A. Khan and H. Dai, "Video Transformer for Deepfake Detection With Incremental Learning," in *Proceedings of the 29th ACM International Conference on Multimedia* (2021), 1821–1828, https://doi.org/10.1145/3474085.3475332.

[69] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring Temporal Coherence for More General Video Face Forgery Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 15044–15054.

[70] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "Deepfake Detection Based on Discrepancies Between Faces and Their

Context," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, no. 10 (2022): 6111–6121, https://doi.org/10.1109/tpami.2021.3093446.

[71] J. Wang, Z. Wu, W. Ouyang, et al., "M2TR: Multi-Modal Multi-Scale Transformers for Deepfake Detection," in *Proceedings of the 2022 International Conference on Multimedia Retrieval* (2022), 615–623, https://doi.org/10.1145/3512527.3531415.

[72] X. Dong, J. Bao, D. Chen, et al., "Protecting Celebrities From Deepfake With Identity Consistency Transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 9468–9478.

[73] Z. Gu, T. Yao, Y. Chen, S. Ding, and L. Ma, "Hierarchical Contrastive Inconsistency Learning for Deepfake Video Detection," in *European Conference on Computer Vision* (Springer, 2022), 596–613.

[74] S. Dong, J. Wang, J. Liang, H. Fan, and R. Ji, "Explaining Deepfake Detection by Analysing Image Matching," in *European Conference on Computer Vision* (Springer, 2022), 18–35.

[75] J. Liang, H. Shi, and W. Deng, "Exploring Disentangled Content Information for Face Forgery Detection," in *European Conference on Computer Vision* (Springer, 2022), 128–145.

[76] W. Zhuang, Q. Chu, Z. Tan, et al., "UIA-ViT: Unsupervised Inconsistency-Aware Method Based on Vision Transformer for Face Forgery Detection," in *European Conference on Computer Vision* (Springer, 2022), 391–407.

[77] C. Miao, Z. Tan, Q. Chu, N. Yu, and G. Guo, "Hierarchical Frequency-Assisted Interactive Networks for Face Manipulation Detection," *IEEE Transactions on Information Forensics and Security* 17 (2022): 3008–3021, https://doi.org/10.1109/tifs.2022.3198275.

[78] P. Yu, J. Fei, Z. Xia, Z. Zhou, and J. Weng, "Improving Generalization by Commonality Learning in Face Forgery Detection," *IEEE Transactions on Information Forensics and Security* 17 (2022): 547–558, https://doi.org/10.1109/tifs.2022.3146781.

[79] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining Efficientnet and Vision Transformers for Video Deepfake Detection," in *International Conference on Image Analysis and Processing* (Springer, 2022), 219–229.

[80] 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* (PMLR, 2020), 3247–3258.

[81] Z. Sun, Y. Han, Z. Hua, N. Ruan, and W. Jia, "Improving the Efficiency and Robustness of Deepfakes Detection Through Precise Geometric Features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 3609–3618.

[82] L. Trinh, M. Tsang, S. Rambhatla, and Y. Liu, "Interpretable and Trustworthy Deepfake Detection via Dynamic Prototypes," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021), 1973–1983.

[83] R. Shao, T. Wu, and Z. Liu, "Detecting and Grounding Multi-Modal Media Manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), 6904–6913.

[84] G. Pang, B. Zhang, Z. Teng, Z. Qi, and J. Fan, "MRE-Net: Multi-Rate Excitation Network for Deepfake Video Detection," *IEEE Transactions on Circuits and Systems for Video Technology* 33, no. 8 (2023): 3663–3676, https://doi.org/10.1109/tcsvt.2023.3239607.

[85] G. Li, X. Zhao, and Y. Cao, "Forensic Symmetry for Deepfakes," *IEEE Transactions on Information Forensics and Security* 18 (2023): 1095–1110, https://doi.org/10.1109/tifs.2023.3235579.

[86] B. Liu, B. Liu, M. Ding, T. Zhu, and X. Yu, "Ti2Net: Temporal Identity Inconsistency Network for Deepfake Detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2023), 4691–4700.

[87] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 3994–4004, https://doi.org/10.1109/cvpr52729.2023.00389.

[88] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang, "ISTVT: Interpretable Spatial-Temporal Video Transformer for Deepfake Detection," *IEEE Transactions on Information Forensics and Security* 18 (2023): 1335–1348, https://doi.org/10.1109/tifs.2023.3239223.

[89] W. Yang, X. Zhou, Z. Chen, et al., "AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake," *IEEE Transactions on Information Forensics and Security* 18 (2023): 2015–2029, https://doi.org/10.1109/tifs.2023.3262148.

[90] S. Fernandes, S. Raj, E. Ortiz, et al., "Predicting Heart Rate Variations of Deepfake Videos Using Neural ODE," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (2019).

[91] L. Li, J. Bao, T. Zhang, et al., "Face X-Ray for More General Face Forgery Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 5001–5010.

[92] "CNN-Generated Images are Surprisingly Easy to Spot. . .for Now," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 8695–8704.

[93] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the Detection of Digital Face Manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 5781–5790.

[94] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN:," in *Construction of a Large-Scale Image Dataset Using Deep Learning With Humans in the Loop* (2015).

[95] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), 3730–3738, https://doi.org/10.1109/iccv.2015.425.

[96] N. Yu, L. S. Davis, and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), 7556–7566.

[97] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *Advances in Neural Information Processing Systems* 33 (2020): 6840–6851.

[98] J. Song, C. Meng, and S. Ermon, in *Denoising Diffusion Implicit Models* (2020).

[99] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, *Score-Based Generative Modeling Through Stochastic Differential Equations* (2020).

[100] A. Q. Nichol and P. Dhariwal, "Improved Denoising Diffusion Probabilistic Models," in *International Conference on Machine Learning* (PMLR, 2021), 8162–8171.

[101] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models," in *2021 IEEE/CVF International Conference on*

Computer Vision (ICCV) (2021), 14347–14356, https://doi.org/10.1109/iccv48922.2021.01410.

[102] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," *Advances in Neural Information Processing Systems* 34 (2021): 8780–8794.

[103] J. Ho and T. Salimans, *Classifier-Free Diffusion Guidance* (2022).

[104] 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* (2022), 10684–10695.

[105] C. H. Wu and F. De la Torre, in *Unifying Diffusion Models' Latent Space, With Applications to Cycle Diffusion and Guidance* (2022).

[106] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver: A Fast Ode Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps," *Advances in Neural Information Processing Systems* 35 (2022): 5775–5787.

[107] A. Nichol, P. Dhariwal, A. Ramesh, et al., in *Glide: Towards Photorealistic Image Generation and Editing With Text-Guided Diffusion Models* (2021).

[108] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, *Hierarchical Text-Conditional Image Generation With Clip Latents* (2022).

[109] M. Hu, Y. Wang, T.-J. Cham, J. Yang, and P. N. Suganthan, "Global Context With Discrete Diffusion in Vector Quantised Modelling for Image Generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 11502–11511.

[110] Y. He, T. Yang, Y. Zhang, Y. Shan, and Q. Chen, *Latent Video Diffusion Models for High-Fidelity Long Video Generation* (2022).

[111] C. Saharia, W. Chan, S. Saxena, et al., "Photorealistic Text-To-Image Diffusion Models With Deep Language Understanding," in *Advances in Neural Information Processing Systems*, ed. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 35 (Curran Associates, Inc, 2022), 36479–36494, https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf.

[112] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum, "Compositional Visual Generation With Composable Diffusion Models," in *European Conference on Computer Vision* (Springer, 2022), 423–439.

[113] J. S. Smith, Y.-C. Hsu, L. Zhang, et al., *Continual Diffusion: Continual Customization of Text-To-Image Diffusion With C-Lora* (2023).

[114] H. Ni, C. Shi, K. Li, S. X. Huang, and M. R. Min, "Conditional Image-To-Video Generation With Latent Flow Diffusion Models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 18444–18455, https://doi.org/10.1109/cvpr52729.2023.01769.

[115] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine Tuning Text-To-Image Diffusion Models for Subject-Driven Generation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 22500–22510, https://doi.org/10.1109/cvpr52729.2023.02155.

[116] C. Meng, R. Rombach, R. Gao, et al., "On Distillation of Guided Diffusion Models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 14297–14306, https://doi.org/10.1109/cvpr52729.2023.01374.

[117] A. Bansal, H.-M. Chu, A. Schwarzschild, et al., "Universal Guidance for Diffusion Models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2023), 843–852, https://doi.org/10.1109/cvprw59228.2023.00091.

[118] M. Kim, F. Liu, A. Jain, and X. Liu, "DCFACE: Synthetic Face Generation With Dual Condition Diffusion Model," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 12715–12725, https://doi.org/10.1109/cvpr52729.2023.01223.

[119] L. Papa, L. Faiella, L. Corvitto, L. Maiano, and I. Amerini, "On the Use of Stable Diffusion for Creating Realistic Faces: From Generation to Detection," in *2023 11th International Workshop on Biometrics and Forensics (IWBF)* (IEEE, 2023), 1–6.

[120] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video Diffusion Models," *Advances in Neural Information Processing Systems* 35 (2022): 8633–8646.

[121] M. Stypułkowski, K. Vougioukas, S. He, M. Zięba, S. Petridis, and M. Pantic, *Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation* (2023).

[122] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood, "Flexible Diffusion Modeling of Long Videos," *Advances in Neural Information Processing Systems* 35 (2022): 27953–27965.

[123] L. Zhang, A. Rao, and M. Agrawala, "Adding Conditional Control to Text-To-Image Diffusion Models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), 3836–3847.

[124] S. Zhao, D. Chen, Y.-C. Chen, et al., "Uni-ControlNet: All-In-One Control to Text-To-Image Diffusion Models," *Advances in Neural Information Processing Systems* 36 (2024).

[125] T. Yi, J. Fang, J. Wang, et al., "GaussianDreamer: Fast Generation From Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 6796–6807, https://doi.org/10.1109/cvpr52733.2024.00649.

[126] H. Chen, M. Xia, Y. He, et al., *VideoCrafter1: Open Diffusion Models for High-Quality Video Generation* (2023).

[127] H. Chen, Y. Zhang, X. Cun, et al., "VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 7310–7320, https://doi.org/10.1109/cvpr52733.2024.00698.

[128] C. Meng, Y. He, Y. Song, et al., in *SDEdit: Guided Image Synthesis and Editing With Stochastic Differential Equations* (2021).

[129] K. Kim, Y. Kim, S. Cho, et al., *DiffFace: Diffusion-Based Face Swapping With Facial Guidance* (2022).

[130] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting Using Denoising Diffusion Probabilistic Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 11461–11471.

[131] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, *DiffEdit: Diffusion-Based Semantic Image Editing with Mask Guidance* (2022).

[132] D. Valevski, M. Kalman, Y. Matias, and Y. Leviathan, *UniTune: Text-Driven Image Editing by Fine Tuning an Image Generation Model on a Single Image* (2022).

[133] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, *Prompt-to-Prompt Image Editing With Cross Attention Control* (2022).

[134] G. Kim, T. Kwon, and J. C. Ye, "DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 2426–2435.

[135] O. Avrahami, D. Lischinski, and O. Fried, "Blended Diffusion for Text-Driven Editing of Natural Images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 18208–18218.

[136] B. Kawar, S. Zada, O. Lang, et al., "Imagic: Text-Based Real Image Editing With Diffusion Models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 6007–6017, https://doi.org/10.1109/cvpr52729.2023.00582.

[137] Y. Shi, C. Xue, J. H. Liew, et al., "DragDiffusion: Harnessing Diffusion Models for Interactive Point-Based Image Editing," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 8839–8849, https://doi.org/10.1109/cvpr52733.2024.00844.

[138] M. Boháček and H. Farid, "A Geometric and Photometric Exploration of GAN and Diffusion Synthesized Faces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), 874–883.

[139] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the Detection of Synthetic Images Generated by Diffusion Models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2023), 1–5.

[140] J. Ricker, S. Damm, T. Holz, and A. Fischer, *Towards the Detection of Diffusion Model Deepfakes* (2022).

[141] Z. Sha, Z. Li, N. Yu, and Y. Zhang, *De-Fake: Detection and Attribution of Fake Images Generated by Text-To-Image Diffusion Models* (2022).

[142] P. Lorenz, R. Durall, and J. Keuper, "Detecting Images Generated by Deep Diffusion Models Using Their Local Intrinsic Dimensionality," in *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (2023), 448–459, https://doi.org/10.1109/iccvw60793.2023.00051.

[143] Z. Wang, J. Bao, W. Zhou, et al., in *Dire for Diffusion-Generated Image Detection* (2023).

[144] R. Ma, J. Duan, F. Kong, X. Shi, and K. Xu, *Exposing the Fake: Effective Diffusion-Generated Images Detection* (2023).

[145] S. Zhou, P. Yang, J. Wang, Y. Luo, and C. C. Loy, "Upscale-a-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 2535–2545, https://doi.org/10.1109/cvpr52733.2024.00245.

[146] X. Chen, Y. Wang, L. Zhang, et al., "Seine: Short-To-Long Video Diffusion Model for Generative Transition and Prediction," in *The Twelfth International Conference on Learning Representations* (2023).

[147] Z. Zhang, B. Wu, X. Wang, et al., "AVID: Any-Length Video Inpainting With Diffusion Model," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 7162–7172, https://doi.org/10.1109/cvpr52733.2024.00684.

[148] L. Ruan, Y. Ma, H. Yang, et al., "MM-diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 10219–10228, https://doi.org/10.1109/cvpr52729.2023.00985.

[149] W. Zhou, X. Zhu, Q.-L. Han, et al., "The Security of Using Large Language Models-A Survey With Emphasis on ChatGPT," *IEEE/CAA Journal of Automatica Sinica* 12, no. 1 (2025): 1–26, https://doi.org/10.1109/JAS.2024.124983.

[150] C. Xu, Y. Qu, Y. Xiang, and L. Gao, "Asynchronous Federated Learning on Heterogeneous Devices: A Survey," *Computer Science Review* 50 (2023): 100595, https://doi.org/10.1016/j.cosrev.2023.100595.

[151] Y. Zhao, Y. Qu, Y. Xiang, M. P. Uddin, D. Peng, and L. Gao, "A Comprehensive Survey on Edge Data Integrity Verification: Fundamentals and Future Trends," *ACM Computing Surveys* 57, no. 1 (2024): 1–34, https://doi.org/10.1145/3680277.

[152] W. Ma, Y. Song, M. Xue, S. Wen, and Y. Xiang, "The "Code" of Ethics: A Holistic Audit of Ai Code Generators," *IEEE Transactions on Dependable and Secure Computing* 21, no. 5 (2024): 4997–5013, https://doi.org/10.1109/TDSC.2024.3367737.

[153] X. Zhu, W. Zhou, Q.-L. Han, W. Ma, S. Wen, and Y. Xiang, "When Software Security Meets Large Language Models: A Survey," *IEEE/CAA Journal of Automatica Sinica* 12, no. 2 (2025): 317–334, https://doi.org/10.1109/JAS.2024.124971.

[154] Z. Yang, K. Zeng, K. Chen, H. Fang, W. Zhang, and N. Yu, "Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 12162–12171, https://doi.org/10.1109/cvpr52733.2024.01156.