Deepfake Generation and Proactive Deepfake Defense: A Comprehensive Survey

Hong-Hanh Nguyen-Le, *Member, IEEE,* Van-Tuan Tran, *Member, IEEE,* Dinh-Thuc Nguyen, *Non-Member* and Nhien-An Le-Khac, *Member, IEEE*

Abstract—**The proliferation of highly realistic deepfakes, powered by Generative Artificial Intelligence (GenAI), presents significant challenges to digital trust and security. This survey provides a comprehensive overview of proactive deepfake detection approaches, including *disruption* and *watermarking* methods. Our survey provides a taxonomy of these strategies based on their existing methodologies and extend the discussion to other perspectives, including imperceptibility, transferability, universality, and robustness. We also explore the associated threat models, considering various adversary objectives and capabilities. Additionally, we review state-of-the-art deepfake generation techniques that provide context for the challenges faced by detection methods.**

Index Terms—**deepfake detection, proactive detection, disruption, watermarking, deepfake generation, generative AI**

I. Introduction

The rapid advancement of Generative Artificial Intelligence (GenAI) has revolutionized the field of digital content creation, demonstrating remarkable capabilities across various modalities. These models have found applications in diverse domains, including image and video generation [1, 2], audio synthesis [3, 4], natural language processing [5, 6], healthcare [7, 8], education [9], and business [10]. The practical utility of GenAI is evident in content generation products developed by both large tech companies and innovative startups, such as ChatGPT[1], Sora[2], HeyGen AI Talking Video Generator[3], ElevenLabs Voice Generator[4], Suno AI Music Generator[5], and Scenario[6].

However, the increasing sophistication of GenAI technologies has raised concerns about privacy and security, particularly regarding the malicious use of this technology for creating deepfakes (DFs). Recent incidents, such as the DF arrest images of Donald Trump [11], the 25 million fraud using DF technology in a video conference call [12] and the artist's style imitation [13], underscore the urgent need to address potential misuse and implement appropriate safeguards. In response, the research community has developed two primary categories of DF defense strategies: **passive** and **proactive**. Passive detection techniques focus on identifying artifacts in DF data to differentiate authentic content from forgeries. These techniques typically employ machine learning/deep learning (ML/DL) algorithms to identify signs of manipulation in images or videos after they have been created and distributed. Conversely, proactive detection approaches aim to address the issue at its source by implementing preventive measures during the content creation or distribution process. Figure 1 illustrates the differences between two types of defense methods.

While passive detectors have shown efficacy against certain types of deepfakes, they face limitations in combating unseen or improved generators that adapt to existing detection methods [14, 15, 16]. As technology progresses, the potential emergence of generators capable of producing virtually indistinguishable synthetic contents poses a significant challenge to the long-term effectiveness of passive detection techniques. In contrast, proactive methods offer a mechanism to block the wide spreading of DFs before causing damage impact, while also empowering content creators and AI model providers with tools to protect their work and maintain output integrity. These proactive methods can be further divided into two main categories:

- **Disruption approaches**: These techniques focus on protecting individuals' data from exploitation or replication by GenAI models. By applying perturbations to images before they are shared online, disruption approaches aim to make it difficult for unauthorized parties to use this data for training or fine-tuning GenAI models.
- **Watermarking approaches**: These methods involve embedding imperceptible and unique signatures into GenAI models' outputs. These watermarks serve as a form of authentication, allowing for the identification of legitimate AI-generated content and the detection of unauthorized manipulations.

This survey paper focuses exclusively on proactive deepfake detection approaches, providing a comprehensive overview of current research, methodologies, and challenges in this emerging field. The survey primarily reviews proactive approaches in the image domain due to the current lack of such approaches in video or audio domains. We also analyze the threat models associated with both disruption and watermarking techniques. Note that, in this survey, we limit our scope to techniques targeting the impersonation of real individuals. We exclude methods that generate entirely new, synthetic images of non-existent people, as these do not involve impersonation and typically serve different purposes than traditional deepfake attacks that aim to impersonate real individuals.

Manuscript received April 19, 2005; revised August 26, 2015.

[1] https://openai.com/index/chatgpt

[2] https://openai.com/index/sora

[3] https://www.heygen.com/

[4] https://elevenlabs.io/

[5] https://sunoai.ai/

[6] https://www.scenario.com/

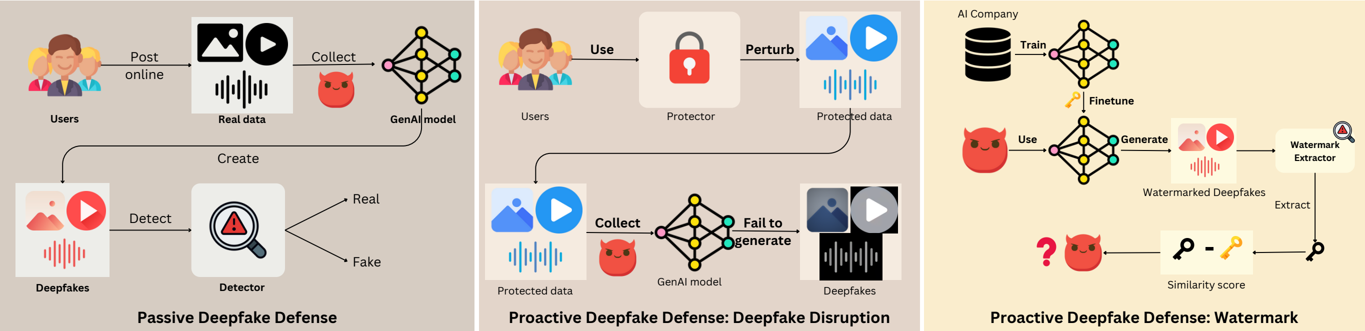

Fig. 1: An overview of passive and proactive deepfake defense approaches

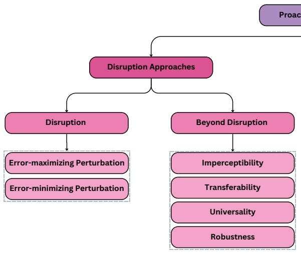

Fig. 2: Taxonomy of proactive anti-DF approaches (in Section IV and V).

A. Contributions

Prior surveys have primarily focused on passive defense approaches [17, 18, 19, 20, 21], while there has been a lack of surveys covering proactive approaches. Our work is distinct from [22], which provides a comprehensive survey of watermark approaches for large language models. In contrast, our survey focuses specifically on proactive approaches, including disruption and watermarking, for combating the generation of DFs using GenAI-based image generation techniques. We particularly emphasize the malicious creation and use of DFs, which represent a significant and growing threat in the digital landscape. To the best of our knowledge, this work represents the first comprehensive survey of proactive approaches for combating DFs. The main contributions of this paper are as follows:

- A novel taxonomy for categorizing proactive DF defense approaches based on their underlying goals and methodologies, enabling a structured analysis of the current state of research in the field of DF detection.
- An exploration of perspectives beyond disruption and watermarking in these approaches, including methods to enhance imperceptibility, transferability, universality, and robustness of DF detectors.
- In-depth analysis of threat models for both disruption and watermarking approaches, considering various adversary objectives, knowledge levels, and capabilities.
- An overview of state-of-the-art (SOTA) DF generation techniques, including face swapping, face reenactment, talking-face video, and artist style imitation. We also analyze crucial techniques that enable high-quality realistic DFs.

B. Organization

Section II provides background information on DFs, generative modeling, adversarial examples and outlines common benchmarks, datasets, and metrics used to evaluate the effectiveness of proactive defense approaches. Section III summarizes recent advances in GenAI-driven DF creation. A comprehensive review of disruption and watermarking strategies is presented in sections IV and V, respectively. In each section, we introduce our taxonomy and discuss properties beyond accuracy detection for these approaches. Figure 2 illustrates more detailed sub-categories of each approach. In Section VI, the definition of threat model and defense model are provided. Finally, we present current challenges and future directions in Section VII and give our conclusion in Section VIII.

II. Backgrounds

In this section, we first review the definition of DF, generative modeling, adversarial examples and then summarize the common datasets and metrics used to evaluate the anti-DF approaches.

A. Concept of Deepfake

Deepfakes (DFs), a portmanteau of ”deep learning” and ”fake”, refer to synthetic content generated using advanced AI techniques, particularly GenAI models [17]. DFs are often used for malicious purposes, such as impersonating individuals, generating non-consensual explicit content or imitating artists’ style, raising serious ethical and societal concerns. DFs are very different from entirely synthetic data. DFs typically manipulate media based on existing data, often combining real footage or images with AI-generated content to create a convincing but false representation. For instance, a deepfake video might superimpose one person’s face onto another’s body, or make it appear as though someone is saying something they never actually said [17]. In contrast, entirely synthetic data are new content generated from scratch without directly manipulating existing media [23]. For example, a GAN trained on a dataset of human faces can create entirely new, non-existent faces that look realistic but do not correspond to any real person.

B. Generative Modeling

Generative models [24] aim to capture the underlying probability distribution of the training data, enabling them to create novel samples that maintain the characteristics of the original data. Given a real data distribution q_{data}, the goal a generative model is to learn the parameters θ such that the model distribution p_θ is close to the data distribution q_{data}. Mathematically, the goal can be defined as:

$$\min_{\theta} d(p_\theta, q_{data}), \tag{1}$$

where $d(.)$ is the distance between probability distributions. Several types of generative models have been developed, each with its unique characteristics and training methodologies:

Autoregressive Models [25]: By the chain rule of probability [24], these models factorize the joint distribution over the n-dimensions as:

$$p_\theta(x) = \prod_{i=1}^{n} p(x_i|x_1, ..., x_{i-1}) = \prod_{i=1}^{n} p(x_i|x_{<i}), \tag{2}$$

where $x_{<i} = [x_1, x_2, \ldots, x_{i-1}]$ denotes the vector of random variables with index less than i.

Flow-based Models [26]: Given a known distribution $p(z)$, an invertible function $f : \mathbb{R}^d \to \mathbb{R}^d$ is applied to obtain the distribution of the resulting random variable $x = f(z)$ through the change of variables rule [26]:

$$p_\theta(x) = p(z) \left| \det \frac{\partial f^{-1}}{\partial x} \right| = p(z) \left| \det \frac{\partial f}{\partial z} \right|^{-1} \tag{3}$$

Latent Variable Models [27]: The idea behind latent variable models is to assume a lower-dimensional latent space and the following generative process:

$$z \sim p_\theta(z);\; x \sim p_\theta(x|z) \tag{4}$$

Energy-based Models [28]: Energy-based models define an energy function $E_\theta(x)$ that assigns low energy to realistic data points and high energy to unrealistic ones. The probability distribution is given by:

$$p_\theta(x) = \frac{e^{-E_\theta(x)}}{\sum_x e^{-E_\theta(x)}} \tag{5}$$

We recommend readers to [29, 24] for better understanding of different types of generative models.

C. Adversarial Examples

Adversarial examples are input data that has been perturbed in order to deceive neural network (NN) classifiers, causing them to misclassify the input. This phenomenon is known as an adversarial attack. Formally, given an input x with true label y, the goal of an adversarial attack is to find an imperceptible perturbation δ such that the perturbed input $x' = x + \delta$ is misclassified by the model f [30, 31].

In an untargeted attack, the optimal adversarial perturbation δ_{adv} is found by maximizing the model's classification loss $\mathcal{L}$:

$$\delta_{\text{adv}} = \arg\max_{\delta \in \Delta} \mathcal{L}(f(x + \delta), y), \tag{6}$$

where Δ denotes the set of allowable perturbations, typically constrained by an ℓ_p-norm bound ϵ such that $||x - x'||_p \leq \epsilon$.

In a targeted attack, the adversary instead seeks to cause the model to misclassify the input as a specific target class y_{target}. The optimal perturbation for a targeted attack minimizes the loss with respect to the target class:

$$\delta_{\text{adv}} = \arg\min_{\delta \in \Delta} \mathcal{L}(f(x + \delta), y_{\text{target}}), \tag{7}$$

D. Datasets and Metrics

1) Datasets.: Proactive approaches for DF defense typically leverage existing datasets for computer vision tasks, such as face recognition or image classification. Table I summarizes common datasets used to evaluate proactive DF approaches. Key datasets include:

- FaceForensics++ (FF++) [32]: 1000 original and manipulated videos, crucial for analyzing manipulation processes.
- CelebFaces Attributes (CelebA) [33]: Large-scale dataset with diverse facial characteristics.
- Flickr-Faces-HQ (FFHQ) [34]: $70,000$ high-quality face images with high-resolution images.
- VGGFace2 [35]: 3.31 million images across 9131 identities, offering variations in pose, age, and ethnicity.
- LFW [36]: $13,233$ real-world images of $5,749$ people in unconstrained environments.
- WikiArt [37]: Over $80,000$ fine-art paintings, crucial for protection methods against style mimicry.
- LSUN [38]: Large-scale images of scenes and objects, valuable for testing in diverse contexts.
- DF Face Swapping Dataset [39]: Custom dataset with $6,274$ images of 78 identities for evaluating disruption methods.

2) Metrics.: The efficacy of proactive DF detection methodologies is evaluated using two primary categories of metrics: Perceptual measurement and Protective measurement metrics.

Perceptual Measurement Metrics. These metrics measure the similarity or difference between the protected/ watermarked sample x' and the original sample x.

Peak Signal-to-Noise Ratio (PSNR). PSNR [40] evaluates the quality of protected images by analyzing pixel-wise variances: $\text{PSNR}(x, x') = -10 \cdot \log_{10}(\text{MSE}(x, x'))$, for $x, x' \in [0, 1]^{c \times h \times w}$, where MSE is the mean squared error.

Structural Similarity Index (SSIM). SSIM [41] quantifies the structural discrepancy between an original image and its protected version: $\text{SSIM}(x, x') = [l(x, y)]^\alpha \ast [c(x, y)]^\beta \ast [s(x, y)]^\gamma$, where $l(\cdot, \cdot)$ compares luminance between x and x', $c(\cdot, \cdot)$ compares contrast between x and x', and $s(\cdot, \cdot)$ compares structure between x and x'.

Fréchet Inception Distance (FID). This metric measures the quality and diversity of protected images [42]. FID computes the Fréchet distance between the Gaussian distribution (m, C) obtained from protected images and that (m_w, C_w) obtained from original images using: $d^2((m, C), (m_w, C_w)) = ||m - m_w||^2 + Tr(C + C_w - 2(CC_w)^{\frac{1}{2}}$, where $Tr(.)$ is the trace of a matrix. One limitation of this metric is that it depends heavily on the choice of neural network model and layer activations used for computing the distributions.

TABLE I: A Summarization of Common Dataset

Dataset	Size	Resolution	Primary Use	Access
FF++ [32]	1000 videos	1920×1080	Facial manipulation	✔
CelebA [33]	$202,599$ images	Varied	Facial manipulation	✔
FFHQ [34]	$70,000$ images	1024×1024	Facial manipulation	✔
VGGFace2 [35]	3.31 million images	Varied	Facial manipulation	✔
LWF [36]	$13,233$ images	250×250	Facial manipulation	✔
WikiArt [37]	$80,000$+ images	Varied	Style manipulation	✔
LSUN [38]	~ 1 million images per category	Varied	Background manipulation	✔
DF Face Swapping Dataset [39]	$6,274$ images	–	Facial manipulation	✘

Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE). BRISQUE [43] is used as a non-reference image quality assessment metric to assess the visual quality and distortion of the protected images. This metric operates directly on the protected image by extracting statistical features that quantify image naturalness from each image independently, then mapping these features to quality scores using a pre-trained model (e.g., support vector regression). A lower BRISQUE score indicates higher image distortion and lower perceptual quality.

CLIP-based genre shift. Shan et al. [44] introduce the CLIP-based genre shift metric to assess the efficacy of techniques designed to prevent AI models from mimicking an artist's unique style. This metric utilizes the genre classification capabilities of a pre-trained CLIP model and computes the proportion of mimicked artworks for which the top three predicted genres are distinct from the original genre of the artist being imitated. A higher genre shift rate is indicative of a more effective protection method, as it implies that the style of the mimicked art has been successfully altered to diverge from the artist's authentic style.

Protective Measurement Metrics. These metrics quantify the effectiveness of proactive DF approaches, particularly the resilience of watermarks and the success rate of protected samples.

Bit Accuracy (BACC). BACC [45] measures the percentage of bits correctly decoded by comparing the extracted bit-string with the pre-defined bit-string: BACC = $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(b_i = b'_i)$, where b is the pre-defined bit-string, b is the decoded bit-string from the watermarked samples, n is the length of bit-string, and $\mathbb{1}$ is the indicator function that equals 1 when the bits match and 0 otherwise. We recommend readers to refer Section V for better understanding the definition of watermark.

Protection Success Rate (PSR). PRS [46] assesses the effectiveness of defense approaches that embed perturbations into users' data. It quantifies the proportion of perturbed samples from which the GenAI model fails to extract information.

III. Deepfake Generation Techniques

A. Current Technologies Enabling Deepfake Generation

1) Conditional Generative Modeling (CGM): Conditional generative modeling (CGM) [49] forms the cornerstone of modern DF generation techniques. CGM extends traditional GenAI models by incorporating additional input conditions, allowing for more controlled and targeted content generation. The primary objective of CGM is to learn the conditional distribution $p_\theta(x|c)$, where x represents the generated output (e.g., manipulated images or video frames) and c denotes the conditioning information. Formally, the training objective for CGMs can be expressed as:

$$\max_{\theta} \mathbb{E}_{x,c}[\log p_\theta(x|c)], \tag{8}$$

where $\mathbb{E}_{x,c}$ is the expectation over both x and c.

In the context of DFs, the conditioning factors c typically encompass target identity, attributions (e.g., facial expressions, age, appearance), and a reference image. GenAI models commonly employed in DF generation can be categorized into three families, each with distinct characteristics and capabilities. Here, we focus on conditional version of these families.

Variational Autoencoders (VAEs): VAEs [27] offer a probabilistic approach to generative modeling by learning a latent representation of the data. Conditional VAEs [50] extend this framework to incorporate conditioning information with the objective function:

$$\max_{\theta,\phi} \mathbb{E}_{q_\phi(z|c)}[\log p_\theta(x|z,c)] - KL(q_\phi(z|x,c) \parallel p_\theta(z|c)), \tag{9}$$

where ϕ is the encoder's parameters, θ os the decoder's parameters, $q_\phi(z|x,c)$ is the approximate posterior distribution, $p_\theta(x|z,c)$ is the likelihood distribution, $p(z|c)$ is the conditional prior distribution of z given c, and KL denotes the Kullback-Leibler divergence.

Generative Adversarial Networks (GANs): GANs have been widely adopted for DF generation due to their ability to produce high-fidelity synthetic content. For the conditional GANs [51, 52, 53], the optimization is formulated as:

$$\begin{aligned}\min_{\mathcal{G}_{(X,C)\to Y}} \max_{\mathcal{D}_Y} & \mathbb{E}_{y,c}[\log \mathcal{D}_Y(y,c)] \\ & + \mathbb{E}_{x,c}[\log(1 - \mathcal{D}_Y(\mathcal{G}_{(X,C)\to Y}(x,c),c))],\end{aligned} \tag{10}$$

where $\mathcal{G}_{(X,C)\to Y}$ maps from the domain X to Y, conditioned on C, and c is the conditional vector. This formulation explicitly incorporates the conditional vector c into both the generator and discriminator, allowing for condition-guided image generation. Recent advancements have developed methods that can control multiple conditions simultaneously or provide explicit control [54, 55, 56, 57, 58].

Diffusion-based Models (DMs): Diffusion models [59, 1] have emerged as a powerful approach for high-quality image synthesis through a process of iterative denoising. Given an input image z_0, a condition c, diffusion algorithms learn a

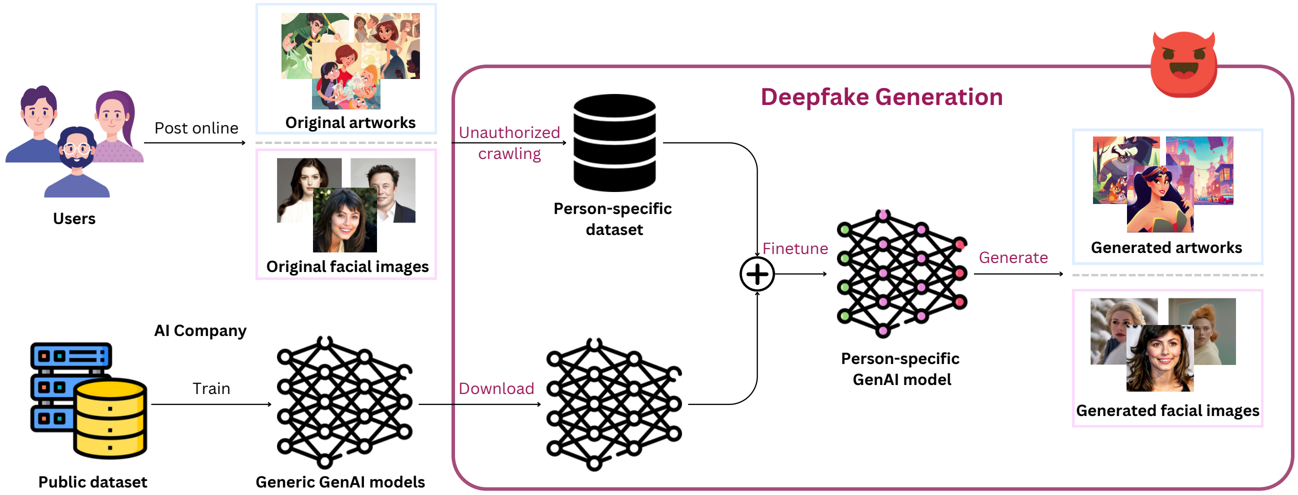

Fig. 3: Overview of the DF generation scenario in the context of Machine-Learning-as-a-Service. The adversary crawls the original images from the victim and uses these to fine-tune the generic GenAI model that is trained and open-sourced by an AI company. Then, he uses this fine-tuned GenAI model to generate DFs. Original and generated images originate from References [47, 48]

network ϵ_θ to predict the noise added to the noisy image z_t with:

$$\min_\theta \mathbb{E}_{z_0,c,\epsilon\sim N(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t,t,c)||_2^2\right], \quad (11)$$

where t represents the number of times noise is added [60] and $||.||$ denotes the squared L_2 norm. To add conditional signals into DMs, the research community has employed training-free techniques like constraining the denoising diffuion process or editing attention layer activations [61, 62, 63, 64, 65]. However, these methods struggle with other problems like depth-to-image or pose-to-image, which require end-to-end (E2E) learning and data-driven solutions. Some recent works presented E2E network architecture that learns conditional controls for large pre-trained DMs [60, 66].

These advancements have not only improved the realism of synthetic content but also expanded the diversity and controllability of generated images. As a result, the line between real and generated images has become increasingly blurred, posing new challenges for DF detection methods that previously relied on identifying imperfections in generated images. This rapid progress underscores the growing need for more robust and sophisticated detection techniques to keep pace with the evolving capabilities of GenAI models. For a more comprehensive understanding of these GenAI models, we direct readers to the published literature [67, 29, 68].

2) Feature Extraction Model (FEM).: FEM is a critical component in the pipeline of GenAI models used for DF generation. FEMs are often implemented as encoders which serve as the initial stage of processing. These models are responsible for transforming raw input data into a compact, latent representation that captures high-level features of the inputs (See Figure 4). Formally, the generation process of GenAI models can be briefly represented as follows:

$$\tilde{x} = \mathcal{G}(E(x)), \quad (12)$$

where E is the FEM, and x is the input. The output of FEM is the latent representation $z = E(x)$ that encodes crucial information such as facial structure, identity, expression, and other relevant attributes. The quality of these extracted features significantly influences the subsequent stages of the DF generation process.

3) Machine-Learning-as-a-Service (MLaaS).: The proliferation of Machine-Learning-as-a-Service (MLaaS) platforms has simplified the process of generating synthetic data. Historically, the creation of high-quality synthetic data required substantial expertise, resources, and time investment. Users need to develop and train their own GenAI models, collect extensive training dataset, and possess significant computational resources for model training. This process was time-consuming, and technically challenging, limiting the number of users capable of producing high-quality synthetic data [70]. The emergence of MLaaS has fundamentally altered this paradigm. Contemporary MLaaS platforms offer key features that dramatically simplify the process of leveraging advanced AI models: pre-trained GenAI models accessible via cloud-based APIs, user-friendly deployment interfaces and scalable computational resources on demand. These services enable users to leverage sophisticated GenAI models without the need for extensive AI expertise or substantial hardware investments [71, 72, 73].

While MLaaS offers numerous benefits for legitimate applications, it also presents significant risks in the context of DF generation. The ease of creating DFs may lead to an increase in the production and distribution of malicious synthetic media. The use of cloud-based services can complicate efforts to trace the origins of DFs. Moreover, as MLaaS platforms continue to improve, the sophistication and realism of generated DFs are likely to increase, presenting an evolving threat landscape. The efficiency and accessibility of MLaaS platforms in the context of DF generation are illustrated by recent research. For example, with only 20 unique pieces of the victim artist's artwork downloaded from online sources, the adversary can generate synthetic artworks mimicking victim's style with impressive accuracy in less than 20 minutes [44]. This example underscores the potential for rapid and targeted misuse of

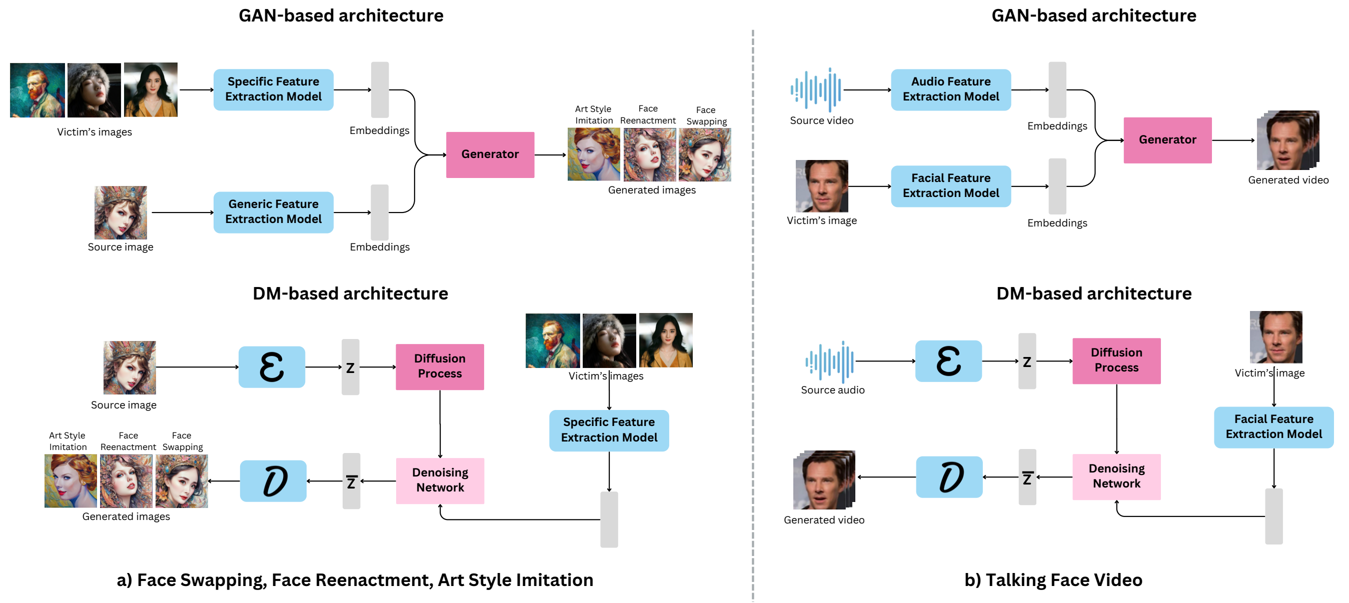

Fig. 4: Deepfake Generation Categories: Face Swapping, Face Reenactment, Art Style Imitation, and Talking Face Video. The top row is GAN-based architecture while the bottom row is DM-based architecture. Original and generated images originate from References [47, 69]

MLaaS platforms in the DF creation. Figure 3 illustrates the process of deepfake generation in the context of MLaaS. It illustrates how adversaries can leverage pre-trained generic GenAI models, typically developed by AI companies using public datasets, and fine-tune them with a relatively small amount of person-specific data obtained through unauthorized crawling. This workflow enables malicious actors to create convincing DFs without the need to build and train models from scratch, significantly reducing the time, expertise, and resources required for DFs generation.

B. State-of-the-art Deepfake Generation

Prior studies in DF detection primarily relied on the inability of GenAI models to perfectly reproduce the natural characteristics of authentic content. For images, this included inconsistencies in facial features, such as differently colored eyes or asymmetric skin tones [74, 75, 76]. However, recent advancements in GenAI models have significantly improved the quality and sophistication of DFs, addressing many previous limitations. We mainly categorize DF generation techniques into four groups: Face Reenactment, Face Swapping, Talking-face Video, and Art Style Imitation. Note that while the synthesis of entirely new DFs is a related technology, it is excluded from this discussion due to the lack of clear attack objectives. This section provides an overview of the current SOTA methods for generating these types of DFs. Figure 4 illustrates the high level design of recent models used for different DF generation.

1) Face Reenactment.: This technology manipulates the target's face to mimic the source's actions by transferring facial expressions and poses from a source actor to a target victim.

Recent research in face reenactment has focused on addressing two significant limitations of previous methods: the need for extensive target data and the challenge of handling large pose variations. Several studies have aimed to reduce the amount of target data required for high-quality reenactment, moving towards one-shot or few-shot learning approaches. To achieve high-fidelity one-shot face reenactment at 1024×1024 resolution, a hybrid latent space decomposition of identity and facial deformation has been propsoed by Oorloff and Yacoob [77]. Other researchers, such as Bounareli et al. [78] discovered controllable directions in GAN latent space for high-quality one-shot reenactment, while Yang et al. [79] utilized hierarchical motion field estimation for real-time one-shot reenactment.

Another line of research has focused on producing high-quality reenactment results under challenging conditions, especially large pose changes and occluded faces. Approaches range from simple methods like feature fusion [80] and pose-adapted shape learning [81] to more complicated methods incorporating 3D landmarks [82] and 3D Morphable Model (3DMM) [83, 84]. Particularly, Hsu et al. [82] incorporated 3D landmark detector to identify 3D landmarks from 2D faces, enabling the labeling of visible and invisible landmarks across various poses. The integration of 3DMM has shown particular promise in handling large pose variations and occlusions. Shang et al. [83] leveraged 3DMM to design a novel joint reconstruction-reenactment framework, providing 3D-aware dense conditions for improve large-pose reenactment. Behrouzi et al. [84] incorporated 3DMM and multi-scale occlusion masks to better handle pose changes and occlusion compared to 2D representations.

The Attack Model. This technology allows attackers to manipulate the facial expressions and movements of a victim in existing images or videos. The attacker could alter a public figure's expressions during a speech to convey different emotions or reactions, potentially changing the perceived context or meaning of the speech. Furthermore, this could be used to manipulate liveness detection systems in video-based authentication, by adding natural movements to static images.

2) Face Swapping.: This form of image manipulation, also known as identity replacement, involves transferring the identity from the source image to the target image while maintaining the target's original pose, expression, and environmental conditions. This is also known as identity replacement. Recent years have witnessed significant advancements in face swapping technology, with research focusing on several key areas to improve the quality and applicability of the results. As GenAI models have evolved, there has been a push towards high-resolution face swapping to create more visually convincing results. Wang et al. [85] used a 3D shape-aware identity extractor and a semantic facial fusion module to preserve the source face shape and generate photo-realistic high-resolution face swapping results. Zhu et al. [86] proposed a module to project faces into an extended version of StyleGAN2 [87]'s latent space to capture more complete facial information at different scales, allowing better preservation of both low-level topology information and high-level semantic details. Xu et al. [88] disentangled the StyleGAN's latent space separately transfer structure attributes (like pose and expression) and appearance attributes (like lighting and skin tone) from the source and target faces. It then aggregates multi-resolution features from the StyleGAN generator with background features from the target image to produce a high-quality swapped face that preserves the source identity while adopting the target's attributes and background. Not only improving high-quality results but Zhu et al. [89] also proposed a reversible autoencoder to project faces into a low-dimensional and reversible latent space, enabling greater flexibility and stability for face-swapping training. Another significant development has been the pursuit of one-shot face swapping. Previous methods often require multiple images of the source and target faces for training, limiting their generation capability. One-shot face swapping aims to perform swapping using only a single image each for the source and target, enabling more flexible and efficient face swapping [90, 86].

Face swapping becomes particularly challenging when the source and target faces have large pose variations or occlusions, leading to artifacts and inconsistencies in the swapped results. To address this, several methods have been proposed to generate more robust and realistic swapped faces under these challenging conditions. Li et al. [91] utilized a pseudo-multi-view training strategy to maintain the consistency of latent code projection when projecting 2D inputs into the latent space of a 3D generative model. Rosberg et al. [92] leveraged intermediate features from a pretrained identity encoder to preserve attributes such as pose, facial expression, and occlusion. Xu et al. [93] augmented local identity-relevant features by using a Transformer to model misaligned cross-scale semantic interactions between local facial regions like lips. Ren et al. [94] proposed to reinforce the disentanglement of identity and non-identity representations without using skip-connection. However, these GAN-based methods often struggle to fully transfer identity and shape during swapping and lack fine-grained control over face attributes and shape during swapping. Zhao et al. [95] reformulated face swapping as conditional inpainting using DMs, which handle limitations in preserving target details, stability, and occlusions. Zhu et al. [89] presented a novel framework called StableSwap that enables stable, controllable, and high-quality face swapping through latent space manipulation and multi-stage feature injection.

The Attack Model. Although face swapping has various applications in entertainment and film production, it also raises significant security and privacy concerns. One primary attack risk is the non-consensual creation of synthetic media for defamation or extortion purposes. In such scenarios, the attacker can swap a victim's face for the body of a porn actress to defame and blackmail the victim [96]. Moreover, this technology has been implicated in the proliferation of disinformation and "fake news". By manipulating images of political figures, celebrities, or other influential individuals, bad actors can craft convincing false narratives or attribute fabricated statements to these personalities [97, 11].

3) Talking-Face Video.: The emergence of talking face synthesis as a distinct category is attributed to recent progress in audio-visual synchronization and video synthesis technologies. The primary objective of this technology is to generate highly realistic and lifelike audio-driven talking face videos from a limited number of static images and a speech audio clip. Earlier approaches faced several key challenges, including (i) a lack of naturalness in facial expressions and head movements, (ii) the requirement for numerous target samples, and (iii) audio-visual misalignment.

To address these limitations, recent research has focused on four main areas of development: one-shot settings, diverse facial expressions, dynamic head movements, and unified approaches combining these aspects. One-shot settings have gained considerable attention due to their practical applications and ability to generalize to unseen subjects. Pioneering work in this direction introduced motion-aware recurrent networks to predict head poses from audio, which is subsequently used to create dense motion fields [98]. However, this method struggled to generate long facial motion sequences with high temporal consistency and audio-lip accuracy, as well as achieving fast training and inference speeds. Ye et al. [99] mitigated this issue by designing an efficient NeRF-based motion-to-video renderer using grid-based embedders and deformable slicing surfaces, enabling fast training and real-time inference. Other researchers have focused on generating high-resolution (1024×1024) videos in a one-shot setting by leveraging pre-trained StyleGAN and disentanglement learning techniques [100].

Dynamic head movements have been recognized as one of the key factors in creating realistic talking face videos. These approaches aim to generate head motions that are not only synchronized with the audio but also reflect natural human behavior. Recent works have utilized disentanglement learning [101] and discrete representation learning [102] to generate stylized talking face videos with diverse head motions. Similarly, diverse facial expressions represent another critical aspect of realistic talking face generation. While some researchers have focused on preserving audio-driven expressions without relying on additional emotion inputs [103], others have enabled flexible control of expressions in talking face generation by using multi-modal inputs [104] and self-supervised learning [105]. To adapt models to unseen faces, Ki and Min [106] proposed a masking technique and applied a few-shot adaptation.

SOTA talking face video generation methods aim to combine one-shot capability, diverse facial expressions, and dynamic head movements in comprehensive approaches. Wang et al. [107] presented a novel E2E framework for extracting comprehensive style representations and generating highly stylized talking head videos with a single reference image. Other methods have introduced diffusion-based models that generate both facial dynamics and head movements in a unified latent space [108, 109, 110]. These advanced techniques demonstrate the potential for creating highly realistic and expressive talking face videos that capture the full range of human conversational behaviors.

The Attack Model. Advancements in this technology have introduced significant potential for malicious use. The primary misuse is impersonation, in which the adversary can synthesize videos depicting individuals articulating statements they never actually made. This is achieved by combining audio snippets with static images of the target individuals. By impersonating public figures, attackers can spread false information and statements, potentially resulting in large-scale manipulation of public discourse and sentiment. Additionally, these technologies can be used for extortion and deception targeting victims' families.

4) Art Style Imitation.: The field of art style generation has evolved significantly, from traditional neural style transfer (NST) techniques to advanced GenAI-driven. Contemporary approaches can replicate an artist's style using open-source txt2img models and a limited set of sample artworks [44]. Early attempts at style transfer methods relied on NST techniques, which combine the content of one image with the artistic style of another. Liu et al. [111] proposed the balance between style transfer and content preservation through multi-level feature alignment. Li et al. [112] introduced an efficient learnable transformation approach, enabling fast style transfer with flexible control. GANs have provided another approach to style transfer, namely image-to-image translation. This approach aims to learn mappings between image domains (e.g., translating a photo into a painting) while preserving the content, offering more generalized and flexible style transfer compared to traditional methods. Notable contributions in this area include two-stream network architecture by Jiang et al. [113] for multi-scale feature extraction, and the latent space encoding method by Richardson et al. [114] for diverse image translations. To address computational constraints, Huang et al. [115] employed unsupervised learning with a novel inversion method for embedding real images into StyleGAN2's latent space.

However, these methods face several challenges: (i) Lack of fine-grained control over the style transfer process; (ii) Depending on large style-specific datasets for training; (iii) Difficulties in preserving original content integrity. The introduction of CLIP enabled more precise and flexible style control through text prompts. Methods like StyleCLIP [116] and CLIPstyler [117] leveraged CLIP's text-image alignment capabilities for text-guided image manipulation and style transfer. The latest advancements in art style generation leverage diffusion-based txt2img generation models like Stable Diffusion [1] and DreamBooth [64]. These models have demonstrated remarkable abilities to generate images in specific artistic styles based on textual descriptions. Textual inversion [118], a technique implemented in many AI-for-Art applications, allows these models to learn new concepts or styles from a small set of example images, enabling fine-grained control over generated images. Particularly, Gal et al. [118] used textual inversion to learn new pseudo-words in the embedding space of a pre-trained txt2img model, which can represent specific user-provided concepts like objects or styles, while Ruiz et al. [64] enabled fine-grained control over generated images by allowing users to specify a unique identifier for the subject in text prompts, which the model learns to associate with the subject's key visual features.

The Attack Model. While these technological advancements offer unprecedented creative possibilities, they also present significant ethical and legal challenges. Malicious actors can now generate AI style mimicking specific artists' styles using as few as 20 unique artworks [44]. This capability raises concerns about copyright infringement and artist impersonation. The potential for generating works that closely imitate copyrighted styles of living artists poses a threat to intellectual property rights [119]. Furthermore, these tools could be misused to create forgeries or fake works attributed to specific artists, potentially damaging reputations and market values [13, 120].

IV. Deepfake Disruption Approaches

The fundamental objective of DF disruption approaches is to **disrupt the adversary's ability to generate DFs** by preventing the GenAI models from effectively extracting information from authentic human data. This concept draws inspiration from research on adversarial examples in image classification [30, 121]. The main idea for this protection is to introduce some tiny perturbations to images that render them *unrecognizable* to GenAI models while remaining *imperceptible* to human observers. In this context, *unrecognizable* implies that the perturbed image cannot be processed as a normal image by the GenAI models, thus restraining models from extracting image features, while *imperceptible* means that added perturbations do not significantly alter the original image semantics, thereby maintaining the image's utility for legitimate purposes. These perturbed images, referred to as protected examples, are designed to safeguard human images from exploitation by GenAI models for DF generation. By transferring original images into protected examples without introducing any noticeable visual artifacts, disruption approaches target that any GenAI models trained on these images fails to produce reasonable-quality images of the target subject (See Figure 6). A defense succeeds when the generated images satisfy one of the criteria: (1) low-quality due to extreme distortion or noticeable artifacts; (2) none or unrecognizable target subjects [122].

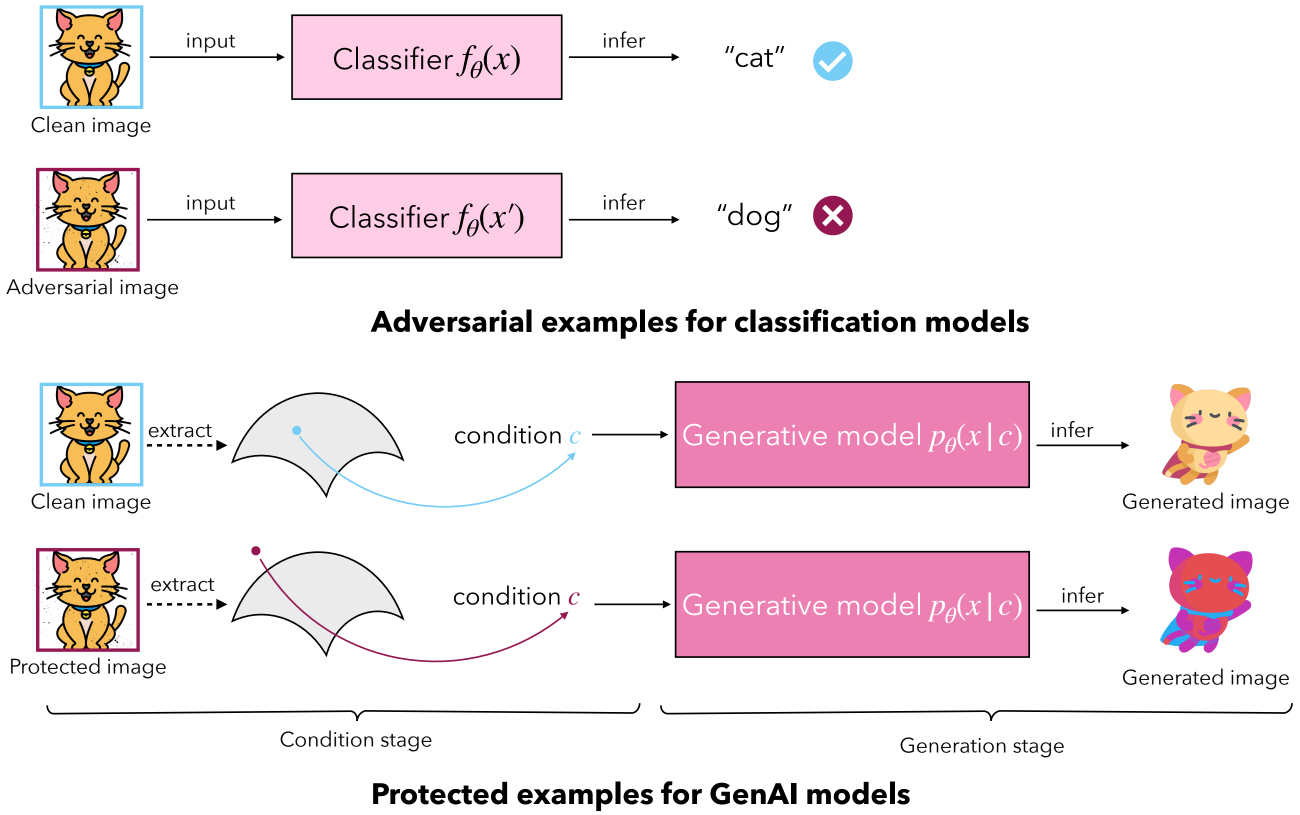

Fig. 5: Difference between adversarial examples and protected examples.

A. Targets of DF Disruption

Although the primary purpose of DF disruption approaches is to disrupt the generation process, creating protected examples for GenAI models is challenging [23]. Unlike classification models, GenAI-based DF generation models exploit input images by generating new images conditioned on the inputs rather than conducting an E2E inference on them. A successful protected example must prevent the extraction of specific features (e.g., identity, style, content) under identifiable conditions by the GenAI model. The Figure 5 illustrates the distinction between adversarial examples in classification models and protected examples in GenAI models. In the GenAI pipeline, an input image's features are extracted during the conditioning stage and subsequently used to generate new images based on these conditions. Protected examples work by misleading the feature extraction process in the conditioning stage, resulting in an out-of-distribution condition. Typically, the feature extraction process employs a feature extraction model (FEM) to capture essential characteristics from the inputs. An effective protected example would lead to poor-quality conditionally generated images. Therefore, to disrupt the generation process, it is crucial to target the FEM.

Based on four DF generation types defined in Section III, we can categorize the disruption targets into three main classes:

- **Attribution FEM**: Attribution FEM extracts features related to specific attributes, such as age, gender, hair color, or emotion. These modules often appear in face manipulation models that aim to modify specific facial attributes. Most manipulation models are conditioned on both the input sample and a target attribution class, allowing for controlled and targeted modifications. For example, in face reenactment, the Attribution FEM might focus on extracting expression-related features from a victim image, such as mouth movements, eye blinks, or eyebrow raises. These extracted features are then used to animate a source image, creating a DF image where the person in the source image appears to be making the expressions from the victim image. By adding imperceptible perturbations into the target images, one could prevent unauthorized manipulation of personal images or videos for DF creation.
- **Style FEM**: Style FEM is commonly used in text2img generation models to extract style-specific features from an artist's original work. These features correspond to the unique artistic style and are crucial for generating images that mimic particular artistic techniques or visual aesthetics. An instance of this would be in models like DALL-E[7] or Midjourney[8], where users can specify an artist's style for image generation. Disrupting the Style FEM could prevent the unauthorized replication of an artist's distinct style in generated images.
- **Facial FEM**: To generate convincing DFs that can impersonate a specific individual, it is necessary to employ an FEM that extracts the victim's identity information or facial characteristics (e.g., facial landmarks) during the generation process. This type of FEM is critical

[7] https://openai.com/index/dall-e-3/
[8] https://www.midjourney.com/home

in face-swapping or talking-face video applications. For example, in a face-swapping DF, the Facial FEM would extract identity from the victim's image. By disrupting this process, one could significantly impair the ability to create realistic impersonations.

B. Defense Scenarios

The effectiveness of disruption approaches can vary significantly depending on the level of information available about the target GenAI models and FEMs. To systematically analyze these approaches, we categorize the defense settings into four distinct scenarios, ranging from the most easiest to the most challenging:

White-box scenario. In this setting, disruption approaches operate under the assumption that complete knowledge about the GenAI models is publicly available, including architectures, hyperparameters, training objectives, and training datasets. This scenario provides the most advantageous conditions for developing disruption techniques, as defenders can leverage detailed knowledge of the model's inner workings to craft highly effective perturbations.

Black-box scenario. The black-box scenario presents a more realistic and challenging setting. In this case, disruption approaches have limited prior knowledge of the model's internal architecture and parameters. The primary source of information available to defenders is the synthetic images generated by the GenAI models.

Uncontrolled scenario. The uncontrolled scenario introduces an additional layer of complexity and represents the most challenging setting for DF disruption. In this advanced scenario. In this scenario, some of the user's clean images are assumed to have been leaked to the public and the adversary has access to a mix of protected and clean images of the target individual. This setting is particularly challenging because the adversary can utilize this mix of images to fine-tune GenAI models to generate reasonable personalized images [122].

C. Catalog of DeepFake Disruption Approaches

In this section, we provide different ways to categorize DF disruption approaches, based on their optimization problems and types of perturbation forms. These categorizations provide a structured framework for understanding and comparing different methods in the field.

Classifying According to Optimization Functions. DF disruption approaches can be distinguished based on the optimization functions used to generate perturbations. Two key categories emerge: **error-maximizing noise** and **error-minimizing noise** methods. Let $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ denote a dataset comprising n samples, where $x_i \in \mathcal{X}$ represents the feature vector of the i-th sample and $y_i \in \mathcal{Y}$ is the corresponding label. The objective of error-maximizing approaches is to maximize the FEM's error or loss function, causing the target FEM Φ to misclassify the label y. This can be formulated as:

$$\delta := \arg\max_{\delta} \mathcal{L}_\theta(\Phi(x+\delta), y), \text{ subject to } |\delta| \leq \epsilon \tag{13}$$

In contrast, error-minimizing methods generate protected examples by minimizing the model's error on inputs:

$$\delta := \arg\min_{\delta} \mathcal{L}_\theta(\Phi(x+\delta), y), \text{ subject to } |\delta| \leq \epsilon \tag{14}$$

The underlying principle is that a smaller loss of the protected examples may deceive the target FEM Φ into perceiving that less information can be learned from those examples. This approach attempts to create perturbations that make the protected examples appear highly confident and "easy" for the model Φ, potentially leading to overfitting or reduced feature extraction.

Classifying According to Perturbation Forms. DF disruption techniques can also be categorized based on the form of perturbations they generate. Two main types are identified: **image-specific perturbations** and **image-agnostic perturbations (person-specific perturbation)**. For n image samples, image-specific perturbations are $x'_i = x_i + \delta_i$, $\delta_i \in \Delta_s = \{\delta_1, \dots, \delta_n\}$. These perturbations are uniquely generated for each individual image. While potentially more precise, this approach may have limited practicality due to the computational overhead of generating a new perturbation for each image. Meanwhile, image-agnostic perturbations are defined as $x'_i = x_i + \delta_{y_i}$, $\delta_{y_i} \in \Delta_p = \{\delta_1, \dots, \delta_K\}$, where K denotes the number of persons. This approach offers two advantages: (i) The perturbation is generated once per person or class, allowing for faster processing of new images; (ii) Compared to image-specific perturbations which require multiple transmissions of new images between the user and the server, person-specific perturbations need only one transmission, potentially reducing the risk of privacy leakage during data transfer [123, 124].

Based on the above observation, we classify the DF disruption approaches based on the optimization functions, including error-maximizing perturbations and error-minimizing perturbations. In Section IV-D and IV-E, we primarily discuss disruption methodologies that aim to generate perturbations by solving the Equation 13 and 14, respectively. Meanwhile, research works that improve imperceptibility of protected images, transferability across target GenAI models, universality of perturbations, and robustness of perturbations against transformations are discussed in Section IV-F. Figure 6 describes the overview of disruption approaches to prevent the GenAI models from exploiting users' data and Table II summarizes disruption approaches.

D. Error-maximizing Perturbation Approaches

Error-maximizing perturbation approaches can be categorized into two main groups based on the optimization techniques employed to solve the equation 13: Optimization-based and Generative-model-based methods.

1) Optimization-based Methods.: These methods typically model the generation of protected images as optimization objectives and employ various techniques to maximize the error in DF generation processes. The approach varies depending on the defense scenario.

In *white-box* settings, several gradient-based methods have shown promise. One widely adopted technique is the Iterative-Fast Gradient Sign Method (I-FGSM) [30], which iteratively

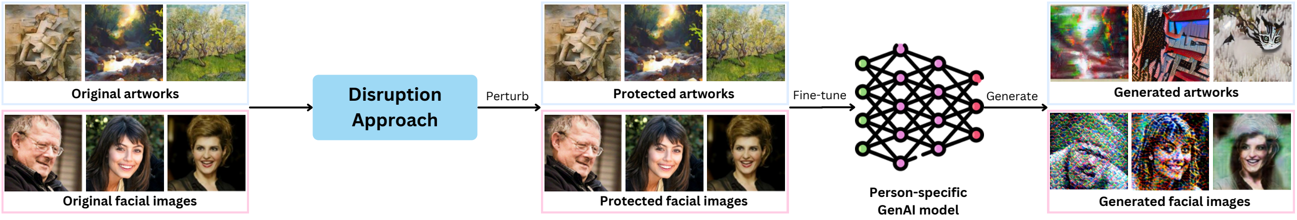

Fig. 6: Overview of disruption approach.

TABLE II: A Summarization of DF Disruption Approaches

Categories		Article	Key Idea	Threat Model			Beyond Disruption			Metrics***			Robustness	
				Target	Knowl.*	Arch.**	I	II	III	S	P	Q	R	PP
Error-maximizing	GAN-based	[125]	GAN from scratch	★	□	G	✔	✘	✘	PSR	L_2	–	✘	✔
		[39]	Dual-GAN from scratch	☆	■	G	✔	✘	✘	ACC	SSIM, FSIM	–	✘	✘
		[46]	Two-stage training	★	□	G	✔	✘	✘	PSR	PSNR, SSIM	–	✘	✔
		[126]	Info-contain. encoder	☆	□	G	✔	✘	✘	PSR	PSNR, SSIM	–	✘	✔
	Optimization-based	[127]	I-FGSM	☆	□	G	✔	✘	✔	NME	SSIM	–	✘	✔
		[128]	I-FGSM + transformation	★	□	G	✔	✔	✘	AIH	–	–	✘	✘
		[129]	PGD	★	□	G	✔	✔	✘	PSR	–	–	✘	✔
		[130]	PGD	★	□	G	✔	✘	✘	PSR	–	–	✘	✘
		[122]	PGD + two-stage training	☆	□	DM	✔	✔	✘	PSR	FDFR, ISM, SER-FQA, BRISQUE	–	✔	✔
		[131]	PGD + multi-task loss	★	□	G	✔	✘	✘	ACC	PSNR, SSIM, L_2	–	✘	✘
		[132]	PDG + Wasserstein distance	★	□	G	✔	✔	✘	PSR	PSNR, SSIM, FID	–	✘	✘
		[133]	Meta-learning	☆	□	DM	✔	✔	✘	–	SDS, IMS, LIQE	–	✘	✔
		[134]	LAB color space	★	■	G	✔	✔	✘	PSR	PSNR, SSIM	–	✘	✔
		[135]	PGD + gradient sliding	★	■	G	✔	✘	✘	PSR	–	QC	✘	✔
		[136]	PCA	★	■	G	✔	✔	✘	PSR	FID	QC	✘	✔
		[137]	Hard model mining	★, ☆	□	G	✔	✔	✘	PSR	L_2	–	✘	✔
		[138]	Compress GAN + AoA alg.	☆ , ★	□	G	✔	✔	✘	PSR	SSIM, PSNR	–	✘	✔
		[44]	Penalty method	✡	□	G, DM	✔	✘	✘	PSR	–	–	✘	✔
		[23]	Monte Carlo estimation + I-FGSM	✡	□	DM	✔	✘	✘	–	FID	–	✘	✔
		[139]	Two-level perturb. fusion + TPE	★	□	G	✔	✔	✘	–	FID	–	✘	✘
Error-minimizing	Optimization-based	[123]	Min-min optimization	☆	□	G	✔	✔	✔	ACC	–	–	✔	✔
		[124]	Min-min-max optimization	☆	□	G	✔	✔	✔	ACC	–	–	✘	✔

☆ Facial FEM ★ Attribution FEM ✡ Style FEM

* This column is the protector's knowledge of target DF generation models. □: white-box, ■: black-box.

** This column indicates the architecture of GenAI models.

*** This column is metrics used to evaluate the performance of methods. S: Protection Success Rate, P: Perceptual Metric, Q: Query numbers, QC: The number of queries.

Beyond Disruption column includes I: Imperceptibility, II: Transferability, III: Universality.

Robustness column denotes the evaluation of the robustness of generated perturbations. R: Perturbation Removing, PP: Post-processing Operations.

updates protected examples by adding small perturbations in the direction of the sign of the loss gradient: $x' = x + \lambda sign(\nabla_x \mathcal{L}_\theta(x, y))$. Sun et al. [127] applied I-FGSM to generate perturbations that disrupt facial landmark extraction, thereby inducing the misalignment in DF face-swapped images. To protect artists from being mimicked by GenAI models, Shan et al. [44] utilized this method to shift artwork representations in the model's feature space towards a chosen target style. This causes models trained on the protected images to generate art in a different style than intended. Liang et al. [23] combined I-FGSM with Monte Carlo estimation for gradients. This method optimizes perturbations across different latent variables sampled from the reverse process of DMs, addressing the challenges posed by the stochastic nature of DMs. Yang et al. [128] enhanced the robustness of generated perturbations by incorporating differentiable random image transformations during the optimization process. Another commonly used technique is Projected Gradient Descent (PGD) [121], which performs gradient-based steps followed by a projection operation to maintain visual similarity between the protected example and the original input: $x_{i+1} = \text{proj}_x(x_i + \lambda sign(\nabla \mathcal{L}_\theta(x, y)))$. Ruiz et al. [129] and Yeh et al. [130] proposed Nullifying and Distoring Attacks to disrupt face manipulation processes.

Recent research has also focused on developing *black-box* approaches that address scenarios where knowledge of target DF generation models is limited or unavailable. Yeh et al. [135] transformed the query space from a unit hypersphere to a scaled hyperellipsoid to improve the query efficiency. However, existing black-box disruption approaches face two primary limitations: (i) The vulnerability of naive adversarial attacks to image restoration techniques [140]; (ii) The need for large amounts of queries to build surrogate models. [136]. Wang et al. [134] addressed the first limitation by generating perturbations in the LAB color space instead of RGB, producing more natural and robust perturbations. Addressing the second limitation, Ruiz et al. [136] leveraged principal component analysis (PCA) to reduce the high-dimensional space of possible perturbations to a lower-dimensional space of principal components, potentially decreasing the number of required queries.

Another work has focused on generating perturbations for real-world online social network (OSN) scenarios in which lossy compression techniques are often applied to reduce the transmission bandwidth and storage space during the image uploads [141]. Qu et al. [138] designed a Compression Approximation GAN (ComGAN) to mimic the perturbation destruction behavior of OSNs and introduced a destruction-aware constraint that guides the ComGAN to learn this behavior. The ComGAN is then integrated into the surrogate model to generate the perturbations that remain effective after compression.

2) Generative-model-based methods.: Generative-model-based methods, on the other hand, generate perturbations using GenAI models. Here, the generator $\mathcal{G}$ takes the original sample x as its input and generates a perturbation $x' = \mathcal{G}(x)$. Then x' is sent to the discriminator $\mathcal{D}$ to distinguish between x and x'. To enable the $\mathcal{G}$ to generate a perturbation that can fool the target model, these approaches often incorporate an adversarial loss into the GAN's loss which represents the distance between the prediction and the target class y_{target}. The loss function can be defined as follows:

$$\min_{G} \max_{D} [\underbrace{\mathcal{L}_{\text{GAN}}(\mathcal{G}, \mathcal{D})}_{\text{GAN loss}} + + \underbrace{\mathcal{L}_{\text{adv}}(x', y_{\text{adv}})}_{\text{Adversarial loss}}], \tag{15}$$

Wang et al. [125] optimized variables in the latent space to produce minimal perturbations in salient facial regions [125]. Dong et al. [39] constructed a framework that includes two generators and two discriminator arranged in a cycle-consistent structure. The first generator is designed to produce adversarial perturbations that can be added to clean images, while the second generator serves the opposite function of removing these perturbations. This bidirectional approach enables the model to learn a reversible transformation between the natural and adversarial image domains. Instead of naively employing a GenAI model to generate perturbations, Huang et al. [46] proposed a two-stage training framework, in which a surrogate model is first trained to imitate the target manipulation model's behavior, and then a perturbation generator is trained to create imperceptible perturbations that disrupt the attribution manipulation process. Zhu et al. [126] designed an encoder to map the original image and a corresponding unique binary code to an protected example that is visually similar to the original image but can disrupt facial manipulation systems.

E. Error-minimizing Perturbation Generation Approaches

Error-minimizing perturbation generation approaches include two-step process. First, given a FEM Φ, the error-minimizing perturbation δ is generated for the input x by solving the bi-level optimization problem:

$$\min_{\theta} \mathbb{E}[\min_{\delta} \mathcal{L}(\Phi(x + \delta), y)] \quad \text{such that} \quad ||\delta||_p \leq \epsilon, \tag{16}$$

where y is the corresponding label. Compared to adversarial training, this formulation represents a min-min bi-level optimization problem. The inner minimization identifies the L_p-norm bounded noise, while the outer minimization optimizes the parameters θ. Subsequently, error-minimizing perturbation is applied to the original input: $x' = x + \arg\min \mathcal{L}(\Phi(x+\delta), y)$.

The concept of error-minimizing perturbation is initially introduced by Huang et al. [123] and later extended by Fu et al. [124] to improve the robustness of generated perturbations. Fu et al. addressed a key limitation of the this approach: its vulnerability to adversarial training. This vulnerability stems from several factors: (i) Error-minimizing perturbations are optimized for standard training loss, which differs from the loss landscape in adversarial training; (ii) Adversarial training produces models less sensitive to small perturbations, potentially negating the effect of error-minimizing perturbations; (iii) Adversarial training makes models more tolerant to higher loss values, diminishing the impact of error-minimizing perturbations. To address these issues, the authors proposed *robust error-minimizing perturbations* generated by solving a min-min-max optimization problem:

$$\min_{\theta} \mathbb{E}[\min_{|\delta^u| \leq \alpha_u} \max_{|\delta^a| \leq \alpha_a} \mathcal{L}(\Phi_{\theta}(x + \delta^u + \delta^a), y)], \tag{17}$$

where α_u ensures the imperceptibility of the generated perturbation, α_a controls the protection level of the perturbation against adversarial training. The underlying principle in Equation 17 is to optimize δ_i^u such that the adversarial example crafted from the protected sample $(x_i + \delta_i^u)$ do not substantially increase the training loss, making it more resilient against various DF generation models.

F. Beyond Disruption

Recent advancements in DF disruption approaches have expanded beyond the primary goal of generating effective perturbations to disrupt GenAI models. In this section, we explore advanced techniques that improve four critical properties: *imperceptibility*, *transferability*, *universality*, and *robustness*.

1) Imperceptibility.: This is a crucial aspect that ensures the applied perturbations remain undetectable by human, preserving the imperceptibility of the original data (e.g., images' natural appearance) and maintaining their utility for legitimate purposes (e.g., sharing on social media).

Norm-constrained perturbation. This is the common technique, which limits the magnitude of perturbations added to the original image using specific norm balls, L_∞, L_2 or L_0 norms. Mathematically, this approach can be expressed as $||\delta||_p \leq \epsilon$, where $|| \cdot ||_p$ denotes the L_p norm. During optimization, perturbations exceeding this constraint are projected back onto the ϵ-ball, often through clipping operations. This technique has been widely adopted in various studies [130, 129, 128, 46, 135, 23, 122, 123, 124, 39, 132, 127, 131, 138] to generate imperceptible perturbations.

Perceptual similarity loss. However, recognizing that L_p norm-based perturbations may not align well with human perception of image similarity [142], researchers have explored alternative methods. One such alternative is the use of perceptual similarity loss, such as Learned Perceptual Image Patch Similarity (LPIPS) loss [143]. LPIPS leverages feature representations learned by convolutional neural networks (CNNs) trained on image recognition tasks, capturing high-level semantic information that is more relevant to human perception than pixel-level differences.

Generative model constraint. This is an additional approach that utilizes the capabilities of generative models to create imperceptible perturbations [39, 125, 46, 126]. Instead of directly adding perturbation to an sample, this approach leverages models trained on original samples to generate perturbations that more closely resemble natural sample variations, thereby enhancing imperceptibility. The approach often incorporates perceptual loss functions into the generative models' objective function. For instance, in a GAN framework, the optimization problem can be formulated as:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} [\underbrace{\mathcal{L}_{\text{GAN}}(\mathcal{G}, \mathcal{D})}_{\text{GAN loss}} + \underbrace{\mathcal{L}_{\text{perceptual}}(x, x')}_{\text{Perceptual loss}} + \underbrace{\mathcal{L}_{\text{task}}(x', y_{\text{adv}})}_{\text{Adversarial loss}}], \tag{18}$$

where $\mathcal{L}_{\text{perceptual}} = \mathbb{E} \max(0, ||\mathcal{G}(x)|| - \epsilon)$. This formulation enables the generation of perturbations that are both visually imperceptible and effective in achieving the desired disruption objective.

2) Transferability.: Transferability refers to the ability of generated perturbations to disrupt multiple target models without prior knowledge of their architectures or parameters.

Input Diversity Strategy. This strategy is a fundamental approach [144], which applies random transformations to the input images at each iteration to create diverse input patterns. This strategy is applied by [127, 128] to improve the transferability of perturbations.

Ensemble Training Strategy: Ensemble training strategy has emerged as a powerful technique to enhance the transferability of protected samples across various DF generation models. This approach involves generating perturbations that can deceive multiple models simultaneously, rather than focusing on a single target. The underlying principle is that by optimizing against diverse models, we can approximate the decision boundaries of potential, inaccessible target models. This strategy increases the likelihood of exploiting common vulnerabilities across various model types, rather than overfitting to the specific characteristics of a single model. Mathematically, the ensembling disruption objective can be formulated as:

$$\max_{x'} \frac{1}{N} \sum_{i=1}^{N} ||\Phi_i(x) - \Phi_i(x')||_2^2, \tag{19}$$

where Φ_i denotes the i^{th} model in the ensemble and N is the number of models and $||.||_2^2$ denotes the squared L2 norm.

While some works [129, 134, 122, 132, 145, 128] simply incorporate the ensemble training strategy into the optimization process, other works introduced additional techniques to enhance transferability. Huang et al. [139] applied a two-level perturbation fusion strategy, consisting of image-level fusion and model-level fusion, to combine perturbations from different images and models. Image-level fusion averages gradients across a batch of images to generate a perturbation that focuses on common facial attributes, while model-level fusion iteratively combines the averaged perturbations from multiple models into a single universal perturbation. The method also incorporates an automatic step size tuning algorithm based on Tree-structured Parzen Estimator to find suitable optimization step sizes for different models, which helps balance the optimization directions across multiple models. Tang et al. [132] employed PGD to optimize a perturbation that maximizes the Wasserstein distance between clean and perturbed features. Compared to MSE, Wasserstein distance can measure the distance between two distributions that do not overlap, potentially leading to more robust perturbations. Liu et al. [133] utilized meta-learning to generate modal-agnostic perturbations. This approach treats the crafted perturbations as ”meta-parameters” that need to generalize across different ”tasks,” where each task corresponds to fooling a specific model at a certain training stage. Van Le et al. [122] developed a two-stage training framework to defend against txt2img synthesis. Their method first trains a surrogate model to imitate the target model's behavior, then optimizes perturbations using PGD on this surrogate model. Guan et al. [137] addressed the issue of perturbation overfitting to more vulnerable models in simple ensemble methods. They designed the hard model mining strategy, which focuses on defending the hardest DF model at each iteration.

3) Universality.: Recent research has also focused on generating person-specific perturbations, known as universal perturbations, which can be consistently applied across multiple data of the same individual. This approach offers advantages in terms of efficiency and scalability compared to sample-specific perturbations. This type of perturbation can be optimized over the entire training dataset or a subset thereof. The main challenge in generating person-specific perturbations lies in accounting for natural variations in data, including differences in poses, illuminations, expressions, and occlusions. Some works [123, 124] proposed a straightforward method for generating universal perturbations for each user by optimizing over the available samples of that user. This approach is formulated as:

$$\delta^* = \max_{\delta} \sum_{i=1}^{n} \sum_{j=1}^{n} d(\Phi_\theta(x_i + \delta), \Phi_\theta(x_j)) \tag{20}$$

4) Robustness.: Robustness is one of crucial factors to preserve the resilience of perturbations against data transformations such as Gaussian filtering. Recent approaches can be classified into: Transformation-based and Noise simulating strategies.

Transformation-based Strategy. Liu et al. [133] incorporated a transformation sampling process when crafting perturbations by using the Expectation over Transformation (EOT) technique [146]. This technique is also applied by Fu et al. [124] to enhance the stability of the generated perturbation against data augmentation.

Noise Simulating Strategy. Those approaches aim to mimic the noise introduced by real-world environments, such as OSN platforms. Wu et al. [141] used a UNet with residual learning and a differentiable JPEG layer $\min_\theta \mathcal{L}_r(J_q(x_i + \Phi_\theta(x_i)), \text{OSN}(x_i))$, where $\mathcal{L}_r$ is a reconstruction loss, and J_q is a differentiable JPEG layer with quality factor q. Qu et al. [138] designed a GAN-based compression approximation network to explicitly model complex OSN compression. A destruction-aware constraint is introduced during the training process to enable the network to learn the ability to destroy perturbations, which is similar to real OSN compression.

V. Watermarking Approaches

Watermarking approaches have emerged as a promising proactive strategy for DF mitigation. These approaches fundamentally modify the GenAI models to embed an identifiable message or signature within the generated content, which can subsequently be extracted and verified using specialized extraction mechanisms. This section explores the objectives, formal definition, key properties, and current watermarking approaches in the context of DF mitigation. In the context of this discussion, outputs produced using watermarking approaches are referred to as *watermarked samples*.

Watermarking approaches serve two primary objectives:

1) **Detection**: whether an image was generated by a specific GenAI model. This involves embedding a random signature into all generated images and subsequently verifying its presence. Attribution is crucial for distinguishing between authentic and synthetic content, particularly in media forensics and content verification scenarios.
2) **Identification**: In the context of MLaaS deployments, watermarking enables tracing generated images back to their source, identifying which specific user generated the image. This capability is crucial for mitigating scenarios where users might exploit GenAI models through APIs to generate DFs for impersonation or misinformation dissemination. Identification can aid in establishing accountability and deterring malicious use of GenAI models.

Formally, a watermark can be defined as a k-bit string $B = \{b_1, b_2, \dots, b_k\} \in \{0,1\}^k$, where k denotes the watermark length, and each b_i as a binary digit (0 or 1). To determine generated images having the embedded messages, a specialized watermark extractor is employed. The watermark extraction process varies based on the objective. (1) **Detection Scenario**: A random signature B is embedded into a synthetic image x' by the GenAI model G. The watermark extractor extracts a watermark B' from x' and compares B and B' using a similarity measure $s = S(B, B')$. For a predetermined threshold λ, if $s < \lambda$, x' is classified as being generated by G. (2) **Attribution Scenario**: When a user u requests to use model G, the provider generates a unique signature B_u for u and embeds it into the model distributed to u. This enables the provider to trace any misuse, such as DF generation, back to the specific user by comparing the extracted message to the user's signature. The choice of similarity measure S and threshold λ is critical and often involves careful calibration to balance false positive and false negative rates. Figure 7 illustrates the process of two scenarios.

To effectively combat DFs, embedded watermarks should exhibit properties: (1) *Undetectability*, which is the ability to embed long bit strings into synthetic images that remain undetectable; (2) *Imperceptibility*, which means the watermark should not introduce perceptible artifacts or significantly alter the visual quality of the synthetic image; (3) *Robustness*, which means the watermark should be robust against common image processing operations and transformations, remaining extractable even after manipulations such as compression, resizing, blurring, or noise adding.

Note that watermarking approaches for DF detection are significantly different from two related areas of research: Synthetic image attribution [147, 148] and Model weight watermarking [149, 150]. The former attributes synthetic images to their source without altering the generator's parameters or embedding any additional information during the generation process, while the latter watermarks the weights of entire models, which is primarily concerned with protecting the intellectual property of the models themselves. Note that existing watermarking approaches are only employed in the image domain.

A. Catalog of Watermarking Approaches

We present two primary classification schemes: one based on the intended defense scenario and another based on the watermark embedding strategy.

Classifying According to Defense Scenarios. Watermarking approaches can be classified into two main categories based on their defensive use cases: **Detection** and **Identification** [45]. In the detection scenario, watermarks are utilized to determine whether an image was generated by authorized GenAI models. The detection process involves extracting a bit string B' from a given image x and comparing it to the pre-defined one B. The detection test is defined as:

$$\begin{cases} \mathcal{M}(B, B') < \lambda, & \text{authourized synthetic image} \\ \mathcal{M}(B, B') \geq \lambda, & \text{authentic image,} \end{cases} \tag{21}$$

where $\mathcal{M}(B, B')$ denotes the number of matching bits between the extracted watermark and the pre-defined one, and $\lambda \in \{0, \dots, k\}$ is a threshold. Regarding the identification scenario, the purpose is to trace the specific GenAI model that generated an image. Each model model i is assigned a unique watermark b_i drawn randomly from $\{0,1\}^k$. The identification process involves comparing the extracted watermark B' to all pre-defined watermarks $\{B_1, \dots, B_N\}$, where N is the number of models. The image is attributed to the model with the highest matching score:

$$\text{argmax}_{i=1\dots N} \mathcal{M}(B_i, B') \tag{22}$$

Classifying According to Watermarking Strategies. Based on strategies to embed watermarks into images generated by GenAI models, existing methods can be categorized into **Training-based** and **Training-free** approaches. Training-based approaches integrate watermarks directly into the GenAI model's training or fine-tuning process. Such approaches typically involve modifying the model's architecture, loss function, or training data to incorporate watermarking objectives. Meanwhile, the second category applied watermarks to pre-trained generators without requiring additional training or extensive fine-tuning.

In this survey, categorize watermarking approaches based on watermarking strategies. Particularly, the subsequent sections provide a more detailed exploration of these watermarking strategies, particularly in Section V-B and V-C. Additionally, Section V-D discusses existing research that aims to improve the imperceptibility and robustness of watermarks. Table III summarizes watermarking approaches.

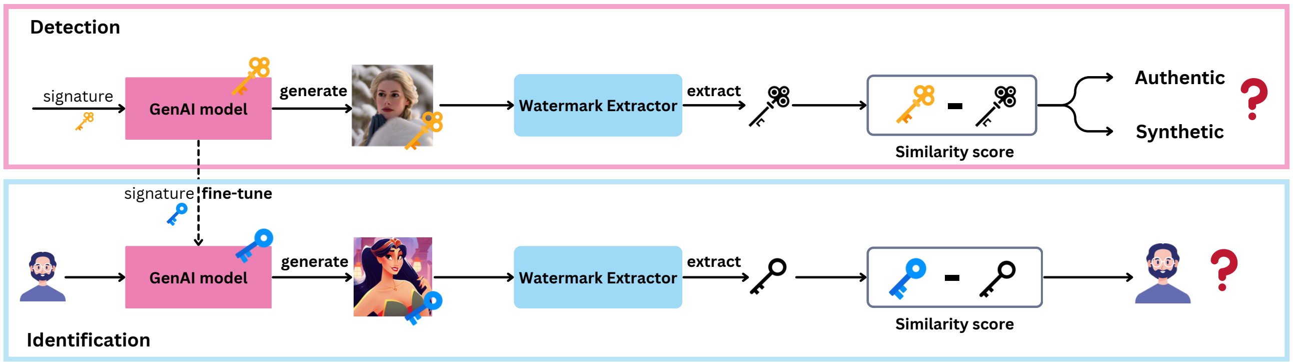

Fig. 7: Overview of watermark extraction process.

B. Training-based Approaches.

Training-based approaches require training the generator $\mathcal{G}$ from scratch or fine-tuning the generator to embed a watermark. In general, these approaches follow a two-stage framework:

1) **Training the Watermarking Encoder-Decoder**: This stage involves training an encoder-decoder network (EDN) to embed and extract watermarks from the training data. This process optimizes the parameters θ and ϕ of the watermark encoder E and watermark decoder D, respectively, to embed k-bit messages into images. The training typically minimizes an objective function that balances the mean squared error (MSE) loss for watermark invisibility and the binary cross-entropy (BCE) loss for bit-wise classification of watermarked message:

$$\min_{\theta,\phi} \mathbb{E}[\mathcal{L}_{\text{MSE}}(x, E_\theta(x, B)) + \mathcal{L}_{\text{BCE}}(D_\phi(E_\theta(x, B)), B)], \tag{23}$$

2) **Training the Watermarking Generative Model**: In this stage, there are two approaches for creating a watermarked generative model: (1) using the watermarking encoder E to create an entirely watermarked dataset and re-train the generative models from scratch; (2) incorporating E into $\mathcal{G}$ with an auxiliary loss function. the GenAI model is trained or fine-tuned using either a watermarked dataset or an augmented loss function that incorporates a watermark-related term supervised by the pre-trained decoder from the first stage.

This approach ensures that the watermarking capability is deeply ingrained in the generative model's parameters, resulting in consistent watermark embedding across all generated outputs. The process leverages the model's learning capacity to balance the dual objectives of realistic image generation and reliable watermark integration.

Several studies have contributed to the development and refinement of this framework. Yu et al. [151] first introduced this two-stage framework for embedding watermarks into GAN-generated images. After training the EDN, the trained encoder is used to create a watermarked dataset which is subsequently used to train the GANs model. However, as the GANs model learns to approximate the distribution of watermarked images, it might lead to watermark degradation in the generated outputs. Fei et al. [152] addressed this limitation by integrating the BCE loss into the GAN's loss that provides a supervised signal to directly train GAN from scratch for watermark embedding task. Yu et al. [153] further improved scalability by incorporating random sampling of watermarks during training, allowing for efficient generation of uniquely watermarked versions of the generator. Recognizing the computational intensity of retraining entire generators, Lukas and Kerschbaum [71] proposed a novel method that enables watermarking pretrained generators without access to training data. Specifically, the authors utilized a pivotal tuning approach to create a copy of the pre-trained generator and optimize it to embed watermarks while preserving output fidelity. Fernandez et al. [45] also applied the same idea for Latent Diffusion Models (LDMs) by fine-tuning the latent decoder based on a pre-trained watermark encoder. To embed watermarks in LDMs, Fernandez et al. focused on modifying only the decoder D of LDMs to incorporate a fixed watermark B, and then fine-tuned it to produce images that contain a specified watermark.

C. Training-free Approaches

Recent developments have introduced a novel category of approaches that can embed watermarks into generated images without additional training or fine-tuning. A notable contribution in this direction is the work by Wen et al. [154], which introduces a technique to watermark outputs of DMs by influencing the entire sampling process. Instead of modifying the pixel space, it applies the watermark to the Fourier transform of the random noise vector before the diffusion process. The use of the Fourier space offers several benefits: (1) Invariance to various transformations (e.g., flips, rotations); (2) More precise control over which frequency components are modified, potentially making the watermark less perceptible in the final image. These methods offer significant advantages in terms of flexibility and efficiency, as they can be readily applied to existing models without the need for retraining or architectural modifications.

D. Beyond Watermarking.

This section reviews research aimed at enhancing the imperceptibility and robustness of watermarks in GenAI models. These properties are crucial for maintaining the integrity and utility of generated content while ensuring effective attribution.

TABLE III: A Summarization of Watermarking Approaches

Categories	Article	Key Idea	Arch.*	Beyond **		Bits◇	Metrics***		Robustness		
				I	II		B	P	N	PP	A
Training-based	[151]	Two-stage + train from scratch	G	✔	✘	100	BACC	FID	✔	✔	✘
	[152]	Two-stage + train from scratch	G	✔	✔	50-100	BACC	PSNR, FID, SSIM	✔	✔	✘
	[153]	Train from scratch + molulate convol.	G	✔	✔	128	BACC, ACC	FID, FPR	✔	✔	✘
	[45]	Two-stage + fine-tune	DM	✔	✔	48	BACC, TPR, FPR	PSNR, FID, SSIM	✔	✔	✔
	[71]	Two-stage + fine-tune	G	✔	✔	40	BACC, ACC	FID	✘	✔	✔
Free-training	[154]	Fourier-space noise	DM	✔	✔	40	AUC, FPR	FID	✘	✔	✘

✘ No ✔ Yes '−' No information
* This column denotes the architecture of GenAI model.
** This column is critical properties of DF Disruption Approaches. I: Imperceptibility, II: Robustness.
*** This column is metrics used to evaluate the performance of methods. B: Bit Accuracy, P: Perceptual Metric.
Robustness column indicates the evaluation of the robustness of generated perturbations. N: Perturbation, PP: Post-processing Techniques, A: Attacks.
◇ The length of watermark

Imperceptibility. Imperceptibility ensures that watermarked samples remain indistinguishable from their non-watermarked counterparts. Initial research in this field, such as that conducted by [151, 152], focused on integrating the watermarking process directly into the generation pipeline of GenAI models. These methods leverage the inherent capabilities of GenAI models to learn and reproduce complex data distributions, effectively incorporating watermarking as an integral component of the image generation process. Subsequent studies incorporated perceptual loss terms into BCE or generator loss functions, employing various perceptual metrics to control distortion between watermarked and non-watermarked samples [45, 71]. Recent advancements have introduced innovative techniques further to ensure the imperceptibility of watermarks in generated data. Yu et al. [153] implemented a consistency loss to ensure that images generated with identical latent codes but different watermarks maintain consistent appearances. Wen et al. [154] embedded the watermark pattern into the Fourier transform of the initial noise vector used in the diffusion process, rather than directly modifying the final generated image. They also designed watermark patterns statistically similar to Gaussian noise to minimize distribution shifts that could potentially impact generation quality.

Robustness. Robust watermarks ensure the embedded message remains detectable even after various manipulations or attacks, including (i) Geometric transformations: rotation, scaling, cropping, and translation; (ii) Processing operations: JPEG compression and Gaussian blur; (iii) Color manipulation: brightness and contrast adjustments, color jitter; (iv) Advanced attacks: overwriting and adversarial perturbations [71]. While these operations can significantly modify an image's appearance, they often preserve its overall semantic content. However, they pose a significant threat to embedded watermarks, potentially distorting or removing them entirely. Researchers have developed several strategies to enhance watermark robustness. One of the more straightforward methods involves training on augmented data. [151, 45]. This approach applies various image transformations during the training of the watermark encoder-decoder network, improving the invariance of watermarked images to these operations. Another strategy proposed by Fei et al. [152], which incorporates a processing layer directly into the GAN training pipeline, applying various image transformations to generated images before watermark decoding. This enables the model to learn more robust watermarking strategies that can withstand common image manipulations. An innovative approach recently introduced involves embedding watermark patterns in the Fourier space of the initial noise vector used in diffusion models [154]. By placing the watermark in a circular mask centered on low-frequency modes, they leveraged the inherent resilience of low-frequency information to operations like JPEG compression and blurring. The robustness property is essential for maintaining the integrity of the watermarking system and its effectiveness in attributing generated images to their source.

Discussion. The emergence of watermarking techniques as a strategy for combating the misuse of AI-generated content while simultaneously fostering the responsible development and application of GenAI models. Training-based watermarking approaches integrate watermarks directly into the generative process, potentially making them more difficult to remove without degrading image quality. However, they suffer from high computational costs and scalability issues, particularly for large-scale models. On the other hand, training-free methods provide efficiency and flexibility, allowing for easy implementation across various pre-trained models without modification. However, the primary limitation of this approach is the lack of comprehensive evaluation. With only a single study currently available in this category, there exists a significant gap in our understanding of its effectiveness, robustness, and adaptability across diverse scenarios and model architectures.

The field of watermarking is in its early stages, with ongoing research efforts primarily focused on striking an optimal balance between imperceptibility, robustness, and computational efficiency. Future work should focus on expanding the scope and depth of evaluation for both training-based and free-training watermarking approaches across a diverse range of GenAI models and potential attack scenarios.

VI. Threat Model

In this section, we define the potential threats that could compromise the security ofproactive DF detectors. This model is crucial for understanding the vulnerabilities of detectors and developing evaluation strategies to assess the detectors. Additionally, we also define the detector's capabilities to combat DFs. We first provide the general threat model for DF detectors, and then define the particular threat models for each approach category.

Threat models against DF detectors can be defined by understanding the following components:

- **Adversary's Objective**: The primary objective of adversaries is to compromise the defense capability of proactive approaches against DFs. For each type of proactive DF approach, there are specific adversary's objectives that are defined later.

- **Adversary's Knowledge**: The knowledge of adversaries about the targeted detector may vary significantly. Such knowledge may include:
  - Architecture, parameters, and objective functions of ML/DL models to build the DF detector;
  - Training dataset;
  - Feature representation of each sample;
  - Form of the decision function;
  - Type of detector's output; e.g., class probabilities, scores, or top-k labels.

  Based on different levels of the adversary's knowledge of the attacked detector system, there are two attack scenarios: (a) **Black-box** attack: The adversary has no internal information about the target detector and can only observe its outputs; (b) **White-box** attack: The adversary has full knowledge of the target detector, including its architecture, parameters, and training dataset. This extensive knowledge enables them to craft highly targeted and potentially more effective attacks against the detection system.
- **Adversary's Capability**: This aspect outlines the actions an adversary can take to compromise the detector. Variations in the adversary's power may include:
  - manipulate training dataset;
  - manipulate test dataset
  - manipulate substitute model to mimic the behavior of the target detector.
- **Adversary's Strategy**: The adversary can implement different strategies to achieve his objective, which are discussed later in Section VI-A and VI-B.

A. Threat Model for Disruption Approaches.

This threat model describes the security landscape for approaches designed to protect individuals' data from being exploited or replicated by GenAI models. The defender is individuals (e.g., artists, photographers, composers, or general users) who seek to protect their images from unauthorized use or exploitation. The entity providing the disruption system or application is referred to as the provider. The provider acts as an intermediary, implementing protective measures on the defender's images to disrupt their use in training GenAI models. The purpose of the adversary is to collect users' data and leverage them for DF generation using GenAI models.

Adversary's Objectives. There are three primary goals of adversaries. **Protected Data Inversion**: Removing perturbations from protected samples to recover the original content, with the assumption of accessing to some unprotected samples from alternative sources. **Sensitive Information Extraction**: Exploiting the users' sensitive information (e.g., identity, artistic style) from protected data, operating under the constraint of no access to unprotected images. **Model Extraction**: Developing techniques to reconstruct or approximate the disruption model/ algorithm by building a surrogate model.

Adversary's Knowledge. Based on the level of knowledge about the disruption system, two adversaries can be considered. The first one is **black-box** adversary, who has limited knowledge of the specific disruption model and its parameters. This adversary relies on observing protected images and may query the publicly available disruption system to gather information. Meanwhile, the **white-box** adversary has full knowledge of the disruption model, including its architecture, parameters, and training dataset.

Adversary's Capabilities and Strategies. Adversaries may employ various strategies depending on their knowledge level, including (1) collecting a large number of unprotected and protected images to fine-tune GenAI models; (2) applying image transformations (e.g., denoising, compression) to attempt the removal of perturbations in protected images; (iii) building a surrogate model to extract or approximate the disruption model.

Defender and Protector's Capabilities. The defender has access to personal computing devices (e.g., laptops) to run the disruption model and add perturbations to their images before sharing them online. Meanwhile, the protector has full access to their own disruption model, including architectures, parameters, and training dataset. The protector's access to users' original images may vary based on the deployment method (server-side or client-side).

B. Threat Model for Watermarking Approaches.

In the context of watermarking approaches for DF detection, the GenAI model provider assumes the role of the defender. The primary purpose of the adversary is to evade the detection process of the watermarking detector.

Adversary's Objectives. Three common goals of adversaries: **Evasion**. Creating DFs that are misclassified as authorized content by deceiving the watermarking detector; **Watermark Extraction**. Stealing the embedded watermarks in order to produce DFs that appear to come from an authorized source; **Watermark Removal**. Removing or corrupting watermarks embedded into synthetic images.

Adversary's Knowledge. The adversary's knowledge about the defense system includes the target GenAI model and the watermarking detector. Note that the secret watermarking key is maintained with strict confidentiality. We consider two types of adversaries based on their level of access to the defense system. The first adversary, the **black-box** adversary, has access to the target generator through APIs. This means he can query the target GenAI model but lacks knowledge of its parameters, training dataset, and watermarking detector. The black-box adversary is constrained by a limited number of queries to the GenAI model, as determined by the provider. The serious adversary is the **white-box** adversary, who has full access to the GenAI model and the detector, meaning that he can manipulate the detector's weights in an attempt to classify their generated images as authorized, or to adjust the GenAI model's weights to eliminate watermarks.

Adversary's Capabilities and Strategies. The adversary can employ various techniques to achieve their objectives, including: (1) applying transformation techniques such as compression, resizing, or filtering to remove or disrupt the embedded watermarks; (2) adding adversarial perturbations into DFs to deceive the detector and evade detection; and (3) estimating the watermark through statistical analysis of multiple outputs.

Defender's Capabilities. The defender has full access to their own GenAI model and the watermarking detector. Given any image, their objective is to (i) determine whether it is DFs or authorized synthetic image and (ii) verify whether it originated from their generator.

VII. Challenges and Future Directions

A. Challenges of Proactive Approaches

This section explores the key obstacles when developing and implementing proactive defense strategies against DFs. By identifying and understanding these challenges, we can better focus our efforts on overcoming these hurdles and improving the robustness of proactive detection techniques.

Lack of robustness evaluation. Current proactive DF detection approaches lack robust evaluation against common image transformations, corruptions, and various attacks, raising concerns about their practical reliability. Key issues include:

- Sensitivity to image transformations and corruptions: While DF disruption and watermarking approaches demonstrate effectiveness in controlled environments, their protection success rate often decrease significantly when protected or watermarked images undergo various modifications (e.g., blurring, scaling, cropping, JPEG compression). These modifications can partially or entirely remove embedded perturbations or watermarks, compromising the integrity of the protection mechanism. This vulnerability enables adversaries to potentially destroy these protective measures by applying standard image processing techniques.
- Sensitive to reconstruction attacks: Chen et al. [140] demonstrated that disruption approaches fail to effectively protect users' images against image restoration methods. Their proposed framework can effectively detects abnormalities induced by adversarial perturbations in DF outputs and subsequently recovers the original desired output.
- Sensitive to image purification techniques: Disruption and watermarking approaches are vulnerable to noise purification techniques. The IMPRESS evaluation platform, introduced by Cao et al. [155], demonstrates the vulnerability of disruption methods that utilize imperceptible perturbations to prevent unauthorized data usage in DMs [44]. For watermarking approaches, diffusion purification can significantly reduce detector performance, which has been explored by Saberi et al. [156].

Lack of training-free approaches for watermarking approaches. Current watermarking approaches for GenAI models predominantly rely on re-training or fine-tuning, presenting several challenges:

- High computational cost: Re-training large-scale GenAI models is resource-intensive and time-consuming [71].
- Model and data specificity: Embedded watermarks are often tied to specific model architectures and training datasets, requiring updates when models or data change.

These limitations highlight the need for free-training watermarking techniques. Currently, the field remains largely unexplored, with the method by Wen et al. [154] being one of the few pioneering approaches in this category.

Limited bit length. Lukas and Kerschbaum [71] have revealed the trade-off between capacity and utility in current watermarking approaches. Capacity refers to the number of bits representing the unique identifier while utility denotes the watermarking impact on the GenAI model's quality and functionality, ideally minimal to maintain realistic and diverse content generation. Limited bit length (typically 40 bits) in watermarks introduces several challenges:

- Restricted unique watermarks: The limited space (2^4) may be inadequate for attributing content to specific models or users in large-scale applications.
- Collision vulnerability: Shorter bit lengths increase the probability of accidental watermark collisions, potentially leading to false positives in detection [157].
- Reduced robustness: Shorter watermarks are more vulnerable to removal or corruption through various attacks or content transformations.

Lack of standardized datasets and benchmarks. The field of proactive DF detection lacks standardized datasets and benchmarks, presenting several challenges:

- Reliance on generic computer vision (CV) datasets: Researchers often use public datasets from general CV tasks, which are not tailored to watermarking and disruption approaches. This hinders accurate assessment of these methods in realistic DF scenarios.
- Inconsistent evaluation: The absence of standardized benchmarks leads to varied metrics, datasets, and GenAI models across studies. This diversity in evaluation methodologies, including differences in image quality, bit accuracy metrics, and imperceptibility measures, complicates fair comparisons between approaches.
- Inadequate robustness assessment: Without comprehensive benchmarks covering a wide range of potential attacks and transformations, it's challenging to adequately assess the robustness of proactive approaches against transformations and adversarial attacks.

B. Future Directions

Based on challenges explored in Section VII-A, we suggest potential research directions for research and development in proactive defense strategies.

Standardized and Dynamic Benchmarks: To address the lack of consistent evaluation protocols, future research should prioritize the creation of comprehensive, standardized benchmarks for proactive approaches. These benchmarks should be designed with expandability in mind, allowing for the incorporation of new GenAI families as they emerge. Key aspects to consider include: (i) create diverse datasets that encompass a wide range of DF types and generation methods; (ii) establish evaluation metrics to ensure fair comparisons between different methods; (iii) develop a modular benchmark structure that allows for easy integration of new generative model families and DF techniques.

Holistic Trustworthiness Evaluation: Future work should shift the current DF detection paradigm toward a more holistic view of trustworthiness [158]. This transformation requires a multifaceted approach that goes beyond mere accuracy metrics. Researchers must develop comprehensive evaluation frameworks that assess multiple aspects of trustworthiness, including adversarial robustness, fairness, explainability, and privacy preservation. These frameworks should not only evaluate each aspect independently but also explore the intricate trade-offs and potential synergies between different trustworthiness dimensions.

Adaptive Bit Allocation for Watermarking Approaches: Future work should explore adaptive bit allocation methods which can dynamically adjust the number of embedded bits based on the content complexity. Research in this direction could involve (i) developing content-aware algorithms that analyze image regions to determine optimal bit allocation and (ii) investigating adaptive encoding schemes that can embed variable-length watermarks efficiently.

Advanced Multimodal Detectors: As talking-face video generation techniques become increasingly sophisticated, there is an urgent need to develop more advanced multimodal detection approaches. Addressing the challenge of detecting DFs that manipulate both audio and visual modalities simultaneously is crucial and should be explored in future research.

VIII. Conclusion

While this survey has focused primarily on proactive DF detection approaches in the image domain, many of the core concepts and techniques can be extended to other modalities such as audio, video, and multi-modals. The fundamental principles of disruption and watermarking approaches, as well as the threat models and evaluation metrics discussed, provide a framework that can be adapted for emerging DF threats across various media types. As GenAI continues to advance, we can expect similar proactive defense strategies to be developed for audio and video DFs, and even multi-modals.

References

[1] 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, pp. 10 684–10 695.

[2] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[3] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[4] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 13 916–13 932.

[5] P. Yu, S. Xie, X. Ma, B. Jia, B. Pang, R. Gao, Y. Zhu, S.-C. Zhu, and Y. N. Wu, "Latent diffusion energy-based model for interpretable text modeling," in *Proceedings of International Conference on Machine Learning (ICML)*, July 2022.

[6] N. Savinov, J. Chung, M. Binkowski, E. Elsen, and A. van den Oord, "Step-unrolled denoising autoencoders for text generation," 2021.

[7] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "Biogpt: generative pre-trained transformer for biomedical text generation and mining," *Briefings in bioinformatics*, vol. 23, no. 6, p. bbac409, 2022.

[8] S. Luo, Y. Su, X. Peng, S. Wang, J. Peng, and J. Ma, "Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9754–9767, 2022.

[9] M. Firat, "What chatgpt means for universities: Perceptions of scholars and students," *Journal of Applied Learning and Teaching*, vol. 6, no. 1, pp. 57–63, 2023.

[10] M. Chui, R. Roberts, and L. Yee, "Generative ai is here: How tools like chatgpt could change your business," *Quantum Black AI by McKinsey*, vol. 20, 2022.

[11] B. Ashley, "Ai-faked images of donald trump's imagined arrest swirl on twitter," March 2023.

[12] H. Chen and K. Magramo, "Finance worker pays out $25 million after video call with deepfake 'chief financial officer'," 2024, accessed on 14 March 2024. [Online]. Available: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

[13] IndustryTrends, "Expert feels lensa ai is pilfering from human art," 12 2022, accessed on 25 July 2024. [Online]. Available: https://www.analyticsinsight.net/ampstories/web-stories/expert-feels-lensa-ai-is-pilfering-from-human-art

[14] C. Dong, A. Kumar, and E. Liu, "Think twice before detecting gan-generated fake images from their spectral domain imprints," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7865–7874.

[15] R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, "Intriguing properties of synthetic images: from generative adversarial networks to diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 973–982.

[16] 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, pp. 1–5.

[17] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.

[18] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu, "Gan-generated faces detection: A survey and new perspectives," *arXiv preprint arXiv:2202.07145*, 2022.

[19] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N.

Khan, "Deepfake detection for human face images and videos: A survey," *Ieee Access*, vol. 10, pp. 18 757–18 775, 2022.

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

[21] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio deepfake detection: A survey," *arXiv preprint arXiv:2308.14970*, 2023.

[22] A. Liu, L. Pan, Y. Lu, J. Li, X. Hu, X. Zhang, L. Wen, I. King, H. Xiong, and P. Yu, "A survey of text watermarking in the era of large language models," *ACM Computing Surveys*, 2024.

[23] C. Liang, X. Wu, Y. Hua, J. Zhang, Y. Xue, T. Song, Z. Xue, R. Ma, and H. Guan, "Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples," in *International Conference on Machine Learning*. PMLR, 2023, pp. 20 763–20 786.

[24] J. M. Tomczak, "Why deep generative modeling?" in *Deep Generative Modeling*. Springer, 2021, pp. 1–12.

[25] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, "Pixelsnail: An improved autoregressive generative model," in *International conference on machine learning*. PMLR, 2018, pp. 864–872.

[26] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *Journal of Machine Learning Research*, vol. 22, no. 57, pp. 1–64, 2021.

[27] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *Advances in neural information processing systems*, vol. 32, 2019.

[28] B. Pang, T. Han, E. Nijkamp, S.-C. Zhu, and Y. N. Wu, "Learning latent space energy-based prior model," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 994–22 008, 2020.

[29] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 11, pp. 7327–7347, 2021.

[30] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[31] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. Ieee, 2017, pp. 39–57.

[32] 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, pp. 1–11.

[33] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[34] 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, pp. 4401–4410.

[35] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[36] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[37] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," *arXiv preprint arXiv:1505.00855*, 2015.

[38] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[39] J. Dong, Y. Wang, J. Lai, and X. Xie, "Restricted black-box adversarial attack against deepfake face swapping," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2596–2608, 2023.

[40] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?" in *2012 Fourth international workshop on quality of multimedia experience*. IEEE, 2012, pp. 37–38.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[42] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[43] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[44] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, "Glaze: Protecting artists from style mimicry by {Text-to-Image} models," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 2187–2204.

[45] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The stable signature: Rooting watermarks in latent diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 466–22 477.

[46] Q. Huang, J. Zhang, W. Zhou, W. Zhang, and N. Yu, "Initiative defense against facial manipulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1619–1627.

[47] Z. Li, M. Cao, X. Wang, Z. Qi, M.-M. Cheng, and

Y. Shan, "Photomaker: Customizing realistic human photos via stacked id embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8640–8650.

[48] B. Andy, "Invasive diffusion: How one unwilling illustrator found herself turned into an ai model," November 2022, accessed on 25 July 2024. [Online]. Available: https://waxy.org/2022/11/invasive-diffusion-how-one-unwilling-illustrator-found-herself-turned-into-an-ai-model/

[49] A. Ajay, Y. Du, A. Gupta, J. B. Tenenbaum, T. S. Jaakkola, and P. Agrawal, "Is conditional generative modeling all you need for decision making?" in *The Eleventh International Conference on Learning Representations*.

[50] A. Mishra, S. Krishna Reddy, A. Mittal, and H. A. Murthy, "A generative model for zero shot learning using conditional variational autoencoders," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 2188–2196.

[51] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-guided face generation using conditional cyclegan," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 282–297.

[52] A. Brock, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[53] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International conference on machine learning*. PMLR, 2017, pp. 2642–2651.

[54] A. Shoshan, N. Bhonker, I. Kviatkovsky, and G. Medioni, "Gan-control: Explicitly controllable gans," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 083–14 093.

[55] J. Zhu, Y. Shen, Y. Xu, D. Zhao, Q. Chen, and B. Zhou, "In-domain gan inversion for faithful reconstruction and editability," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[56] M. Kowalski, S. J. Garbin, V. Estellers, T. Baltrušaitis, M. Johnson, and J. Shotton, "Config: Controllable neural face image generation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 299–315.

[57] Y. Shen, P. Luo, J. Yan, X. Wang, and X. Tang, "Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 821–830.

[58] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "Stylerig: Rigging stylegan for 3d control over portrait images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6142–6151.

[59] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 21 696–21 707, 2021.

[60] 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, pp. 3836–3847.

[61] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *International Conference on Machine Learning*. PMLR, 2022, pp. 16 784–16 804.

[62] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6007–6017.

[63] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 511–22 521.

[64] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 500–22 510.

[65] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4296–4304.

[66] Q. Wang, X. Bai, H. Wang, Z. Qin, and A. Chen, "Instantid: Zero-shot identity-preserving generation in seconds," *arXiv preprint arXiv:2401.07519*, 2024.

[67] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," *arXiv preprint arXiv:2303.04226*, 2023.

[68] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[69] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9299–9306.

[70] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in neural information processing systems*, vol. 34, pp. 852–863, 2021.

[71] N. Lukas and F. Kerschbaum, "{PTW}: Pivotal tuning watermarking for {Pre-Trained} image generators," in *32nd USENIX Security Symposium (USENIX Security*

23), 2023, pp. 2241–2258.

[72] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, and Y. Pan, "Generative adversarial networks: A survey toward private and secure applications," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–38, 2021.

[73] Scenario.GG, "Ai-generated game assets," 2022.

[74] S. Hu, Y. Li, and S. Lyu, "Exposing gan-generated faces using inconsistent corneal specular highlights," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2500–2504.

[75] L. Nataraj, T. M. Mohammed, S. Chandrasekaran, A. Flenner, J. H. Bappy, A. K. Roy-Chowdhury, and B. Manjunath, "Detecting gan generated fake images using co-occurrence matrices," *arXiv preprint arXiv:1903.06836*, 2019.

[76] G. Li, X. Zhao, and Y. Cao, "Forensic symmetry for deepfakes," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1095–1110, 2023.

[77] T. Oorloff and Y. Yacoob, "Robust one-shot face video re-enactment using hybrid latent spaces of stylegan2," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 947–20 957.

[78] S. Bounareli, C. Tzelepis, V. Argyriou, I. Patras, and G. Tzimiropoulos, "One-shot neural face reenactment via finding directions in gan's latent space," *International Journal of Computer Vision*, pp. 1–31, 2024.

[79] K. Yang, K. Chen, D. Guo, S.-H. Zhang, Y.-C. Guo, and W. Zhang, "Face2face ρ: Real-time high-resolution one-shot face reenactment," in *European conference on computer vision*. Springer, 2022, pp. 55–71.

[80] S. Bounareli, C. Tzelepis, V. Argyriou, I. Patras, and G. Tzimiropoulos, "Hyperreenact: one-shot reenactment via jointly learning to refine and retarget faces," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7149–7159.

[81] G.-S. J. Hsu, J.-Y. Zhang, H. Y. Hsiang, and W.-J. Hong, "Pose adapted shape learning for large-pose face reenactment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7413–7422.

[82] G.-S. Hsu, C.-H. Tsai, and H.-Y. Wu, "Dual-generator face reenactment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 642–650.

[83] J. Shang, Y. Zeng, X. Qiao, X. Wang, R. Zhang, G. Sun, V. Patel, and H. Fu, "Jr2net: joint monocular 3d face reconstruction and reenactment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2200–2208.

[84] T. Behrouzi, A. Shahroudnejad, and P. Mousavi, "Maskrenderer: 3d-infused multi-mask realistic face reenactment," *Pattern Recognition*, p. 110891, 2024.

[85] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, and R. Ji, "Hififace: 3d shape and semantic prior guided high fidelity face swapping," *arXiv preprint arXiv:2106.09965*, 2021.

[86] Y. Zhu, Q. Li, J. Wang, C.-Z. Xu, and Z. Sun, "One shot face swapping on megapixels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4834–4844.

[87] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[88] Y. Xu, B. Deng, J. Wang, Y. Jing, J. Pan, and S. He, "High-resolution face swapping via latent semantics disentanglement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7642–7651.

[89] Y. Zhu, W. Zhao, Y. Tang, Y. Rao, J. Zhou, and J. Lu, "Stableswap: Stable face swapping in a shared and controllable latent space," *IEEE Transactions on Multimedia*, 2024.

[90] Q. Li, W. Wang, C. Xu, Z. Sun, and M.-H. Yang, "Learning disentangled representation for one-shot progressive face swapping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[91] Y. Li, C. Ma, Y. Yan, W. Zhu, and X. Yang, "3d-aware face swapping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 705–12 714.

[92] F. Rosberg, E. E. Aksoy, F. Alonso-Fernandez, and C. Englund, "Facedancer: Pose-and occlusion-aware high fidelity face swapping," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 3454–3463.

[93] C. Xu, J. Zhang, M. Hua, Q. He, Z. Yi, and Y. Liu, "Region-aware face swapping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7632–7641.

[94] X. Ren, X. Chen, P. Yao, H.-Y. Shum, and B. Wang, "Reinforced disentanglement for face swapping without skip connection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 665–20 675.

[95] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu, "Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8568–8577.

[96] J. Vincent, "New ai deepfake app creates nude images of women in seconds," 6 2019, accessed on 20 July 2024. [Online]. Available: https://www.theverge.com/2019/6/27/18760896/deepfake-nude-ai-app-women-deepnude-non-consensual-pornography

[97] S. Adam and M. Paul, "The people onscreen are fake. the disinformation is real." July 2023, accessed on 25 July 2024. [Online]. Available: https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake

[98] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2head: Audio-driven one-shot talking-head generation with natural head motion," *arXiv preprint*

arXiv:2107.09293, 2021.

[99] Z. Ye, J. He, Z. Jiang, R. Huang, J. Huang, J. Liu, Y. Ren, X. Yin, Z. Ma, and Z. Zhao, "Geneface++: Generalized and stable real-time audio-driven 3d talking face generation," *arXiv preprint arXiv:2305.00787*, 2023.

[100] F. Yin, Y. Zhang, X. Cun, M. Cao, Y. Fan, X. Wang, Q. Bai, B. Wu, J. Wang, and Y. Yang, "Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan," in *European conference on computer vision*. Springer, 2022, pp. 85–101.

[101] Y. Ma, S. Wang, Z. Hu, C. Fan, T. Lv, Y. Ding, Z. Deng, and X. Yu, "Styletalk: One-shot talking head generation with controllable speaking styles," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1896–1904.

[102] S. Tan, B. Ji, Y. Ding, and Y. Pan, "Say anything with any style," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 5088–5096.

[103] S. Tan, B. Ji, and Y. Pan, "Emmn: Emotional motion memory network for audio-driven emotional talking face generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 146–22 156.

[104] C. Xu, J. Zhu, J. Zhang, Y. Han, W. Chu, Y. Tai, C. Wang, Z. Xie, and Y. Liu, "High-fidelity generalized emotional talking face generation with multi-modal emotion space learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6609–6619.

[105] Z. Sun, Y. Xuan, F. Liu, and Y. Xiang, "Fg-emotalk: Talking head video generation with fine-grained controllable facial expressions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 5043–5051.

[106] T. Ki and D. Min, "Stylelipsync: Style-based personalized lip-sync video generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 841–22 850.

[107] S. Wang, Y. Ma, Y. Ding, Z. Hu, C. Fan, T. Lv, Z. Deng, and X. Yu, "Styletalk++: A unified framework for controlling the speaking styles of talking heads," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[108] S. Xu, G. Chen, Y.-X. Guo, J. Yang, C. Li, Z. Zang, Y. Zhang, X. Tong, and B. Guo, "Vasa-1: Lifelike audio-driven talking faces generated in real time," *arXiv preprint arXiv:2404.10667*, 2024.

[109] H. Wei, Z. Yang, and Z. Wang, "Aniportrait: Audio-driven synthesis of photorealistic portrait animation," *arXiv preprint arXiv:2403.17694*, 2024.

[110] L. Tian, Q. Wang, B. Zhang, and L. Bo, "Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions," *arXiv preprint arXiv:2402.17485*, 2024.

[111] S. Liu, T. Lin, D. He, F. Li, M. Wang, X. Li, Z. Sun, Q. Li, and E. Ding, "Adaattn: Revisit attention mechanism in arbitrary neural style transfer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6649–6658.

[112] X. Li, S. Liu, J. Kautz, and M.-H. Yang, "Learning linear transformations for fast image and video style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3809–3817.

[113] L. Jiang, C. Zhang, M. Huang, C. Liu, J. Shi, and C. C. Loy, "Tsit: A simple and versatile framework for image-to-image translation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 206–222.

[114] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2287–2296.

[115] J. Huang, J. Liao, and S. Kwong, "Unsupervised image-to-image translation via pre-trained stylegan2 network," *IEEE Transactions on Multimedia*, vol. 24, pp. 1435–1448, 2021.

[116] 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, pp. 2085–2094.

[117] G. Kwon and J. C. Ye, "Clipstyler: Image style transfer with a single text condition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 062–18 071.

[118] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[119] M. Heikkilä, "This artist is dominating AI-generated art. And he's not happy about it." August 2024, accessed on 3 April 2024. [Online]. Available: https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/

[120] C. Erik, "The alt-right manipulated my comic. then a.i. claimed it." December 2022, accessed on 25 July 2024. [Online]. Available: https://www.nytimes.com/2022/12/31/opinion/sarah-andersen-how-algorithim-took-my-work

[121] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[122] T. Van Le, H. Phung, T. H. Nguyen, Q. Dao, N. N. Tran, and A. Tran, "Anti-dreambooth: Protecting users from personalized text-to-image synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2116–2127.

[123] H. Huang, X. Ma, S. M. Erfani, J. Bailey, and Y. Wang, "Unlearnable examples: Making personal data unexploitable," in *International Conference on Learning Representations*.

[124] S. Fu, F. He, Y. Liu, L. Shen, and D. Tao, "Robust unlearnable examples: Protecting data privacy against adversarial learning," in *International Conference on Learning Representations*.

[125] X. Wang, J. Huang, S. Ma, S. Nepal, and C. Xu, "Deepfake disrupter: The detector of deepfake is my friend," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 920–14 929.

[126] Y. Zhu, Y. Chen, X. Li, R. Zhang, X. Tian, B. Zheng, and Y. Chen, "Information-containing adversarial perturbation for combating facial manipulation systems," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2046–2059, 2023.

[127] P. Sun, Y. Li, H. Qi, and S. Lyu, "Landmark breaker: obstructing deepfake by disturbing landmark extraction," in *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2020, pp. 1–6.

[128] C. Yang, L. Ding, Y. Chen, and H. Li, "Defending against gan-based deepfake attacks via transformation-aware adversarial faces," in *2021 international joint conference on neural networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[129] N. Ruiz, S. A. Bargal, and S. Sclaroff, "Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 236–251.

[130] C.-Y. Yeh, H.-W. Chen, S.-L. Tsai, and S.-D. Wang, "Disrupting image-translation-based deepfake algorithms with adversarial attacks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 2020, pp. 53–62.

[131] Z. Fang, Y. Yang, J. Lin, and R. Zhan, "Adversarial attacks for multi target image translation networks," in *2020 IEEE International Conference on Progress in Informatics and Computing (PIC)*. IEEE, 2020, pp. 179–184.

[132] L. Tang, D. Ye, Z. Lu, Y. Zhang, and C. Chen, "Feature extraction matters more: An effective and efficient universal deepfake disruptor," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[133] Y. Liu, C. Fan, Y. Dai, X. Chen, P. Zhou, and L. Sun, "Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 219–24 228.

[134] R. Wang, Z. Huang, Z. Chen, L. Liu, J. Chen, and L. Wang, "Anti-forgery: Towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations," *Proceedings of International Joint Conference on Artificial Intelligence*, 2022.

[135] C.-Y. Yeh, H.-W. Chen, H.-H. Shuai, D.-N. Yang, and M.-S. Chen, "Attack as the best defense: Nullifying image-to-image translation gans via limit-aware adversarial attack," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 188–16 197.

[136] N. Ruiz, S. A. Bargal, C. Xie, and S. Sclaroff, "Practical disruption of image translation deepfake networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, 2023, pp. 14 478–14 486.

[137] W. Guan, Z. He, W. Wang, J. Dong, and B. Peng, "Defending against deepfakes with ensemble adversarial perturbation," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 1952–1958.

[138] Z. Qu, Z. Xi, W. Lu, X. Luo, Q. Wang, and B. Li, "Df-rap: A robust adversarial perturbation for defending against deepfakes in real-world social network scenarios," *IEEE Transactions on Information Forensics and Security*, 2024.

[139] H. Huang, Y. Wang, Z. Chen, Y. Zhang, Y. Li, Z. Tang, W. Chu, J. Chen, W. Lin, and K.-K. Ma, "Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 989–997.

[140] Z. Chen, L. Xie, S. Pang, Y. He, and B. Zhang, "Magdr: Mask-guided detection and reconstruction for defending deepfakes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9014–9023.

[141] H. Wu, J. Zhou, J. Tian, J. Liu, and Y. Qiao, "Robust image forgery detection against transmission over online social networks," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 443–456, 2022.

[142] Z. Zhao, D. Dua, and S. Singh, "Generating natural adversarial examples," *arXiv preprint arXiv:1710.11342*, 2017.

[143] C. Laidlaw, S. Singla, and S. Feizi, "Perceptual adversarial robustness: Defense against unseen threat models," in *International Conference on Learning Representations (ICLR)*, 2021.

[144] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2730–2739.

[145] Z. Yu, S. Zhai, and N. Zhang, "Antifake: Using adversarial audio to prevent unauthorized speech synthesis," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 460–474.

[146] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 284–293.

[147] T. Bui, N. Yu, and J. Collomosse, "Repmix: Repre-

sentation mixing for robust attribution of synthesized images," in *European Conference on Computer Vision*. Springer, 2022, pp. 146–163.

[148] T. Yang, J. Cao, Q. Sheng, L. Li, J. Ji, X. Li, and S. Tang, "Learning to disentangle gan fingerprint for fake image attribution," *arXiv preprint arXiv:2106.08749*, 2021.

[149] Y. Adi, C. Baum, M. Cisse, B. Pinkas, and J. Keshet, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in *27th USENIX security symposium (USENIX Security 18)*, 2018, pp. 1615–1631.

[150] A. Bansal, P.-y. Chiang, M. J. Curry, R. Jain, C. Wigington, V. Manjunatha, J. P. Dickerson, and T. Goldstein, "Certified neural network watermarks with randomized smoothing," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1450–1465.

[151] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz, "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 14 448–14 457.

[152] J. Fei, Z. Xia, B. Tondi, and M. Barni, "Supervised gan watermarking for intellectual property protection," in *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2022, pp. 1–6.

[153] N. Yu, V. Skripniuk, D. Chen, L. S. Davis, and M. Fritz, "Responsible disclosure of generative models using scalable fingerprinting," in *International Conference on Learning Representations*.

[154] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, "Tree-rings watermarks: Invisible fingerprints for diffusion images," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[155] B. Cao, C. Li, T. Wang, J. Jia, B. Li, and J. Chen, "Impress: evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[156] M. Saberi, V. S. Sadasivan, K. Rezaei, A. Kumar, A. Chegini, W. Wang, and S. Feizi, "Robustness of ai-image detectors: Fundamental limits and practical attacks," *arXiv preprint arXiv:2310.00076*, 2023.

[157] H. Zhang, B. L. Edelman, D. Francati, D. Venturi, G. Ateniese, and B. Barak, "Watermarks in the sand: Impossibility of strong watermarking for generative models," *arXiv preprint arXiv:2311.04378*, 2023.

[158] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy ai: From principles to practices," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–46, 2023.