*ECAI 2023*
*K. Gal et al. (Eds.)*


*doi:10.3233/FAIA230558*

# GAN-Generated Faces Detection: A Survey and New Perspectives

**Xin Wang [1,3], Hui Guo [1], Shu Hu [2], Ming-Ching Chang [3] and Siwei Lyu [1]**

[1] University at Buffalo, SUNY, USA. {xwang264, hguo8, siweilyu}@buffalo.edu
[2] Indiana University–Purdue University Indianapolis, USA. shuhu@cmu.edu
[3] University at Albany, SUNY, USA. {xwang56, mchang2}@albany.edu

**Abstract.** Generative Adversarial Networks (GAN) have led to the generation of very realistic face images, which have been used in fake social media accounts and other disinformation matters that can generate profound impacts. Therefore, the corresponding GAN-face detection techniques are under active development that can examine and expose such fake faces. In this work, we aim to provide a comprehensive review of recent progress in GAN-face detection. We focus on methods that can detect face images that are generated or synthesized from GAN models. We classify the existing detection works into four categories: (1) deep learning-based, (2) physical-based, (3) physiological-based methods, and (4) evaluation and comparison against human visual performance. For each category, we summarize the key ideas and connect them with method implementations. We also discuss open problems and suggest future research directions.

## 1 Introduction

The development of Generative Adversarial Networks (GANs) [32] enables generating high-realistic human faces images that are visually difficult to discern from real ones [53, 54, 55], some examples[1] are shown in Figure 1. GAN-generated faces (GAN-faces) can be easily used in creating fake social media accounts [89, 90, 40, 107] for malicious purposes that cause significant social concerns. For example, a high school student created a fake candidate by using a GAN-generated face in a voting event that tricked Twitter into obtaining a coveted blue checkmark, thereby verifying the authenticity of the fake candidacy [89]. This fake candidate passing verification could set up donation channels to absorb public funds, which not only damages property-related laws but also diminishes election integrity. Furthermore, the fake social media accounts used GAN-faces as profile images which also generate serious negative social impacts [90, 79]. For example, these fake accounts can be associated with numerous high-level executives within a company, and if they were to post comments about the company's financial situation, it could cause significant disruption in the stock market. This is because the false information they spread can mislead investors and cause them to make incorrect financial decisions, leading to significant losses.

Automatic detection of GAN-faces is of emerging need [85], so numerous detection approaches have been developed to combat the malicious use of GAN-faces. However, effective GAN-face detection is still a complex and difficult problem, which typically suffers from two major challenges. First, an accurate and flexible GAN-face detection method should be able to expose the large variation of GAN-face images synthesized or generated from numerous GAN models, while remaining robust to adversarial attacks. Secondly, the decision process and the detection result should be **explainable to human users**, especially for non-AI experts, instead of only fitting to specific datasets via complex deep networks.

**Figure 1.** *Examples of GAN faces generated by StyleGAN [54] (**left**), StyleGAN2 [55] (**middle**), and StyleGAN3 [52] (**right**).*

In this paper, we focus the scope of the survey on the detection of **GAN-based entire face synthesis**[2], which was a significant milestone of the Artificial Intelligence Generated Content (AIGC) [11]. And this is the right time to deal with the massive use of generative AI [24]. We start our survey by chronologically summarizing major GAN-face generation milestones (§ 2) as well as GAN-face detection methods with highlights of important breakthroughs along with in Figure 2. Early GAN-face detection methods are mainly Deep Learning (DL)-based methods [22, 72], *etc.*; see § 3.1. Although they achieve promising performance in practice, it is difficult to explain the under-taking mechanism or decisions being made.

The above limitations are overcome by approaches reasoning upon physical cues (§ 3.2) or physiological cues (§ 3.3) that are explainable in nature. Recent works in this category distinguish GAN-faces by exploring the inadequacy of the GAN synthesis models in repre-

[1] https://thispersondoesnotexist.com

[2] It is different from face manipulation, which only manipulates the existing face images, instead of generation from scratch.


**Figure 2.** *A brief chronology for GAN-faces generation and detection works.* ***Generation:*** *The initial GAN model is proposed in 2014 and can only generate 32×32 faces. After 2017 the series of StyleGAN models can generate high-realistic faces that are hard to spot from human eyes.* ***Detection:*** *The earliest detection techniques are mainly based on DNN in 2018. Due to their limitation of performance and interpretability, methods based on physical and physiological cues are developed in 2019 ∼ 2021. Since StyleGAN2 generated faces are very difficult to discern from human eyes, human visual performance on GAN-generated faces is under active investigation since 2021. The listed methods represent milestones and breakthroughs in the chronology. See § 3 for complete survey.*

senting human faces and their corresponding relations in the physical world [112, 61, 75]. For example, [42] inspect the inconsistency of the corneal specular highlights between the two eyes. However, these methods work under strict assumptions such as frontal portrait faces or a clearly visible reflector in the eyes. To eliminate these limitations and explore more robust models, [36] introduce a physiological-based method by examining pupil shape inconsistencies. As the human eye provides the optics and photoreception for the visual system, the pupil should generally be circular on the eye surface or appear to be elliptical in the image when viewed with an orientation. The key idea is that physiological inconsistency artifacts between the eyes (*e.g.* difference from comparing the boundary of pupil shapes) can be identified to distinguish GAN-faces from real faces.

An important aspect of the GAN-face detection in contrast to other AI problems (such as image classification) is that *human performance for GAN-face detection is much worse than AI algorithmic methods*. As shown in [84], human accuracy for GAN-face detection is around 50%∼60%, which shows that topics on improving or accommodating human performance are essential. We provide a comprehensive discussion in § 3.4 on the topic of human visual performance for GAN-face detection.

The datasets are the driving force behind the rapid development of GAN models and GAN-face detection methods. We survey popular datasets and major evaluation metrics in § 4. For completeness, we also list other related surveys in § 5. In the foreseeable future, there are a number of critical problems that are yet to be resolved for existing GAN-face detection methods. With the development of the GAN models, it is thus important to anticipate such new developments and improve the detection methods accordingly. We discuss future research opportunities in § 6.

The contribution of this paper is summarized in the following:

- To the best of our knowledge, this work is the first comprehensive review that discusses different types of GAN-face detection methods. We particularly include the explainable methods that provide interpretability of the decision process and results that ease human understanding.
- We organize and summarize the vast literature on GAN-face detection into four categories: (1) deep learning-based methods, (2) physical-based methods, (3) physiological-based methods, and (4) human visual performance.
- Human visual performance of recognizing GAN-faces is important, especially for people to check for their social networking and possible security or privacy violations. We provide a comprehensive discussion on human visual performance and strategies for checking against fake GAN-generated faces.
- We propose several issues associated with existing state-of-the-art methods and discuss future research directions.

## 2 GAN Generation of Highly Realistic Faces

We next provide a brief summary of mainstream methods for generating high-quality faces that most GAN-face detection works are targeting. Further details on the various kinds of GANs can be found in the surveys of [44, 111].

In the past five years, numerous GAN models (*e.g.*, PG-GAN [53], BigGAN [8], StyleGAN [54], StyleGAN2 [55], etc.) have been developed to synthesis and create realistic-looking face images with diversity from random noise input. These GANs can effectively encode rich semantic information in the intermediate features [4] and latent space [31, 45, 102] for high-quality face image generation. Moreover, these GANs can generate fake face images with various attributes, including various ages, expressions, backgrounds, and viewing angles. However, due to the lack of inference functions or encoders in GANs, such manipulations in latent space are only applicable to images generated from GANs, not to any given real images.

To address the above issue, GAN inversion methods can invert a given image back into its latent space of a pre-trained GAN model [111]. The GAN generator can then reconstruct the image accurately from the inverted code in approximation. This inversion method plays a key role in bridging real and fake face image domains. Therefore, it can significantly improve the quality of the generated face images and be applied widely in state-of-the-art GAN models including StyleGAN2 [55], StyleGAN3 [52], InterFaceGAN [102], and Image2StyleGAN++ [1].

## 3 GAN-face Detection Methods

We organize existing GAN-face detection literature into four categories. Although there exist similarities of various methods *e.g.* across categories, we organize them primarily by their motivations and key ideas. Table 1 summarizes mainstream GAN-face detection methods with the datasets used and performance comparison.

### *3.1 Deep Learning-based Methods*

Deep learning-based GAN-face detection methods extract signal-level features to train Deep Neural Network (DNN) classifiers to dis-

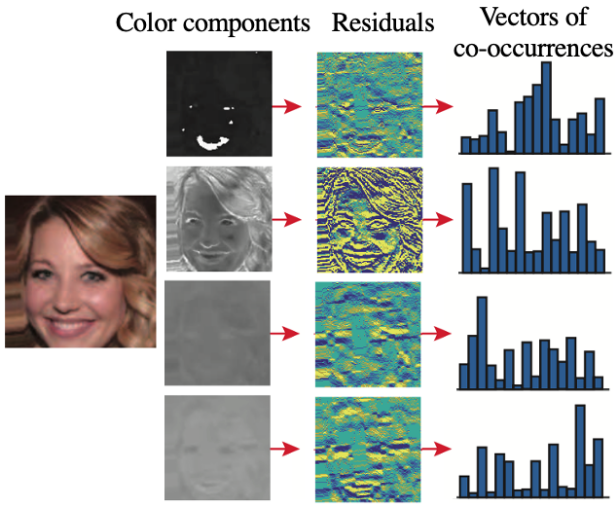


**Figure 3.** *The features of capturing color image statistics for training the classifier [62].*

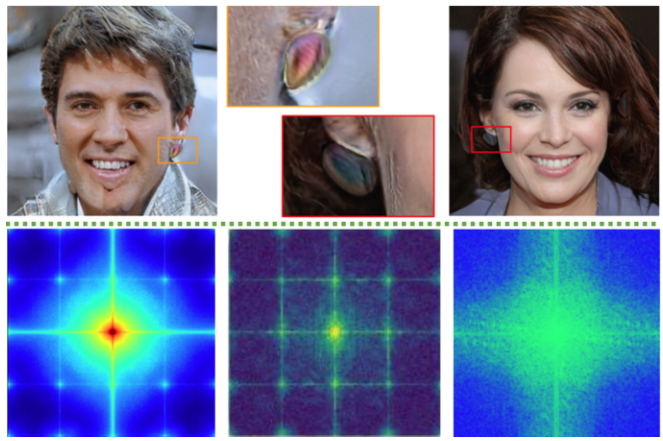

**Figure 4.** ***Top:*** *Visible color artifacts of GAN image [34].* ***Bottom:*** *Invisible artifacts of GAN image, averaged Fourier spectrum [34], frequency analysis [110], frequency spectrum [23].*

tinguish fake faces from real ones in an end-to-end learning framework [47, 26, 9].

The earliest work of [22] employed VGG-Net [104] for GAN-face detection. To train the network, real faces are collected from the CelebA face dataset [69], and fake faces are generated using DC-GANs [97] and PG-GAN [53], where the VGG-16 architecture is used with pre-train weights of VGG-Face [10]. [78] found that signals in the residual field can serve as effective features to distinguish real and GAN-faces. They first processed the input faces with high-pass filters, and the resulting residuals were fed into deep networks for GAN-face detection. [62] identified GAN-faces by analyzing the chrominance color components. They first extracted a feature set to capture color image statistics (See Figure 3), then use the concatenated features to train a GAN-face classifier. Similarly, [15] found that both the luminance and chrominance cues are useful for improving GAN-face detection. More recently, [27] used a dual-channel CNN to reduce the impact of many widely-used image post-processing operations. The deep CNN of their network extracts features of the pre-processed images, and the shallow CNN extracts features from the high-frequency components of the original image.

**GAN-face detection in real-world scenarios.** The work [43] developed a framework for evaluating detection methods under cross-model, cross-data, and post-processing evaluations, to examine features produced from commonly-used image pre-processing methods. More recently, many variants of feature-based models have been studied [110, 30, 68, 16]. However, the detection results from all these feature-based methods are not explainable, so it is unclear why the decision was given to any input face.

**One-shot, incremental and advanced learning.** A one-shot GAN-face detection method was studied recently in [71]. Scene understanding is applied to determine out-of-context objects that appeared in the GAN-faces to distinguish GAN-faces from the real ones. The work [73] applied incremental learning for GAN-faces image detection, where the key idea is to detect and classify new GAN-generated faces without decreasing the performance on existing ones.

**Difficulty Analysis.** More difficulty analysis and systemic evaluations using state-of-the-art DNNs for GAN-face detection are investigated in [34, 109, 110, 46, 23], both visible and invisible artifacts are analyzed in these works (See Figure 4). For example, [110] find that the CNN-generated images share some common systematic flaws, resulting in them being surprisingly easy to spot for now. To investigate *Are GAN-generated images easy to detect?* [34] conducted the study to analyze the performance of the existing GAN-faces detection methods on different datasets and using different metrics. On the country, they concluded that we are still very far from having reliable tools for GAN image detection.

Unfortunately, all aforementioned methods in this subsection can not provide explainable results. To overcome this shortcoming, an attention-based method was proposed in [38] to spot GAN-generated faces by analyzing eye inconsistencies. Specifically, this model learned to identify inconsistent eye components by localizing and comparing the iris artifacts. Visual results from [38] showed a clear difference between the attention maps of the irises from the GAN-faces and real ones. For GAN-faces the attention map highlighted the artifact regions on the irises, and for real faces, there is no significant concentration of the attention map. However, the attention map still cannot provide enough explainability to understand the behavior of the learned model.

In summary, Deep Learning-based methods achieved impressive performance on GAN-face detection [35]. However, it is difficult to explain or interpret the decision process of the learned model as a black box. Nonetheless, fake face detection in the real-world favors explainability, alongside from the overall accuracy. Particularly, people do care more for use cases such as *"This picture looks like someone I know, and if the AI algorithm tells it is fake or real, then what is the reasoning and should I trust?"*

## 3.2 *Physical-based Methods*

Physical-based methods identify GAN-faces by looking for artifacts or inconsistencies among the face and the physical world, such as the illumination and reflections in perspective.

The early work of [49] analyzed the internal camera parameters and light source directions from the perspective distortion of the specular highlights of the eyes to reveal traces of image tampering. Recently, [75] identified early versions of GAN-faces [53] based on an observation that the specular reflection in the eyes of GAN-faces is either missing or appearing as a simple white blob. However, such artifacts have been largely corrected in recent GAN-faces such as StyleGAN2. The method of [42] looked for inconsistency between the two eyes to identify GAN-generated faces. Specifically, the corneal specular highlights of the eyes are detected and aligned for

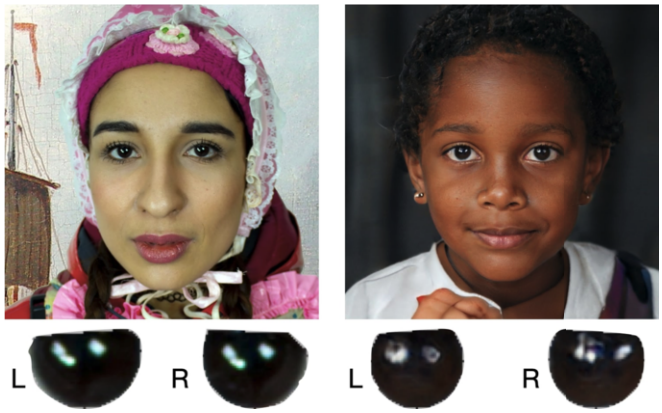


**Figure 5.** ***Top:*** *Corneal specular highlights for a real human face (**left**) and a GAN-face (**right**) [42].* ***Bottom:*** *The corneal regions are isolated and scaled for better visibility. Note that the corneal specular highlights for the real face have strong similarities while those for the GAN-faces are different.*

pixel-wise Intersection of Union (IoU) comparison. As shown in Figure 5, the assumption is that real human eyes captured by a camera under a portrait setting should exhibit a strong resemblance between the corneal specular highlights between the two eyes. In contrast, this assumption is not true for GAN synthesized eyes, where inconsistencies include different numbers, different geometric shapes, or different relative locations of the specular highlights. However, this method operates on strong assumptions of the frontal portrait pose, far away lighting source(s), and the existence of the eye specular highlights. When these assumptions are violated, false positives may increase significantly.

In summary, the physical-based detection methods are more robust to adversarial attacks, and the predicted results afford intuitive interpretations to human users [42].

### 3.3 Physiological-based Methods

Physiologically-based methods investigate the semantic aspect of the human faces [18], including cues such as symmetry, iris color, pupil shapes, *etc.*, where the identified artifacts are used for exposing GAN-faces.

Early works of [72, 113, 76] indicated that StyleGAN [53] generated faces contain obvious artifacts including asymmetric faces and inconsistent iris colors [75]. [112] found that GAN can generate facial parts (*e.g.*, eyes, nose, skin, mouth) with a great level of realistic details, yet there is no explicit constraint over the locations of these parts on the face. In other words, the facial parts of GAN-faces may not appear to be coherent or natural-looking, when compared to real faces. They indicated that these abnormalities in the configuration of facial parts in GAN-faces could be revealed using the locations of the facial landmark points (*e.g.*, tips of the eyes, nose, and mouth), which can be effectively detected using automatic algorithms. The normalized locations of these facial landmarks can be used features to train a classifier to identify GAN-faces. However, GAN-face generation has also improved on the other hand. Face images generated by StyleGAN2 have improved greatly in quality and are free of obvious physiological artifacts [53, 54, 55]. And the synthesis process of GAN-faces is further optimized in StyleGAN3. It exhibits a more natural transformation hierarchy of different scales of features. They are fully equivariant to translation and rotation, which further improved the physiological consistency of the generated faces.

A relatively new physiological-based GAN-face detection method

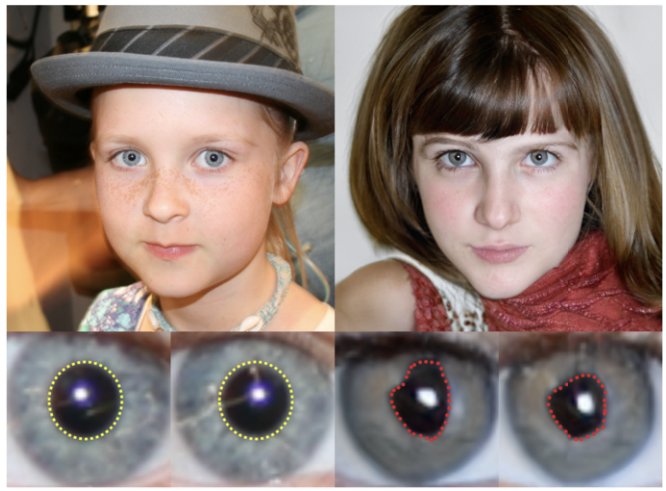

**Figure 6.** *Pupils of real (**left**) human face and GAN-face (**right**) [36]. Note that the pupils for the real eyes have strong circular shapes (yellow) while those for the GAN-generated pupils are with irregular shapes (red).*

is proposed in [36], motivated by a simple observation that GAN-faces exhibit a common artifact of irregular pupil shapes. Specifically, pupils from real human faces should appear to be a smooth circle or ellipse; in contrast, pupils from GAN-faces can appear with irregular shapes or boundaries (See Figure 6). This artifact is universal for all known GAN models up to date (including PG-GAN [53], StyleGAN3 [52], and SofGAN [13]), and this artifact occurs in eyes from the synthesized humans and animals. One fundamental reason for the existence of such artifacts in GAN-generated faces is due to their lack of understanding of human eye anatomy, particularly the geometry and shape of the pupils. The method of [36] first localize the eyes and segment out the pupil region. Next, an ellipse model is parametrically fit to the pupil boundary. Boundary IoU [17] is then calculated between the extracted pupil mask and the fitted ellipse to estimate the "circularness" of the pupils. However, false positive can arise in rare cases of non-elliptical pupils in real faces due to diseased or infected eyes.

In summary, physiological-based method comes with stronger interpretability. However, like other forensic approaches, environmental constraints such as occlusion and visibility of the eye from the face image is still a major limitation.

### 3.4 Human Visual Performance

Although many automatic GAN-face detection algorithms have been developed, human visual performance in identifying and exposing GAN-faces has not been investigated sufficiently. Compared with other AI problems such as image recognition, GAN-face detection is a much more challenging problem for human eyes. Thus, it is important to study how well human eyes can identify GAN-faces and the related social impacts and ethical issues [3].

Standard metrics for evaluating the effectiveness of automatic algorithms in detecting GAN-faces include ROC analysis and Precision-Recall. While these metrics can be applied to study human perceptual performance, they are not directly suitable in reflecting the true deceptiveness of the highly realistic GAN-faces for the general public. Human performance is largely biased, and with weak but proper hints (such as looking for the correct physiological cues), human performance in identifying fake faces can boost greatly [57].

An early work [58] conducted a study to measure the human abil-

| Paper | Category | Method | Real Face (#test) | GAN Face (#test) | Performance |
|---|---|---|---|---|---|
| [84] | Human | Visual | FFHQ (400) | StyleGAN2 (400) | Acc: 0.5∼0.6 |
| [58] | Human | Visual | FFHQ (150) | StyleGAN2, etc. (150) | Acc: 0.26∼0.8 |
| [22] | DL | CNN | CelebA (200) | PGGAN, DCGAN (200) | Acc: 0.80 |
| [19] | DL | CNN | CelebA (1250) | PCGAN (1250) | Acc: 0.98 |
| [78] | DL | CNN | CelebA-HQ (15K) | PGGAN (15K) | Acc: 0.99 |
| [80] | DL | CNN | CelebA (500) | StarGAN (4498) | Acc:0.99 |
| [27] | DL | CNN | CelebA-HQ (7K) | PGGAN (7K) | Acc: 0.98 |
| [73] | DL | Incremental Classifier | - | StarGAN (2.4K), etc. | Acc: 0.815∼1 |
| [71] | DL | Out of context object detection | - | StyleGAN (100) | Acc: 0.80 |
| [109] | DL | DNN | FFHQ (1K) | StyleGAN2 (1K), etc. | Acc: 0.88∼0.991 |
| [62] | DL | Disparities in Color Components | CelebA-HQ, FFHQ (50K) | StyGAN, ProGAN (50K) | Acc: 0.997 |
| [110] | DL | CNN | FFHQ (1K) | StyleGAN (1K) | Acc: 0.84 |
| [43] | DL | ForensicTransfer | FFHQ (3K), etc. | StyleGAN (3K), ProGAN (3K), etc. | Acc: 0.01∼1 |
| [30] | DL | CNN | CelebA (164), CelebA-HQ (1.5K) | StarGAN (1476), ProGAN (3.7K) | Acc: 0.6768∼0.849 |
| [68] | DL | CNN | FFHQ (10K), CelebA-HQ (10K) | StyleGAN (10K), PGGAN (10K), etc. | Acc: 0.9854∼0.991 |
| [14] | DL | Xception | FFHQ (7K) | LGGF (14K) | Acc: 0.99 |
| [15] | DL | Improved Xception | CelebA (202,60) | PGGAN (202,60) | Acc: 0.713∼0.977 |
| [34] | DL | CNN | RAISE (≤7.8K) | StyleGAN2 (3K), ProGAN (3K), etc. | Acc: 0.928∼0.999 |
| [16] | DL | CNN | FFHQ (20K) | StyleGAN (20K), etc. | Acc: 0.9895∼1 |
| [38] | DL | Residual Attention | FFHQ (748) | StyleGAN2 (750) | AUC: 1 |
| [88] | DL | Cross-Co-Net | FFHQ (4K) | StyleGAN2 (4K) | Acc: 0.998 |
| [23] | DL | Enhanced spectrum based CNN | FFHQ (3.2K) | StyleGAN, StyleGAN2 (3.2K) | Acc: 0.95 ∼ 0.96 |
| [108] | DL | Siamese Network | FFHQ, CelebA-HQ (10K) | ProGAN, StyleGAN3 (20K) | AUC: 0.996 ∼ 1 |
| [42] | Physic | Corneal specular highlight | FFHQ (500) | StyleGAN2 (500) | AUC: 0.94 |
| [75] | Physiology | Eye color | CelebA (1K) | ProGAN (1K), Glow (1K) | AUC: 0.70∼0.85 |
| [112] | Physiology | Landmark locations | CelebA (≥50K) | PGGAN (25K) | AUC: 0.9121∼0.9413 |
| [36] | Physiology | Irregular pupil shape | FFHQ (1.6K) | StyleGAN2 (1.6K) | AUC: 0.91 |

**Table 1.** *Summary of GAN-face detection methods with the corresponding datasets, statistics and performance scores. The gray rows highlight those where* individual *predicted results of the method are* **explainable** *to humans. Note that datasets used in the works are self-collected and can contain different subsets across papers. So the performance scores do not represent fair comparisons.*

ity to recognize fake faces. Their dataset consists of 150 real faces and 150 GAN faces. Real faces are selected from the Flickr-Faces-HQ (FFHQ) dataset, and GAN-Faces are generated from state-of-the-art GANs, including PG-GAN, StyleGAN, and StyleGAN2. The 630 participants sequentially completed 34 tasks to distinguish 30 faces each time. Those faces were randomly selected in equal portions from each category. Results showed that participants had lost the ability to judge newer GAN-faces. Accuracy is not impacted when the test speeds up or the participants have seen similar synthetic faces produced by the generators before.

A recent work [84] examined people's ability to discriminate GAN-faces from real faces. Specifically, 400 StyleGAN2 faces and 400 real faces from the FFHQ dataset are selected with large diversity across the genders, ages, races, *etc.*, and two sets of experiments are conducted. In the first set of experiments, 315 participants were shown a few examples of GAN-faces and real faces, and around 50% of accuracy is obtained. In the second set of experiments, 170 new participants were given a tutorial consisting of examples of specific artifacts in the GAN-faces. Participants were also given feedback afterward. However, it was found that such training and feedback only improve a little bit of average accuracy. Therefore, this work concluded that the StyleGAN2 faces are realistic enough to fool both naive and trained human observers, more extended studies are summarized in [87], the experiment 3 is conducted to further investigate whether synthetic faces activate the same judgements of trustworthiness. A perception of trustworthiness could also help distinguish real from GAN-faces. Their experimental results are shown in Figure 7. However, no information on what synthesis artifacts are provided for participant training in this study. We believe there is still space to improve human capability in discerning GAN-faces if sufficient hints are provided, including physiological cues (*e.g.* pupil shapes [36]) and dataset statistics (*e.g.* GAN-faces are usually trained with FFHQ

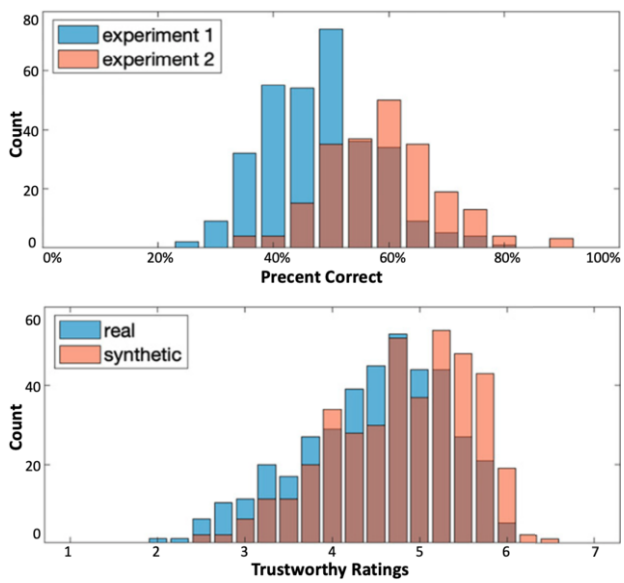


**Figure 7.** *Human visual performance [87]. **Top:** Average performance of experiments 1 and 2, the accuracy is around 50%. In experiment 2, the training and feedback improves average performance a little bit. **Bottom:** Trustworthy ratings for experiment 3, a rating of 1 corresponds to the lowest trust.*

samples that are biased toward portrait faces and celebrity styles).

GANs are under active development, so it is expected that the difficulty of discerning GAN-faces will continue to increase. It is important to find generic and consistent cues for human eyes to effectively distinguish GAN-faces. Typically, useful cues are generally universal for exposing other types of AI tampered faces, including morphed faces, swapped faces, painting faces. Recently, an open platform to

study whether a human can distinguish AI-synthesized faces from real faces visually by using the cues is developed in [37]. The discovery of such cues can also be leveraged for improving the GAN face synthesis algorithm to produce faces that are even harder to distinguish for human eyes.

In summary, there is no doubt that the studies of human visual performance are invaluable to research detection techniques as well as a better understanding of the insufficient of the GAN-faces.

## 4 Datasets and Performance Evaluation

With the rapid development of AI discriminative and generative models, many human facial datasets have been constructed. Among these datasets, real face images are mainly collected from the FFHQ dataset [54], CelebA [69], CelebA-HQ [53], RAISE [20] *etc.*. Synthesized face images are collected using state-of-the-art GAN models and LGGF [14].

Early GAN-faces datasets are mainly comprised of PGGAN, and recent datasets are typically based on StyleGAN2. NVIDIA has recently curated a StyleGAN3 generated set at https://github.com/NVlabs/stylegan3-detector that can be used to evaluate GAN-face detection performance. Table 1 list mainstream datasets for GAN-face detection. Note that datasets used for each work are self-collected and can contain different subsets across papers. This is due to that only specific subsets are relevant to individual methods. For example, in [36], only face images with visible eye pupils are used for training and evaluation.

As GAN-face detection is a binary classification problem, **evaluation metrics** are typically based on Accuracy, Precision-Recall, ROC analysis, and AUC. To the best of our knowledge, a sufficiently large-scale benchmark dataset for empirical evaluation of GAN-face detection is still lacking.

## 5 Related Surveys

Although the scope of this survey is on the detection of *GAN-based entire face synthesis*, the GAN-face detection task is closely related to other fake face detection tasks [82, 25]. For completeness, we also discuss other surveys in the related fields.

**GAN-face Detection**. A related survey in [66] only discussed DL-based GAN-face detection works and ignored other significant non-DL-based works. Their survey neglects the interpretability issues, which is crucial for applying DL-based methods for detecting GAN-faces in practice.

**Morphed-face Detection**. Morphed-face detection aims to detect images that have been merged by two or more face images [95]. Morphed-face detection is a challenging problem due to the complex nature of the morphing techniques used to create these images. Therefore, the development of effective and accurate morphed-face detection methods is of utmost importance to prevent potential harm caused by these fraudulent activities [101]. The current surveys provide an overview of the recent advances towards both the generation and the detection of morphing face [100, 98]. However, the motivation for generating morphed faces is usually different from GAN-face, these morphed faces are often created for identity theft and profile impersonation.

**Manipulated-face Detection**. Manipulated-face detection involves the identification of images that have been manipulated or tampered with by AI algorithms [91]. The increasing availability of AI algorithms has made it easier to create manipulated images for fraudulent purposes, such as face swapping [48], and facial manipulation [67].


**Figure 8.** *Video conferencing DeepFakes detection [39]. **Left:** A video call attendant is being actively authenticated with the live patterns shown on the screen. **Right:** A real person's cornea will produce an image of the pattern shown on the screen while a real-time DeepFake cannot.*

The current surveys [83, 106] indicate that the development of reliable and robust manipulated-face detection methods is essential to safeguard the authenticity and integrity of digital media and prevent harm caused by the misuse of manipulated images.

As previously mentioned, it's important to note that face manipulation and GAN-based face generation are distinct techniques. While GANs generate faces from scratch, face manipulation involves altering or modifying existing face images using various techniques.

**DeepFakes Detection**. DeepFakes detection involves detecting and identifying images, videos, audio, and text that have been generated or manipulated using artificial intelligence techniques [81, 114, 103]. DeepFakes are often created to deceive and manipulate viewers by inserting fake information into real events or spreading misinformation or creating fake news [105, 92, 5]. The use of DeepFakes poses a significant threat to the authenticity and credibility of social media [6], making it essential to develop reliable and effective DeepFakes detection methods [77, 21, 51].

The DeepFakes detection is primarily focused on identifying and combating the spread of fake news and misinformation [115, 70, 116, 59]. The current surveys can help prevent the potential harm caused by the misuse of DeepFakes and ensure that the information presented is accurate and trustworthy [106, 50]. Furthermore, other surveys have been conducted with a different perspective compared to existing survey papers, for example, the survey [56] mostly focuses on the audio deepfakes that are overlooked in most of the previous surveys, [74] focus on the audio-visual DeepFakes generation and detection, and the benchmarks [2, 65, 94, 60].

Moreover, the COVID pandemic [63] has led to the wide adoption of online video calls. The increasing reliance on video calls provides opportunities for new impersonation attacks by fraudsters using the advanced real-time DeepFakes [29]. Video conferencing DeepFakes poses new challenges to detection methods, which have to run in real-time as a video call is ongoing. More recently, the video conferencing DeepFakes detection methods (Figure 8) are also developed [39], the topic of video conferencing DeepFakes detection has not been covered in the previous surveys.

In summary, DeepFakes detection is a broad topic. Although GAN-face detection is often included in related surveys, it may lack a detailed analysis of the methods used to detect GAN-generated faces. The in-depth analysis and discussion of the GAN-face detection methods in our survey, such as analyzing the explainable features that differentiate real and fake faces, can help researchers develop more effective detection techniques. By further investigating the GAN-face detection methods, we can better understand the challenges and limitations of DeepFakes detection and develop more accurate and reliable methods for detecting GAN-based DeepFakes.

## 6 Future Directions

After reviewing existing methods of GAN-face detection with identified advantages and limitations, we next discuss future research directions that are promising for developing forensic algorithms that will be more effective, interpretable, robust, and extensible.

### 6.1 Against the Evolution of GAN models

Although the existing GAN models can not generate perfect fake faces due to known vulnerabilities, more powerful GAN models are under active development and certainly will come out in the near future. We anticipate that the known artifacts of GAN-faces (*e.g.* inconsistent corneal specular highlights [42], irregular pupil shapes [36], symmetry inconsistencies such as different earrings, *etc.*) can be fixed by incorporating relevant constraints to existing GAN models; however, how best to effectively enforce such constraints are still open questions. More powerful deep neural network architectures, training tricks, and larger training data will continue to push the state-of-the-art GAN models. For example, StyleGAN3 [52] presents a comprehensive overhaul of all signal processing aspects of StyleGAN2 to improve the texture and 3D modeling of the GAN-generated faces. The demands for searching for effective cues for exposing new GAN-faces and developing more powerful GAN-face detection methods continue to rise.

**Low-power demands.** In addition, computationally effective GAN-face detectors that can run on edge devices are of practical importance. Since GAN-faces can directly cause concerns and impacts regarding identities and social networks, forensic analytics should ideally be able to run on smartphones. Research on how best to migrate high FLOPS GPU models toward mobile applications has practical needs.

### 6.2 How to Develop Good Interpretation Methods

One critical disadvantage of many GAN-face detection methods is that they do not afford interpretability for the predicted results. Methods based on the widely-used attention mechanism [38] can not provide an interpretable explanation of the prediction results. Although the attention heat map highlights pixels that the network predicts, the mechanism can not tell *why* these pixels are selected that improves performance. Furthermore, although the current physical [42] and physiological-based methods [36] can provide interpretability of their predicted results, their assumptions are *per-cue* based (such as the iris or pupil inconsistencies) that might not be extensible to future GAN models that are specifically designed. How best to develop an end-to-end mechanism that can effectively leverage physical and physiological cues for GAN-face detection is still an open research question.

**Learning multiple cues.** From the numerous GAN-detection methods being surveyed, we observe that methods depending on a single cue or a few cues cannot retain performance, extensibility, and explainability at a time when dealing with complex real-world challenges such as occlusions and noisy data. It is difficult for features drawn from a single cue to cover multiple characteristics or artifacts. So how best to improve the generalization of the learning system, and how best to integrate or fuse the learning of multiple cues into a unified framework will be the key. Ensemble learning [99], multi-model/task learning and knowledge distillation [33] are directions that future GAN-face detection models can benefit.

### 6.3 Robust to Adversarial Attack

As DNNs are widely used in GAN-face detection (either as a component or as the main model), DNNs are known to be vulnerable against *adversarial attacks*, which are based on intentionally designed perturbations or noises that are particularly effective and harmful to the DNNs. With the increasing effectiveness of adversary attack technologies [41], research efforts start to focus on attacking fake face detectors particularly instead of focusing on general classifiers. Anti-forensics methods for evading fake detection via adversarial perturbations have been studied including [12, 28]. These methods of attacking fake image detectors usually generate adversarial perturbations to perturb almost the entire image, which is redundant and can increase the perceptibility of perturbations. [64] introduced a sparse attacking method called Key Region Attack to disrupt the fake image detection by determining key pixels to make the fake image detector only focus on these pixels. Their adversarial perturbation appears only on key regions and is hard for humans to distinguish. In general, future GAN-face detection methods need to be cautious in dealing with adversary attacks.

### 6.4 Imbalanced Distribution of Data

In the real world, real faces usually significantly outnumber GAN-generated faces in online applications. The data distribution for GAN-face detection is very imbalanced. Thus, the performance of GAN-face detection methods trained on balanced datasets may degrade when used for real-world applications, *e.g.* high accuracy but low sensitivity for spotting GAN-faces in practice.

As an initial effort, the method of [38, 96] addresses the imbalance learning issues by maximizing the ROC-AUC via approximation and relaxation of the AUC using Wilcoxon-Mann-Whitney (WMW) statistics. Experimental results showed the robustness of the model learned from imbalanced data. Looking forward, how best to deal with learning from extremely imbalanced data in real-world settings is an open question.

### 6.5 Handling Mixtures with Other Fake Faces

As face image tampering technologies continue to develop, including Face Morphing [86], Face swapping [93], Diffusion synthesized faces [7], *etc.*, GAN-face detection forensics should be robust enough to deal with the mixture of face faking or synthesis methods. In addition to the *detection* of GAN-faces, the *attribution* (find out what tools were used in the generation and the source where the faces come from) and *characterization* (find out the purpose of the generation and if the intention is malicious) are with growing importance. The DARPA Semantic Forensic (SemaFor) program https://www.darpa.mil/program/semantic-forensics of the U.S. is an ongoing effort that addresses these issues.

## 7 Conclusion

This paper presents a comprehensive, up-to-date review of GAN-face detection methods. We have reviewed the state-of-the-art models from multiple perspectives as well as provided details of major approaches. Although GAN-face detection has made notable progress recently, there is still significant room for improvement. Detecting GAN-faces in real-world settings remains challenging and with high demand, and we have discussed future research directions. The surveyed techniques and cues can also benefit the detection of other fake face-generation tools such as face morphing and swapping.

## Acknowledgements

This work is supported by the National Science Foundation, Project SaTC-2153112.

## References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka, 'Image2StyleGAN++: How to edit the embedded images?', in *CVPR*, (2020).
[2] Enes Altuncu, Virginia NL Franqueira, and Shujun Li, 'Deepfake: Definitions, performance metrics and standards, datasets and benchmarks, and a meta-review', *arXiv preprint arXiv:2208.10913*, (2022).
[3] Sarah Barrington and Hany Farid, 'A comparative analysis of human and ai performance in forensic estimation of physical attributes', *Scientific Reports*, **13**(1), 4784, (2023).
[4] David Bau, Hendrik Strobelt, William Peebles, et al., 'Semantic photo manipulation with a generative image prior', *ACM TOG*, (2019).
[5] Kratika Bhagtani, Amit Kumar Singh Yadav, Emily R Bartusiak, Ziyue Xiang, Ruiting Shao, Sriram Baireddy, and Edward J Delp, 'An overview of recent work in media forensics: Methods and threats', *arXiv preprint arXiv:2204.12067*, (2022).
[6] Matyáš Boháček and Hany Farid, 'Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms', *Proceedings of the National Academy of Sciences*, **119**(48), e2216035119, (2022).
[7] Matyáš Boháček and Hany Farid, 'A geometric and photometric exploration of gan and diffusion synthesized faces', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 874–883, (2023).
[8] Andrew Brock, Jeff Donahue, and Karen Simonyan, 'Large scale GAN training for high fidelity natural image synthesis', *arXiv:1809.11096*, (2018).
[9] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang, 'End-to-end reconstruction-classification learning for face forgery detection', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113–4122, (2022).
[10] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman, 'VGGFace2: A dataset for recognising faces across pose and age', in *FG*. IEEE, (2018).
[11] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun, 'A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt', *arXiv preprint arXiv:2303.04226*, (2023).
[12] Nicholas Carlini and Hany Farid, 'Evading deepfake-image detectors with white-and black-box attacks', in *CVPRW*, (2020).
[13] Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu, 'SofGAN: A portrait image generator with dynamic styling', *ACM transactions on graphics*, (2021).
[14] Beijing Chen, Xingwang Ju, et al., 'Locally GAN-generated face detection based on an improved Xception', *Information Sciences*, **572**, (2021).
[15] Beijing Chen, Xin Liu, Yuhui Zheng, et al., 'A robust GAN-generated face detection method based on dual-color spaces and an improved Xception', *TCSVT*, (2021).
[16] Beijing Chen, Weijin TAN, et al., 'Distinguishing between natural and gan-generated face images by combining global and local features', *Chinese Journal of Electronics*, (2022).
[17] Bowen Cheng, Ross Girshick, Piotr Dollár, et al., 'Boundary IoU: Improving object-centric image segmentation evaluation', in *CVPR*, (2021).
[18] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin, 'Fakecatcher: Detection of synthetic portrait videos using biological signals', *IEEE transactions on pattern analysis and machine intelligence*, (2020).
[19] L Minh Dang, Syed Ibrahim Hassan, et al., 'Deep learning based computer generated face identification using convolutional neural network', *Applied Sciences*, (2018).
[20] Duc-Tien Dang-Nguyen et al., 'Raise: A raw images dataset for digital image forensics', in *ACM multimedia systems conference*, (2015).
[21] Debayan Deb, Xiaoming Liu, and Anil K Jain, 'Unified detection of digital and physical face attacks', in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–8. IEEE, (2023).
[22] Nhu-Tai Do, In-Seop Na, and Soo-Hyung Kim, 'Forensics face detection from GANs using convolutional neural network', in *ISITC*, (2018).
[23] Chengdong Dong, Ajay Kumar, and Eryun 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*, pp. 7865–7874, (2022).
[24] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al., 'Art and the science of generative ai', *Science*, **380**(6650), 1110–1111, (2023).
[25] Jianwei Fei, Yunshu Dai, Peipeng Yu, Tianrun Shen, Zhihua Xia, and Jian Weng, 'Learning second order local anomaly for general face forgery detection', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20270–20280, (2022).
[26] Tao Fu, Ming Xia, and Gaobo Yang, 'Detecting gan-generated face images via hybrid texture and sensor noise based features', *Multimedia Tools and Applications*, (2022).
[27] Yong Fu, Tanfeng Sun, Xinghao Jiang, Ke Xu, and Peisong He, 'Robust GANs-face detection based on dual-channel CNN network', in *CISP-BMEI*. IEEE, (2019).
[28] Apurva Gandhi and Shomik Jain, 'Adversarial perturbations fool deepfake detectors', in *IJCNN*, (2020).
[29] Candice R. Gerstner and Hany Farid, 'Detecting real-time deep-fake videos using active illumination', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 53–60, (June 2022).
[30] Michael Goebel, Lakshmanan Nataraj, Tejaswi Nanjundaswamy, et al., 'Detection, attribution and localization of GAN generated images', *arXiv:2007.10466*, (2020).
[31] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola, 'GANalyze: Toward visual definitions of cognitive image properties', in *ICCV*, (2019).
[32] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al., 'Generative adversarial nets', in *NeurIPS*, (2014).
[33] Jianping Gou, Baosheng Yu, et al., 'Knowledge distillation: A survey', *IJCV*, (2021).
[34] Diego Gragnaniello, Davide Cozzolino, et al., 'Are GAN-generated images easy to detect? a critical analysis of the state-of-the-art', in *ICME*, (2021).
[35] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi, 'Exploiting fine-grained face forgery clues via progressive enhancement learning', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 735–743, (2022).
[36] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu, 'Eyes tell all: Irregular pupil shapes reveal GAN-generated faces', *ICASSP*, (2022).
[37] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu, 'Open-eye: An open platform to study human performance on identifying ai-synthesized faces', *MIPR*, (2022).
[38] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu, 'Robust attentive deep neural network for exposing GAN-generated faces', *IEEE Access*, (2022).
[39] Hui Guo, Xin Wang, and Siwei Lyu, 'Detection of real-time deepfakes in video conferencing with active probing and corneal reflection', *ICASSP*, (2023).
[40] Travis Hartman and Raphael Satter, 'These faces are not real', in *https://tmsnrt.rs/3rsneCO*, (2020).
[41] Shu Hu, Lipeng Ke, Xin Wang, and Siwei Lyu, 'Tkml-ap: Adversarial attacks to top-k multi-label learning', in *ICCV*, pp. 7649–7657, (2021).
[42] Shu Hu, Yuezun Li, and Siwei Lyu, 'Exposing GAN-generated faces using inconsistent corneal specular highlights', in *ICASSP*. IEEE, (2021).
[43] Nils Hulzebosch, Sarah Ibrahimi, and Marcel Worring, 'Detecting CNN-generated facial images in real-world scenarios', in *CVPR Workshops*, (2020).
[44] Abdul Jabbar, Xi Li, and Bourahla Omar, 'A survey on generative adversarial networks: Variants, applications, and training', *ACM Computing Surveys*, (2021).
[45] Ali Jahanian, Lucy Chai, and Phillip Isola, 'On the "steerability" of generative adversarial networks', in *ICLR*, (2019).
[46] Hyeonseong Jeon, Youngoh Bang, Junyaup Kim, and Simon S Woo, 'T-GD: Transferable GAN-generated images detection framework', *ICML*, (2020).
[47] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi, 'Frepgan: robust deepfake detection using frequency-level perturba-

tions', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1060–1068, (2022).
[48] Wei Jiang and Wentao Dong, 'Facke: a survey on generative models for face swapping', *arXiv preprint arXiv:2206.11203*, (2022).
[49] Micah K. Johnson and Hany Farid, 'Exposing digital forgeries through specular highlights on the eye', in *Information Hiding*, volume 4567 of *LNCS*, (2008).
[50] Felix Juefei-Xu, Run Wang, Yihao Huang, et al., 'Countering malicious deepfakes: Survey, battleground, and horizon', *arXiv:2103.00218*, (2021).
[51] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu, 'Countering malicious deepfakes: Survey, battleground, and horizon', *International Journal of Computer Vision*, **130**(7), 1678–1734, (2022).
[52] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, 'Alias-free generative adversarial networks', *NeurIPS*, **34**, (2021).
[53] Tero Karras et al., 'Progressive growing of GANs for improved quality, stability, and variation', *ICLR*, (2018).
[54] Tero Karras, Samuli Laine, and Timo Aila, 'A style-based generator architecture for generative adversarial networks', in *CVPR*, (2019).
[55] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, 'Analyzing and improving the image quality of StyleGAN', in *CVPR*, (2020).
[56] Zahra Khanjani, Gabrielle Watson, and Vandana P Janeja, 'How deep are the fakes? focusing on audio deepfake: A survey', *arXiv preprint arXiv:2111.14203*, (2021).
[57] Ali Khodabakhsh, Raghavendra Ramachandra, et al., 'Subjective evaluation of media consumer vulnerability to fake audiovisual content', in *QoMEX*, (2019).
[58] Federica Lago, Cecilia Pasquini, Rainer Böhme, et al., 'More real than real: A study on human visual perception of synthetic faces', *arXiv:2106.07226*, (2021).
[59] Trung-Nghia Le, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen, 'Robust deepfake on unrestricted media: Generation and detection', in *Frontiers in Fake Media Generation and Detection*, 81–107, Springer, (2022).
[60] Chuqiao Li, Zhiwu Huang, Danda Pani Paudel, Yabin Wang, Mohamad Shahbazi, Xiaopeng Hong, and Luc Van Gool, 'A continual deepfake detection benchmark: Dataset, methods, and essentials', in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1339–1349, (2023).
[61] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang, 'Detection of deep network generated images using disparities in color components', *arXiv:1808.07276*, (2018).
[62] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang, 'Identification of deep network generated images using disparities in color components', *Signal Processing*, **174**, (2020).
[63] Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, et al., 'Using artificial intelligence to detect covid-19 and community-acquired pneumonia based on pulmonary ct: evaluation of the diagnostic accuracy', *Radiology*, **296**(2), E65–E71, (2020).
[64] Quanyu Liao, Yuezun Li, Xin Wang, Bin Kong, Bin Zhu, Siwei Lyu, et al., 'Imperceptible adversarial examples for fake image detection', *ICIP*, (2021).
[65] Chenhao Lin, Jingyi Deng, Pengbin Hu, Chao Shen, Qian Wang, and Qi Li, 'Towards benchmarking and evaluating deepfake detection', *arXiv preprint arXiv:2203.02115*, (2022).
[66] Xin Liu and Xiao Chen, 'A survey of GAN-generated fake faces detection method based on deep learning', *JIHPP*, **2**(2), (2020).
[67] Yunfan Liu, Qi Li, Qiyao Deng, Zhenan Sun, and Ming-Hsuan Yang, 'Gan-based facial attribute manipulation', *arXiv preprint arXiv:2210.12683*, (2022).
[68] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr, 'Global texture enhancement for fake face detection in the wild', in *CVPR*, (2020).
[69] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, 'Deep learning face attributes in the wild', in *ICCV*, (December 2015).
[70] Asad Malik, Minoru Kuribayashi, Sani M Abdullahi, and Ahmad Neyaz Khan, 'Deepfake detection for human face images and videos: A survey', *IEEE Access*, **10**, 18757–18775, (2022).
[71] Hadi Mansourifar and Weidong Shi, 'One-shot GAN generated fake face detection', *arXiv:2003.12244*, (2020).
[72] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi, 'Do GANs leave artificial fingerprints?', in *MIPR*. IEEE, (2019).
[73] Francesco Marra, Cristiano Saltori, et al., 'Incremental learning for the detection and classification of GAN-generated images', in *WIFS*. IEEE, (2019).
[74] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik, 'Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward', *Applied Intelligence*, **53**(4), 3974–4026, (2023).
[75] Falko Matern, Christian Riess, and Marc Stamminger, 'Exploiting visual artifacts to expose deepfakes and face manipulations', in *WACVW*. IEEE, (2019).
[76] Scott McCloskey and Michael Albright, 'Detecting GAN-generated imagery using color cues', *arXiv:1812.08247*, (2018).
[77] Yisroel Mirsky and Wenke Lee, 'The creation and detection of deepfakes: A survey', *ACM Computing Surveys (CSUR)*, **54**(1), 1–41, (2021).
[78] Huaxiao Mo, Bolin Chen, and Weiqi Luo, 'Fake faces identification via convolutional neural network', in *ACM IH&MMSEC*, (2018).
[79] Shivansh Mundra, Gonzalo J Aniano Porcile, Smit Marvaniya, James R Verbus, and Hany Farid, 'Exposing gan-generated profile photos from compact embeddings', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 884–892, (2023).
[80] Lakshmanan Nataraj, Tajuddin M. Mohammed, et al., 'Detecting GAN generated fake images using co-occurrence matrices', *Electronic Imaging*, (2019).
[81] Thanh Thi Nguyen, Cuong M Nguyen, et al., 'Deep learning for deepfakes creation and detection: A survey', *arXiv:1909.11573*, (2022).
[82] Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao, 'Core: Consistent representation learning for face forgery detection', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12–21, (2022).
[83] Ahmad Nickabadi, Maryam Saeedi Fard, Nastaran Moradzadeh Farid, and Najmeh Mohammadbagheri, 'A comprehensive survey on semantic facial attribute editing using generative adversarial networks', *arXiv preprint arXiv:2205.10587*, (2022).
[84] Sophie Nightingale et al., 'Synthetic faces: how perceptually convincing are they?', *Journal of Vision*, (2021).
[85] Sophie Nightingale and Hany Farid, 'Synthetic faces are more trustworthy than real faces', *Journal of Vision*, **22**(14), 3068–3068, (2022).
[86] Sophie J Nightingale, Shruti Agarwal, and Hany Farid, 'Perceptual and computational detection of face morphing', *Journal of Vision*, **21**(3), (2021).
[87] Sophie J Nightingale and Hany Farid, 'Ai-synthesized faces are indistinguishable from real faces and more trustworthy', *Proceedings of the National Academy of Sciences*, (2022).
[88] Ehsan Nowroozi, Mauro Conti, and Yassine Mekdad, 'Detecting high-quality gan-generated face images using neural networks', *arXiv preprint arXiv:2203.01716*, (2022).
[89] Donie O'Sullivan, 'A high school student created a fake 2020 US candidate. Twitter verified it', in *CNN Business, https://cnn.it/3HpHfzz*, (2020).
[90] Donie O'Sullivan, 'How fake faces are being weaponized online', in *CNN Business*, (2020).
[91] Samay Pashine, Sagar Mandiya, Praveen Gupta, and Rashid Sheikh, 'Deep fake detection: Survey of facial manipulation detection solutions', *arXiv preprint arXiv:2106.12605*, (2021).
[92] Leandro A Passos, Danilo Jodas, Kelton AP da Costa, Luis A Souza Júnior, Danilo Colombo, and João Paulo Papa, 'A review of deep learning-based approaches for deepfake content detection', *arXiv preprint arXiv:2202.06095*, (2022).
[93] Ivan Perov, Daiheng Gao, et al., 'Deepfacelab: A simple, flexible and extensible face swapping framework', *arXiv:2005.05535*, (2020).
[94] Minh Tam Pham et al., 'A dual benchmarking study of facial forgery and facial forensics', *arXiv preprint arXiv:2111.12912*, (2021).
[95] Erion-Vasilis Pikoulis, Zafeiria-Marina Ioannou, et al., 'Face morphing, a modern threat to border security: Recent advances and open challenges', *Applied Sciences*, (2021).
[96] Wenbo Pu, Jing Hu, Xin Wang, Yuezun Li, Shu Hu, et al., 'Learning a deep dual-level network for robust deepfake detection', *Pattern Recognition*, (2022).
[97] Alec Radford, Luke Metz, et al., 'Unsupervised representation learning with deep convolutional generative adversarial networks',

*arXiv:1511.06434*, (2015).
[98] Sushma Venkatesh Raghavendra Ramachandra Kiran Raja and Christoph Busch, 'Face morphing attack generation & detection: A comprehensive survey'.
[99] Omer Sagi and Lior Rokach, 'Ensemble learning: A survey', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **8**(4), (2018).
[100] Iman Saleem and Bahja Khudair Shukr, 'Techniques and challenges for generation and detection face morphing attacks: A survey', *Iraqi Journal of Science*, 385–404, (2023).
[101] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt, and Christoph Busch, 'Face recognition systems under morphing attacks: A survey', *IEEE Access*, **7**, 23012–23026, (2019).
[102] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou, 'Interpreting the latent space of GANs for semantic face editing', in *CVPR*, (2020).
[103] Kaede Shiohara and Toshihiko Yamasaki, 'Detecting deepfakes with self-blended images', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18720–18729, (2022).
[104] Karen Simonyan and Andrew Zisserman, 'Very deep convolutional networks for large-scale image recognition', *arXiv:1409.1556*, (2014).
[105] P Swathi and Saritha Sk, 'Deepfake creation and detection: A survey', in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 584–588. IEEE, (2021).
[106] Ruben Tolosana, Ruben Vera-Rodriguez, et al., 'Deepfakes and beyond: A survey of face manipulation and fake detection', *Information Fusion*, **64**, (2020).
[107] James Vincent, 'A spy reportedly used an AI-generated profile picture to connect with sources on LinkedIn', in *https://bit.ly/35BU215*, (2020).
[108] Jun Wang, Benedetta Tondi, and Mauro Barni, 'An eyes-based siamese neural network for the detection of gan-generated face images', *Frontiers in Signal Processing*, (2022).
[109] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, et al., 'Fakespotter: A simple yet robust baseline for spotting AI-synthesized fake faces', *IJCAI*, (2020).
[110] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros, 'CNN-generated images are surprisingly easy to spot... for now', in *CVPR*, (2020).
[111] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang, 'GAN inversion: A survey', *arXiv:2101.05278*, (2021).
[112] Xin Yang, Yuezun Li, Honggang Qi, and Siwei Lyu, 'Exposing GAN-synthesized faces using landmark locations', in *ACM Workshop on IHMMSec*, (2019).
[113] Ning Yu, Larry S Davis, and Mario Fritz, 'Attributing fake images to GANs: Learning and analyzing GAN fingerprints', in *ICCV*, (2019).
[114] Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu, 'A survey on deepfake video detection', *Iet Biometrics*, **10**(6), 607–624, (2021).
[115] Tao Zhang, 'Deepfake generation and detection, a survey', *Multimedia Tools and Applications*, **81**(5), 6259–6276, (2022).
[116] Sm Zobaed, Fazle Rabby, et al., 'Deepfakes: Detecting forged and synthetic media content using machine learning', *Artificial Intelligence in Cyber Security*, 177–201, (2021).