International Journal of Computer Vision (2022) 130:1678–1734
https://doi.org/10.1007/s11263-022-01606-8

Countering Malicious DeepFakes: Survey, Battleground, and Horizon

Felix Juefei-Xu[1] · Run Wang[2] · Yihao Huang[3] · Qing Guo[4,5] · Lei Ma[6] · Yang Liu[5,7]

Received: 27 February 2021 / Accepted: 11 March 2022 / Published online: 4 May 2022

Abstract

The creation or manipulation of facial appearance through deep generative approaches, known as *DeepFake*, have achieved significant progress and promoted a wide range of benign and malicious applications, e.g., visual effect assistance in movie and misinformation generation by faking famous persons. The evil side of this new technique poses another popular study, i.e., *DeepFake detection* aiming to identify the fake faces from the real ones. With the rapid development of the DeepFake-related studies in the community, both sides (i.e., DeepFake generation and detection) have formed the relationship of battleground, pushing the improvements of each other and inspiring new directions, e.g., the evasion of DeepFake detection. Nevertheless, the overview of such battleground and the new direction is unclear and neglected by recent surveys due to the rapid increase of related publications, limiting the in-depth understanding of the tendency and future works. To fill this gap, in this paper, we provide a comprehensive overview and detailed analysis of the research work on the topic of DeepFake generation, DeepFake detection as well as evasion of DeepFake detection, with more than 318 research papers carefully surveyed. We present the taxonomy of various DeepFake generation methods and the categorization of various DeepFake detection methods, and more importantly, we showcase the battleground between the two parties with detailed interactions between the adversaries (DeepFake generation) and the defenders (DeepFake detection). The battleground allows fresh perspective into the latest landscape of the DeepFake research and can provide valuable analysis towards the research challenges and opportunities as well as research trends and future directions. We also elaborately design interactive diagrams (http://www.xujuefei.com/dfsurvey) to allow researchers to explore their own interests on popular DeepFake generators or detectors.

Keywords DeepFake Generation · DeepFake Detection · Face · Misinformation · Disinformation · DeepFakes

Communicated by Jian Sun.

Project lead: Felix Juefei-Xu.

✉ Run Wang
wangrun@whu.edu.cn

✉ Qing Guo
tsingqguo@ieee.org

[1] Alibaba Group, Sunnyvale, CA, USA

[2] Key Laboratory of Aerospace Information Security and Trust Computing, School of Cyber Science and Engineering, Wuhan University, Wuhan, China

[3] East China Normal University, Shanghai, China

[4] College of Intelligence and Computing, Tianjin University, Tianjin, China

[5] Nanyang Technological University, Singapore, Singapore

[6] Alberta Machine Intelligence Institute (AMII), University of Alberta, Edmonton, AB, Canada

[7] Zhejiang Sci-Tech University, Hangzhou, China

1 Introduction

> If you know the enemy and know yourself, you need not fear the result of a hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat. If you know neither the enemy nor yourself, you will succumb in every battle.
> *The Art of War*
> *Sun Tzu*

Ever since digital visual media came along, there has always been a need to manipulate them for various purposes. Usually, such digital media manipulation requires domain expertise and is quite time and effort consuming, such as using professional software like Adobe Photoshop (Adobe 2021c) for editing a photograph, or Adobe Lightroom (Adobe 2021b) for retouching it. In the sound and voice domain, similar professional software is available for carrying out various types of signal manipulation such as using Adobe Audition (Adobe 2021a), or Auto-Tune (Antares Audio Technologies

2021), etc. In the domain of motion pictures, the manipulation of videos oftentimes require very sophisticated theatrical visual effects (VFX) in the post-processing, and when it comes to recreating animated faces with realistic facial muscle movements and expressions, motion capture techniques with the help of high-speed tracking of markers are usually adopted, such as in James Cameron's Avatar (IMDb 2021) movie.

With the advances of deep generative models such as autoregressive models (Van Oord et al. 2016; Van den Oord et al. 2016), variational autoencoders (VAE) (Kingma and Welling 2013, 2019), normalizing flow models (Rezende and Mohamed 2015), and generative adversarial networks (GAN) (Goodfellow et al. 2014a), anyone can now produce a realistically looking face whose identity does not exist in the world, or perform facial manipulations, such as identity swap, in a video with a high level of realism. The AI- or deep learning-based face image and video manipulation is what the community refers to as the DeepFake. In contrast, the notion of CheapFake is recently coined to encompass non-AI ("cheap") manipulations of multimedia content (Aneja et al. 2021). The low barriers to entry and wide accessibility of pre-trained high-performance DeepFake generator are what the problem is. DeepFake, when used maliciously, is a pressing and tangible threat to the integrity of media information available to us.

In this survey, we follow the widely adopted conventions and define DeepFake as *the creation and the manipulation of facial appearance (attributes, identity, expression) through deep generative approaches*, and it can be classified into the following 4 categories: (1) entire face synthesis, (2) attribute manipulation, (3) identity swap, and (4) expression swap (*aka* reenactment), as depicted in Fig. 1. Here, the facial attributes (such as hair color, facial hair, skin tone, eyewear, etc.) exclude identity and expression as attributes. DeepFake manipulations may exhibit different risk levels and the risk level highly depends on the type of specific applications and somewhat subjectively depending on the actual use case. Here, we provide some examples that may pose risks based on different DeepFake categories. The identity swap, altering the hard biometrics (Jain et al. 2007) (pertains to identity information) of a subject, poses risks to a wide range of safety-critical scenarios since the identity information is tampered. The expression swap and attribute manipulation, although only tampers with the soft biometrics (Jain et al. 2007) (pertains to facial attributes) of the subject, may also pose risks to certain applicable scenarios where people can easily verify the identity of the subject, but not what she/he says, such as in political elections, e.g., the DeepFaked Obama video (NPR 2020). Generically speaking, the entire face synthesis may seem to pose a lower risk than the three categories mentioned above since it is not based on the manipulation of hard or soft biometrics of the subject. However, depending on the applications, even the entire face synthesis can become risky and troublesome, imagining the swarm of fake accounts on social media platforms. In short, all four aforementioned DeepFake modalities can potentially pose high level of risks and need to be addressed properly. In terms of the popularity ranking measured by how often the categories are tested by DeepFake detectors according to the surveyed literature, the identity swap ranks the highest for the popularity, and the expression swap is the least attempted category. For attribute manipulation and entire face synthesis, usually the DeepFake generation process does not require a target face, and one can tune the desired facial attributes through adjusting the latent vector during the deep generative modeling. For identity swap, the target can either be a video sequence or simply a single face image, with the former renders better swap results. For expression swap, the target is usually in the form of a video sequence. Although it is technically manageable to use just a single face image as the target, the result will look weird since the expression won't be changing throughout the entire DeepFake video. From top to bottom, the four panels in Fig. 1 illustrate the four categories of DeepFakes. Four examples are shown for both 'identity swap' and 'expression swap', with each example associated with a target, real, and DeepFake sequences of 5 frames. Within each panel, the two examples in the top row show the DeepFake manipulation that is pretty subtle, which demonstrate the minuscule manipulations that some DeepFakes can present, and the two examples in the bottom row show more drastic DeepFake manipulations. Across 'identity swap' and 'expression swap', as a comparison, one example is shown in both scenarios and is highlighted by ●, to showcase the difference in the DeepFake frames for these two modalities coming from the same 'target' and 'real' sources. Readers are encouraged to zoom in on the image. Actual full-resolution videos are available on the project website (http://www.xujuefei.com/dfsurvey) to better illustrate the DeepFake phenomenon. For 'attribute manipulation' and 'entire face synthesis' on the bottom panel, both real and DeepFakes are shown.

The DeepFake technology itself, in our opinion, is neutral and can be used both benignly and maliciously. We first discuss some of the positive and benign uses of DeepFake technology. For example, Synthesia (2021) uses DeepFake technology to provide cost-effective synthetic training videos for companies during the Covid-19 pandemic when it is getting harder and more expensive to shoot corporate training videos with real actors. On the same line of thought, there has been an increase in a digital avatar or virtual assistance by means of DeepFake technology (Pinscreen 2021) to be used for e.g., video conferencing scenarios (Wang et al. 2021). DeepFake can also be used for assisting the facial visual effects in movie and TV show production for re-creating a role appearance for some celebrities that may have passed

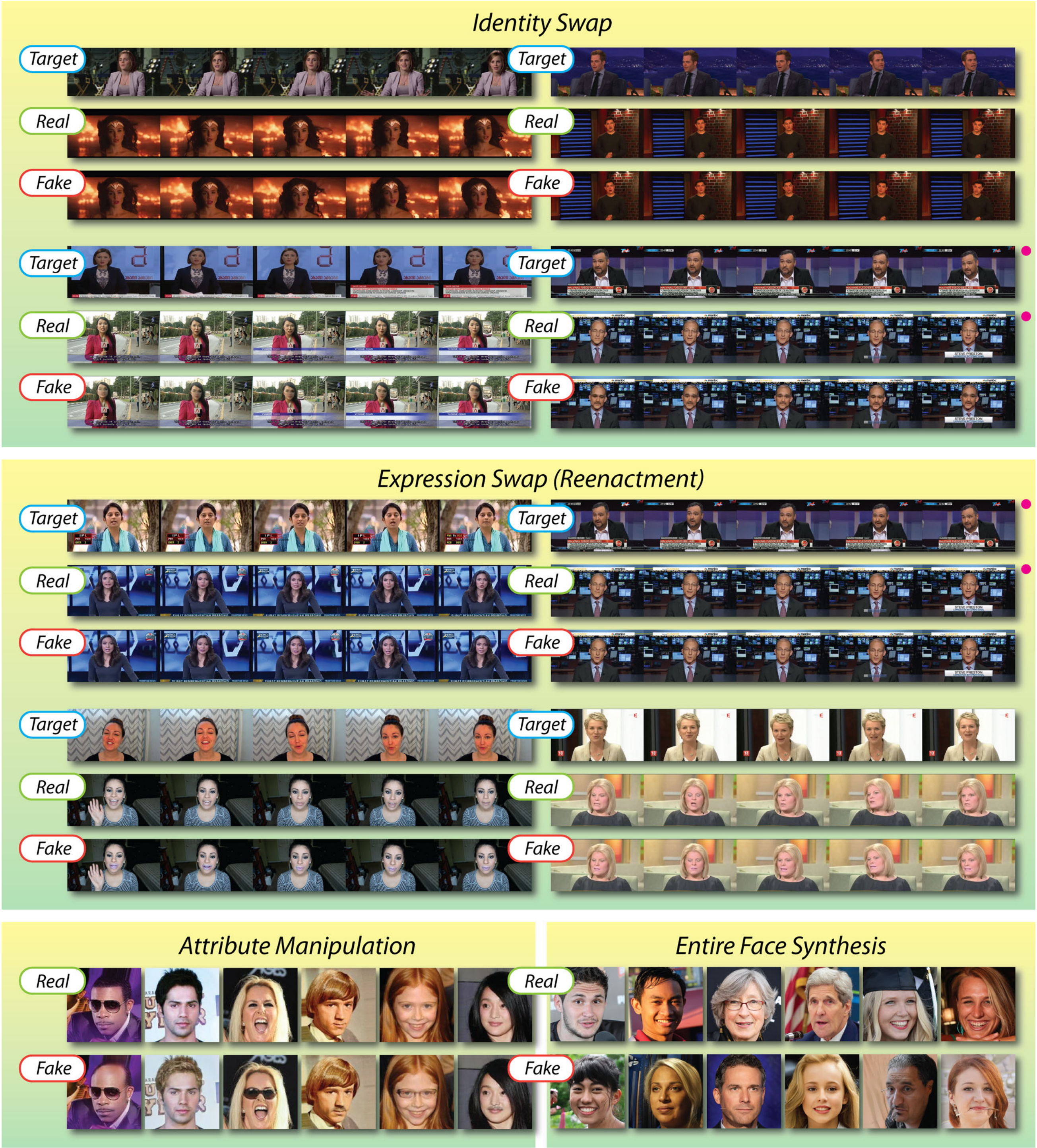


Fig. 1 From top to bottom, the four panels illustrate the four categories of DeepFakes. Four examples are shown for both 'identity swap' and 'expression swap', with each example associated with a target, real, and DeepFake sequences of 5 frames. Within each panel, the two examples in the top row show the DeepFake manipulation that is pretty subtle, which demonstrate the minuscule manipulations that some DeepFakes can present, and the two examples in the bottom row show more drastic DeepFake manipulations. Across 'identity swap' and 'expression swap', as a comparison, one example is shown in both scenarios and is highlighted by ●, to showcase the difference in the DeepFake frames for these two modalities coming from the same 'target' and 'real' sources. Readers are encouraged to zoom in on the image. Actual full-resolution videos are available on the project website (http://www.xujuefei.com/dfsurvey) to better illustrate the DeepFake phenomenon. For 'attribute manipulation' and 'entire face synthesis' on the bottom panel, both real and DeepFakes are shown. In terms of popularity as being attempted by DeepFake detectors according to the survey results, the ranking is as the following: identity swap > entire face synthesis > attribute manipulation ≫ expression swap

away, or for paying tribute to the lost ones in a memorial concert. Sometimes, creative scenes can be made by using DeepFake to join together celebrities across geographic and generation boundaries. For example, Pinscreen (2021) has used DeepFake technology to bring Prime Minister Mark Rutte, and Queen Máxima of The Netherlands, as well as other Dutch celebrities to a live TV broadcast. We have also seen a surge in popularity of DeepFake being used in consumer smartphone applications for everyday entertainment purposes, especially targeted for making viral videos on social media platforms, such as Zao (2021), Reface (Reface 2021), Facebrity (Facebrity 2021), etc.

The aforementioned cases are benign DeepFakes. Powerful as the technology is, there is a thin line between good and evil depending on what the content is, as well as what the intent is, and it is easy to cross over. DeepFake can be maliciously capitalized by bad actors to cause real harm. For example, when it is applied to politicians and fueled with targeted misinformation or disinformation, it can really sway people's opinions and can lead to detrimental outcomes such as manipulated and interfered election without people even knowing about it. To this date, many politicians and state leaders have been misrepresented by DeepFakes. Recently in September of 2020, DeepFake videos featuring Russian President Vladimir Putin and North Korean leader Kim Jong-un appeared on social media stating the message that "America doesn't need any election interference from them; it will ruin its democracy by itself". This was a campaign put forward by the nonpartisan advocacy group RepresentUs to protect voting rights during the then upcoming US presidential election, and the goal of the videos is to shock Americans into understanding the fragility of democracy (MIT TR 2020). In April 2018, a DeepFake of Barack Obama was created by comedian Jordan Peele in collaboration with BuzzFeed which served as a public service announcement (PSA) to increase awareness of DeepFakes (BuzzFeed 2018). During the most recent Christmas holidays, a DeepFake Queen Elizabeth II was shown dancing across TV screens as part of a British broadcaster's warning against the proliferation of misinformation (CNN 2020). Not just world leaders, celebrities or even average people can also fall victims of malicious DeepFakes. In fact, some of the earliest infamous use cases of malicious DeepFakes have been DeepFake pornography often featuring female celebrities. Financial fraud as a result of DeepFake or looming fake accounts on social media platforms created realistically by using DeepFakes are all examples of malicious use of DeepFake technology.

Lawmakers and governing bodies across the world are responding to the proliferation of malicious DeepFakes with new policies, regulations, and laws. Take the United States for example, in December 21, 2018, US Senator Ben Sasse introduced a bill "S.3805—Malicious Deep Fake Prohibition Act of 2018" (115th Congress 2018) that "establishes a new criminal offense related to the creation or distribution of fake electronic media records that appear realistic". In June 12, 2019, US Congresswoman Yvette Clarke introduced the "H.R.3230 Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019", also known as the "DEEP FAKES Accountability Act". The act aims at "combating the spread of disinformation through restrictions on deep-fake video alteration technology" (116th Congress 2019a). In July 9, 2019, US Senator Rob Portman introduced the bill "S.2065—Deepfake Report Act of 2019" (116th Congress 2019b) that requires "the Science and Technology Directorate in the Department of Homeland Security to report at specified intervals on the state of digital content forgery technology. Digital content forgery is the use of emerging technologies, including artificial intelligence and machine learning techniques, to fabricate or manipulate audio, visual, or text content with the intent to mislead". State-wide legislature entities have also responded by proposing counter-measures of DeepFakes such as the "Nonconsensual pornography law" in Virginia (The Verge 2019b), a law "to criminalize publishing and distributing DeepFake videos intended to harm a candidate or influence results within 30 days of an election" in Texas (Texas 2019), and a similar law in California (California 2019). Mirroring this new California law on political ads, the Chinese government "makes it a criminal offense to publish deepfakes or fake news without disclosure" (The Verge 2019a). Many more countries followed suit.

Social media platforms are also actively taking measures to tackle synthetic and manipulated media on their platform. For example, Twitter signals viewers that a tweet contains manipulated media content such as DeepFakes by placing a tag on the tweet and providing a link to credible news articles debunking the hoax (Twitter Blog 2019). Facebook fosters the development of high-performance DeepFake detection tools by hosting the DeepFake Detection Challenge (DFDC) (Dolhansky et al. 2020) in December 2019 with 2114 worldwide participants generating more than 35,000 models.

In the computer vision community, the study of DeepFake has certainly gained traction in recent years. Figure 2 shows the year-by-year number of papers on the topic of DeepFakes from its inception in 2016, and we will detail the paper collection schema in Sect. 2. As shown in Fig. 2, around 78% of the papers appeared in the last two years, indicating the trending research interests revolved around the topic of DeepFakes.

This paper seeks to further raise the awareness of the danger of the emerging DeepFake technology by surveying the state-of-the-art DeepFake literature. We organize the literature according to three aspects: the generation of DeepFakes in Sect. 3 (such as the aforementioned four main DeepFake categories, other generation methods not covered, and datasets), detection of DeepFakes in Sect. 4 (such as methods based on spatial, frequency, and biological cues, as well

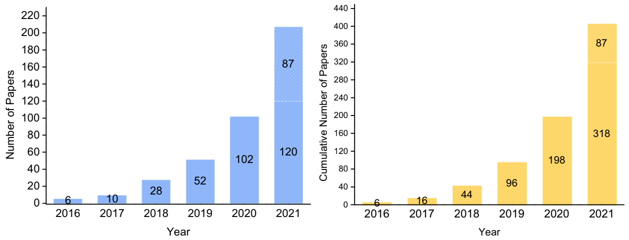


Fig. 2 This figure shows a growing trend in the number of papers in the DeepFake field in recent years. The papers are collected according to the criteria introduced in Sect. 2.1, including arXiv, conference, and journal articles. They are categorized by the year of the last updated version. (L) The number of published DeepFake related papers year over year since its inception in 2016. (R) The cumulative number of published DeepFake related papers year over year since its inception in 2016. It shows that over 78% papers were published in the last 2 years. The bars with shadow are the projected number of published papers

as other detection methods not covered), and the evasion of DeepFake detection in Sect. 6. More importantly, we present the battleground between DeepFake generation methods and DeepFake detection methods in Sect. 5 where we analyze the tightly-knit interactions between the two parties and how the later DeepFake detectors outcompete the earlier ones, supplemented by statistics and analyses through large-scale visualizations. In Sect. 7, we recapitulate the findings after surveying each DeepFake-related topics and identify existing challenges and research opportunities while being forward-looking with discussions on the horizons and how the future generation technology surrounding DeepFakes may evolve into.

1.1 Importance of the Survey of the Battleground

Unlike many other sub-fields of computer vision, the emerging domain of DeepFake generation and detection is, by nature, competitive. The rapid development of the defenders (i.e., DeepFake detectors) are further accelerated by the push from the adversaries (i.e., DeepFake generators), and vice versa. When a novel method, say from the DeepFake generator side, is being developed, the authors will naturally want their method to penetrate some latest DeepFake detectors, and the resulting generator, will later be attempted to be blocked by some newer detectors down the line. Just within last year or two, we have seen huge progresses made alternately on both sides of the battleground, both aiming to outcompete the other side.

That is exactly why understanding each individual DeepFake generation and detection method, albeit important, may not be enough, and it does not tell the whole story. The collective understanding of the interplay and the interactivity of the methods both within each side and across two sides will bring fresh and clearer view of the landscape, and will bring new knowledge and insight for the development of the next-generation defenders and adversaries.

For these very reasons mentioned above, we have taken the effort to survey, extract, tabulate, and finally construct the battleground landscape between the DeepFake generation methods and DeepFake detection methods that have been proposed so far. By analyzing which of the DeepFake generation methods are attempted by each of the DeepFake detection method from the battleground, we are able to present to the readers the visualization of one aspect of the battleground using a Sankey diagram in Fig. 11 with interactive diagrams available on the project website for better interactive probing. Readers can simply select one node on either side of the battleground, and all the highlighted paths will connect the corresponding opponents on the other side. Similarly, by analyzing which baseline DeepFake detection methods that each detector has benchmarked against, as observed from the battleground, we are able to showcase and trace back the technical evolution of each detector using a chord diagram in Fig. 14 with interactive diagrams available on the project website. Readers can select one node on the interactive chord diagram, and the highlighted paths will connect all the corresponding baseline detection methods benchmarked by the selected detector.

Some collective knowledge are quite unique to the battleground and are not conspicuous if we only analyze individual methods. For example, from the battleground landscape in Figs. 11 and 14, we can identify important trends on both sides, as well as algorithmic hot zones where seminal papers are indicated by busy nodes. We can also locate and discover where the major battles are fought, which nodes are the uprising feuds and which nodes are becoming obsolete algorithmically. With the color-coded paths indicating various types of detection methods (i.e., spatial-based, frequency-based, biological signal-based, etc.), we are able to provide an insightful understanding of what types of detection methods are being attempted on which particular generation methods. From the battleground, we are able to systemically extract the performance scores of each DeepFake detection method on every generation method it has evaluated on. Of course, the same generation method will be attempted by multiple detectors, each with a detection accuracy scores. By knitting the entire network and sorting the rankings, we are able to provide some strength measurement for each DeepFake generation method by means of Elo rating (Wikipedia 2021a), as can be seen in Table 6. Maybe more surprisingly, we may be able to identify the paths in the woods that are less traveled. From a practical point of view, if a practitioner is just entering the field, we are quite confident that the battleground presented in this survey paper will serve as an asset to both help identify a research direction more effectively, and to help understand the status quo more comprehensively. We provide more detailed analysis in Sect. 5.

As the battleground landscape serving as the cornerstone, we strive to continue building a DeepFake survey that is both content-rich and content-distinctive especially in terms of the following important traits such as the timeliness, the scale, the detailedness of the textual, tabular, and visual presentation, the thoroughness of the technical evolution, battleground analyses, and horizon analyses.

There has been previous work that surveyed or discussed aspects of the literature on topics related to DeepFake generation and detection. However, our survey that uniquely manifest the battleground landscape between the adversaries and the defenders still stands out. Mirsky and Lee (2021) pivoted their survey to the DeepFake generation aspect with detailed model architecture charts for each individual DNN used for DeepFake generation methods the authors have surveyed, which is both informative and illustrative. However, less attention is paid to the DeepFake detection aspect, the technical evolution of both the generation and detection methods, and interplay and the battle between the two in their survey. Neekhara et al. (2021) provided a practical perspective that focuses on the adversarial threats to DeepFake detection. By studying the commonalities between various DeepFake detection methods by interpreting the model decisions using gradient-based saliency maps, the authors can create adversarial examples that are highly transferable across different DeepFake detection methods, revealing the vulnerabilities of the DeepFake detectors. Verdoliva (2020) discussed the interplay between multimedia forensics and DeepFakes. From the visual media integrity verification point of view, the authors have provided a detailed discussion on how the conventional and modern media forensics methods are conducted for general-purpose image and video manipulation. Later, they discuss how some of the methods can be applied to DeepFake detection. With the majority of the surveyed papers being conventional media forensics approaches, the overlapping is insignificant. In addition, there are also a few relatable DeepFake surveys first published in late 2019 and early 2020 (Nguyen et al. 2019d; Tolosana et al. 2020; Lyu 2020). As far as we know, although these earlier surveys shared similar themes, they were still deficient in terms of the following important aspects such as timeliness, scale, and detailedness of the survey, as well as thorough analyses on the technical evolution together with discussions and analyses of the horizons. Most importantly, the battleground landscape between the DeepFake generation and detection methods were not covered in these prior surveys. Here we list some of the comparisons between ours and prior surveys in Table 1.

In summary, this paper differentiates itself from the earlier survey papers with the following unique features and important contributions:

Table 1 Comparisons with prior survey papers on DeepFake-related topics

Prior surveys	Timeliness	Scale	Detailedness	Technical evolution	Battleground analyses	Horizon analyses
Mirsky and Lee (2021)	Sep 13, 2020	193*	High	Briefly	✗	Very briefly
Verdoliva (2020)	Jan 18, 2020	265*	Medium	Briefly	✗	Briefly
Nguyen et al. (2019d)	Apr 26, 2021	162*	Medium	Briefly	✗	Briefly
Tolosana et al. (2020)	Jun 18, 2020	200*	Medium	Briefly	✗	Briefly
Lyu (2020)	Mar 11, 2020	34*	Low	Briefly	✗	Briefly
Ours	Aug 1, 2021 ✓	318† ✓	High ✓	Thoroughly ✓	✓	Thoroughly ✓

Our survey papers is more informative and competitive in terms of the timeliness, scale, and detailedness, as well as the detailed manifestation of the technical evolution, battleground analyses, and horizon analyses

*Total reference count, including auxiliary papers

† Excluding auxiliary papers

(1) *Timeliness* The field of DeepFake generation and detection is fast growing. The paper collects and surveys the most up-to-date research work that shows the state-of-the-art performance in DeepFake generation and detection.

(2) *Scale* The paper provides by far the largest scale and the most comprehensive survey of over 318 research papers on the topics of DeepFake generation, DeepFake detection, and evasion of DeepFake detection, with detailed categorizations and analyses.

(3) *Detailedness* This survey utilizes many graphic visualizations and diagrams (Sankey diagrams, fishbone diagrams, chord diagrams, etc.), as well as many very detailed long tables to best illustrate and highlight the properties, interactions, characteristics, evolution, and important traits of the technical methods surveyed and discussed. The diagrams and long tables may serve as a starting point for quick lookup and method comparisons, and the accompanying text provides in-depth discussion as a complement.

(4) *Technical evolution analyses* In addition to the detailed methodological introduction of each individual DeepFake generation and detection methods following the taxonomy, this paper uniquely manifests the technical evolution among the aforementioned methods, providing a more comprehensive and clearer picture of the evolutionarily technological landscape of the state-of-the-art DeepFake generation methods and detection methods.

(5) *Battleground analyses* There exists an adversarial and battling nature between DeepFake generation methods and DeepFake detection methods. Each party progresses by outcompeting the other side. The paper uniquely captures the tightly-connected interactivities between DeepFake generation and detection methods as well as among various detection methods themselves, revealing evidence for research hot zones and trends for future topics.

(6) *Horizon analyses* By virtue of the detailed surveys and analyses of the battleground landscape, the paper exposes challenges, identifies open research problems, and hints promising future research directions on the topics of DeepFake generation and detection.

Figure 3 depicts the tree diagram of the paper structure. For enhanced readability, here we provide suggestions for different types of readers and practitioners. For those who are new to the field and want to get up to speed quickly, it is advised to first go through Sects. 3.9 and 4.5 for the summary of the DeepFake generation and detection methods, along with the technical evolution highlights in Sects. 3.2.1, 3.3.1, 3.4.1, 3.5.1 and 4.1.6, 4.2.3, 4.3.4, respectively. Then the readers can move on to Sect. 5 for the Battleground, Sect. 6 for the Evasion methods, and Sect. 7 for the Horizon. For those who are already in the field and are interested in the technical details, it is advised to first go through all the subsections in Sects. 3 and 4, and then move on to Sects. 5 and 6. For those who are already in the field and want to identify latest technical trend in the DeepFake generation and detection literature, or in a particular direction, it is advised to first go through Sect. 5 in detail, and then pay attention to Sects. 3.2.1, 3.3.1, 3.4.1, 3.5.1 and Sects. 4.1.6, 4.2.3, 4.3.4 for the technical evolution, and finally move on to individual sections.

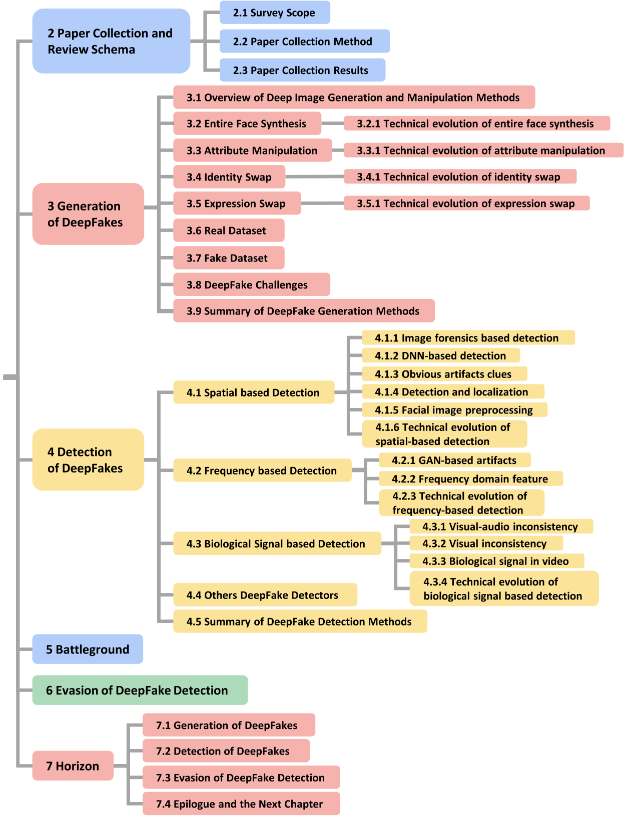

2 Paper Collection and Review Schema
2.1 Survey Scope
2.2 Paper Collection Method
2.3 Paper Collection Results
3 Generation of DeepFakes
3.1 Overview of Deep Image Generation and Manipulation Methods
3.2 Entire Face Synthesis
3.2.1 Technical evolution of entire face synthesis
3.3 Attribute Manipulation
3.3.1 Technical evolution of attribute manipulation
3.4 Identity Swap
3.4.1 Technical evolution of identity swap
3.5 Expression Swap
3.5.1 Technical evolution of expression swap
3.6 Real Dataset
3.7 Fake Dataset
3.8 DeepFake Challenges
3.9 Summary of DeepFake Generation Methods
4 Detection of DeepFakes
4.1 Spatial based Detection
4.1.1 Image forensics based detection
4.1.2 DNN-based detection
4.1.3 Obvious artifacts clues
4.1.4 Detection and localization
4.1.5 Facial image preprocessing
4.1.6 Technical evolution of spatial-based detection
4.2 Frequency based Detection
4.2.1 GAN-based artifacts
4.2.2 Frequency domain feature
4.2.3 Technical evolution of frequency-based detection
4.3 Biological Signal based Detection
4.3.1 Visual-audio inconsistency
4.3.2 Visual inconsistency
4.3.3 Biological signal in video
4.3.4 Technical evolution of biological signal based detection
4.4 Others DeepFake Detectors
4.5 Summary of DeepFake Detection Methods
5 Battleground
6 Evasion of DeepFake Detection
7 Horizon
7.1 Generation of DeepFakes
7.2 Detection of DeepFakes
7.3 Evasion of DeepFake Detection
7.4 Epilogue and the Next Chapter

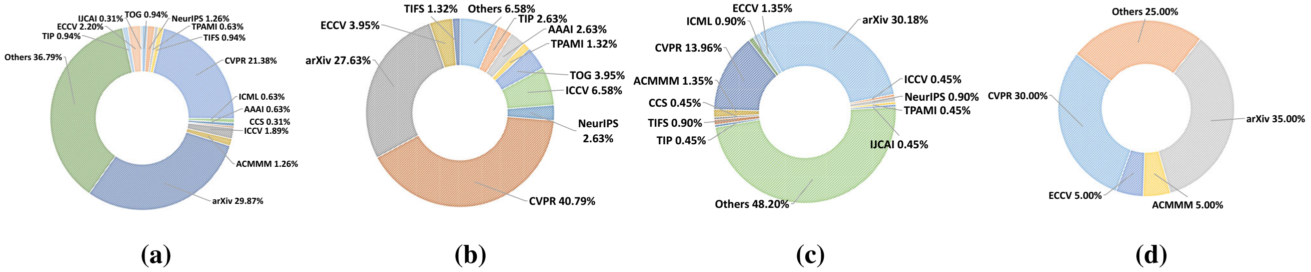
Fig. 3 Tree diagram showing the paper structure

2 Paper Collection and Review Schema

This section covers the survey scope, survey methodology, and paper collection results.

2.1 Survey Scope

The paper focuses on the technical aspect of DeepFakes via surveying related research papers on the topics of DeepFake generation, DeepFake detection, and evasion of DeepFake detection. The social and ethical aspects regarding DeepFakes, although briefly touched in this paper, are not the focus of this survey. A more loosely defined DeepFake could mean voice, gesture, body, and any type of media manipulation.



(a) (b) (c) (d)

Fig. 4 **a** For the papers in DeepFake research area, we can find that a large amount of papers are from Others and arXiv. The papers published in top conferences and journals only account for a third of the total. Furthermore, a lot of top papers are published in CVPR, which accounts for about half the published top papers. **b** For DeepFake generation methods, we can find that a large amount of the papers are from CVPR and arXiv. Two thirds of the generation papers are published in top conferences and journals. **c** For DeepFake detection methods, we can find that a large amount of the papers are from Others and CVPR. The published top papers only make up a small part of the total. This suggests that progress in DeepFake detection is not enough. **d** For DeepFake evasion methods, we can find that the volume of articles is not large and a large amount of the papers are from arXiv. One fourth of the papers are published in top conferences

In this survey paper, we focus solely on the topics of facial DeepFakes.

2.2 Paper Collection Method

To collect research papers on DeepFakes across different research areas as comprehensive as possible, we first collect the papers from a Github repository[1] which lists more than 100 papers about DeepFakes generation, DeepFake detection, and evasion of DeepFake detection. Then, we apply keywords matching to search DeepFake papers from two popular scientific databases (Google Scholar and DBLP) and arXiv where the newest papers are posted. The keywords are listed below.

- DeepFake/fake/editing/edit + facial image/face/swapping /video
- synthesis/GAN-synthesized/AI-synthesized
- manipulation/forgery/tampered face + detection

To ensure a more comprehensive and accurate survey, we also manually browse the recent three years publications in top-tier conferences and the corresponding workshops to avoid the limitations of keywords matching. Additionally, for DeepFake generation papers, we mainly collect the methods which have been mentioned in the previous DeepFake detection papers.

2.3 Paper Collection Results

Overall, we have collected 318 papers from Google Scholar and arXiv. The papers mainly include DeepFake generation, DeepFake detection, and evasion of DeepFake detection topics. Figure 4 shows the distribution of papers published in different research venues. Here we categorize the papers of top conferences and journals to specific classification (i.e., one of CVPR, ICCV, ECCV, TPAMI, etc.). Meanwhile, we bucket other published papers in non-top conferences and journals into the “Others” category. For unpublished papers with exposure on arXiv, we define them as the “arXiv” category.

For DeepFake generation methods, we can find that a large amount of the papers are from CVPR and arXiv. Two-thirds of the generation papers are published in top conferences and journals. This shows that a large proportion of DeepFake generation methods have gone through strict peer-review process and are relatively trustworthy. For DeepFake detection methods, however, we can find that a mass of the papers are from arXiv and Others. The published top-tier papers only make up a small part of the total. This suggests that progress in DeepFake detection is slower compared to that of the DeepFake generators. For DeepFake evasion methods, we can find that the volume of articles is small and a large amount of the papers are from arXiv, ECCV and Others. Half of the papers are published in top conferences. This shows that there is considerable improvement in DeepFake evasion research area.

In summary, for the papers in DeepFake research area, we can find that a large number of papers are from Others and arXiv. The papers published in top conferences and journals only account for a third of the total. Furthermore, a lot of top papers are published in CVPR, which accounts for about two-thirds the published top papers.

3 Generation of DeepFakes

In the research area of DeepFake generation, there are two parts to focus on: generation methods and datasets. We first

[1] https://github.com/clpeng/Awesome-Face-Forgery-Generation-and-Detection.

provide an overview of the general DeepFake techniques and introduce the methods which can be seen as DeepFake generation methods in a broader sense (e.g., style transfer, inpainting, super resolution, etc.) in Sect. 3.1. This gives readers an understanding of the general DeepFake generation techniques.

Then we focus on face appearance-related DeepFake methods which are most anticipated and influential in the field. For DeepFake generation methods, according to the consensus in the DeepFake field (Mirsky and Lee 2021; Verdoliva 2020; Tolosana et al. 2020), there are mainly four categories based on their function: entire face synthesis, attribute manipulation, identity swap, and expression swap as depicted in Fig. 1. We introduce these methods in Sects. 3.2–3.5. The other important part is the dataset. We highlight the major real image/video datasets which are used in the generation methods above and the fake image/video datasets generated by them. The content is introduced in Sects. 3.6 and 3.7, followed by DeepFake challenges in Sect. 3.8. The various DeepFake generation methods are summarized in Sect. 3.9. The highlights of the technical evolution of DeepFake generation methods discussed in Sects. 3.2.1, 3.3.1, 3.4.1, and 3.5.1.

3.1 Overview of Deep Image Generation and Manipulation Methods

In this section, we aim to give readers an understanding of the general deep image generation techniques which can be seen as the DeepFake technique in a broader sense.

The methods such as style transfer (Chen et al. 2018c; Yao et al. 2019), image inpainting (Yu et al. 2018, ?), rendering (Chen and Koltun 2017; Park et al. 2019; Liu et al. 2021a), super resolution (Dai et al. 2019; Guo et al. 2020b; Liu et al. 2020a; Mei et al. 2020), fusion (Lin et al. 2019), de-identification (Sun et al. 2018; Li and Lin 2019; Maximov et al. 2020), etc. share some of the technical similarities of DeepFake generation. However, these methods are not the focus of this survey. Instead, we mainly pay attention to the face appearance-related DeepFake methods.

As introduced in Sects. 3.2–3.5, existing DeepFake generation methods mainly consist of four types (i.e., entire face synthesis, attribute manipulation, identity swap, expression swap), depending on the tasks' requirements. To achieve a comprehensive survey, we detail the technological evolution of the four types, respectively. Note that, we focus on introducing their intuitive idea and categorizing their optimization methods. To make it clear, we use Fig. 5 to show the overall evolution of the four types of DeepFake generation methods, respectively.

3.2 Entire Face Synthesis

Definition 1 Entire face synthesis aims to generate non-existent fake face image $\mathbf{x_f}$ from random vector $\mathbf{v}$ with neural network $\phi(\cdot)$. That is $\mathbf{x_f} = \phi(\mathbf{v})$.

For entire face synthesis tasks, GANs and VAEs are both feasible neural networks $\phi(\cdot)$. However, according to the surveys (Verdoliva 2020; Nguyen et al. 2019d; Tolosana et al. 2020; Lyu 2020), GANs are the mainstream baseline technique. Many famous and popular entire face synthesis techniques such as PGGAN, StyleGAN, etc. are GAN-based and are able to generate high-quality DeepFake images. Compared with GANs, VAEs usually generate less realistic faces (i.e., being blurred). The reason why the images generated by VAEs tend to be blur is that the training principle makes VAEs assign a high probability to training data points, which cannot ensure that blurry data points are assigned to a low probability (Huang et al. 2018). Since the DeepFake images generated by VAE are not realistic enough, this section mainly introduces the GAN-related works.

Using GANs for entire face synthesis is actually a kind of distribution mapping. The GANs learn the mapping from random distribution to human face distribution. Existing state-of-the-art methods can stably generate high-resolution images. which is benefited from the continuous improvement of the GAN network and training procedure. However, the current methods still suffer from the training difficulty (e.g., mode collapse problem of GAN training procedure). Furthermore, the generated images are not realistic enough due to the lack of general knowledge of face distribution (e.g., facial symmetry).

As shown in the entire face synthesis part of Fig. 1, the fake images are very realistic and it is hard to distinguish real images from fake ones. Existing works mainly focus on improving the training stability, resolution, and controllable face attribute.

The classical examples are deep convolutional GAN (DCGAN) (Radford et al. 2015), Wasserstein GAN (WGAN) (Arjovsky et al. 2017), progressive growing GAN (PGGAN) (Karras et al. 2017), and style-based GAN (StyleGAN) (Karras et al. 2019).

The very first work which combines convolutional neural network (CNN) and GAN is a deep convolutional generative adversarial network (DCGAN) (Radford et al. 2015). It focuses on unsupervised learning and has comparable performance in image classification tasks with the pre-trained discriminator. The generator of it can easily manipulate lots of the semantic properties (i.e., manipulate attribute of a human face) of generated images profile from its interesting vector arithmetic properties.

Two years later, there has been an explosion of in-depth research on GANs. Some GANs put emphasis on the sta-

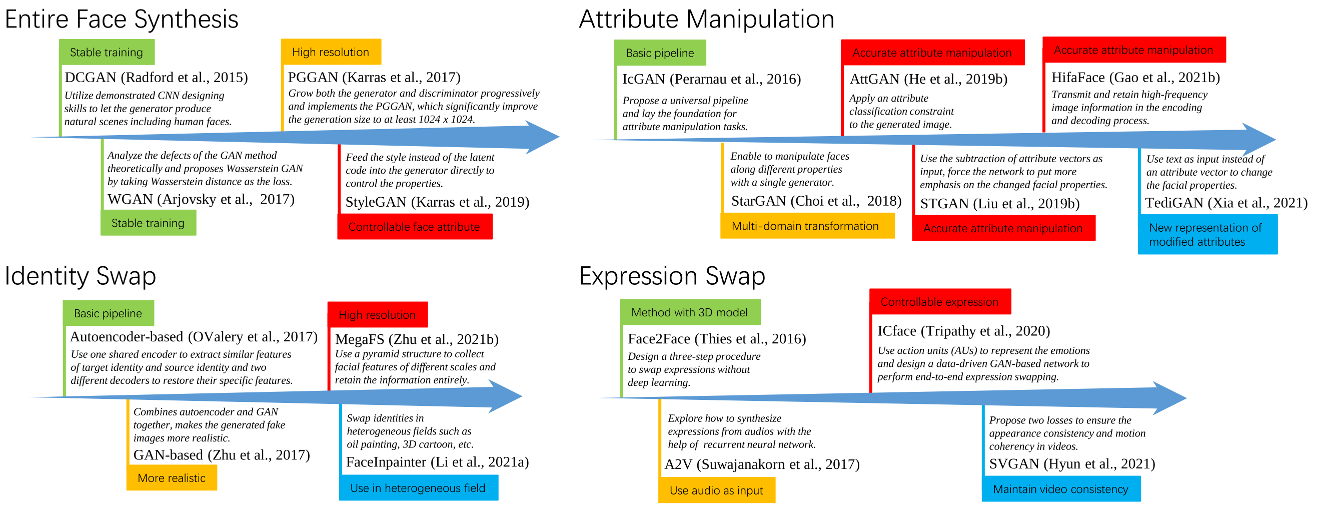

Entire Face Synthesis
Stable training
DCGAN (Radford et al., 2015)
Utilize demonstrated CNN designing skills to let the generator produce natural scenes including human faces.
High resolution
PGGAN (Karras et al., 2017)
Grow both the generator and discriminator progressively and implements the PGGAN, which significantly improve the generation size to at least 1024 x 1024.
Analyze the defects of the GAN method theoretically and proposes Wasserstein GAN by taking Wasserstein distance as the loss.
WGAN (Arjovsky et al., 2017)
Stable training
Feed the style instead of the latent code into the generator directly to control the properties.
StyleGAN (Karras et al., 2019)
Controllable face attribute
Attribute Manipulation
Basic pipeline
IcGAN (Perarnau et al., 2016)
Propose a universal pipeline and lay the foundation for attribute manipulation tasks.
Accurate attribute manipulation
AttGAN (He et al., 2019b)
Apply an attribute classification constraint to the generated image.
Accurate attribute manipulation
HifaFace (Gao et al., 2021b)
Transmit and retain high-frequency image information in the encoding and decoding process.
Enable to manipulate faces along different properties with a single generator.
StarGAN (Choi et al., 2018)
Multi-domain transformation
Use the subtraction of attribute vectors as input, force the network to put more emphasis on the changed facial properties.
STGAN (Liu et al., 2019b)
Accurate attribute manipulation
Use text as input instead of an attribute vector to change the facial properties.
TediGAN (Xia et al., 2021)
New representation of modified attributes
Identity Swap
Basic pipeline
Autoencoder-based (OValery et al., 2017)
Use one shared encoder to extract similar features of target identity and source identity and two different decoders to restore their specific features.
High resolution
MegaFS (Zhu et al., 2021b)
Use a pyramid structure to collect facial features of different scales and retain the information entirely.
Combines autoencoder and GAN together, makes the generated fake images more realistic.
GAN-based (Zhu et al., 2017)
More realistic
Swap identities in heterogeneous fields such as oil painting, 3D cartoon, etc.
FaceInpainter (Li et al., 2021a)
Use in heterogeneous field
Expression Swap
Method with 3D model
Face2Face (Thies et al., 2016)
Design a three-step procedure to swap expressions without deep learning.
Controllable expression
ICface (Tripathy et al., 2020)
Use action units (AUs) to represent the emotions and design a data-driven GAN-based network to perform end-to-end expression swapping.
Explore how to synthesize expressions from audios with the help of recurrent neural network.
A2V (Suwajanakorn et al., 2017)
Use audio as input
Propose two losses to ensure the appearance consistency and motion coherency in videos.
SVGAN (Hyun et al., 2021)
Maintain video consistency

Fig. 5 The evolution of DeepFake generation techniques with a fishbone diagram for each DeepFake generation type

bility of the GAN training. The groundbreaking work is Wasserstein-GAN (WGAN) (Arjovsky et al. 2017). In the first published GANs, the procedure requires researchers to carefully maintain a balance between generator and discriminator. The mode dropping phenomenon also occurs frequently. To solve these hot potatoes, WGAN has theoretically minimized a reasonable and efficient approximation of the expectation-maximization (EM) distance, which only needs a few optimization designs on the original GANs.

There are many types of research based on the WGAN. Gradient penalty WGAN (WGAN-GP) (Gulrajani et al. 2017) has indicated that WGAN sometimes still generates poor samples or fails to converge. The reason is that WGAN uses weight clipping to enforce a Lipschitz constraint. To improve the weight clipping operation, they have proposed to penalize the norm of the gradient of the discriminator with respect to its input fake image. The new designs train stably when generating high-quality home images. Simply using Wasserstein probability can not simultaneously satisfy sum invariance, scale sensitivity, and unbiased sample gradients. To improve it, Cramer GAN (CramerGAN) (Bellemare et al. 2017) has combined the best of the Wasserstein and Kullback–Leibler divergences to propose the Cramér distance. The CramerGAN performs significantly better than the WGAN. Boundary equilibrium GAN (BEGAN) (Berthelot et al. 2017) is also an improved version of WGAN (Arjovsky et al. 2017). To further balance the power of the discriminator against the generator, they have suggested pairing an equilibrium enforcing method with a loss derived from the Wasserstein distance together. They also have proposed a new way to control the trade-off between image diversity and visual quality.

Some other works focus on how to generate high-resolution images. The resolution of the images generated by them is at least 1024 × 1024. Meanwhile, the images are detailed, and it is quite difficult to distinguish between the genuine and the fake, which is very amazing. PGGAN (Karras et al. 2017) is the very first and famous work that proposes an effective method to generate high-resolution images. The resolution of the generated images is 1024 × 1024. It has proposed to progressively grow both the image resolution of the generator and discriminator. The images are starting from a low resolution and being detailed step by step with the new layers added in the model. This method is very reasonable in that it can speed up the training as well as greatly stabilize the GAN. However, the training procedure is still not good enough that some of the generated images are far from real. BigGAN (Brock et al. 2018) has attempted to generate high-resolution diverse images from datasets such as ImageNet (Deng et al. 2009). They have applied orthogonal regularization to enforce the generator to be satisfied with a simple “truncation trick”. Thus, the user can control the trade-off between image fidelity and variety by reducing the variance of the generator’s input.

To control the properties of generated images elaborately, StyleGAN (Karras et al. 2019) has proposed a new design to automatically learn the unsupervised separation of high-level attributes such as pose and human identity. The architecture also leads to stochastic variation in the generated images (e.g., freckles, hair). Furthermore, it enables intuitive, scale-specific control of the synthesis. StyleGAN2 (Karras et al. 2020) has exposed several typical artifacts of StyleGAN and has proposed changes in both model architecture and training methods to address them. In particular, they have encouraged good conditioning in the mapping from latent codes to images by the new design of generator normalization, progressive growing, and generator regularization.

Different from the previous methods which use the GAN framework, generative flow (Glow) (Kingma and Dhariwal 2018) is a flow-based generative model that uses an invertible 1×1 convolution. The method is based on the theory that a generative model optimized towards the plain log-likelihood objective has the ability to generate efficient realistic-looking synthesis and manipulate large images.

3.2.1 Technical Evolution of Entire Face Synthesis

To summarize, a straightforward way for the entire face synthesis is to regard it as an image generation task. Goodfellow et al. (2014a) propose the generative adversarial network (GAN) and the trained generator is able to produce meaningful examples, e.g., handwritten numbers and human faces. However, the generated examples are usually with low resolution and artifacts. Moreover, the networks are unstable to train (Goodfellow et al. 2014a). Then, to address the issues of GANs, Radford et al. (2015) propose the deep convolutional generative adversarial network (DCGAN), which utilizes demonstrated CNN designing skills to let the generator produce natural scenes including human faces. As a result, DCGAN is identified as the early work for DeepFake (Radford et al. 2015).

After that, some similar GAN-based methods are proposed. Nevertheless, they usually encounter difficulties (e.g., non-convergence, gradient vanishing, collapsed mode) in the training procedure. To solve these difficulties and generate images effectively and stably, Arjovsky et al. (2017) analyzes the defects of the GAN method theoretically and proposes Wasserstein GAN by taking Wasserstein distance as the loss. Wasserstein distance cures the main training problems of GANs including DCGAN with relaxed requirements on the balance between discriminator and generator, the designs of network architecture, and reduced mode dropping phenomenon (Arjovsky et al. 2017). As a result, Wasserstein GAN usually generates more natural and higher quality fake faces than DCGAN.

Although achieving significant progress, the GAN-based methods could not synthesize fake faces with high resolution. In particular, the fake images generated by the previous methods are less than 256×256. The capability of generating high-resolution images is particularly important for real-world applications since the low-resolution fake faces can be easily identified. To address this issue, Karras et al. (2017) propose to grow both the generator and discriminator progressively and implement the PGGAN, which significantly improves the generation size to at least 1024×1024. Specifically, the main challenges of generating high-resolution fake images stem from two aspects. First, it is easy to distinguish the fake images from the real ones with the discriminator when the resolution is high, which magnifies the vanishing gradient problem of the generator during GAN training. Second, due to the memory limitation, generating high-resolution images leads to smaller batch sizes, affecting the stability of training. To solve these problems, PGGAN proposes to progressively increase the image resolution of the generated images of the generator and the discriminator in the training procedure with the spatial resolution of the generator and discriminator being 4×4 pixels at the beginning. As the training advances, they incrementally add layers to the generator and discriminator and increase the spatial resolution of the generated images. As a result, the method is able to produce high-resolution fake faces.

The above methods are able to generate natural and realistic fake faces but cannot control the properties we want to be fake. To complement the capability, based on PGGAN, Karras et al. (2019) propose StyleGAN that feeds the style instead of the latent code into the generator directly. Specifically, StyleGAN transforms the latent code to ‘style’ code via a nonlinear mapping network. Then, the style code is used to conduct the adaptive instance normalization after each convolution. In addition, Gaussian noise images are embedded after each convolution layer for stochastic detail generation. The StyleGAN is able to generate high-resolution images with higher image quality and wider detailed variations.

Overall, the technical evolution of the entire face synthesis mainly follows the development of the GAN, which aims to solve the challenges that may arise in real-world applications, e.g., high-resolution and style-guided generations. For the entire face synthesis, we think the improvements of training stability, resolution, and controllable facial attributes are the major technical characteristics that have been evolved throughout the years, and we highlight seminal works DCGAN, WGAN, PGGAN, StyleGAN in the top left panel of Fig. 5, showcasing the improvements in different angles.

3.3 Attribute Manipulation

Definition 2 Attribute manipulation aims to modify facial properties $\mathbf{P}$ of a real face image $\mathbf{x_r}$ to generate a new fake image $\mathbf{x_f}$ with neural network $\phi(\cdot, \cdot)$. That is $\mathbf{x_f} = \phi(\mathbf{x_r}, \mathbf{P})$.

Using GANs for attribute manipulation is actually a kind of latent space editing. The key point is the quality of the GAN inversion technique. With a better attribute disentangle technique, the GANs for attribute manipulation can achieve more accurate attribute control. Existing state-of-the-art methods [e.g., HifaFace (Gao et al. 2021d)] can perform accurate face editing while maintaining rich details of non-editing areas. However, the current methods are still limited by the labels in the training dataset. That is, it is difficult to control the attributes that do not exist in the label of the training dataset.

As shown in the attribute manipulation part of Fig. 1, the real images are modified with facial attributes such as bald, blond hair, eyeglasses, etc. Existing works mainly focus on improving attribute manipulation accuracy.

Attribute manipulation is also known as face editing, which can not only modify simple face attributes such as hair color, bald, smile, but also retouch complex attributes like gender, age, etc. The classical examples are StarGAN (Choi et al. 2018) and selective transfer GAN (STGAN) (He et al. 2019b).

Invertible conditional GAN (IcGAN) (Perarnau et al. 2016) is the earliest attempt in GAN-based facial attribute manipulation. Based on an extension of the idea of conditional GAN (cGAN) (Mirza and Osindero 2014), they have evaluated encoders to map a real image into a latent space and a conditional representation, which allows the reconstruction and modification of arbitrary attributes of real human face images. The expression generative adversarial network (ExprGAN) (Ding et al. 2018) has added an expression controller module that can learn an expressive and compact expression code to the encoder-decoder network. The expression controller module enables it to edit photo-realistic facial expressions with controllable expression intensity.

Previous studies can only perform image-to-image translation for two domains, which is cumbersome and time-consuming. To be more efficient, StarGAN (Choi et al. 2018) has designed a single model to perform image-to-image translations for multiple domains. It allows simultaneous training of multiple different-domain datasets within a single network. As an improvement, StarGAN2 (Choi et al. 2020) simultaneously satisfies two properties in image-to-image translation: diversity of generated images as well as scalability over multiple domains. To represent diverse styles of a specific domain, they have replaced StarGAN's domain label with their domain-specific style code. To adapt the style code, they have proposed two modules: a mapping network and a style encoder. The style code can be extracted from a given reference image with a style encoder while the mapping network can transform random Gaussian noise into a style code. Utilizing these style codes, the generator learns to successfully synthesize diverse images over multiple domains.

Although StarGAN is effective, due to the limitation of the content of the datasets, it can only generate a discrete number of expressions. To address this limitation, GAN animation (GANimation) (Pumarola et al. 2018) has introduced a novel GAN conditioning method based on action units (AU) annotations. It defines the human expression with a continuous manifold of the anatomical facial movements. The magnitude of activation of each AU can be controlled independently. Different AUs can also be combined with each other with this method.

Most of the previous work inevitably changes the attribute irrelevant regions. To solve this problem, spatial attention GAN (SaGAN) (Zhang et al. 2018) propose a module to only change the attribute-specific region and keep the other area unchanged. This work properly exploits the attention mechanism to ensure a better face editing effect, which shows the feasibility of the attention mechanism in face manipulation.

Previous methods have attempted to establish an attribute-independent latent representation for further attribute editing. However, since the facial attributes are relevant, requesting for the invariance of the latent representation to the attributes is excessive. Therefore, simply forcing the attribute-independent constraint on the latent representation not only restricts its representation ability but also may result in information loss, which is harmful to attribute editing. To solve this problem, facial Attribute editing (AttGAN) (He et al. 2019b) has removed the strict attribute-independent constraint from the latent representation. It just applies the attribute classification constraint to the generated image to guarantee the correctness of attribute manipulation. Meanwhile, it groups attribute classification constraint, reconstruction learning, and adversarial learning together for high-quality facial attribute editing. The model supports direct attribute intensity control on multiple facial attribute editing within a single model.

Considering that the specific editing task is only related to the changed attributes instead of all target attributes, as an improvement of AttGAN, STGAN (Liu et al. 2019) has selectively taken the difference between target and source attribute vectors as the input of the model. Furthermore, they have enhanced attribute editing by adding a selective transfer unit that can adaptively select and modify the encoder feature to the encoder-decoder.

Mask-guided portraiting editing (MaskPE) (Gu et al. 2019) proposes a unique way to manipulate face attributes. They use a face parsing mask to guide the generation of face attributes. The main idea is to separately embed five facial components (i.e., left eye, right eye, mouth, skin & nose, and hair) into latent codes based on face parsing masks. Then they can modify any facial component independently.

Due to the lack of paired images during training, previous methods typically use cycle consistency to keep the non-editing attributes unchanged. However, even if the cycle consistency is satisfied, images may still be blurry and lose rich details from input images for that the generator tends to find a tricky way (i.e., encodes the rich details of the input image into the output image in the form of hidden signals) to satisfy the constraint of cycle consistency. To solve this problem, Gao et al. (2021d) propose high-fidelity arbitrary face editing (HifaFace) to maintain rich details (e.g., wrinkles) of non-editing areas. Their work has two improvements. The first is that they directly feed the high-frequency information of the input image into the end of the generator with wavelet-based skip-connection, which relieves the pressure of the generator to synthesize rich details. The second is that they use another high-frequency discriminator as a complement to the image-level discriminator to encourage the image to have rich details.

Text-guided diverse image generation and manipulation GAN (TediGAN) (Xia et al. 2021) is a special network for multi-modal image generation and manipulation with textual descriptions. They map the image and text into a common embedding space to learn text-image matching. The method allows the user to edit the appearance of different attributes interactively.

HistoGAN (Afifi et al. 2021) chooses a special angle to manipulate images. They use color histograms to manipulate the color blending of the images only, which is a very targeted research content.

3.3.1 Technical Evolution of Attribute Manipulation

In contrast to the entire face synthesis, the attribute manipulation of the face is usually regarded as the image manipulation task that changes facial properties (e.g., hair's color and style, facial hair, etc.) of input faces. The task can be tackled by incorporating encoder-decoder and GANs.

Perarnau et al. (2016) start the first work for attribute manipulation, which is denoted as IcGAN. This work proposes a universal pipeline and lays the foundation for attribute manipulation tasks. Specifically, IcGAN first encodes real images into the latent space and then changes the latent codes corresponding to different facial properties. After that, it decodes the changed latent codes to fake face images. Although effective, IcGAN can be time-consuming when it aims to perform the manipulation of multiple facial attributes since each attribute is addressed via an independent deep model.

To allow flexible GANs and avoid high time consumption, Choi et al. (2018) propose the StarGAN and design a network that enables to manipulate faces along with different properties with a single generator. The intuitive idea is to encode the real image and its respective source domain label via the generator and produce the fake image. At the same time, the discriminator is designed to classify the real or fake faces and identify the domain. As a result, the learned GAN cannot only contain the semantic representation of facial properties and the respective domain information.

Although face attribute editing is available, the desired attribute variation is specified as the input for the encoder while the latent representation is not constrained, which may result in information loss and lead to over-smooth and distorted generation. To alleviate these drawbacks, He et al. (2019b) propose the AttGAN that applies an attribute classification constraint to the generated image. As a result, the correct change of attributes can be guaranteed.

In addition to the AttGAN, Liu et al. (2019) implement a better generator (STGAN), forcing the network to emphasize the desired changed facial properties while preserving the other property areas. To this end, they use the subtraction of attribute vectors (i.e., source vector—target vector) to replace the source vector as the inputs of the decoder. Moreover, they propose a novel architecture named the selective transfer unit to improve attribute manipulation ability and image quality. As a result, this work can improve attribute manipulation accuracy as well as perception quality.

Although AttGAN and STGAN do not raise obvious variations on the undesired face attributes, they may lead to the variations of details (e.g., wrinkles) in those undesired areas. This is caused by the cycle consistency during training. Specifically, due to the lack of paired images during training, AttGAN and STGAN use the cycle consistency to avoid the changing of the undesired attributes. However, these generators map the details of the input image to a new one via hidden signals to achieve the cycle consistency that cannot be guaranteed. To solve this problem, Gao et al. (2021d) propose high-fidelity arbitrary face editing (HifaFace) to maintain rich details of undesired attribute areas. The main idea is to transmit and retain high-frequency image information in the encoding and decoding process. There are two main improvements. First, they directly feed the high-frequency information of the input image into the end of the generator with wavelet-based skip-connection, which relieves the pressure of the generator to synthesize rich details. Second, they add a high-frequency discriminator as a complement to encourage the image to have rich details, which prevents the generator from finding a trivial solution for cycle consistency. As a result, the generated images have rich details with higher fidelity.

Recently, Xia et al. (2021) propose a new design to use text as input instead of an attribute vector to change the facial properties of real images. The main idea of them is to map text and images to the same semantic latent space. Thus they can use text to replace attribute vectors. They extend the availability and diversity of attribute manipulation tasks.

In summary, recent works for attribute manipulation mainly focus on how to change the desired face attributions effectively while preserving other areas via a single generator. For the attribute manipulation, we think the basic pipeline and the improvements of multi-domain transformation, accurate attribute manipulation, and new representation of modified attributes are the major technical characteristics that have been evolved throughout the years, and we highlight seminal works IcGAN, StarGAN, AttGAN, STGAN, HifaFace, TediGAN in the top right panel of Fig. 5, showcasing the improvements in different angles.

3.4 Identity Swap

Definition 3 Identity swap aims to replace the identity of source image $\mathbf{x_s}$ by the identity $\mathbf{t_i}$ of target image $\mathbf{x_t}$ with neural network $\phi(\cdot,\cdot)$ and generate a new fake image $\mathbf{x_f}$. That is $\mathbf{x_f} = \phi(\mathbf{x_s}, \mathbf{t_i})$.

As shown in the identity swap part of Fig. 1, the images in the fake videos have uneven qualities. Existing works mainly focus on improving the realism and resolution of the image.

In general, the architectures used for these functions mainly fall into two categories: autoencoder-based and GAN-based. The classical works are cycle-consistent GAN (CycleGAN) (Zhu et al. 2017, 2021b).

The methods which make the concept of DeepFake, especially identity swapping, become widely known are methods based on autoencoder. The autoencoder-based methods (OValery 2017) have no specific name or architecture. However, as they are all based on autoencoder, their pipeline is similar. The methods use one shared encoder and two independent decoders. The encoder and one of the decoders are trained by source identity while the encoder and the other decoder are trained by target identity. When the model is well trained, the encoder has the ability to extract the common features of source and target identities while the decoder records the specific features. At inference time, the image of the source identity goes through the encoder and the opposite decoder, producing a realistic swap.

Nowadays, GAN-based methods are the mainstream in identity swap. The first work of the GAN-based method was CycleGAN (Zhu et al. 2017) proposed in 2017. In previous works, the absence of paired examples is always the limitation in image transformation tasks. CycleGAN has artfully solved this problem. Define a source domain X and a target domain Y, it builds a mapping $G : X \rightarrow Y$ which is highly under-constrained and similarly constructs an inverse mapping $F : Y \rightarrow X$. Then the cycle consistency loss which enforces $F(G(X)) \approx X$ (and vice versa) is the optimization target of the model. Through this circulation, there is no need for paired samples. Meanwhile, although not mentioned in the paper, the framework of CycleGAN can be used for identity swap easily. Faceswap-GAN (Lu 2018) is the implementation of CycleGAN which provides an identity swap functionality. It simply adds the adversarial loss and perceptual loss to encoder architecture.

Face swapping GAN (FSGAN) (Nirkin et al. 2019) is a subject agnostic method that doesn't rely on the training of pairs of faces. It is also the first to simultaneously adjust the pose, expression, and identity variations for both a single image and a video sequence.

The research in identity swap has been stagnated for a long time until the appearance of FaceShifter (Li et al. 2020b). It proposes a two-stage procedure for high fidelity and occlusion-aware face-swapping. Unlike many existing face-swapping works that leverage only limited information from the target image, FaceShifter generates the swapped face by thoroughly and adaptively exploiting the information of the target image.

Appearance optimal transport (AOT) (Zhu et al. 2020) has formulated appearance mapping as an optimal transport problem. They have proposed an AOT model to formulate it in both latent and spatial space. In particular, a relighting module is designed to simulate the optimal transport plan. The optimization target is minimizing the Wasserstein distance of the learned features in the latent space, which enables better performance and less computation than conventional optimization.

Information disentangling and swapping network (InfoSwap) (Gao et al. 2021a) aims to extract the most expressive information for identity representation. The main idea is to formulate the learning of disentangled representations as optimizing an information bottleneck trade-off. The information bottleneck principle provides a guarantee that in the latent space, areas scored as identity-irrelevant indeed contribute little information to predict identity.

Megapixel level face swapping (MegaFS) (Zhu et al. 2021b) has proposed the first one-shot ultra-high-resolution face swapping method. To overcome the information loss in the encoder, they use a hierarchical representation face encoder (HieRFE) to find the complete face representation. Then they use a face transfer module (FTM) to control multiple attributes synchronously without explicit feature disentanglement. The contributions are ground-breaking.

FaceInpainter (Li et al. 2021a) proposes a controllable face inpainting network under heterogeneous domains (i.e., oil painting, 3D cartoons, pencil drawing, exaggerated drawing, etc.). The framework has two stages. In the first stage, they use a styled face inpainting network (SFI-Net) to map the identity and attribute properties to the swapped face. The second stage contains a joint refinement network (JR-Net) that refines the attributes and identity details, generating occlusion-aware and high-resolution swapped faces with visually natural fused boundaries.

3.4.1 Technical Evolution of Identity Swap

Identity swap is usually achieved by conducting replacement on the identity-related features and decoding these features to the image level. As a result, the identity of the input face image (i.e., the source identity) can be changed to the desired one (i.e., the target identity). Specifically, the general pipeline is implemented via the autoencoder (OValery 2017) that contains one shared encoder and two independent decoders. It first uses the encoder to extract features of the source and target identities, respectively, and get their respective latent codes. Then, the method uses the two independent decoders to reconstruct the source and target images, respectively. During the identity swap, the latent code of the source identity is fed to the decoder of the target identity. As a result, the decoded face is swapped. Under this pipeline, one of the key problems is how to extract or select faces' features as the latent code for replacement. Another problem is how to make the generated images more realistic.

To improve the fidelity of the generated face images, Lu (2018) combines the autoencoder and GAN. Compared with the naive autoencoder-based methods, it adopts GAN's advantages while making the fake images more realistic due to the supervision of the discriminator. The above methods do well on low-resolution images but cannot generate high-resolution swapped faces. This issue stems from the compressed representations whose information is insufficient for high-quality face generation during the swapping. To alleviate this drawback, MegaFS (Zhu et al. 2021b) proposes the hierarchical representation face encoder (HieRFE), which uses a pyramid structure to collect facial features under different scales and retain the information entirely. As a result, the method can focus on processing the high-level semantic information while retaining the low-level details after identity swap at the megapixel level.

The previous methods have achieved great progress on photorealistic images. However, their capability of addressing source and target images with heterogeneous materials (e.g., oil painting, 3D cartoon, etc.) is less effective due to the different textures of the source and target images. A recent work (Li et al. 2021a) designs a two-stage framework to solve the issue. After explicitly disentangling the foreground (i.e., face and neck) from the background (e.g., hair, clothes, etc.) in the source identity, the first stage is to combine the attribute codes of the source identity and identity code of the target identity with the fixed background extracted from source identity. As a result, the swapped face contains the target identity with source background and attributes. However, the segmented background and the generated foreground cannot be well integrated under some complex scenes. Then, the second stage is to refine the coarse result from the first stage to make the source and the target background consistent at the fusion boundary.

Overall, the technical evolution of identity swap mainly focuses on the better separation of identity and attribute features of the source and target images and how to fuse them. For the identity swap, we think that the basic pipeline and the improvements of realistic, resolution, and heterogeneous fields are the major technical characteristics that have been evolved throughout the years, and we highlight seminal works autoencoder-based method, GAN-based method, MegaFS, and FaceInpainter in the bottom left panel of Fig. 5, showcasing the improvements in different angles.

3.5 Expression Swap

Definition 4 Expression swap aims to replace the expression of source image $\mathbf{x_s}$ by the expression $\mathbf{t_e}$ of target image $\mathbf{x_t}$ with neural network $\phi(\cdot, \cdot)$ and generate a new fake image $\mathbf{x_f}$. That is $\mathbf{x_f} = \phi(\mathbf{x_s}, \mathbf{t_e})$.

As shown in the expression swap part of Fig. 1, usually the mouth of the real images are changed. Existing works mainly focus on improving the diversity of input source and video consistency.

Expression swap is also known as face reenactment. The classical examples are ICface (Tripathy et al. 2020) and SVGAN (Hyun et al. 2021).

Face2Face (Thies et al. 2016) has proposed a three-step procedure. It first uses a global non-rigid model-based bundling approach to reconstruct the shape identity of the target human based on a prerecorded training sequence. Then it uses a transfer function to efficiently exploit deformation transfer in the low-dimensional semantic space. At last, the image-based mouth synthesis approach exploits the best matching mouth shapes offline sample sequence to generate a realistic mouth.

A2V (Suwajanakorn et al. 2017) has used a recurrent neural network to train a model that can map from raw audio features of Obama's weekly address footage to mouth shapes. It is a cross-modal method that leverages the pronunciation features of the target person to synthesize the correct lip shapes for given audio content. It doesn't need an original video as expression-driven material. To match the input audio track, they have synthesized high-quality mouth texture and composited it with proper 3D pose matching to change what he appears to be saying.

Pose-controllable audio-visual system (PC-AVS) (Zhou et al. 2021a) is another state-of-the-art cross-modal method. Previous audio-driven talking human face synthesis methods fail to model head pose, one of the key factors for talking faces to look natural. This is because pose information can rarely be inferred from audios. To solve this problem, PC-AVS introduces extra pose source video to compensate only for head motions and successfully disentangle the representations of talking human faces into the spaces of speech content, head pose, and identity respectively.

Previous cross-modal methods only put emphasis on the lip motions and ignore the implicit ones such as head poses and eye blinks that have a weak correlation with the input audio. To model these implicit relationships, face implicit attribute learning generative adversarial network (FACIAL-GAN) (Zhang et al. 2021a) integrates the phonetics-aware, context-aware, and identity-aware information to synthesize the 3D face animation with realistic motions of lips, head poses, and eye blinks.

Previous works may lose detailed information of the target leading to a defective output. To solve this problem, MarioNETte (Ha et al. 2020) has proposed a few-shot face reenactment framework that preserves the information of target identity even in situations where the facial characteristics of the source identity are far from the target. It has also introduced landmark transformation to cope with the varying facial characteristics of different people.

Interpretable and controllable face reenactment network (ICface) (Tripathy et al. 2020) has proposed a two-stage neural network face animator which can control the pose and expressions of a given face image. The face animator is a data-driven and GAN-based system that is suitable for a large number of identities.

Self-supervised video GAN (SVGAN) (Hyun et al. 2021) first puts emphasis on exploiting the discriminator of the GAN. They hypothesize two prominent constraints for realistic videos: consistency of appearance and coherency of motion. With these constraints, GANs are more likely to generate realistic videos. In other words, they have well defined what constraints should synthesized videos satisfy first.

Wang et al. (2021) propose a one-shot neural talking-head synthesis approach. The method uses unsupervised learning to decompose for key features of an image: appearance feature, canonical keypoint, head pose, expression deformation. With the appearance feature and canonical keypoint of the source image, and synthesized with the head pose and expression deformation, a new fake image can be created. This work clearly disassembles the face information and reasonably exploits them.

Most of the DeepFake detection methods did not take expression swap as the main detection objective. In our opinion, there are several reasons. As we can see from the previous description, expression swap has a similar technique to identity swap. Thus most of the detection methods are not specifically designed for them and only a few detection methods consider detecting expression swaps. On the other hand, it usually needs the coordination of audio to achieve a better display effect in the expression swap. Only the detection methods which simultaneously take images and audio into account are designed for this problem. The swapped expression strongly depends on the source video or image. The audio-video coordination opens the door for detection algorithm to tackle this problem from multiple angles, reducing the difficulty of this problem. Therefore, as we mainly investigate the DeepFake generation methods that are mentioned by the DeepFake detection methods, we take expression swap as an extension of identity swap in the survey.

3.5.1 Technical Evolution of Expression Swap

In contrast to identity swap, expression swap is to replace the features of the mouth in the source image and produce a new face with the same identity but a different expression.

The early work Face2Face (Thies et al. 2016) designs a three-step procedure to swap expressions without deep learning. Specifically, Face2Face first strips the identity information from the source identity. Then, it transfers the source identity's expression to the target one. Finally, it synthesizes a realistic target mouth region. The whole pipeline is reasonable but requires complex 3D face models and considerable efforts to capture all the subtle movements in the face.

In addition, to swap the expression according to the given images, recent works also explore how to synthesize expressions from audios. For example, A2V (Suwajanakorn et al. 2017) successfully synthesized fake videos of Obama (i.e., the 44th president of the United States) according to the given audio. To this end, Suwajanakorn et al. (2017) use recurrent neural networks to map audios to the sparse shape of the mouth (i.e., 18 lip fiducials). Then, they generate photo-realistic mouth texture based on the generated lip fiducials. As a result, the expression swap based on the audios is achieved.

To bypass the explicit 3D model fitting, a straightforward way is to learn a deep model implicitly via large-scale data. However, there are two problems. First, it is hard to collect expression and pose representation that is independent of the identity feature. Second, such an implicit model usually lacks interpretability and does not easily allow Hence, it is difficult to synthesize diverse face attributes from the other faces. To solve these problems, ICFA (Tripathy et al. 2020) proposes to use action units (AUs) (Friesen and Ekman 1978) to represent the emotions. The AUs represent the activations of 17 facial muscles and each combination of them can produce different facial expressions. The advantages of AU-based expression representation are as follows. First, it is a relatively straightforward and flexible way to extract expressions from any facial image. Second, this representation is fairly independent of the identity-specific characteristics of the face. As a result, the first problem can be solved. To swap the expression to the target face, ICface eliminates the expression of the input face first, which is done by mapping the input image to a neutral state representing zero AU values. Then, it uses a conditional GAN to take the neutral image and the previous facial attribute vector (i.e., AUs) as input to generate expression, which solves the second problem. The model is also interpretable in that the facial attribute vector can be manually defined.

Previous works focus on the expression swap of images. However, the expression swap on videos needs to maintain the consistency of the face across frames, which is much more difficult than swapping on images. To achieve a more realistic expression swap on video, Hyun et al. (2021) propose the SVGAN that clearly defines two constraints (i.e., appearance contrastive loss & temporal structure loss) that should be satisfied in the video synthesis. The appearance contrastive loss makes the discriminator learn the representations of appearance which is invariant throughout time in videos. On the other hand, temporal structure loss forces the discriminator to figure out whether the video is coherent or not in temporal ordering. These two losses ensure the appearance of consistency and motion coherency in videos. They achieve a good effect by simply adding these two constraints to the discriminator.

Overall, the development of expression swap techniques follows the requirements of real-world scenarios, such as the audios as guidance, controllable expression, and temporal consistency across video frames). For the expression swap, we think the improvements of the diversity of input sources, controllable expression, and video consistency are the major technical characteristics that have been evolved throughout the years, and we highlight seminal works A2V, ICface, SVGAN in the bottom right panel of Figure 5, showcasing the improvements in different angles.

3.6 Real Dataset

Real datasets are required for supervised training of DeepFake detectors. Here we introduce popular real datasets. Please note that the following datasets are real image datasets for that the independent real video dataset is infrequent and most of the real face video datasets used in generating the fake datasets are collected by them from YouTube or shot by them with the actors invited by them. We record the information in Table 5 as introduced in Sect. 3.7. The datasets are introduced in ascending order by their release dates.

CASIA-WebFace (Yi et al. 2014) has proposed a semi-automatic way to collect a lot of face images from the Internet. The dataset contains 10,575 subjects and 494,414 images, which is both diverse and large.

CelebA (Liu et al. 2015) is constructed by labeling images selected from a famous face dataset: CelebFaces (Sun et al. 2013). CelebA contains ten thousand identities, each of which has twenty images, a total of 200,000 images. Each image in CelebA is annotated with forty face attributes and five key points by a professional labeling company, which is extremely abundant and useful.

VGGFace (Parkhi et al. 2015) consists of 2,622 identities with 2.6 million images. The famous VGGNet (Parkhi et al. 2015) is trained by this dataset.

MegaFace (Kemelmacher-Shlizerman et al. 2016) includes more than 690K different individuals with one million photos. They have established MegaFace challenge which evaluates how face recognition algorithms perform under the perturbation of a very large number of "distractors" (i.e., individuals that are not in the probe set).

LSUN (Yu et al. 2015) is the only non-face real dataset discussed by us, for that it is widely used by fake generation methods. It contains around one million labeled images for each of 10 scene categories and 20 object categories.

To develop face recognition technologies, Microsoft Celeb (MS-Celeb-1M) (Guo et al. 2016) has collected 10 million face images of nearly 100,000 individuals from the Internet.

VGGFace2 (Cao et al. 2018) contains 3.31 million images of 9,131 subjects. Images are harvested from the Internet and have large variations in pose, age, illumination, ethnicity, and profession.

Karras et al. (2019) has collected a new high-resolution dataset of human faces, Flickr-Faces-HQ (FFHQ). It contains 70,000 high-quality images at 1024 × 1024 resolution. The dataset includes vastly more variation than CelebA-HQ (Karras et al. 2017) in terms of age, ethnicity, and image background. It also has much more accessories such as eyeglasses, sunglasses, hats, etc.

All the above real face datasets can generate DeepFake dataset with three categories (i.e., entire face synthesis, identity swap, expression swap). The datasets which have the abundant label information, especially face attributes are superior ones for attribute manipulation. For example, CelebA and CelebA-HQ are the most usually used real face dataset to generate attribute manipulation images.

3.7 Fake Dataset

Fake image/video datasets are an important benchmark for testing the performance of existing DeepFake generation methods. With the development of fake generation methods, the quality and fidelity of fake datasets are getting higher and higher. Here we introduce popular fake datasets. The datasets are introduced in ascending order by their release dates.

The UADFV dataset (Li et al. 2018b) consists of 98 videos, with 49 real videos from YouTube and 49 synthesized videos, which are made using the FakeAPP (FaceApp 2021).

The DeepFake-TIMIT dataset (Korshunov and Marcel 2018) consists of 620 DeepFake videos of 32 subjects. In DeepFake-TIMIT, each subject has 20 DeepFake videos. 10 videos are of size 64 × 64 while the other 10 videos are of size 128 × 128. The synthesized videos are generated using faceswap-GAN (Lu 2018).

DeepFakes Detection Challenge Preview (DFDC Preview) (Dolhansky et al. 2019) dataset consisting of 5K videos with two facial modification algorithms. The actors are of different gender, skin-tone, age, etc. They record videos with arbitrary backgrounds thus bringing visual variability.

Google DFD (Dufour and Gully 2019) contains over 3,000 manipulated videos from 28 actors in various scenes. The videos are generated from hundreds of real videos by using publicly available DeepFake generation methods.

FaceForensics++ (Rossler et al. 2019) is a famous fake video dataset consisting of 1,000 original video sequences that have been manipulated with four automated face manipulation methods: DeepFakes, Face2Face, FaceSwap, and NeuralTextures. The videos are generated from 977 trackable YouTube videos. The people in most of the videos are frontal faces.

Celeb-DF (Li et al. 2020e) has presented a large-scale challenging DeepFake video dataset, which contains 5639 high-quality DeepFake videos of celebrities generated using an improved synthesis process.

Diverse Fake Face Dataset (DFFD) (Dang et al. 2020) has collected a large-scale dataset that contains numerous types of facial forgeries. Among all images and video frames, 47.7% are from male subjects, 52.3% are from females, and the majority of samples are in the age of 21–50 years. They utilize FFHQ, CelebA, and source frames from Face-Forensics++ as the real face samples. For facial identity and expression swap, they use all the video clips from FaceForensics++. They have adopted two methods FaceAPP (FaceApp 2021) and StarGAN (Choi et al. 2018) to generate attribute manipulated images. Recent works such as PGGAN (Karras et al. 2017) and StyleGAN (Karras et al. 2017) are used for face image synthesis.

FakeCatcher (Ciftci et al. 2020a) has collected over 140 online videos, up to 30 GB. Unlike most of the fake datasets, it includes "in the wild" videos, independent of the generative model, resolution, compression, content, and context.

iFakeFaceDB (Neves et al. 2020) is a fake image dataset for the study of synthetic face manipulation detection. It contains about 87,000 synthetic face images generated by the StyleGAN model (Karras et al. 2019) and transformed with the GANprintR (Neves et al. 2020) approach. All images are of size 224 × 224.

Facebook has constructed an extremely large face video dataset to enable the training of detection models. They organized a famous DeepFake Detection Challenge (DFDC) (Dolhansky et al. 2020) Kaggle competition. The DFDC is a publicly-available face swap video dataset, with 128,514 videos, over 100,000 total clips sourced from 3426 actors. They use various face swap methods with two kinds of augmentations (distractor and augmenter). Distractor means overlaying various kinds of objects (including images, shapes, and text) onto a video while augmenter means applying geometric and color transforms, frame rate changes, etc., onto a video.

Dong et al. (2020) have built a large-scale DeepFake detection dataset "Vox-DeepFake", which has a total of 2 million real videos and fake videos. Compared to existing datasets, it has better quality and diversity in terms of identities and video content. Furthermore, they have supplied the explicit reference identity information for each real/fake video, which is more informative than previous datasets.

DeeperForensics-1.0 (Jiang et al. 2020b) has represented the largest face forgery detection dataset by far, with 60,000 videos constituted by a total of 17.6 million frames. There are 50,000 original collected videos and 10,000 manipulated videos including 100 actors. In particular, 55 of them are males and 45 of them are females. The actors have four typical skin tones: white, black, yellow, brown. All faces are clean without glasses or decorations. Unlike previous data collection in the wild, they build a professional indoor environment for a more controllable data collection and add a mixture of perturbations to videos making the dataset better imitate real-world scenarios. The perturbation is added by systematically applying seven types of distortions (compression, blurry, noise, etc.) to the fake videos at five intensity levels. They also propose DeepFake variational auto-Encoder (DF-VAE) as a new end-to-end face swap method.

As the previous fake datasets were filmed with limited actors in limited scenes, and the fake videos are generated with a few popular DeepFake software, the diversity of the dataset is scarce. In contrast, wild DeepFake can have many persons in one scene, and the scenes vary significantly. Meanwhile, wild DeepFake may even be generated by combinations of DeepFake software. Furthermore, the fake videos in the fake dataset may not be well processed for that the face regions in them often have perceptible distortions. To provide a more realistic DeepFake dataset, Zi et al. (2020) collect their dataset WildDeepfake, which contains 7314 face sequences extracted from 707 DeepFake videos collected completely from the internet. WildDeepfake is able to test the effectiveness of DeepFake detectors against real-world DeepFake.

ForgeryNet (He et al. 2021) has tried to build an extremely large face forgery dataset designed for four tasks: image forgery classification, spatial forgery localization, video forgery classification, temporal forgery localization. This fake dataset contains 2.9 million images and 221,247 videos, which is the largest one among the fake datasets. It also provides 15 manipulation approaches with more than 36 mix-perturbations on over 5400 subjects. The dataset surpasses the other fake datasets both in scale and diversity.

Pu et al. (2021) also pay attention to whether the existing detection methods can effectively adapt to the DeepFake videos in the wild. They build a fake dataset DF-W, which contains 1869 fake videos collected from YouTube, Bilibili (2010) and Reddit.

Most of the fake datasets put emphasis on collecting videos in which only exist one manipulated person. However, the existing detection methods fail to detect the multi-person videos effectively. Thus, Zhou et al. (2021b) build a large dataset FFIW$_{10K}$, which comprises 10,000 high-quality fake videos and real videos, with an average of three human faces in each frame. This fake dataset is more challenging and points out the future research direction of the detection methods.

Similar to FFIW$_{10K}$, OpenForensics (Le et al. 2021) also take care of the capability of the DeepFake detection methods on multi-person images. It contains 45,473 real images and 70,325 fake images, a total of 115,325 images with 334,126 faces in the images. It is worth mentioning that OpenForensics not only introduce multi-face forgery detection task but also propose segmentation in-the-wild task. For these two tasks, they provide face-wise rich annotations such as forgery category, bounding box, segmentation mask, forgery boundary, and general facial landmarks. The abundant annotations make OpenForensics the first dataset that supports the DeepFake localization task, which is meaningful for multi-media forgery forensics. Furthermore, OpenForensics can also be used for general object detection and segmentation tasks, which shows its versatility.

As shown in Table 6, according to the citation metric, FaceForensics and FaceForensics++ are the datasets with the highest citations. According to the Elo rating score, DFDC is the dataset with the highest Elo score. As shown in Fig. 11, FaceForensics++ is the most commonly detected fake dataset for facial appearance swapping detection task while PGGAN is the most commonly detected DeepFake technique for the entire face synthesis detection task. We suggest the researchers put more emphasis on these datasets.

3.8 DeepFake Challenges

In recent years, there have been two famous DeepFake challenge: DeepFake Detection Challenge (DFDC) (Dolhansky et al. 2020) and DeeperForensics Challenge 2020 (Jiang et al. 2021a). The dataset DeeperForensics-1.0 (Jiang et al. 2020b) and DFDC (Dolhansky et al. 2020) used by these two challenges are significantly larger than the previous datasets. They have 100,000 videos and 100,000,000 numbers of frames.

The DeepFake Detection Challenge was hosted on the Kaggle platform[2] by Facebook. During the course of the challenge, 2,114 teams participated. All final evaluations were tested on a private dataset, using a single V100 GPU. Submissions had to run over 10,000 videos within 90 hours. Of all of the scores on the private test set, 60% of submissions had a log loss lower than or equal to 0.69, which is similar to the score if one were to predict a probability of 0.5 for every video. In contrast, the best models achieved very good detection performance on DFDC videos. In the top-5 winning solutions of DFDC, all of them were image-based detection methods. Three of the five methods used EfficientNet (Tan and Le 2019) as the backbone model.

The DeeperForensics Challenge 2020 is hosted on the CodaLab platform[3] in conjunction with ECCV 2020. During the course of the challenge, a total of 115 participants registered for the competition, and 25 teams made valid submissions. Similar to DFDC, the DeeperForensics Challenge uses binary cross-entropy loss (BCELoss) to evaluate the performance of detection models. The evaluation is conducted on a private test set, containing 3000 videos. A total of two online evaluations (each with 7.5 hours of runtime limit) are allowed. Top-3 winning solutions achieved promising performance. Two of them used EfficientNet as the backbone model and all of them used augmentation in model training.

From the challenges, we can find two key points for building a powerful model. First, the backbone selection of the forgery detection models is important. The high-performance winning solutions are based on the state-of-the-art (SOTA) EfficientNet. Second, applying appropriate data augmentations may better simulate real-world scenarios and boost the model performance.

3.9 Summary of DeepFake Generation Methods

DeepFake generation methods have developed rapidly in recent years. Across the four main categories (entire face synthesis, attribute manipulation, identity swap, expression swap) and “other” generation methods, the high quality of the generated images has made it extremely hard for human eyes to distinguish between real and fake. Meanwhile, more and more real image datasets and fake image datasets also promote the development of generation and detection of the DeepFake research field.

However, we still think there’s a large space to improve for the DeepFake generation methods. For example, the resolution of the generated images, manipulable face properties, the continuity of the video, etc., could be further improved, which is introduced in detail in Sect. 7.

To demonstrate the DeepFake generation methods, real datasets, and fake datasets in detail, we build four tables.

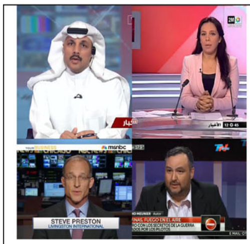
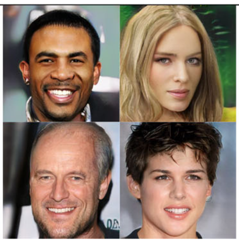
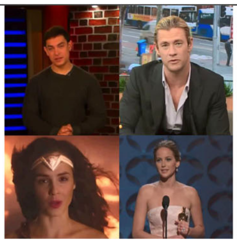
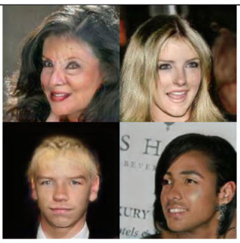
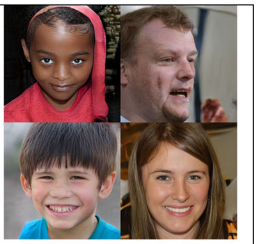
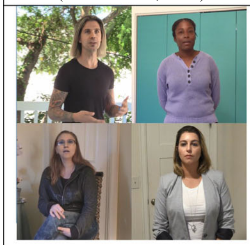
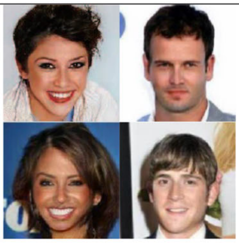
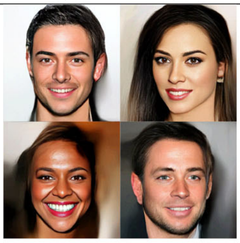
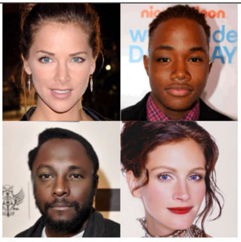
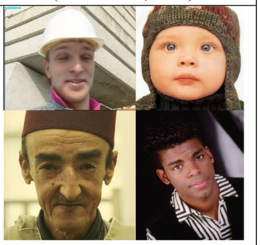
Table 2 shows the images of top-5 generation methods based on battleground (Figure 11) and Elo rating. In the first row, we show the DeepFake generation methods which are attempted the most by DeepFake detection methods according to Figure 11. In the second row, we show the DeepFake generation methods which have the highest Elo rating according to Table 6. To give a more complete impression of the generation methods, Table 3 shows the low-quality images of the GANs included in table 2. Table 4 and 5 mainly introduce the information about real datasets and fake datasets. For real datasets, we have collected the number of images they contain and the diversity of the subjects. For the fake dataset, we have collected the number of images/videos they contain and show the number of real/fake ones clearly.

Table 6 is information statistics of the DeepFake generation methods, real datasets, and fake datasets. It contains the release time, the first author, the citations/days, the citation per day, the abbreviation of the method name, the resolution of images, the Elo rating, and the project URL. For each of the DeepFake generation methods, real datasets and fake datasets, we sort them in ascending chronological order. For the resolution of images in each real dataset, fake dataset, and generation method, we have collected the value from their

[2] https://www.kaggle.com/c/deepfake-detection-challenge.

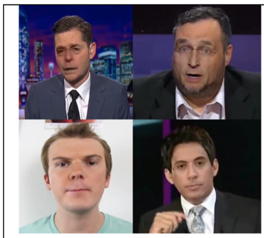
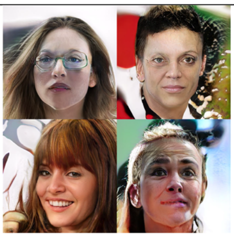
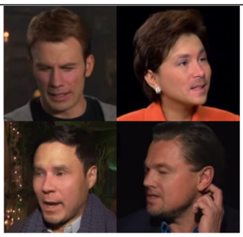
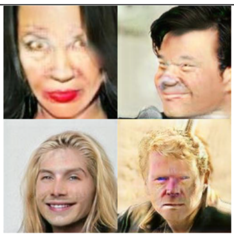
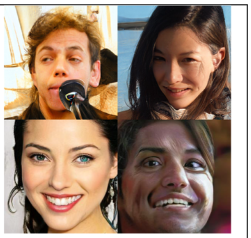
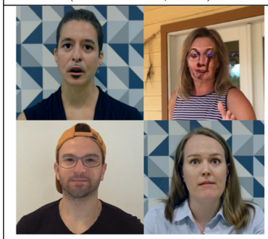
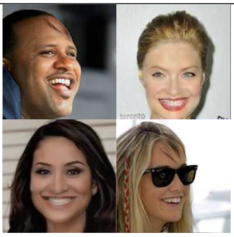
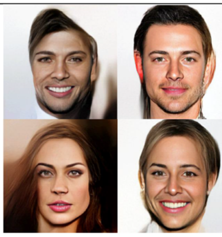
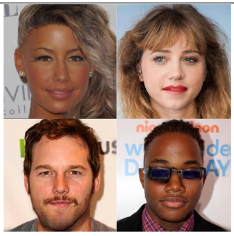
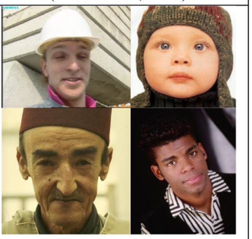
[3] https://competitions.codalab.org/competitions/25228.

Table 2 Top-5 generation methods based on battleground (top) and Elo rating (Wikipedia 2021a) (bottom) separately

				
FaceForensics++ (Rossler et al., 2019)	PGGAN (Karras et al., 2017)	Celeb-DF (Li et al., 2020e)	StarGAN (Miyato et al., 2018)	StyleGAN (Karras et al., 2019)
				
DFDC (Dolhansky et al., 2020)	GDWCT (Cho et al., 2019)	Glow (Kingma and Dhariwal, 2018)	SC-FEGAN (Jo and Park, 2019)	SAN (Dai et al., 2019)

It is worth noting that the "SAN" method is special because it is inherently a super-resolution method. As can be observed, the latest DeepFake generation methods are able to produce highly realistic facial images that are immensely hard to tell apart from real ones using human perception

Table 3 Some lower-quality image examples (that show more visible artifacts) generated by the DeepFake generation methods mentioned in Table 2, with an exception of the SAN (Dai et al. 2019) method, which is a super-resolution technique and super-resolution is not easy to produce low-quality images

				
FaceForensics++ (Rossler et al., 2019)	PGGAN (Karras et al., 2017)	Celeb-DF (Li et al., 2020e)	StarGAN (Miyato et al., 2018)	StyleGAN (Karras et al., 2019)
				
DFDC (Dolhansky et al., 2020)	GDWCT (Cho et al., 2019)	Glow (Kingma and Dhariwal, 2018)	SC-FEGAN (Jo and Park, 2019)	SAN (Dai et al., 2019)

Thus we show the same images as in Table 2

paper. For those which have no exact value, we download the dataset and calculate the resolution of images/videos. For the value which has "aver" label, the resolution is calculated by taking the average resolution of dozens of images/videos. For the value which has "align" label, the resolution is recorded by the common resolution used in the DeepFake research field. For FakeCatcher, we could not find its resources and use "N/A" to label it. As shown in Table 6, according to citation, Elo rating and time, the top-3 models used for each type of DeepFake are DCGAN, GDWCT, StyleGAN2 (entire face synthesis), StarGAN, AttGAN, HifaFace (attribute manipulation), FaceForensics++, DFDC, OpenForensics (identity swap), Face2Face, SVGAN (expression swap).

Table 4 Information of real datasets

Database	Subjects	Images	Img per subject
CASIA-WebFace (Yi et al. 2014)	10,575	494,414	46.75
CelebA(Liu et al. 2015)	10,000	200,000	20
VGGFace2 (Cao et al. 2018)	9131	3,310,000	362.50
FFHQ (Karras et al. 2019)	Unknown	70,000	Unknown
VGGFace (Parkhi et al. 2015)	2622	2,600,000	991.61
Ms-Celeb-1M (Guo et al. 2016)	100,000	10,000,000	100
MegaFace (Kemelmacher-Shlizerman et al. 2016)	690,572	4,700,000	6.81
LSUN (Yu et al. 2015)	30	1,000,000	33333.33

Table 5 Information of fake datasets

Database	Total images/videos	Real images	Fake images	Real videos	Fake videos
UADFV (Li et al. 2018b)	98	–	–	49	49
DeepFake-TIMIT (Korshunov and Marcel 2018)	620	–	–	–	620
DFDC Preview (Dolhansky et al. 2019)	5214	–	–	1131	4113
FaceForensics++ (FF++) (Rossler et al. 2019)	5000	–	–	1000	4000
Celeb-DF (Li et al. 2020e)	6229	–	–	590	5639
FakeCatcher (Ciftci et al. 2020a)	142	–	–	–	142
DFFD (Dang et al. 2020)	303,039	58,703	240,336	1000	3000
Google DFD (Dufour and Gully 2019)	3431	–	–	363	3068
DFDC (Dolhansky et al. 2020)	128,154	–	–	23,954	104,500
DeeperForensics (Jiang et al. 2020b)	60,000	–	–	50,000	10,000
Vox-DeepFake (Dong et al. 2020)	2,171,215	–	–	1,125,429	1,045,786
WildDeepFake (Zi et al. 2020)	7314	–	–	3805	3509
ForgeryNet (He et al. 2021)	3,117,309	1,438,201	1,457,861	99,630	121,617
DF-W (Pu et al. 2021)	1869	–	–	–	1869
FFIW_{10K} (Zhou et al. 2021b)	20,000	–	–	10,000	10,000
OpenForensics (Le et al. 2021)	115,325	45,473	70,325	–	–

Elo rating (Wikipedia 2021a) is widely used in chess and competitive games for calculating the players' ranking. For chess players, they may have different playing styles, which makes ranking difficult. Elo rating can give a relatively objective ranking according to the historical record. Similar to chess players, DeepFake generation methods also have different styles. Thus using the Elo rating to rank the detection difficulty of the generation method is also suitable. Although Elo rating can not give a very accurate ranking, it can give an efficient, intuitive, and objective ranking that is purely based on every single one-on-one battle between a detector and a generator.

This paper mainly focuses on the battleground between DeepFake generation and DeepFake detection. For this purpose, we need to point out the DeepFake generation methods that are most difficult to detect by the existing DeepFake detection methods. Under the battleground, the desired metric should satisfy several requirements: (1) the metric can reflect the historical performance of a DeepFake generation method, (2) the metric should be flexible to reflect that a DeepFake generation method can be evaluated by more than one DeepFake detection methods, and across many research papers, (3) the metric should be as objective as possible to evaluate different DeepFake types (i.e., entire face synthesis, attribute manipulation, identity swap, expression swap). Existing metrics for evaluating deep generative approaches such as PSNR, SSIM are suitable for the tasks that ground truth image is available for the prediction image to compare with. The metrics such as FID, Inception score are suitable for comparing different image feature distributions. It is obvious that these metrics that evaluate the image quality or distributional closeness do not satisfy the three requirements mentioned above. To meet this challenge, we apply the Elo rating, a performance-related metric that is widely used in the ranking of chess or Go players. The Elo rating metric can simultaneously satisfy the three requirements.

For the calculating of Elo rating in DeepFake generation methods, we first set the Elo score of all the DeepFake gener-

Table 6 Summary of the paper information of DeepFake generation methods and related datasets until August 1st, 2021

Time	Author	Citation	Days	Citation per day	Method/dataset	Resolution of images	Elo rating/ rank	Project URL (◇ 1st party project unavailable, 3rd party listed)
2014.11.28	Yi et al. (2014)	1391	2438	0.57	CASIA-WebFace	Align (250,250,3)	N/A	http://cbsr.ia.ac.cn/english/CASIA-WebFace-Database
2015.09.07	Parkhi et al. (2015)	4270	2155	1.98	VGGFace	Aver (569,395,3)	N/A	http://www.robots.ox.ac.uk/~vgg/data/vgg_face/
2015.09.24	Liu et al. (2015)	3956	2138	1.85	CelebA	Aver (613,507,3)	N/A	http://mmlab.ie.cuhk.edu.hk/projects/CelebA
2016.06.04	Yu et al. (2015)	772	1884	0.41	LSUN	Align (256,256,3)	N/A	http://www.yf.io/p/lsun
2016.06.27	Kemelmacher-Shlizerman et al. (2016)	636	1861	0.34	MegaFace	(112,112,3)	N/A	http://megaface.cs.washington.edu/
2016.07.27	Guo et al. (2016)	1044	1831	0.57	Ms-Celeb-1M	Align (112,112,3)	N/A	http://www.msceleb.org/
2018.05.15	Cao et al. (2018)	1209	1378	0.88	VGGFace2	Aver (241,234,3)	N/A	http://zeus.robots.ox.ac.uk/vgg_face2/
2019.03.29	Karras et al. (2019)	1994	856	2.33	FFHQ	(1024,1024,3)	N/A	http://github.com/NVlabs/ffhq-dataset
2018.03.24	Rössler et al. (2018)	177	1226	0.14	FaceForensics	Aver (628,998,3)	1319/78	http://niessnerlab.org/projects/roessler2019FaceForensicspp
2018.06.11	Li et al. (2018b)	292	1147	0.25	UADFV	Aver(459,405,3)	1311/80	◇ http://cutt.ly/Xh5dpZu
2018.12.20	Korshunov and Marcel (2018)	171	955	0.18	DeepFake-TIMIT	(384,512,3)	1386/66	http://www.idiap.ch/dataset/deepfaketimit
2019.09.24	Dufour and Gully (2019)	24	677	0.03	Google DFD	(1080,1920,3)	1342/75	http://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection
2019.10.23	Dolhansky et al. (2019)	123	648	0.19	DFDC Preview	(1080,1920,3)	1470/6	http://ai.facebook.com/datasets/dfdc/
2019.10.29	Rossler et al. (2019)	430	642	0.67	FaceForensics++	Aver (628,998,3)	1170/83	http://github.com/ondyari/FaceForensics
2020.06.14	Li et al. (2020e)	117	413	0.28	Celeb-DF	Aver (498,907,3)	1463/7	http://www.cs.albany.edu/~lsw/celeb-deepfakeforensics
2020.06.14	Dang et al. (2020)	65	413	0.16	DFFD	(1024,1024,3)	1418/16	http://cvlab.cse.msu.edu/dffd-dataset
2020.07.19	Ciftci et al. (2020a)	73	378	0.19	FakeCatcher	N/A	1428/13	http://cs.binghamton.edu/~ncilsal2/DeepFakesDataset/
2020.10.12	Zi et al. (2020)	9	293	0.03	WildDeepFake	N/A	1400/24	http://github.com/deepfakeinthewild/deepfake-in-the-wild
2020.10.28	Dolhansky et al. (2020)	48	277	0.17	DFDC	(1080,1920,3)	***1621/1***	http://ai.facebook.com/datasets/dfdc
2020.12.07	Dong et al. (2020)	0	237	0	Vox-DeepFake	(224,224,3)	1443/10	N/A
2020.12.11	Jiang et al. (2020b)	47	233	0.2	DeeperForensics	aver(560,856,3)	1430/11	http://liming-jiang.com/projects/DrF1/DrF1
2021.03.11	He et al. (2021)	0	143	0	ForgeryNet	aver(670,1145,3)	1400/24	http://yinanhe.github.io/projects/forgerynet
2021.03.18	Zhou et al. (2021b)	0	136	0	FFIW$_{10K}$	N/A	1400/24	http://github.com/tfzhou/FFIW
2021.06.18	Pu et al. (2021)	0	44	0	DF-W	aver(751,1181,3)	1400/24	http://github.com/jmpu/webconf21-deepfakes-in-the-wild

Table 6 continued

Time	Author	Citation	Days	Citation per day	Method/dataset	Resolution of images	Elo rating/ rank	Project URL (◇ 1st party project unavailable, 3rd party listed)
2021.07.30	Le et al. (2021)	2	2	1	OpenForensics	(512,512,3)	1400/24	http://sites.google.com/view/ltnghia/research/openforensics
2016.01.07	Radford et al. (2015)	***9725***	2033	**4.78**	DCGAN	(64,64,3)	1457/8	http://github.com/Newmu/dcgan_code
2016.06.19	FaceSwap (2016)	3	1869	0	FaceSwap	(720,1280)	1400/24	http://github.com/MarekKowalski/FaceSwap/
2016.06.27	Thies et al. (2016)	868	1861	0.47	Face2Face	(720,1280,3)	1413/20	http://gvv.mpi-inf.mpg.de/projects/MZ/Papers/DemoF2F/page
2016.11.13	Mao et al. (2016)	173	1722	0.10	LSGAN	(112,112,3)	1345/74	http://github.com/xudonmao/LSGAN
2016.11.19	Perarnau et al. (2016)	432	1716	0.25	IcGAN	(64,64,3)	1306/81	http://github.com/Guim3/IcGAN
2017.05.30	Bellemare et al. (2017)	195	1524	0.13	CramerGAN	(160,160,3)	1400/24	http://github.com/mbinkowski/MMD-GAN
2017.05.31	Berthelot et al. (2017)	989	1523	0.65	BEGAN	(128,128,3)	1291/79	◇ http://github.com/carpedm20/BEGAN-tensorflow
2017.07.10	Iizuka et al. (2017)	1089	1483	0.73	G&L	(256,256,3)	1400/24	http://github.com/satoshiiizuka/siggraph2017_inpainting
2017.07.10	Suwajanakorn et al. (2017)	533	1483	0.36	A2V	(1080,1920,3)	1400/24	http://github.com/supasorn/synthesizing_obama_network_training
2017.07.22	Shu et al. (2017)	203	1471	0.14	NFE	(64,64,3)	1400/24	http://github.com/zhixinshu/NeuralFaceEditing
2017.07.28	Chen and Koltun (2017)	676	1465	0.46	CRN	(512,1024,3)	1406/20	http://github.com/CQFIO/PhotographicImageSynthesis
2017.09.13	Ding et al. (2018)	106	1418	0.07	ExprGAN	(128,128,3)	1417/17	http://github.com/HuiDingUMD/ExprGAN
2017.10.22	Zhu et al. (2017)	***9156***	1379	***6.64***	CycleGAN	(256,256,3)	1400/24	http://junyanz.github.io/CycleGAN/
2017.12.06	Arjovsky et al. (2017)	**7289**	1334	**5.46**	WGAN	(64,64,3)	1368/72	http://github.com/martinarjovsky/WassersteinGAN
2017.12.25	Gulrajani et al. (2017)	*5147*	1315	*3.91*	WGAN-GP	(128,128,3)	1399/62	http://github.com/igul222/improved_wgan_training
2018.01.28	Lu (2018)	N/A	1281	N/A	Faceswap-GAN	(256,256,3)	1400/24	http://github.com/shaoanlu/faceswap-GAN
2018.02.16	Miyato et al. (2018)	2246	1262	1.78	SNGAN	(128,128,3)	1400/24	http://github.com/niffler92/SNGAN
2018.02.26	Karras et al. (2017)	*3208*	1252	2.56	PGGAN	(1024,1024,3)	1335/76	http://github.com/tkarras/progressive_growing_of_gans
2018.03.21	Bińkowski et al. (2018)	340	1229	0.28	MMDGAN	(160,160,3)	1400/24	http://github.com/mbinkowski/MMD-GAN
2018.03.21	Yu et al. (2018)	1029	1229	0.84	ContextAtten	(512,680,3)	1430/12	http://github.com/JiahuiYu/generative_inpainting

Table 6 continued

Time	Author	Citation	Days	Citation per day	Method/dataset	Resolution of images	Elo rating/ rank	Project URL (◇ 1st party project unavailable, 3rd party listed)
2018.04.18	Natsume et al. (2018)	52	1201	0.04	RSGAN	(128,128,3)	1400/24	N/A
2018.06.19	Chen et al. (2018a)	425	1139	0.37	SITD	(4000,6000,3)	1352/73	http://github.com/cchen156/Learning-to-See-in-the-Dark
2018.07.10	Kingma and Dhariwal (2018)	1110	1118	0.99	Glow	(256,256,3)	**1511/3**	http://github.com/openai/glow
2018.09.08	Pumarola et al. (2018)	346	1058	0.33	GANimation	(128,128,3)	1390/64	http://github.com/albertpumarola/GANimation
2018.09.08	Zhang et al. (2018)	92	1058	0.09	SaGAN	(128,128,3)	1400/24	http://github.com/elvisyjlin/SpatialAttentionGAN
2018.09.21	Choi et al. (2018)	1796	1045	1.72	StarGAN	(256,256,3)	1386/65	http://github.com/yunjey/stargan
2019.02.25	Brock et al. (2018)	1941	888	2.19	BigGAN	(256,256,3)	1446/9	http://github.com/ajbrock/BigGAN-PyTorch
2019.04.04	Chen et al. (2018c)	59	850	0.07	GatedGAN	(128,128,3)	1368/71	http://github.com/xinyuanc91/Gated-GAN
2019.04.28	Thies et al. (2019)	300	826	0.36	Neural-Texture	(512,512,3)	1400/24	◇ http://github.com/SSRSGJYD/NeuralTexture
2019.05.20	He et al. (2019b)	259	804	0.32	AttGAN	(384,384,3)	1426/15	http://github.com/LynnHo/AttGAN-Tensorflow
2019.06.09	Cho et al. (2019)	45	784	0.06	GDWCT	(256,256,3)	***1532/2***	http://github.com/WonwoongCho/GDWCT
2019.06.16	Karras et al. (2019)	1994	777	*2.57*	StyleGAN	(1024,1024,3)	1384/68	http://github.com/NVlabs/stylegan
2019.06.16	Liu et al. (2019)	116	777	0.15	STGAN	(384,384,3)	1390/63	http://github.com/csmliu/STGAN
2019.06.16	Dai et al. (2019)	383	777	0.49	SAN	(392,392,3)	*1486/5*	http://github.com/daitao/SAN
2019.06.16	Yao et al. (2019)	48	777	0.06	AAMS	(512,512,3)	1400/24	http://github.com/JianqiangRen/AAMS
2019.06.16	Chen et al. (2019c)	30	777	0.04	HomointerpGAN	(128,128,3)	1335/77	http://github.com/yingcong/HomoInterpGAN
2019.06.16	Gu et al. (2019)	38	777	0.05	MaskPE	(256,256,3)	1400/24	http://github.com/cientgu/Mask_Guided_Portrait_Editing
2019.10.29	Li et al. (2019b)	27	642	0.04	IMLE	(256,512,3)	1319/79	http://github.com/zth667/Diverse-Image-Synthesis-from-Semantic-Layout
2019.10.29	Jo and Park (2019)	102	642	0.16	SC-FEGAN	(512,512,3)	*1489/4*	http://github.com/run-youngjoo/SC-FEGAN
2019.10.29	Lin et al. (2019)	53	642	0.08	CocoGAN	(384,384,3)	1400/24	http://github.com/hubert0527/COCO-GAN
2019.11.05	Park et al. (2019)	821	635	1.29	GauGAN	(256,512,3)	1383/69	http://github.com/NVlabs/SPADE
2020.03.03	Durall et al. (2020)	46	516	0.09	WUCGAN	(1024,1024,3)	1400/24	http://github.com/cc-hpc-itwm/UpConv
2020.04.26	Choi et al. (2020)	220	462	0.48	StarGANv2	(512,512,3)	1400/24	http://github.com/clovaai/stargan-v2
2020.05.20	Petrov et al. (2020)	36	438	0.08	DeepFaceLab	(448,448,3)	1427/14	http://github.com/iperov/DeepFaceLab
2020.06.16	Karnewar and Wang (2020)	48	411	0.12	MSGGAN	(1024,1024,3)	1385/67	http://github.com/akanimax/msg-stylegan-tf
2020.06.16	Li et al. (2020b)	14	411	0.03	FaceShifter	(256,256,3)	1402/21	http://lingzhili.com/FaceShifterPage/

Table 6 continued

Time	Author	Citation	Days	Citation per day	Method/dataset	Resolution of images	Elo rating/ rank	Project URL (◇ 1st party project unavailable, 3rd party listed)
2020.06.16	Li et al. (2020b)	83	411	0.20	Image2StyleGAN++	(1024,1024,3)	1400/24	N/A
2020.06.16	Li et al. (2020b)	218	411	0.53	InterFaceGAN	(1024,1024,3)	1400/24	http://github.com/genforce/interfacegan
2020.06.16	Li et al. (2020b)	35	411	0.08	GANLocalEditing	(1024,1024,3)	1400/24	http://github.com/cyrilzakka/GANLocalEditing
2020.10.22	Viazovetskyi et al. (2020)	21	283	0.07	StyleGAN 2	(1024,1024,3)	1415/18	http://github.com/NVlabs/stylegan2
2020.11.05	Zhu et al. (2020)	6	269	0.02	AOT	aver(628,998,3)	1400/24	http://github.com/zhuhaozh/AOT
2020.12.03	Noroozi (2020)	1	241	0	slcGAN	(256,256,3)	1400/24	N/A
2020.12.05	Jung and Keuper (2020)	4	239	0.02	SpectralGAN	(256,256,3)	1400/24	http://github.com/steffen-jung/SpectralGAN
2020.12.16	Liu et al. (2021a)	0	228	0	STIGAN	(1024,1024,3)	1400/24	http://github.com/odegeasslbc/Self-Supervised-Sketch-to-Image-Synthesis-PyTorch
2020.12.17	Esser et al. (2021)	30	227	0.13	TTGAN	(1024,1840,3)	1400/24	http://compvis.github.io/taming-transformers/
2020.12.23	Jiang et al. (2020a)	4	221	0.02	FFL	(256,256,3)	1400/24	http://github.com/EndlessSora/focal-frequency-loss
2021.06.18	Afifi et al. (2021)	1	44	0.02	HistoGAN	(1024,1024,3)	1400/24	http://github.com/mahmoudnafifi/HistoGAN
2021.06.18	Wang et al. (2021)	13	44	0.30	vid2vid	(1024,1024,3)	1400/24	http://nvlabs.github.io/face-vid2vid
2021.06.18	Tripathy et al. (2021)	1	44	0.02	FACEGAN	(256,256,3)	1400/24	http://tutvision.github.io/FACEGAN
2021.06.18	Zhu et al. (2021b)	0	44	0	MegaFS	(1024,1024,3)	1400/24	http://github.com/zyainfal/One-Shot-Face-Swapping-on-Megapixels
2021.06.18	Gao et al. (2021a)	0	44	0	InfoSwap	(512,512,3)	1400/24	http://github.com/GGGHSL/InfoSwap-master
2021.06.18	Hyun et al. (2021)	0	44	0	SVGAN	(64,64,3)	1400/24	N/A
2021.06.18	Xia et al. (2021)	4	44	0.09	TediGAN	(1024,1024,3)	1400/24	http://github.com/IIGROUP/TediGAN
2021.06.18	Li et al. (2021a)	0	44	0	FaceInpainter	(512,512,3)	1400/24	N/A
2021.06.18	Zhang et al. (2021b)	1	44	0.02	DLN	(224,224,3)	1400/24	N/A
2021.06.18	Gao et al. (2021d)	0	44	0	HifaFace	(256,256,3)	1400/24	N/A

We mainly show the time, citation, days (from time of exposure), citation per day, method/dataset name, resolution, Elo rating/rank (Wikipedia 2021a) and project URL. From top to bottom, the table is comprised of three sections, and we show real datasets, fake datasets and generation methods in order. The selected time is when the camera-ready version of the paper is released, thus may be later than the time of the citation. For each of them, the papers are sorted by time. We also highlight the top-5 methods of citation, citation per day, and Elo rating. For resolution, we use “aver” to label the dataset which has images/videos of different sizes. For these datasets, we sample the dataset and show the average resolution. We also use “align” to label the datasets which usually be used after alignment and give the resolution of them. In total, 83 DeepFake generation methods are listed in the table ation methods to 1,400. Second, for each DeepFake detection method, we collected the generation method detected by them and sort the generation methods by the detection accuracy/AUC (e.g., if the detection accuracy on generation method A is lower than that on B, then we consider that the quality of A is higher than B, recorded as (A $>$ B)). Third, for the generation rank in each DeepFake detection method, we generate the strong and weak relationship between each pair of them (e.g., if A $>$ B $>$ C, then we generate three relationships A $>$ B, B $>$ C, A $>$ C). Fourth, we update the Elo score of each generation method by the strong and weak relationships. Consider the score of generation methods A and B are **score_A** and **score_B**, if A $>$ B, then the updated score **score_A**$'$ and **score_B**$'$ are calculated by below formula (1)–(4). At last, we sort the generation methods by their Elo scores. Elo rating helps us to build strong and weak relationships with limited game information of different DeepFake generation methods. However, the Elo rating system has its inherent shortcoming. If a generation method only appears once but beats high-score competitors, its score will go up a lot. On the other hand, if a generation method has not been detected for a long time, its score will remain unchanged and cannot reflect its true difficulty.

$$\textbf{score\_A\_adjust} = \frac{1}{1 + 10^{\frac{\textbf{score\_B}-\textbf{score\_A}}{400}}} \tag{1}$$

$$\textbf{score\_B\_adjust} = \frac{1}{1 + 10^{\frac{\textbf{score\_A}-\textbf{score\_B}}{400}}} \tag{2}$$

$$\textbf{score\_A}' = \textbf{score\_A} + 32 * (1 - \textbf{score\_A\_adjust}) \tag{3}$$

$$\textbf{score\_B}' = \textbf{score\_B} - 32 * \textbf{score\_B\_adjust} \tag{4}$$

4 Detection of DeepFakes

In recent years, studies are continuously working on developing various techniques to identify whether a still image or video is synthesized with AI (especially manipulated with GANs and its variants) or produced naturally with a camera. Generalize to unseen synthesized techniques, robust against various attacks (e.g., adversarial attacks, image/video transformations), and providing explainable detection results are three critical factors for a detector practicality deployed in the wild. In this section, we mainly review recent studies on DeepFake detection based on their extracted features (e.g., spatial (Sect. 4.1), frequency (Sect. 4.2), and biological signals (Sect. 4.3)) and introduce their performance on the aforementioned three essential factors. In Sect. 4.4, we detail the methods that can not be classified into the three typical DeepFake detection methods. To better present the DeepFake detection methods to readers, we use three tables (Table 7) to summarize the existing DeepFake detection methods, a chord diagram (Fig. 14) to show the performance among various DeepFake detectors in Sect. 4.5, and a fishbone diagram (Fig. 10) to present the evolution of three typical detection techniques.

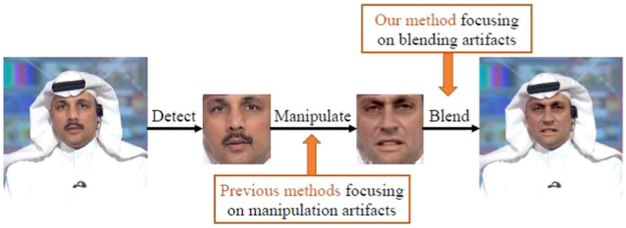

Our method focusing on blending artifacts
Detect
Manipulate
Blend
Previous methods focusing on manipulation artifacts

Fig. 6 The difference between real and fake from the spatial domain, especially the discrepancies across the blending the boundary (Li et al. 2020c)

4.1 Spatial based Detection

Recently, detecting DeepFakes on the spatial domain is the most popular techniques adopted by existing studies. They observe various visible or invisible artifacts on the spatial domain for distinguishing real and fake. Figure 6 shows the potential of working on spatial domain for DeepFake detection.

4.1.1 Image Forensics based Detection

The traditional forensics-based techniques inspect the disparities in pixel-level, which is investigated by recent studies for DeepFake detection. They provide explainable clues in the detection and introduce the differences between real and fake. However, these works suffer the robustness issues when the images or videos are manipulated by simple transformations.

Li et al. (2020a) observe that the differences between synthesized faces and real faces are revealed in the chrominance components, especially in the residual domain. They propose to train a one-class classifier on real faces by leveraging the differences in the chrominance components for tackling the unseen GANs. However, their performance against perturbation attacks like image transformations is unknown.

Photo response non uniformity (PRNU) pattern is a noise pattern in a digital image caused by the light sensor in camera, which could be applied for distinguishing DeepFakes from authentic videos (Koopman et al. 2018). Others explore utilizing the co-occurrence matrices for differentiating real and fake faces (Nataraj et al. 2019). The insight behind these works is obvious, but their effectiveness in tackling challenging high-quality DeepFakes is not clear.

Similarly, in tackling the fake videos, researchers also borrow the ideas from the traditional video forensic by leveraging the local motion features captured from real videos to spot the abnormality of manipulated videos (Wang et al.

Table 7 Summary of existing DeepFake detection methods

Time	Author	Method	Classifier	Performance		Databases	Multimedia		B/L	Capabilities		
				Worst	Best		Img	Vid		G	R	E
2017.08.04	Zhang et al. (2017)	Type I5	SVM, RF, MLP	ACC: 0.654	ACC: 0.9355	Self-built	✓	✗	N/A	✗	✗	✗
2018.03.29	Zhou et al. (2017)	Type I2	CNN	N/A	AUC: 0.927	Self-built	✓	✗	B	✗	✗	✗
2018.04.10	Marra et al. (2018)	Type I2	N/A	N/A	N/A	Self-built	✓	✗	N/A	✗	✓	✗
2018.06.11	Li et al. (2018b)	Type III2	CNN	N/A	AUC: 0.99	UADFV	✗	✓	N/A	✗	✗	✓
2018.06.20	Mo et al. (2018)	Type I2	CNN	N/A	ACC: 0.994 PGGAN	Self-built	✓	✗	N/A	✗	✗	✗
2018.08.29	Koopman et al. (2018)	Type I1	N/A	N/A	N/A	Self-built	✗	✓	N/A	✗	✗	✗
2018.09.04	Afchar et al. (2018)	Type I2	CNN	ACC: 0.891	ACC: 0.984	FF++	✗	✓	N/A	✗	✗	✗
		http://github.com/DariusAf/MesoNet										
2018.10.15	Tariq et al. (2018)	Type I2	CNN	ACC: 0.9399	ACC: 0.9999	Self-built	✓	✗	T	✗	✗	✗
2018.10.18	Hsu et al. (2018)	Type I2	CNN	ACC: 0.818 WGAN-GP	ACC: 0.947 LSGAN	Self-built	✓	✗	B	✗	✗	✗
2018.11.12	Li et al. (2018a)	Type IV1	N/A	N/A	N/A	Self-built	✓	✗	N/A	✓	✗	✗
2018.11.27	Güera and Delp (2018)	Type III2	RNN	ACC: 0.967	ACC: 0.971	Self-built	✗	✓	N/A	✗	✗	✓
2018.12.20	Korshunov and Marcel (2018)	Type III1	CNN	N/A	EER: 0.0333	DeepFake-TIMIT	✗	✓	B	✗	✗	✓
2019.01.07	Matern et al. (2019)	Type III2	KNN, MLP, LR	AUC: 0.843 Glow	AUC: 0.866	FF	✓	✗	N/A	✗	✗	✓
2019.02.26	Chen et al. (2019d)	Type IV1	SVM	ACC: 1.0	ACC: 1.0	Self-built	✓	✗	N/A	✗	✓	✗
2019.03.28	Marra et al. (2019a)	Type II1	N/A	ACC: 0.99	ACC: 0.998	Self-built	✓	✗	N/A	✓	✗	✗
2019.03.30	Yang et al. (2019b)	Type III2	SVM	AUC: 0.83	AUC: 0.9413	Self-built	✓	✓	T	✗	✗	✓
2019.05.12	Nguyen et al. (2019b)	Type I2	CNN	ACC: 0.8333	ACC: 0.9933	FF	✗	✓	T	✗	✓	✗
2019.05.12	Yang et al. (2019a)	Type III2	SVM	AUC: 0.843	AUC: 0.89	MFC, UADFV	✗	✓	N/A	✗	✗	✓
2019.05.16	Sabir et al. (2019)	Type I2	RNN	ACC: 0.9843	ACC: 0.9959	FF++	✗	✓	T	✗	✗	✗
2019.05.22	Li and Lyu (2019)	Type III2	CNN	ACC: 0.932	ACC: 0.999	UADFV, DeepFake-TIMIT	✗	✓	B	✗	✗	✓
2019.06.17	Nguyen et al. (2019a)	Type I2	CNN	ACC: 0.5232	ACC: 0.9277	FF++	✗	✓	P	✓	✗	✗
2019.06.18	Agarwal et al. (2019)	Type I3	SVM	AUC: 0.93	AUC: 1	Self-built	✗	✓	N/A	✗	✗	✓
2019.06.18	Fernandes et al. (2019)	Type III3	LSTM, VAE	AUC: 0.93	AUC: 1	Self-built	✓	✓	N/A	✗	✗	✓
2019.06.18	Amerini et al. (2019)	Type I2	CNN	ACC: 0.7546	ACC: 0.8161	FF++	✗	✓	N/A	✗	✗	✗
2019.08.16	Yu et al. (2019b)	Type II1	CNN	ACC:0.9766	ACC: 0.9950	Self-built	✓	✗	B	✓	✗	✓
		https://github.com/ningyu1991/GANFingerprints										
2019.09.01	Dang et al. (2019)	Type I2	CNN	AUC: 0.90	AUC: 0.934	Self-built	✓	✗	B	✗	✗	✗
2019.09.22	He et al. (2019a)	Type I2	CNN	ACC: 0.9987	ACC: 1.0	Self-built	✓	✗	P	✗	✓	✗
2019.09.22	McCloskey and Albright (2019)	Type II1	SVM	AUC: 0.61	AUC: 0.92	Self-built	✓	✗	N/A	✗	✗	✗

Table 7 continued

Time	Author	Method	Classifier	Performance		Databases	Multimedia		B/L	Capabilities		
				Worst	Best		Img	Vid		G	R	E
2019.10.03	Nataraj et al. (2019)	Type I1	CNN	ACC: 0.9937 StarGAN	ACC: 9971 cycleGAN	Self-built	✔	✘	B	✔	✔	✘
2019.10.06	Marra et al. (2019b)	Type I2	Multi-task Incremental Learning	ACC: 0.6771	ACC: 0.9937	Self-built	✔	✘	B	✘	✘	✘
2019.10.12	Songsri-in and Zafeiriou (2019)	Type I4	CNN	N/A	ACC: 0.9926	FF++	✔	✔	T	✘	✘	✘
2019.10.15	Zhang et al. (2019)	Type II1	CNN	ACC: 0.786	ACC: 1	Self-built	✔	✘	B	✔	✔	✔
		https://github.com/ColumbiaDVMM/AutoGAN										
2019.10.29	Nguyen et al. (2019c)	Type I2	CNN	N/A	ACC: 0.9311	FF++	✘	✔	T	✘	✘	✘
2019.11.11	Sohrawardi et al. (2019)	Type I3	LSTM	ACC: 0.86	ACC: 0.95	FF++	✘	✔	P	✘	✔	✘
2019.11.17	Fernando et al. (2019)	Type I2	HMN	ACC: 0.8412	ACC: 0.9997	FF++	✔	✔	B	✔	✔	✘
2019.11.27	Cozzolino et al. (2018)	Type I2	CNN	ACC: 0.7062	ACC: 1.0	Self-built	✔	✘	B	✔	✘	✘
2019.12.10	Xuan et al. (2019)	Type I2	CNN	ACC: 0.6055	ACC: 0.9545	Self-built	✔	✘	N/A	✔	✘	✘
2019.12.12	Li et al. (2019a)	Type I4	CNN	ACC: 0.9835	AUC: 0.9918	FF++	✘	✔	T	✘	✘	✘
2019.12.21	Yu et al. (2019a)	Type I2	CNN	ACC: 0.805	ACC: 0.981	FF++	✔	✔	T	✘	✘	✘
2020.01.03	Hsu et al. (2020)	Type I2	CNN	Pre.: 0.929 DCGAN	Pre.: 0.988 WGAN	Self-built	✔	✘	P	✘	✘	✘
2020.01.21	Kumar et al. (2020)	Type I2	CNN	ACC.: 0.9120	ACC.: 0.9996	FF	✘	✔	B	✘	✔	✘
2020.02.11	Tarasiou and Zafeiriou (2020)	Type I2	CNN	ACC: 0.8876	ACC: 0.9803	Google DFD, FF++, Celeb-DF	✘	✔	N/A	✘	✘	✘
2020.03.03	Durall et al. (2020)	Type II2	CNN	ACC: 0.85	ACC: 0.90	FF++	✘	✔	N/A	✘	✘	✔
2020.03.04	Durall et al. (2019)	Type II2	SVM, K-Means, LR	ACC: 0.9	ACC: 1	FF++	✔	✔	N/A	✘	✘	✘
		https://github.com/cc-hpc-itwm/DeepFakeDetection										
2020.03.19	Liu et al. (2020b)	Type I2	CNN	ACC: 0.9854	ACC: 0.991	Self-built	✔	✘	B	✔	✔	✘
2020.03.19	Kumar et al. (2020)	Type I2	CNN	AUC: 0.8273	AUC: 0.997	FF++, Celeb-DF	✘	✔	B	✘	✘	✘
2020.03.27	Mansourifar and Shi (2020)	Type I2	CNN	N/A	ACC: 0.81	Self-built	✔	✘	P	✘	✘	✘
2020.04.06	Dogonadze et al. (2020)	Type I2	CNN	N/A	ACC: 0.748	FF++	✔	✔	B	✘	✘	✘
		https://github.com/Megatvini/DeepFaceForgeryDetection/										
2020.04.16	Bonettini et al. (2021a)	Type II2	RF	ACC: 0.9813	ACC: 1	Self-built	✔	✘	B	✘	✔	✘
2020.04.16	Bonettini et al. (2021b)	Type I2	CNN	AUC: 0.8785	AUC: 0.9444	DFDC, FF++	✘	✔	T	✘	✘	✘
		github.com/polimi-ispl/icpr2020dfdc										
2020.04.19	Li et al. (2020c)	Type I4	HRNet (Sun et al. 2019)	AUC: 0.7115	AUC: 0.9540	Celeb-DF, DFDC, FF++, Google DFD	✘	✔	B	✔	✘	✔
2020.04.22	Guarnera et al. (2020a)	Type I5	KNN, SVM, LDA	ACC: 0.8840 GDWCT	ACC: 0.9981 StyleGAN2	Self-built	✔	✘	N/A	✘	✘	✔
2020.04.29	Agarwal et al. (2020a)	Type I3	CNN	AUC: 0.824	AUC: 0.989	Celeb-DF, DFDC-Preview	✘	✔	B	✔	✘	✘
2020.05.04	Wu et al. (2020)	Type I2	CNN+RNN	ACC: 0.9011	ACC: 0.9857	FF++	✘	✔	B	✘	✘	✘

Table 7 continued

Time	Author	Method	Classifier	Performance		Databases	Multimedia		B/L	Capabilities		
				Worst	Best		Img	Vid		G	R	E
2020.05.12	Hulzebosch et al. (2020)	Type I5	CNN	ACC: 0.983	ACC: 0.998	Self-built	✓	✗	N/A	✓	✓	✓
2020.05.17	Sambhu and Canavan (2020)	Type I2	CNN	ACC: 0.993	ACC: 0.996	FF++	✓	✓	T	✗	✗	✗
2020.05.19	Ding et al. (2020)	Type I2	CNN	ACC: 0.9197	ACC: 1	Self-built	✓	✗	N/A	✗	✗	✗
2020.06.01	Chugh et al. (2020)	Type III1	N/A	ACC: 0.916	ACC: 0.983	DFDC, DeepFake-TIMIT	✗	✓	B	✗	✗	✓
2020.06.09	Huang et al. (2022)	Type I4	CNN	N/A	N/A	Self-built	✓	✗	N/A	✓	✓	✓
2020.06.16	Mas Montserrat et al. (2020)	Type I2	RNN, CNN	-	log-likelihood err: 0.321	DFDC	✗	✓	B	✓	✗	✗
2020.06.16	Agarwal et al. (2020b)	Type III1	CNN	ACC: 0.928	ACC: 0.97	Self-built	✗	✓	N/A	✗	✗	✓
2020.06.16	Tursman et al. (2020)	Type III2	Hierarchical Clustering	ACC: 0.60	ACC: 0.98	Self-built	✗	✓	T	✗	✗	✓
2020.06.16	Wang et al. (2020e)	Type I2	CNN	Pre.: 0.529	Pre.: 1.0	FF++	✓	✓	P	✓	✓	✗
		https://github.com/peterwang512/CNNDetection										
2020.06.16	Khalid and Woo (2020)	Type I2	One-class VAE	ACC: 0.712	ACC: 0.982	FF++	✗	✓	B	✗	✗	✗
2020.06.19	Bai et al. (2020)	Type II2	CNN	ACC: 0.9319	ACC: 0.9768	Self-built	✓	✗	B	✗	✗	✗
2020.06.26	Frank et al. (2020)	Type II2	CNN	ACC: 0.9780	ACC: 0.9991	Self-built	✓	✗	T	✗	✓	✓
		https://github.com/RUB-SysSec/GANDCTAnalysis										
2020.06.26	de Lima et al. (2020)	Type I2	CNN	ACC: 0.7625	ACC: 0.9826	Celeb-DF	✗	✓	B	✗	✗	✗
		https://github.com/oidelima/Deepfake-Detection										
2020.06.28	Trinh et al. (2021)	Type I2	DNN	ACC: 0.9037	ACC: 0.9625	FF++, Celeb-DF	✗	✓	B	✗	✗	✓
		github.com/loc-trinh/DPNet										
2020.07.02	Tolosana et al. (2020)	Type IV1	N/A	N/A	N/A	FF++, Celeb-DF, DFDC, UADFV	✗	✓	N/A	✗	✗	✗
2020.07.15	Pishori et al. (2020)	Type IV1	LSTM, Eye Blink,Grayscale Histograms	ACC: 0.8167	ACC: 0.8571	DFDC	✗	✓	B	✗	✗	✓
2020.07.16	Wang et al. (2020d)	Type I5	CNN	ACC: 0.682 DFDC ACC: 0.88 StarGAN	ACC: 0.985 FF++ ACC: 0.986 PGGAN	Celeb-DF FF++, DFDC	✓	✓	B	✓	✗	✓
2020.07.19	Ciftci et al. (2020a)	Type III3	CNN	ACC: 0.9107	ACC: 0.96	FF++, Celeb-DF, FF, FakeCatcher	✗	✓	B	✓	✗	✓
2020.07.20	Goebel et al. (2020)	Type I4	CNN	N/A	ACC: 0.9916	Self-built	✓	✗	B	✓	✓	✓
2020.07.30	Bonomi et al. (2020)	Type I3	SVM	ACC: 0.8595	ACC: 0.9024	FF++	✗	✓	B	✓	✓	✓
2020.08.01	Mittal et al. (2020)	Type III1	N/A	ACC: 0.844	ACC: 0.966	DFDC, DeepFake-TIMIT	✗	✓	B	✓	✗	✓
		https://gamma.umd.edu/deepfakes/										
2020.08.07	Guarnera et al. (2020b)	Type I3	Random Forest	ACC: 0.7219	ACC: 0.9964	Self-built	✓	✓	B	✗	✓	✓
2020.08.10	Jeon et al. (2020b)	Type I2	SVM, RF, MLP	ACC: 0.654	ACC: 0.9355	Self-built	✓	✓	N/A	✗	✗	✗

Table 7 continued

Time	Author	Method	Classifier	Performance		Databases	Multimedia		B/L	Capabilities		
				Worst	Best		Img	Vid		G	R	E
		github.com/cutz-j/FDFtNet										
2020.08.10	Jeon et al. (2020a)	Type I2	CNN	AUC: 0.7969	AUC: 0.9839	Self-built	✔	✘	B	✔	✘	✘
		https://github.com/cutz-j/T-GD										
2020.08.11	Li et al. (2020d)	Type I2	S-MIL	ACC: 0.7535	ACC: 1.0	Celeb-DF, FF++, DFDC	✘	✔	B	✘	✘	✘
2020.08.11	Wang et al. (2020a)	Type I1	AdaBoost	ACC: 0.565	ACC: 0.991	FF++	✘	✔	B	✔	✔	✔
2020.08.24	Chai et al. (2020)	Type I3	CNN	Pre.: 0.9138	Pre.: 1	FF++	✔	✔	T	✔	✘	✔
		https://github.com/chail/patch-forensics										
2020.08.26	Qi et al. (2020)	Type III3	CNN	ACC: 0.641	ACC: 1.0	FF++, DFDC	✘	✔	B	✘	✔	✔
2020.08.26	Ciftci et al. (2020b)	Type III2	CNN	ACC: 0.8662	ACC: 0.9466	FF++, Celeb-DF	✘	✔	T	✔	✘	✔
2020.08.27	Nirkin et al. (2020)	Type I3	CNN	AUC: 0.66	AUC: 0.997	Celeb-DF, FF++, DFDC	✘	✔	B	✔	✘	✔
2020.08.31	Li et al. (2020a)	Type I1	one-class	ACC: 0.9795	ACC: 1	Self-built	✔	✘	B	✔	✘	✔
2020.09.04	Masi et al. (2020)	Type I2	RNN	AUC: 0.8659	AUC: 0.9912	Celeb-DF, FF++, DFDC	✘	✔	B	✔	✘	✘
2020.09.13	Feng et al. (2020)	Type I2	CNN	AUC: 0.999	AUC: 0.999	UADFV, Celeb-DF, FF++	✘	✔	B	✔	✘	✘
2020.09.16	Tariq et al. (2020)	Type I2	Convolutional LSTM	ACC: 0.7412	ACC: 0.995	FF++	✘	✔	B	✔	✘	✘
2020.09.20	Du et al. (2019)	Type IV	CNN	ACC: 0.5905	ACC: 0.9991	Self-built	✔	✔	B	✔	✘	✘
2020.10.02	Barni et al. (2020)	Type II2	CNN	-	ACC: 0.997	Self-built	✔	✘	N/A	✔	✔	✔
		https://github.com/ehsannowroozi/FaceGANdetection										
2020.10.12	Hu et al. (2021)	Type III2	N/A	N/A	AUC: 0.94	Self-built	✔	✘	N/A	✘	✘	✔
2020.10.22	Ganiyusufoglu et al. (2020)	Type I2	3D CNN	Pre.: 0.9429	Pre.: 0.9929	FF++, DFDC	✘	✔	T	✔	✘	✘
2020.10.24	Dang et al. (2020)	Type I4	CNN	ACC: 0.712	ACC: 0.984	Celeb-DF, UADFV, DFFD	✔	✘	B	✔	✘	✔
2020.10.27	Qian et al. (2020)	Type II2	CNN	ACC: 0.9043	ACC: 0.9999	FF++	✘	✔	B	✘	✔	✘
2020.10.27	Yu et al. (2020b)	Type II2	CNN	ACC: 0.9895	ACC: 0.9975	FF++, DFDC	✔	✔	B	✔	✘	✘
2020.10.28	Chen and Yang (2021)	Type I2	CNN	ACC: 0.9575	ACC: 0.9998	FF++, Celeb-DF	✔	✔	B	✔	✘	✔
2020.11.16	Bondi et al. (2020)	Type IV1	CNN	AUC: 0.922	AUC: 0.998	Celeb-DF, FF++, DFDC	✘	✔	T	✔	✔	✘
2020.11.19	Zhu et al. (2021a)	Type I2	CNN	ACC: 0.8731	ACC: 0.9972	FF++, DFDC, Google DFD	✘	✔	B	✔	✔	✔
2020.11.19	Wang et al. (2020f)	Type I3	CNN	ACC: 0.9438	ACC: 0.9935	FF++	✘	✔	P	✔	✘	✔
2020.12.04	Cozzolino et al. (2020)	Type I2	CNN	AUC: 0.863	AUC: 0.960	FF++, DFDC, Google DFD	✘	✔	T	✘	✔	✘
2020.12.07	Dong et al. (2020)	Type IV1	CNN	AUC: 0.9061	AUC: 0.9854	Google DFD, FF++, Celeb-DF, Vox-DeepFake	✘	✔	B	✔	✔	✔
2020.12.07	Pu et al. (2020)	Type IV1	CNN	F1: 0.9014	F1: 0.9963	Self-built	✔	✘	P	✔	✔	✔

Table 7 continued

Time	Author	Method	Classifier	Performance		Databases	Multimedia		B/L	Capabilities		
				Worst	Best		Img	Vid		G	R	E
2020.12.08	Kukanov et al. (2020)	Type IV1	CNN	EER: 7.16	EER: 6.03	DeepFake-TIMIT, FF++	✔	✘	P	✘	✘	✘
2020.12.14	Hernandez-Ortega et al. (2020)	Type III3	Convolutional attention net	ACC: 0.944	ACC: 0.987	Celeb-DF, DFDC-Preview	✘	✔	P	✘	✘	✔
2020.12.14	Haliassos et al. (2021)	Type III2	CNN	AUC: 0.735	AUC: 0.997	DeeperForensics, Celeb-DF, FF++, DFDC	✔	✘	B	✔	✔	✔
2020.12.16	Zhao et al. (2020)	Type I3	CNN	AUC: 0.9438	AUC: 0.9998	Google DFD, Celeb-DF, DFDC, FF++, DFDC-Preview, DeeperForensics	✘	✔	P	✔	✔	✔
2020.12.16	Guo et al. (2020c)	Type I2	CNN	ACC: 0.9102	ACC: 0.9852	FF++	✔	✔	T	✔	✔	✘
2020.12.19	Sun et al. (2020a)	Type I2	CNN	AUC: 0.681	AUC: 0.738	FF++, Celeb-DF	✘	✔	B	✘	✔	✔
2021.06.18	Zhao et al. (2021)	Type I2	N/A	AUC:0.904	AUC:0.993	Celeb-DF FF++, DFDC	✔	✔	B	✔	✔	✔
		https://github.com/yoctta/multiple-attention										
2021.06.18	Liu et al. (2021b)	Type II2	CNN	AUC: 0.828	AUC: 0.953	Celeb-DF, FF++, DFDC	✔	✔	B	✔	✔	✔
2021.06.18	Agarwal and Farid (2021)	Type III1	N/A	AUC: 0.70	AUC: 0.98	Self-build	✘	✔	N/A	✔	✘	✔
2021.06.18	Schwarcz and Chellappa (2021)	Type I2	CNN	AUC: 0.586	AUC: 0.965	Celeb-DF, FF++, DFDC	✔	✔	N/A	✔	✘	✔
2021.06.18	Wang and Deng (2021)	Type I2	CNN	AUC: 0.934	AUC: 1.0	DFFD, Celeb-DF	✔	✔	T	✔	✘	✔
2021.06.18	Li et al. (2021b)	Type II2	CNN	ACC: 0.890	ACC: 0.994	FF++	✔	✔	B	✔	✔	✔
2021.06.18	Luo et al. (2021)	Type II2	CNN	AUC: 0.497	AUC: 0.995	DeeperForensics, FF++ Google DFD, DFDC, Celeb-DF	✔	✔	B	✔	✔	✔
2021.06.18	Kim et al. (2021)	Type I2	N/A	ACC:0.731	ACC:0.870	FF++	✔	✔	B	✔	✔	✘
		https://github.com/alsgkals2/FReTAL										
2021.06.18	Chandrasegaran et al. (2021)	Type II2	KNN	ACC:0.677	ACC:0.999	Self-build	✔	✔	N/A	✔	✔	✔
		https://keshik6.github.io/Fourier-Discrepancies-CNN-Detection										

We mainly show the time, author, the method type, employed classifier, the performance (e.g., worst and best performance), evaluation databases, evaluation object (e.g., image or video), compared baselines, and the detection capabilities in DeepFake detection. P, T, and B are employed for representing the baselines. P represents the method simply employs peer works for comparison, T indicates the method simply adopts some simple CNN models for comparison, B denotes the method leverages both peer works and conventional CNN models for comparison. G, R, and E are adopted for representing the capabilities of detection method. G represents whether the method has the generalization capabilities in tackling unseen DeepFakes, R indicates whether the method is robust against various attacks, especially image/ video transformations, E denotes whether the method provides explainable detection results. In total, there are 117 methods listed in the table

2020a). Leveraging the image/video forensic techniques is a direct idea for fighting against DeepFake by focusing on the low-level features, but they are not practical to be deployed in the wild where DeepFakes suffer known and unknown degradations.

4.1.2 DNN-based Detection

These methods are totally data-driven by utilizing existing or designing new DNN-based models by extracting spatial features to improve the effectiveness and generalization ability of detection. However, these DNN-based detection methods all suffer from the adversarial attacks with additive noises and all the studies failed in evaluating their effectiveness in tackling adversarial noise attacks (Carlini and Farid 2020). The existing studies by leveraging DNN to identify DeepFakes can be classified into the following three categories.

Improving Generalization Abilities Conventional DNNs have been widely applied in detecting fake faces, but they will overfit to specific manipulation types and suffer the transferability issues where the capabilities of unseen manipulation methods are lacking. Thus, motivated by the social perception and social cognition processes of the human brain, a novel hierarchical memory network (HMN) is employed for detecting fake faces to address the transferability issues and improve the effectiveness in tacking unknown GANs (Fernando et al. 2019).

Gram-Net (Liu et al. 2020b) improves the robustness and generalization ability to existing CNNs in discriminating synthesized fake faces by leveraging global texture features. Experimental results indicate that Gram-Net shows strong robustness against perturbation attacks like downsampling, JPEG compression, blur, and noise. Additionally, Gram-Net claims its generalization ability in tackling different GANs, which has shown promising applications in the wild.

Wang et al. (2020e) observe that a binary classifier by leveraging a simple ResNet-50 as the backbone which is a pre-trained model with ImageNet shows strong generalization capabilities in detecting GAN-synthesized still images. Their classifier is merely trained on PGGAN database and could be well generalized to other GAN architectures like StyleGAN, StarGAN, etc. Experiments show that they are robust against perturbation attacks by incorporating various data augmentation strategies into the training process. Obviously, detecting still images synthesized by SOTA GANs like StyleGAN is not hard.

OC-FakeDect (Khalid and Woo 2020) proposes training on real faces with a one-class variational autoencoder (VAE) to detect synthesized fake faces. OC-FakeDect is vastly different from all the existing detectors which is highly dependent on collected fake faces by using binary classifiers. They claim that the approach has a good performance in generalizing across different DeepFake methods, but their robustness against perturbation attacks is unclear.

Investigating the Artifact Clues In order to focus on the intrinsic forensics clues, image preprocessing by using smoothing filtering or noise is employed for destroying the low-level unstable artifacts in GAN-synthesized images (Xuan et al. 2019). Investigating the intrinsic clues could significantly improve the generalization ability of the CNN model in identifying unknown GANs.

CLRNet (Tariq et al. 2020) proposes a convolutional long short-term memory (LSTM) based residual network for capturing the temporal information in consecutive frames. Transfer learning is employed for improving the generalization ability in tackling fake videos created with different synthetic methods.

SSTNet (Wu et al. 2020) incorporates the spatial, steganalysis, and temporal features for detecting DeepFakes. Specifically, a deep model XceptionNet is employed for extracting spatial features, a simplified XceptionNet with a constraint on the conventional filter to learn statistical characteristics of image pixels for steganalysis feature extraction, recurrent neural network (RNN) is applied for extracting temporal features. SSTNet reveals that combing multi-modal features from a wide range of levels is a promising idea for developing powerful DeepFake detectors.

Instead of using various CNN models for addressing general object recognition tasks, researchers observed that the face recognition systems focus on learning the representation of faces. Thus, Nhu et al. (2018) employ a deep face recognition system to extract the face representation for building a binary classifier to detect real and fake faces. DPNet (Trinh et al. 2021) captures the temporal dynamic features with a carefully designed DNN to build an *interpretable* DeepFake detection framework to explain why a video is predicted as fake or real.

In DeepFake videos, the temporal artifacts across frames indicate the abnormal face in the video. A recurrent convolutional network is exploited to capture the temporal discrepancies in fake videos (Sabir et al. 2019). de Lima et al. (2020) have employed the temporal information by leveraging various typical CNNs for DeepFake detection.

Empowering CNN Models FDFtNet (Jeon et al. 2020b) provides a reusable fine-tuning network to improve the capabilities of existing CNN models (e.g., SqueezeNet, ShallowNetV3, ResNetV2, and Xception) in detecting fake images effectively. A fine-tune transformer (FTT) is designed with self-attention to extract different features from the image, then an MBblockV3 adopts different convolution and structure techniques to extract features. FDFtNet outperforms the baselines by using various CNNs. However, their robustness in tacking with unseen GANs and the evaluation on perturbation attacks is still unclear.

Inspired by the advances of deep learning, various DNN-based approaches are continuously proposed for distinguishing synthesized fake faces, such as deep transfer learning (Ding et al. 2020; Dogonadze et al. 2020; Jeon et al. 2020a), customized CNN (Dang et al. 2019; Marra et al. 2020; Cozzolino et al. 2018), CNN with local features (Tarasiou and Zafeiriou 2020), CNN with optical flow (Amerini et al. 2019), ensembled CNNs (Bonettini et al. 2021b; Tariq et al. 2018), light-weight CNN (Sambhu and Canavan 2020), 3D CNNs (Ganiyusufoglu et al. 2020), a two-stream CNN using RGB space and multi-scale retinex space (Chen et al. 2019a), two-stream Faster R-CNN with features from the RGB image and noise features by using steganalysis (Zhou et al. 2018), two-stream neural network with GoogLeNet for observing artifacts and a patch-based triplet network for capturing local noise residual (Zhou et al. 2017), multistream deep learning network for capturing the artifacts by using Face2Face reenactment (Kumar et al. 2020), incorporating CNN with image segmentation (Yu et al. 2019a), capsule networks (Nguyen et al. 2019b, c), pairwise learning (Hsu et al. 2020), metric learning (Kumar et al. 2020), incremental learning (Marra et al. 2019b), multiple instance learning (Li et al. 2020d), few-shot learning (Mansourifar and Shi 2020), enhanced MesoNet (Kawa and Syga 2020; Afchar et al. 2018), a combination with CNN and RNN (Güera and Delp 2018; Mas Montserrat et al. 2020), DNN with contrastive loss (Hsu et al. 2018), multi-attentional network (Zhao et al. 2021).

Though, numerous studies are working on proposing various DNN-based detection methods to discern fake faces. However, they are not robust to be deployed in dealing with real-world scenarios according to a recent study (Hulzebosch et al. 2020). Leveraging the power of CNN models as the backbone is a promising idea for detecting DeepFakes in the wild, however, the biggest obstacle is that the DNN models are susceptible to adversarial noise attacks.

4.1.3 Obvious Artifacts Clues

Due to the limitation of existing AI techniques, the generated DeepFakes exhibit some obvious artifacts which could be leveraged for detection by using some simple DNN models. Chai et al. (2020) investigate that local patches have redundant artifacts which could be used for differentiating fake faces. A fully convolutional approach is applied for training classifiers to focus on image patches. This approach can be well generalized across different network architectures, image datasets, etc. The discrepancy between faces and their context is another artifact clue for detecting fake faces (Nirkin et al. 2020). A face identification network is trained by using the face region to identify the person, while a context recognition network is trained by utilizing the face context like hair, ear to identify the person. Two vectors from the aforementioned two networks are compared for detecting the identity-to-identity discrepancies. This approach also has a good generalization ability across GANs. For each individual that is speaking, their facial and head movements are always in distinct pattern (Agarwal et al. 2019, 2020a). This could be exploited to protect celebrities with large historic training data. These approaches simply leverage the artifact clues for detection without introducing any new DNN models, thus they will be invalid when the GAN is updated or the artifacts are fixed in the new version.

4.1.4 Detection and Localization

Beyond DeepFake detection, some researchers are working on locating the manipulated regions which provides evidence for forensics and inspires future work to develop more powerful DeepFake detectors by focusing on the manipulated regions. FakeLocator (Huang et al. 2022) investigates the architecture of existing GANs and observed that the imperfection of upsampling methods exhibits obvious clues for detection and forgery localization where the manipulated area could be precisely marked. They employ an encoder-decoder network to extract the fake texture with devised gray-scale prediction map for better detection and localization. FakeLocator performs well across different GANs and shows strong generalization capabilities in unknown synthetic techniques. Furthermore, the robustness against perturbation attacks (e.g., compression, blur, noise, and low-resolution) is also promising. Locating the manipulated area provides clear explanations why the image is identified as fake.

Dang et al. (2020) also study the localization of forgery area in fake faces by estimating an image-specific attention map. However, the estimation of the attention map fails to work in a totally unsupervised manner. The inverse intersection non-containment (IINC), a novel metric, is proposed for evaluating the performance of facial forgery localization. They also claim that forgery detection can work well in both seen and unseen synthetic techniques. The evaluation of the robustness against perturbation attacks is still lacking. The proposed attention map predicts the manipulated pixels, which provides a direct decision for determining fake faces.

Multi-task learning could also be used for classifying and locating the manipulated facial images (Nguyen et al. 2019a). Formulating fake forensics as a segmentation task to localize the manipulated region in synthesized faces is another interesting idea in fighting against DeepFakes (Li et al. 2019a; Chen and Yang 2021). Combing deep learning and co-occurrence matrices could also be used for the detection, attribution, and localization of GAN images (Goebel et al. 2020).

Face X-ray (Li et al. 2020c) observes that a manipulated facial image always blends into an existing background image. Thus, the discrepancies across the blending the boundary could be used as a signal for detecting manipulated fake faces. Face X-ray is designed for working as both DeepFake detection and manipulated region localization.

Songsri-in and Zafeiriou (2019) release the first forensic localization dataset with labeled corresponding binary masks. The dataset contains real facial images, generated facial images, and partially manipulated facial images. ManTra-Net (Wu et al. 2019) proposes an end-to-end fully convolutional network for addressing various types of image forgery, such as splicing, copy-move, removal, enhancement, and even unknown types. ManTra-Net formulates the localization problem as local anomaly detection and has a board of applications than the existing image forgery localization methods.

4.1.5 Facial Image Preprocessing

Some studies propose preprocessing the facial images before sending them to binary classifiers for discrimination. These works hope that the preprocessed DeepFakes could expose their fake textures to simple classifiers, such as, shallow neural networks, conventional machine learning models (e.g. SVM, KNN).

FakeSpotter (Wang et al. 2020d) observes that the layer-by-layer neuron behaviors provide more subtle features for capturing the differences between real and fake faces. This work provides a new insight for spotting fake faces by monitoring third-party DNN-based neuron behaviors, which could be extended to other fields like fake speech detection (Wang et al. 2020c). Experimental results also show its robustness against four common perturbation attacks and its capabilities in detecting DeepFake videos. However, the generalization ability of unseen techniques is still unclear. FakeSpotter simply receives facial images as input, thus the detection result is lacking explainability.

The EM algorithm is employed for extracting the local features to represent the convolutional traces in the generated facial images (Guarnera et al. 2020a). Then, some naive classifiers like K-nearest neighbors (KNN), support vector machine (SVM), and latent Dirichlet allocation (LDA) could easily classify real and fake faces. Actually, some dimension reduction algorithms like T-SNE could non-linearly separate the real and fake faces. However, the robustness against perturbation attacks and generalization ability in different GANs are unclear.

In the real-world scenario, videos always suffer various degradations such as compression, blurring, etc. ARENnet (Guo et al. 2020c) aims at highlighting the tampered artifacts by suppressing the image content to build a practical DeepFake detector. An adaptive residuals extraction network (AREN) is designed for suppressing image content to learn prediction residuals via an adaptive convolution layer. Then, a fake face detector ARENnet is constructed by integrating AREN with CNN to deal with the fake videos suffering degradations. ARENnet claims the robustness against perturbation attacks and generalization ability in unseen GANs.

Chen and Yang (2021) also study to improve the quality of training dataset by employing dataset preprocessing techniques to remove the false face detected in videos. They observe that preprocessing the training dataset can significantly increase the detection performance in comparison with baselines.

Zhang et al. (2017) detect the key points in facial images and applied a descriptor to represent them for capturing the local information, then a linear classifier is applied for effective detection. This approach could be integrated into face verification systems to enhance their security.

Studies have shown that the preprocessed DeepFakes could obviously improve the detection performance. However, attackers can use other preprocessing techniques to remove the artifacts which could be used for DeepFake detection, which poses potential threats to the community.

4.1.6 Technical Evolution of Spatial-Based Detection

In this subsection, we introduce the evolution of the spatial-based DeepFake detection techniques and present the strength and weakness in detecting DeepFakes as well. Due to the low quality faces generated by the early DeepFakes, researchers first investigate the differences of real and fake faces in the spatial domain since 2017. Investigating on the spatial domain is a straightforward idea for distinguishing real and fake faces, which could borrow ideas from the traditional digital media forensics.

The spatial based DeepFake detection methods aim at leveraging the power of DNN models to capture the subtle differences between real and fake in the spatial domain. The detection task can be simply formulated as a binary classification problem. Most of these studies are working towards two directions, observing more visual artifacts and developing powerful DNN models which could work in an end-to-end manner.

The simple approach for detecting DeepFakes is employing the traditional image forensics techniques by inspecting the disparities at the pixel-level, such as studying the PRNU pattern which caused by the light sensor in camera (Koopman et al. 2018) and the chrominance components (Li et al. 2020a) on real and fake faces. However, such solution usually suffers from performance decline when the DeepFakes' quality is degraded. Due to the significant progress of DNN models in various cutting-edge fields, powerful DNN models are designed or employed for spotting DeepFakes, such as FDFtNet (Jeon et al. 2020b) or ResNet-50 (Wang et al. 2020e). In addition, some researchers employ shallow machine learning models like KNN and SVM to detect DeepFakes with handcraft features. For example, Guarnera et al. (2020a) apply

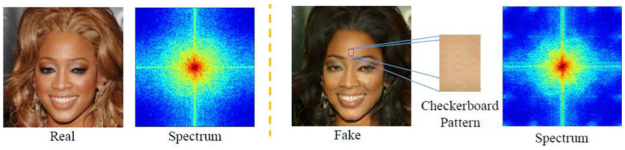


Fig. 7 The difference between real and fake from the frequency domain, especially noticing the difference in their spectra (Zhang et al. 2019)

EM to extract local features and use KNN and SVM for classification.

Beyond DeepFake detection, Huang et al. (2022) aim to localize the GAN-based manipulated region with gray-scale prediction map, which is helpful for fine-grained DeepFake forensics.

The early DeepFakes present abnormal visual artifacts like discrepancy between faces and their context (Nirkin et al. 2020), providing effective artifact clues for detection. We believe more and more interesting studies on the spatial domain will be proposed by our community.

Overall, the spatial based DeepFakes detection is one of the most popular solutions. It works well when DeepFakes exhibit obvious visual artifacts. However, it will be not a promising way when the DeepFakes become realistic in the near future. Nevertheless, there are two critical challenges via the spatial based solution to fight against DeepFakes, i.e., poor generalization capability against unknown synthetic techniques and low robustness to adversarial attacks (Carlini and Farid 2020).

4.2 Frequency based Detection

Beyond distinguishing real and fake from the spatial domain, some studies are working on exploiting the differences between real and fake from the frequency domain. Figure 7 represents the potentials of employing the frequency artifacts for detecting DeepFakes, where the GAN-based manipulation introduces invisible artifacts in the frequency domain.

4.2.1 GAN-Based Artifacts

Instead of examining the visual artifacts, some researchers are working on investigating the imperfection design of existing GANs, which provides obvious signals for differentiating real and fake faces. They are normally working on the frequency domain, but there are generalized artifacts of existing GANs.

McCloskey and Albright (2019) investigate the architecture of the generator model and observed that the internal value of the generator is normalized which limits the frequency of saturated pixels. Then, a simple SVM-based classifier is trained to measure the frequency of saturated and under-exposed pixels in each facial image for discriminating fake faces.

AutoGAN (Zhang et al. 2019) identifies a unique artifact in GANs which is introduced due to the upsampling design of common GAN pipelines. Then, a GAN simulator is proposed for simulating the artifacts without accessing pre-trained GANs to improve the generalization ability of existing detectors. The artifacts are manifested as replications of spectra in the frequency domain. Finally, a classifier is trained by using the frequency spectrum for discriminating GAN-synthesized fake faces. They claim that the observed GAN-based artifacts could generalize well in unseen synthetic techniques with similar architectures. However, their robustness against perturbation attacks is not explored.

Yu et al. (2019b) first introduce the GAN fingerprints for classifying the images as real or synthesized with GANs. The GAN fingerprints could be further utilized for predicting the source of images. Study has shown that small differences in GAN training could result in distinct GAN fingerprints. However, the fingerprints could be easily destroyed by simple perturbation attacks like blur, JPEG compression, etc. Other studies (Marra et al. 2019a) also leverage the GAN fingerprints for discriminating GAN-synthesized fake faces. The GAN artifacts are a promising clue for detection, however the artifacts could be easily corrupted with some simple image transformations like shallow reconstruction with principal component analysis (PCA) (Huang et al. 2020b).

4.2.2 Frequency Domain Feature

The differences between real and synthesized fake faces could also be revealed in the frequency domain. Here, we mainly introduce the studies simply employing frequency domain as features for differentiating real and fake. These methods are often failed in tackling unknown GAN-synthesized DeepFakes, which are vast different from the aforementioned methods in Sect. 4.2.1.

Frank et al. (2020) comprehensively investigate the artifacts revealed in the frequency domain across different GAN architectures and datasets. They observe that severe artifacts are introduced due to the upsampling techniques in GANs. Experiments demonstrate that a classifier with a simple linear model and a CNN-based model could both achieve promising results on the entire frequency spectrum. Furthermore, the classifier trained on the frequency domain is robust against common perturbation attacks (e.g. blurring, cropping) and tackles the future unseen GANs (Frank et al. 2020).

FGPD-FA (Bai et al. 2020) extracts three types of features (e.g., statistical, oriented gradient, and blob) in the frequency domain for differentiating real and fake faces. F^3-Net (Qian et al. 2020) considers two complementary frequency-aware clues for detection, namely frequency-aware pattern from frequency-aware image decomposition, and local frequency statistics. Specifically, discrete cosine transform (DCT) is applied for frequency-domain transformation. Finally, a two-stream collaborative learning framework collaboratively learns the two frequency clues and achieves considerable performance in low-quality DeepFake video detection.

Barni et al. (2020) propose exploiting the inconsistencies among spectral bands, then a CNN model is trained with cross-band co-occurrence matrices and pixel co-occurrence matrices for discriminating real and fake faces. Yu et al. (2020b) explore the channel difference image (CDI) and spectrum image (SI) to work as intrinsic clues for distinguishing images generated with a camera and manipulated with AI techniques. Octave convolution (OctConv) (Chen et al. 2019b) is leveraged for capturing the frequency domain to learn the intrinsic feature from CDI and SI to determine fake faces. These two intrinsic clues are claimed to generalize well in unseen manipulations. To improve the transferability of face forgery detection method across unseen synthetic techniques, Liu et al. (2021b) combine spatial image and phase spectrum to capture the up-sampling artifacts in existing GANs for aiding detection. Masi et al. (2020) propose a two-branch network for amplifying the artifacts in the synthesized faces by combing clues from the color domain and frequency domain. Furthermore, the two-branch network has claimed a good performance across datasets. Actually, leveraging the frequency domain to distinguish real and fake faces is widely applied in recent studies (Durall et al. 2019; Bonettini et al. 2021a; Guarnera et al. 2020c).

4.2.3 Technical Evolution of Frequency-Based Detection

In this subsection, we introduce the evolution of the frequency-based DeepFake detection techniques and present the strength and weakness in detecting DeepFakes as well. Beyond detecting DeepFakes via the spatial information, exploring the artifacts of DeepFakes in the frequency domain is another effective solution. The frequency based DeepFake detection methods identify DeepFakes via artifacts in the frequency domain, benefiting higher generalization ability.

The frequency based DeepFake detection methods mainly rely on two kinds of information, i.e., *the artifacts in the spectra introduced by GAN* and *frequency domain features of real or fake faces*. For the first kind solution, researchers observe that the GAN-synthesized facial images exhibit obvious artifacts in the spectra, which provides effective clues for detection with high generalization. Figure 7 visualizes the spectra maps of real and fake face (Zhang et al. 2019), respectively. Then, a series of studies try to mine GAN-based artifacts of fake faces effectively and achieve better generalization capabilities when addressing unknown synthetic techniques. In particular, McCloskey and Albright (2019) measure the frequency of saturated and under-exposed pixels to discriminate real and fake faces. Zhang et al. (2019) identify the artifacts introduced by GAN due to the common used upsampling operation. The above two methods claim their competitive generalization capabilities in unknown DeepFake techniques. For the second solution, features in the frequency domain is leveraged as clues for detection. For example, Frank et al. (2020) employ the entire frequency spectrum as features for differentiating fake. Barni et al. (2020) exploit the inconsistencies among spectrum bands to discriminate fake.

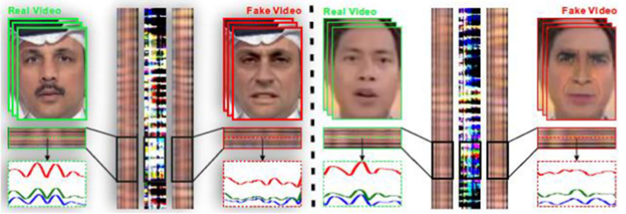


Fig. 8 The difference between real and fake from the biological signal domain, especially the colorful motion-magnified spatial-temporal (MMST) maps between them (Qi et al. 2020)

The frequency based DeepFake detection method could generalize well on unknown synthetic techniques, but they are not robust to various image degradations, such as image compression, reconstruction, etc. (Huang et al. 2020b). Thus, they are less practical in the real-wold scenario where known and unknown image degradations are common. As a result, more robust frequency based methods should be developed, which is critical for detectors to be further deployed in the wild.

4.3 Biological Signal based Detection

Real still facial images and videos are produced with cameras, which are natural compared to the synthesized fake faces. The biological signal exhibits a clear signal for distinguishing real and fake. In general, the biological signals are existed in both real videos and synthesized fake videos. However, the biological signals revealed in the real faces videos are natural and realistic. Unfortunately, in the fake videos, the biological signals are generated with low-quality and most of time the perceptual biological signals are disappeared, such as the consistency between visual and audio. Figure 8 shows the sample of biological signals which could served as clues for DeepFake detection. These biological signals can be classified into the following categories.

4.3.1 Visual-Audio Inconsistency

For DeepFake video, combining visual and audio to identify the inconsistency in fake faces is a new insight for distinguishing DeepFakes. These methods can well explain why the video is fake. A Siamese network is employed for modeling the visual and audio in videos with a combination of two triplet loss functions for measuring the similarity (Mittal et al. 2020). Specifically, one loss function is designed for computing the similarity between visual and audio, the other loss function is devised for calculating the effect cues, like perceived emotion. Experiments show that it outperforms conventional DNN-based methods in detecting fake videos. Lip-sync is a typical DeepFake by generating a person's mouth to be consistent with a person's speech. With the basic insight that the dynamics of mouth shape are sometimes inconsistent with a spoken phoneme due to the highly compelling circumstances. Specifically, the lips have to be closed when spoken some words that begin with M, B, P. However, this is violated in fake videos. Researchers leveraged this clue for detecting lip-sync DeepFakes (Agarwal et al. 2020b). Modality dissonance score (MDS) is proposed for measuring the audio-visual dissonance in videos (Chugh et al. 2020). Specifically, MDS is based on *contrastive loss* which enforces the distance between visual and audio to be closer for real, and further for synthesized fake video. Additionally, MDS can be used for temporal forgery localization which identifies the tampered segment in the video. However, Korshunov and Marcel (2018) investigate several baselines for evaluating existing studies in DeepFake detection including lip-sync inconsistency detection. They observe that detecting the inconsistency of lip-sync is not effective for fighting DeepFakes. They also release a public dataset for the community.

4.3.2 Visual Inconsistency

Visual inconsistency indicates that the synthesized faces are not natural, especially the shape, facial features, and landmarks of faces. Li and Lyu (2019) observe that the synthesized fake faces are always in fixed sizes due to the limitation of computation resources and the production time of DeepFake algorithms. The fixed size of synthesized faces leaves artifacts in warping to match the source face, which can be employed for DeepFake detection. Then, a CNN model is trained for detecting the artifacts. The lack of eye blinking is another telltale sign for exposing DeepFakes (Li et al. 2018b). A CNN combined with a recursive neural network is trained for distinguishing the eye state. The mismatched facial landmarks in fake faces are invisible to human eyes, but they can be easily revealed from head poses estimated from 2D landmarks (Yang et al. 2019a, b). A naive SVM classifier is finally trained for capturing the differences between estimated head poses, which is further employed for DeepFake detection. Visual artifacts such as eyes, teeth, facial contours will be an important clue for exposing DeepFakes (Matern et al. 2019). The inconsistent corneal specular highlight between two eyes is another clue for exposing the GAN-synthesized faces (Hu et al. 2021). This inconsistency is mainly due to the lack of physical/physiological constraints in the existing popular GANs. These methods are all based on the observation that the fake faces exhibit obvious artifacts to human eyes, especially the inconsistencies that appeared in the face compared with real faces. They provide strong guarantees to explain the decision in distinguishing real or fake, but they will be invalid when advanced GANs are proposed. Furthermore, their robustness against perturbation attacks is unclear.

4.3.3 Biological Signal in Video

Biological signals in the video are not easily replicable. In FakeCatcher, six different biological signals are extracted to exploit the spatial and temporal coherence for authenticating real videos taken by the camera (Ciftci et al. 2020a). Studies have shown that the heart rate could be used for detecting fake videos, however, obtaining the heart rate from videos is a time-consuming task. Fernandes et al. (2019) use neural ordinary differential equation (Neural-ODE) (Chen et al. 2018b) trained on the original videos to predict the heart rate of testing videos. DeepRhythm also exposes DeepFake videos by monitoring the heartbeat rhythms (Qi et al. 2020). Specifically, they develop motion-magnified spatial-temporal representation (MMSTR) to video for highlighting the heart rhythm signals. Finally, a dual-spatial-temporal attentional network is designed for detecting fake video based on the output of MMSTR. DeepFakesON-Phys (Hernandez-Ortega et al. 2020) also leverages heart rate for DeepFake detection by using remote photoplethysmography (rPPG) to illustrate the presence of blood flow by observing the subtle color changes in human skins. A convolutional attention network (CAN) is proposed for extracting the spatial and temporal information from video frames to detect DeepFake video. Beyond the DeepFake detection, PPG could be used for discovering the generative model which is used for generating DeepFake (Ciftci et al. 2020b). In detecting DeepFake videos, the biological signal exposed by the heart rate provides promising clues for detection. This will be a promising idea for dealing with future advanced GANs, since the subtle biological characteristics are a challenge for synthesis.

4.3.4 Technical Evolution of Biological Signal based Detection

In this subsection, we introduce the evolution of the biological signal based DeepFake detection techniques and present the strength and weakness in detecting DeepFakes as well. With the rapid development of deep synthesis techniques, the fake images would be perfectly synthesized without exposing any artifacts in both the spatial and frequency domains in the near future, which would pose more challenges for DeepFake detection. Recently, some researchers are working to explore the biological signals in the facial videos to serve as effective fake indicators since the signals are not natural and unrealistic in fake videos.

The existing biological signal based DeepFake detection methods utilize biological signals that are broken and could not be easily replicated by state-of-the-art DeepFake techniques. Early works study the irregular eye blinking (Li et al. 2018b), the mismatch facial landmarks (Yang et al. 2019a, b), and the fixed size of synthesized faces (Li and Lyu 2019), etc. Nevertheless, the above visual inconsistences could be easily removed in the advanced DeepFakes. In addition to the visual information, the audio of the video is also an important clue for DeepFake detection. Specifically, the inconsistency between visual and audio is common in fake videos. However, Korshunov and Marcel (2018) observed that the simple lip-sync is not enough for accurate DeepFake detection. Then, some studies are working on how to measure the similarity between visual and audio (Mittal et al. 2020; Chugh et al. 2020) and further explore strong visual-audio inconsistency signals for DeepFake detection (Agarwal et al. 2020b). Moreover, some works also exploit the subtle color changes in human skins introduced by the normal heartbeat to authentic real videos (Qi et al. 2020; Hernandez-Ortega et al. 2020). We believe that more and more interesting and robust biological signals will be observed for discriminating DeepFakes in the wild.

Overall, In the near future, the DeepFake could be realistic where the spatial and frequency based detection methods could hardly exhibit noticeable and detectable artifacts by human eyes and machines. As a result, the biological signals would be a more effective solution for fighting against DeepFake that could be deployed in the real world. Nevertheless, the solution might be invalid when the biological signals are enhanced manually, and exploring more informative biological signals would be the most promising one for the future detection.

4.4 Other DeepFake Detectors

Besides the aforementioned three types, some studies cannot be classified into any of them. Here, we introduce them with an independent subsection. Fraga-Lamas and Fernández-Caramés (2020) provide a comprehensive overview by leveraging distributed ledger technologies (DLT) to combat digital deception. Hasan and Salah (2019) also leverage blockchain to trace and track the source of multimedia which provides insight for combating DeepFake videos. Instead of a focus on the multimedia self, FakeET (Gupta et al. 2020) explores to leverage the user behavior clues for DeepFake detection, specifically the eye-gaze. Tolosana et al. (2020) explore the role of different facial regions in contributing to the DeepFake detection. They find that the artifacts which exist in the specific facial region could improve the detection performance by a large margin than the entire face. Similarly, Du et al. (2019) observe that concentrating on the forgery region could help for DeepFake detection. Maurer (2000) approaches the DeepFake detection as a hypothesis testing problem and presents a generalizable statistical framework based on the information-theoretic study of authentication.

Type ID	Name
Type I1	Image Forensics based Detection
Type I2	DNN-based Detection
Type I3	Obvious Artifacts Clues
Type I4	Detection and Localization
Type I5	Facial Image Preprocessing
Type II1	GAN-based Artifacts
Type II2	Frequency Domain
Type III1	Visual-audio Inconsistency
Type III2	Visual Inconsistency
Type III3	Biological Signal in Video
Type IV1	Others

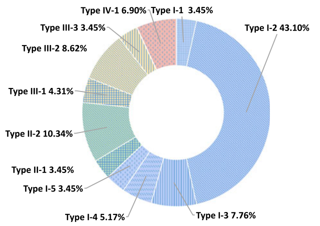


Fig. 9 (L) Summary of various types of DeepFake detection methods, including the type ID and the name of each type. (R) The proportion of various types of DeepFake detection methods in our collected DeepFake detection papers

4.5 Summary of DeepFake Detection Methods

In this section, we use a long table and a chord diagram to summarize the existing DeepFake detection methods and a fishbone diagram to show the evolution of the three DeepFake detection techniques.

Tables 7, tabulate the summary of DeepFake detection methods, where Fig. 9 gives the meaning of type ID and the proportions. In these tables, we mainly show the method type, the adopted classifier, the claimed performance, compared baselines, and its capabilities with regard to the generalization capabilities in tackling unseen DeepFakes, the robustness against various attacks, and whether provides explainable detection results.

In analyzing the Tables 7, we can gather the following interesting findings. Due to the powerful capabilities of DNN model, CNN models are served as the most popular backbone in the DeepFake detection classifiers. However, linear machine learning models like KNN are rarely employed in detection. In employing the evaluation metrics, ACC and AUC are the two popular metrics for evaluating the performance of DeepFake detection methods. Compared with DeepFake videos, the still images are easier to be detected by various DeepFake detectors. Researchers tend to evaluate their method on public DeepFake videos, rather than build their own synthesized-images datasets for evaluation due to the lack of public fake image datasets. The existing studies claimed their effectiveness in detecting DeepFakes with high confidence, however most of them failed in evaluating their effectiveness in tackling unseen DeepFakes and their robustness against image/video transformations, which is critical for a detector deployed in the wild. Additionally, these methods failed in providing evidence to introduce the differences between real and fake, thus the explainability is limited in existing studies.

Figure 10 presents the milestone studies of DeepFake detection with a fishbone diagram. In investigating the three classical DeepFake detection techniques, we observed that there are two critical challenges that should be addressed for the future DeepFake detection techniques. The first challenge is that the fake textures for DeepFake detection might be corrupted or intentionally removed. The second one is that the quality of synthesized images would be further improved with the development of synthetic techniques. As a result, the community should develop more robust models against various degradations to capture the subtle differences between real and fake faces and investigate more long-standing clues to detect unknown DeepFake synthetic techniques.

5 Battleground

In the previous two sections, we have discussed recent advances in DeepFake generation methods (Sect. 3) and DeepFake detection methods (Sect. 4), respectively. The two parties naturally form a battleground, where the "offenders" or the "adversaries" are the DeepFake generation methods, and the "defenders" are the DeepFake detection methods. By illustrating and visualizing the battleground, we hope to gain insights and knowledge about the most current battling landscape and interactions between DeepFake generation and detection methods. We believe that incremental but continuous scientific progresses can be made through the competition between adversaries and defenders, and new observations can be obtained when defeating the other side. It is the unceasing battling between the two parties that will most likely make the meaningful progress to push the field (i.e., high-fidelity generation of DeepFakes as well as high-performance detection of DeepFakes) forward possible.

Among all 318 DeepFake-related papers surveyed so far, we have kept the important ones in tables across Sects. 3 and 4. As previously tabulated in Table 6, we have surveyed 83 DeepFake generation methods in Sect. 3. As tabulated in Table 7, we have surveyed 117 DeepFake detection methods in Sect. 4. In order to create a full map of the battleground, for each of the DeepFake detection methods, we aim to know which DeepFake generation method the detector attempted to counter, i.e., to perform DeepFake detection on. In the Sankey diagram (Wikipedia 2021b) shown in Fig. 11, we have chronologically arranged various surveyed DeepFake generation methods (including datasets) on the left column and the surveyed DeepFake detection methods on the right column. A curve connecting the node A on the left and the node B on the right means that DeepFake detection method B has evaluated and reported detection results on the DeepFake generation method A in its paper. After all the nodes are connected by traversing the 83 × 117 generation-detection relationships, Fig. 11 now presents the status of the DeepFake generation-detection battleground. As the out degree for each node shown in the figure, we can tell how popular each DeepFake generation method or DeepFake detection method is. For example, the FaceForensics++ has a large out degree, which means that it is evaluated by a large number of

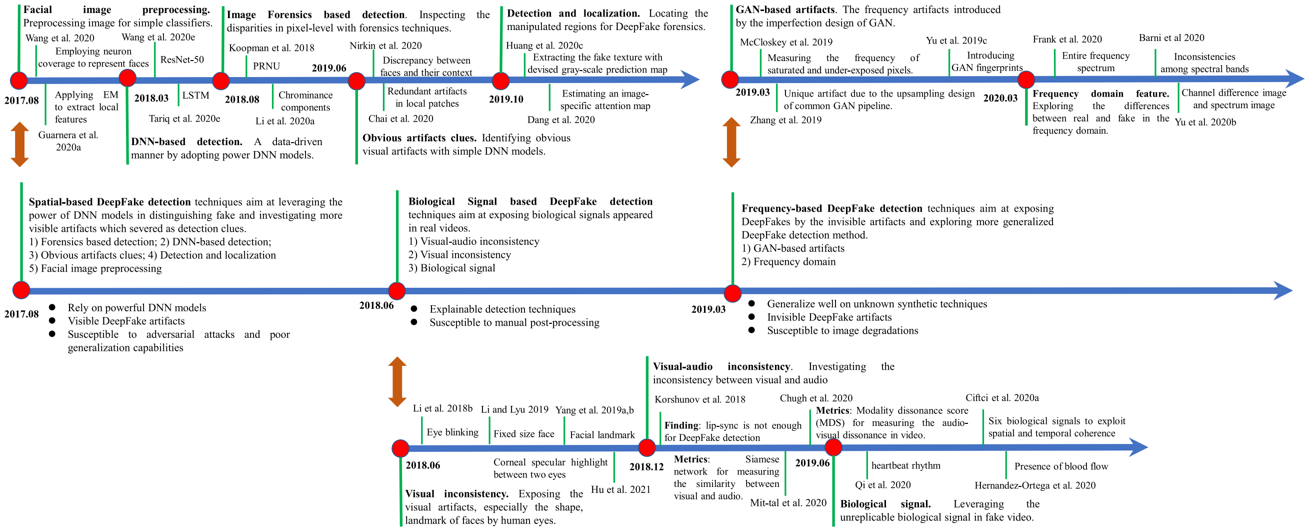

Facial image preprocessing. Preprocessing image for simple classifiers.
Wang et al. 2020
Employing neuron coverage to represent faces
Wang et al. 2020e
ResNet-50
2017.08
Applying EM to extract local features
Guarnera et al. 2020a
2018.03
LSTM
Tariq et al. 2020e
DNN-based detection. A data-driven manner by adopting power DNN models.
Image Forensics based detection. Inspecting the disparities in pixel-level with forensics techniques.
Koopman et al. 2018
PRNU
2018.08
Chrominance components
Li et al. 2020a
2019.06
Nirkin et al. 2020
Discrepancy between faces and their context
Redundant artifacts in local patches
Chai et al. 2020
Obvious artifacts clues. Identifying obvious visual artifacts with simple DNN models.
Detection and localization. Locating the manipulated regions for DeepFake forensics.
Huang et al. 2020c
Extracting the fake texture with devised gray-scale prediction map
2019.10
Estimating an image-specific attention map
Dang et al. 2020
GAN-based artifacts. The frequency artifacts introduced by the imperfection design of GAN.
McCloskey et al. 2019
Measuring the frequency of saturated and under-exposed pixels.
Yu et al. 2019c
Introducing GAN fingerprints
2019.03
Unique artifact due to the upsampling design of common GAN pipeline.
Zhang et al. 2019
2020.03
Frank et al. 2020
Entire frequency spectrum
Barni et al 2020
Inconsistencies among spectral bands
Frequency domain feature. Exploring the differences between real and fake in the frequency domain.
Channel difference image and spectrum image
Yu et al. 2020b
Spatial-based DeepFake detection techniques aim at leveraging the power of DNN models in distinguishing fake and investigating more visible artifacts which severed as detection clues.
1) Forensics based detection; 2) DNN-based detection;
3) Obvious artifacts clues; 4) Detection and localization
5) Facial image preprocessing
2017.08
● Rely on powerful DNN models
● Visible DeepFake artifacts
● Susceptible to adversarial attacks and poor generalization capabilities
Biological Signal based DeepFake detection techniques aim at exposing biological signals appeared in real videos.
1) Visual-audio inconsistency
2) Visual inconsistency
3) Biological signal
2018.06
● Explainable detection techniques
● Susceptible to manual post-processing
Frequency-based DeepFake detection techniques aim at exposing DeepFakes by the invisible artifacts and exploring more generalized DeepFake detection method.
1) GAN-based artifacts
2) Frequency domain
2019.03
● Generalize well on unknown synthetic techniques
● Invisible DeepFake artifacts
● Susceptible to image degradations
Li et al. 2018b
Eye blinking
Li and Lyu 2019
Fixed size face
Yang et al. 2019a,b
Facial landmark
2018.06
Corneal specular highlight between two eyes
Hu et al. 2021
Visual inconsistency. Exposing the visual artifacts, especially the shape, landmark of faces by human eyes.
Visual-audio inconsistency. Investigating the inconsistency between visual and audio
Korshunov et al. 2018
Finding: lip-sync is not enough for DeepFake detection
2018.12
Metrics: Siamese network for measuring the similarity between visual and audio.
Mit-tal et al. 2020
Chugh et al. 2020
Metrics: Modality dissonance score (MDS) for measuring the audio-visual dissonance in video.
2019.06
heartbeat rhythm
Qi et al. 2020
Ciftci et al. 2020a
Six biological signals to exploit spatial and temporal coherence
Presence of blood flow
Hernandez-Ortega et al. 2020
Biological signal. Leveraging the unreplicable biological signal in fake video.

Fig. 10 The evolution of DeepFake detection techniques with a fishbone diagram. In the main fishbone, the weakness and strengths of each detection method are presented as well. For each DeepFake detection method in the sub-fishbone diagram, the milestone studies are added for presenting the significant progress, especially their novelty on technical, the problem addressed, and new insight for defending DeepFakes

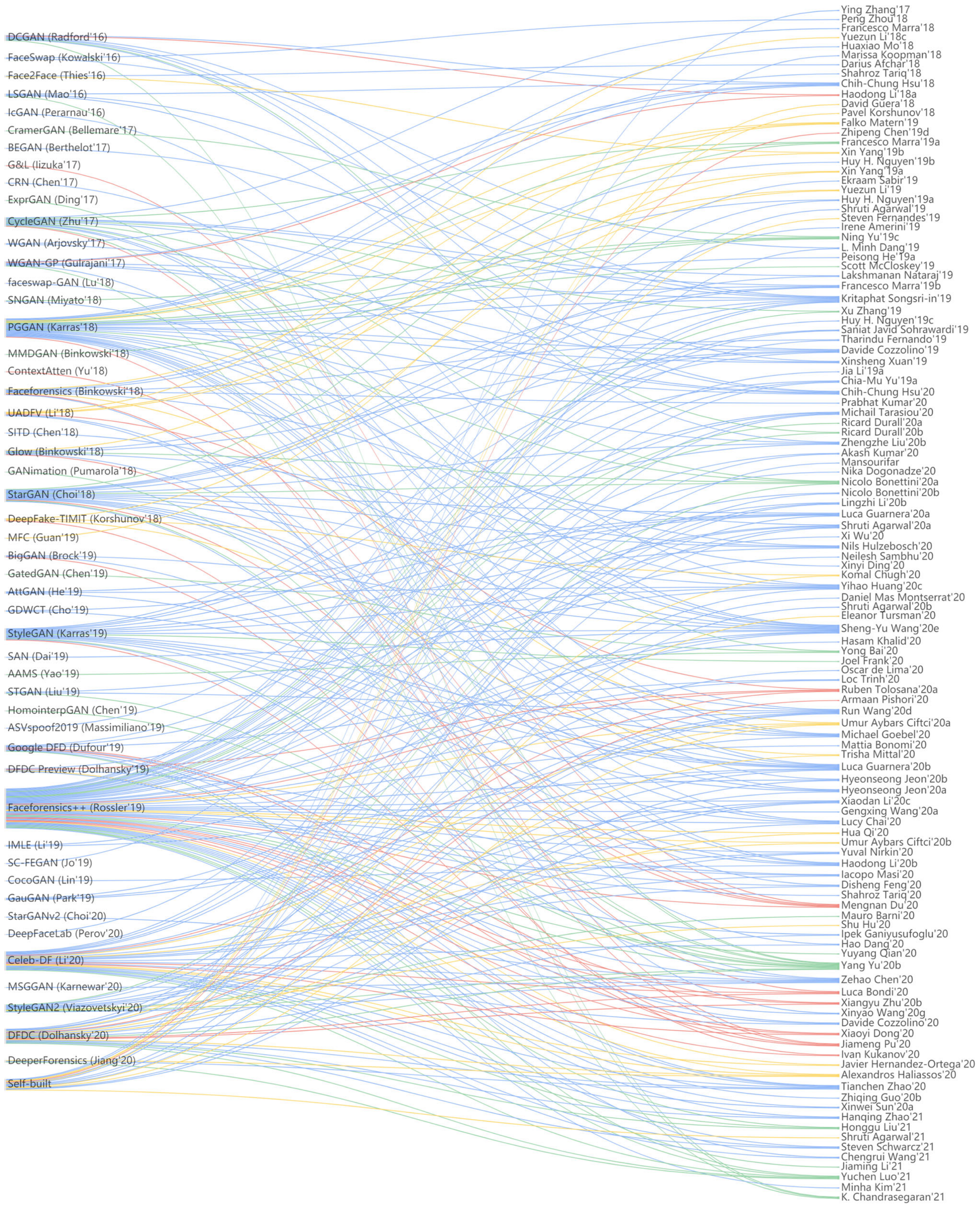


Fig. 11 Battleground diagram between DeepFake generation and detection. The Sankey diagram shows the interactions between various DeepFake detection methods (right column) and various DeepFake generation methods (left column). Both of the generation and detection methods are sorted by the release time and labeled with the corresponding years (same as the order in Tables 6, 7). Four colors represent the different types of detection methods introduced in Tables 7: Blue is Type-I (spatial based) methods, green is Type-II (frequency based) methods, yellow is Type-III (biological signal based) methods, and red is Type-IV (others) methods. Interactive diagram is available at http://www.xujuefei.com/dfsurvey

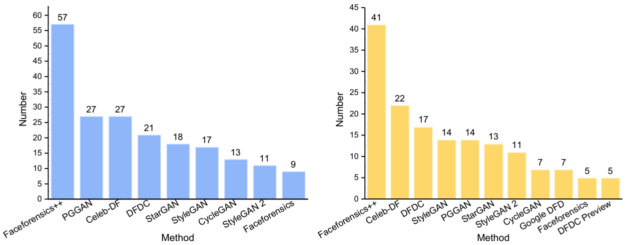


Fig. 12 (L) Top-9 most popular DeepFake generation methods or datasets based on the battleground. (R) 2020's Top-11 most popular DeepFake generation methods or datasets based on the battleground

DeepFake detection methods. Similarly, a big clustered connections on the right side indicates that a particular DeepFake detection method (e.g., Sheng-Yu Wang's method) has been evaluated extensively across various DeepFake generation methods. The colorful curves represent the different types of detection methods introduced in Table 7 as well as tabulated in Fig. 9: Blue is Type-I (i.e., spatial based) methods, green is Type-II (i.e., frequency-based) methods, yellow is Type-III (i.e., biological signal based) methods, and i.e., red is Type-IV (others) methods. The 'self-built' methods on the left-side bottom of the battleground represent the nameless methods addressed by the detection methods on the right side.

Figure 12 shows the top-9 most popular DeepFake generation methods or datasets based on the topology of the battleground figure, as well as the top-11 most popular DeepFake generation methods or datasets in the year 2020 alone. As expected, systemically organized DeepFake datasets such as FaceForensics++ (Rossler et al. 2019), Celeb-DF (Li et al. 2020e), DFDC (Dolhansky et al. 2020) as well as open-sourced high-fidelity face generation methods such as PGGAN (Karras et al. 2017), StarGAN (Choi et al. 2018), StyleGAN (Karras et al. 2019), etc., are on the top of the list.

Based on the above discussion about the most popular or most widely evaluated DeepFake generation methods or datasets, we have made some interesting observations: (1) the surveyed DeepFake detectors perform more detection experiments on DeepFake images than on DeepFake videos; (2) only a tiny portion of the surveyed detection methods work on both DeepFake image and video detection tasks; (3) for those detectors for both DeepFake image and video detection, most of them focus on the latest high-fidelity image-based DeepFakes while on the less state-of-the-art video-based DeepFakes, although both modalities are concurrently accessible. This can be partially attributed to the fact that video-based DeepFake datasets are more scarce, and/or the latest ones are much more challenging to tackle.

We try to capture this phenomenon through the Sankey diagram in Fig. 13, where only 10 of the surveyed 117 DeepFake detection methods have attempted the DeepFake detection on both the image and video modalities. A curve connecting a node A on the left column and a node B on the right column means that a particular DeepFake detector evaluated on the image-based DeepFake generation method A has also been evaluated on video-based DeepFake generation method B, as reported in its paper.

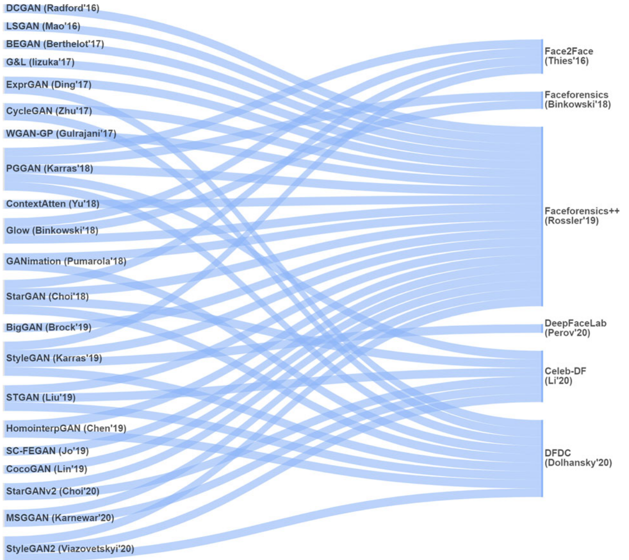


Fig. 13 Relation pairs of the image- and video-based DeepFake generation methods that are simultaneously evaluated by some DeepFake detection methods. Interactive diagram is available at http://www.xujuefei.com/dfsurvey

Moreover, we try to understand for a particular DeepFake detection method listed on the right column of Fig. 11 , which previously published detectors has it benchmarked against. Figure 14 presents a chord diagram to show the 'competition' among later detectors and earlier ones. In the chord diagram, each node represents a DeepFake detection method, and a link connecting a node A and a node B means that the method A has been compared with the method B as a baseline in A's paper, which infers that the method A comes after B. We also notice that many of the DeepFake detectors are benchmarked against common machine learning (ML) based classifiers such as KNN and logistic regression, or popular DNNs such as the ResNet (He et al. 2016), etc. Therefore, we also list out 30 popular ML-based methods in Fig. 14, and a link between a DeepFake detection method A and an ML-based method B can be established when the method B is compared by the A's paper. We provide an interactive diagram[4] to facilitate the interpretation of the graph. The top-5 popular baselines adopted in the evaluation are XceptionNet (Chollet 2017), Afchar et al. (2018), Nguyen et al. (2019a), ResNet (He et al. 2016), and Yang et al. (2019b). XceptionNet, ResNet, and VGG are the top-3 CNN models that are employed as the baselines for comparison. In particular, XceptionNet is the most popular baseline and more

[4] http://www.xujuefei.com/dfsurvey.

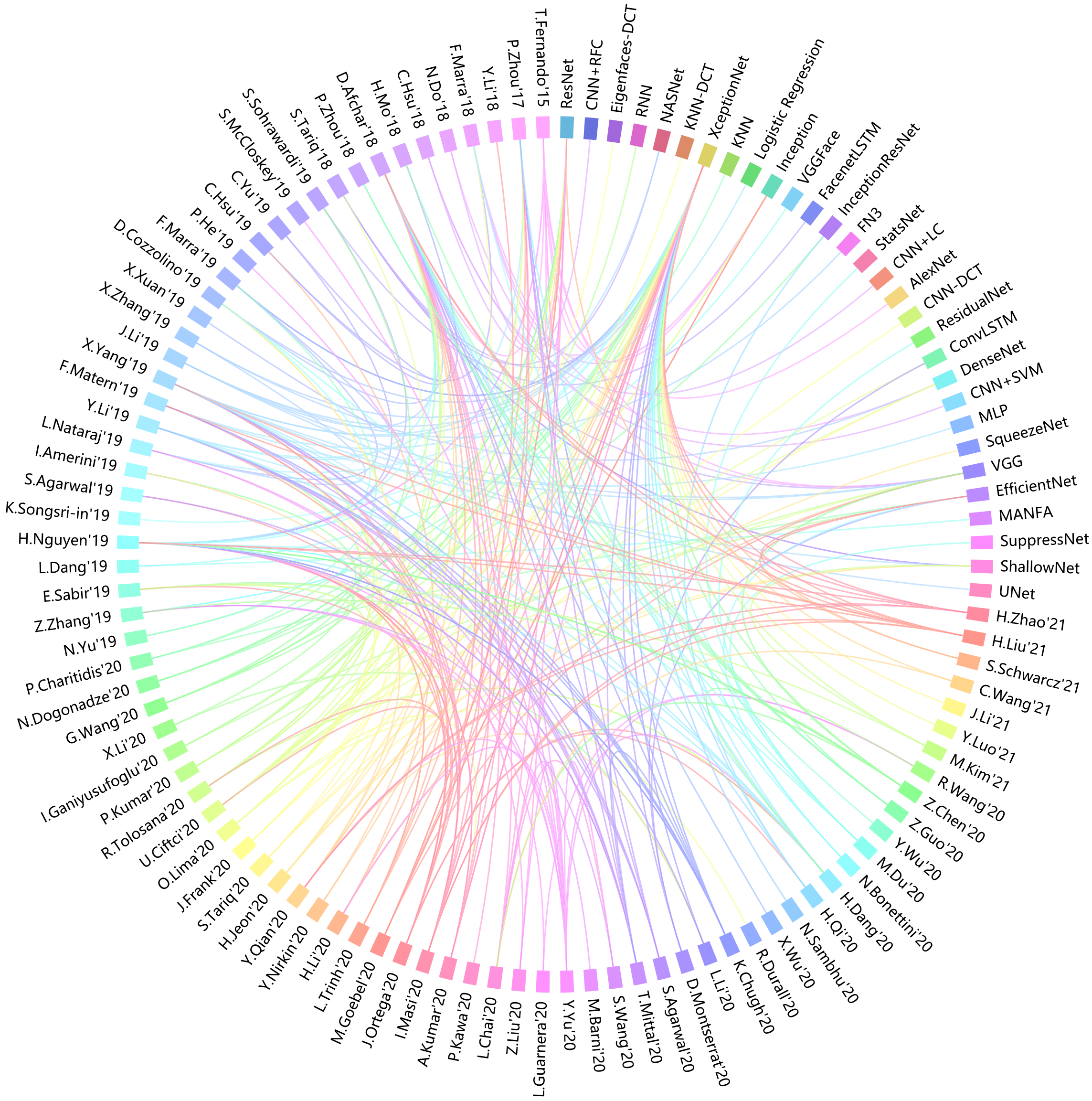


Fig. 14 A chord diagram represents the comparison among the existing detection methods. The node indicates the method for DeepFake detection and the link represents that one of the work is served as the baseline in the evaluation. The baselines include typical CNN models and the works with/without the peer review. An interactive diagram is available at http://www.xujuefei.com/dfsurvey

than one-third studies compare with it. Figure 15 (L) shows the top-11 most popular DeepFake detection methods chosen as baselines and Fig. 15 (R) shows the top-10 most popular ML-based methods chosen as baselines by various DeepFake detectors (See Fig. 14). We also identify the DeepFake detection methods that conduct the most extensive comparison experiments, that is, the number of baselines used by these methods are ranked in the top 9 according to the chord diagram. Figure 16 (L) shows the top-9 DeepFake detection methods that benchmark against the most number of baselines, and Fig. 16 (R) shows the top-8 DeepFake detection methods that benchmark against the most number of baselines in 2020.

Another way is to measure the popularity of the DeepFake generation and detection methods through the citation count as well as citations normalized by the number of days since exposure. Figure 17 (L) shows the top-10 DeepFake generation methods or datasets based on their citations. Fig. 17 (R) shows the top-10 DeepFake generation methods or datasets based on citations normalized by the number of days since exposure. Similarly, Fig. 18 (L) shows the top-10 DeepFake detection methods based on citations and Fig. 18 (R) presents the top-10 DeepFake detection methods based on citations normalized by the number of days since exposure. In addition, Fig. 19 shows the top-10 DeepFake generation methods

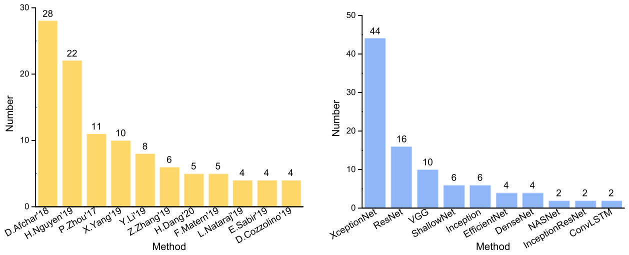


Fig. 15 (L) Top-11 most popular DeepFake detection methods chosen as baselines. (R) Top-10 most popular ML-based methods chosen as baselines

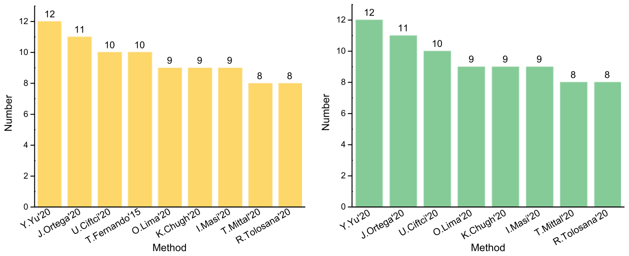


Fig. 16 (L) Top-9 DeepFake detection methods that benchmark against the most number of baselines. (R) Top-8 DeepFake detection methods that benchmark against the most number of baselines in 2020

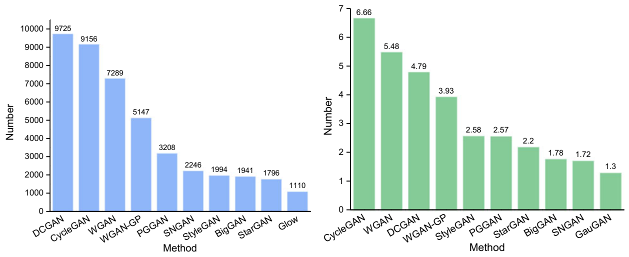


Fig. 17 (L) Top-10 DeepFake generation methods or datasets based on citations. (R) Top-10 DeepFake generation methods or datasets based on normalized citations

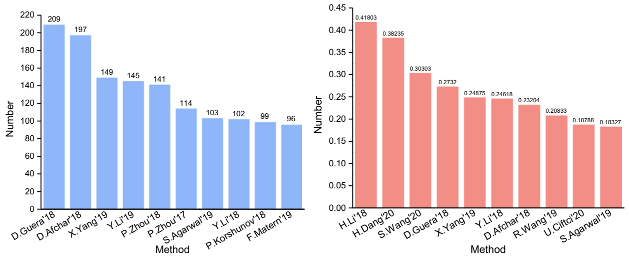


Fig. 18 (L) Top-10 DeepFake detection methods based on citations. (R) Top-10 DeepFake detection methods based on normalized citations

or datasets based on Elo rating with a default score set to 1400.

Regarding the citation of DeepFake detection methods reported in Figs. 18 and 11, it is actually difficult to identify some seminal milestone papers, although some papers have received more popularity than others. This phenomenon can be attributed to multiple factors and is a double-edged sword. The field of DeepFake detection is relatively new, thus it may take more time for any milestone papers to stand out. The lack of milestone papers can also be a positive indicator that the current state-of-the-art researches are multi-threaded and do not anchor on a few seminal works. Whether we are able to witness some new research hot zones emerge as the time goes by, the field is poised to progress at a fast pace.

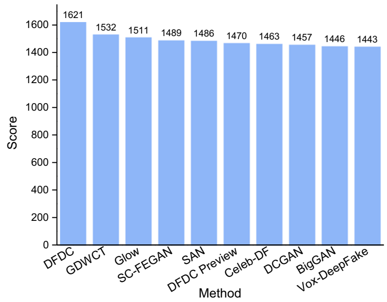


Fig. 19 Top-10 DeepFake generation methods or datasets based on Elo rating (Wikipedia 2021a). Default score is 1400

6 Evasion of DeepFake Detection

With the rapid development of DeepFake detectors, researchers start paying attention to design methods to evade the fake faces being detected. Specifically, given a real or fake face, evasion methods map it to a new one that cannot be correctly classified by the state-of-the-art DeepFake detectors, hiding the fake faces from being discovered. An exemplar pipeline of the evasion of DeepFake detection is shown in Fig. 20. We can roughly divide all methods into three types.

The *first* type is based on the adversarial attack. For example, Carlini and Farid (2020) add imperceptible adversarial perturbations to the fake/real faces and show that even the state-of-the-art DeepFake detectors are vulnerable to both white-box and black-box attacks (Carlini and Wagner 2017; Brown et al. 2017) with significant accuracy reduction on the public datasets (Wang et al. 2020e; Frank et al. 2020). Similarly, Gandhi and Jain (2020) use the fast gradient sign method (Goodfellow et al. 2014b) and C&W attacks (Carlini and Wagner 2017) to fool DeepFake detectors. Then, they propose two methods with the Lipschitz regularization (Woods et al. 2019) and deep image prior (Ulyanov et al. 2018) to improve the adversarial robustness of DeepFake detectors. Neekhara et al. (2021) further study the adversarial attack-based evasion methods on the more challenging DeepFake Detection Challenge (DFDC) dataset (Dolhansky et al. 2020) and find that the input-preprocessing steps, as well as face detection methods across DeepFake detectors, make the adversarial transferability difficult. Then, they implement a high transferability attack method based

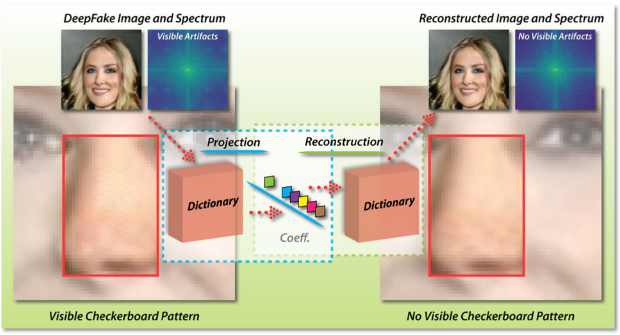

DeepFake Image and Spectrum
Visible Artifacts
Reconstructed Image and Spectrum
No Visible Artifacts
Projection
Reconstruction
Dictionary
Coeff.
Dictionary
Visible Checkerboard Pattern
No Visible Checkerboard Pattern

Fig. 20 Evasion of DeepFake detection via shallow reconstruction (Huang et al. 2020b)

on the universal adversarial perturbations to alleviate the challenges. In general, the adversarial attack-based methods inevitably introduce noise to the face images, leading to quality reduction.

The *second* type of methods focus on removing the fake traces in the frequency domain. Recent works on the detection of DeepFake images have pointed out that they are actually easily distinguishable by artifacts in their frequency spectra. Thus, some generation methods attempt to repair the flaw in the generation procedure. Durall et al. (2020) show that CNN-based generative deep networks with common up-sampling methods cannot reproduce spectral distributions of the real or natural training data, making the fake or generated images easily identified. To alleviate this drawback, they propose a novel spectral regularization objective term for training the GANs. Jiang et al. (2020a) also note this phenomenon and find that narrowing the frequency domain gap can improve the image synthesis quality further. To this end, they propose a frequency domain optimization target (i.e., focal frequency loss). The proposed loss enforces the model to dynamically focus on the frequency components that are hard to synthesize by down-weighting the easy frequencies. As a result, the method enhances the synthesis quality significantly. Jung and Keuper (2020) identify a straightforward solution for this issue by equipping the generative models with a spectral discriminator, thus the new trained GANs can generate images with realistic frequency spectra. These methods mainly focus on the mismatching between real and fake faces in the frequency domain while neglecting other potential factors that may make fake faces be identified easily.

The *third* kind of methods regard evasion as a general image generation process and use advanced image filtering or generative models to mislead DeepFake detectors. Huang et al. (2020b) demonstrate that the DeepFake detectors can be easily evaded via the shallow reconstruction based on sparse coding and dictionary-based reconstruction. In addition to the non-deep-learning solution, Huang et al. (2020a, 2021a) propose to fool the DeepFake detectors by first adding the deliberate noise to destroy the fake trace in the frequency domain and then reconstructing the clear counterpart via a deep kernel prediction network. Besides, Neves et al. (2020) remove the 'fingerprints' in the fake faces through a pre-trained GAN model, which can spoof the DeepFake detectors while maintaining the visual quality of the fake faces. In contrast to the above solutions of adding extra modules for evading DeepFake detection, Osakabe et al. (2021) propose to enhance the CycleGAN (Zhu et al. 2017) by equipping the fixed convolutional layers to remove the checkerboard artifacts.

7 Horizon

In this section, we touch upon the challenges and opportunities for future research directions surrounding DeepFake generation and DeepFake detection methods, as well as the evasion of DeepFake detection. The segmented discussions will be followed by a bird's-eye view comment of the entire DeepFake research field moving forward in the epilogue.

7.1 Generation of DeepFakes

We have surveyed and tabulated more than 91 papers published through the peer-review process or posted on arXiv on the topic of DeepFake generation and the datasets tasked for the DeepFake detection. The observed findings and challenges can shed some light on the future work in creating more realistic and detection-evasive DeepFakes. A much improved DeepFake generation method will in turn push forward the development of the DeepFake detection method.

- *Lacking ultra high-resolution images* Since PGGAN proposed a method to generate high-resolution (1024 × 1024) images, the new methods on synthesizing full fake images haven't been innovating towards higher resolution images. With the development of high-definition display resolution of phones or computers, 1024 × 1024 resolution may not enough in the near future.
- *Limited properties of face manipulation methods* The attribute manipulation methods can only change the properties given by the training set. Thus, the properties provided by these attribute manipulation methods are somewhat limited. An attribute manipulation method that is independent of the training set properties is desired.
- *Less consideration of video continuity* The identity swap and expression swap usually ignore the continuity of videos. They do not take physiological signals such as eye blink frequency, heart beat frequency into consideration.
- *Lacking diversified DeepFake datasets* The latest fake dataset are obsessed with being large scale. Most of them

only expand the diversity of the content-related factors such as gender and age of the subject, the place where the face photo is taken, the illumination condition, etc. The diversity in video quality such as various resolutions, various compression degrees, or other degradations commonly found in videos, etc., have not been fully taken into account. Furthermore, the DeepFake generation method used by these fake dataset are somewhat limited, which may fall short when tasked to demonstrate the diversity of different generation methods. The latest DeeperForensics-1.0 (Jiang et al. 2020b) dataset has been a good attempt in this regard by incorporating diverse perturbations such as Gaussian blue, added noise, JPEG compression, contrast change, etc. However, at the moment these perturbations are artificially added image-level degradations during post-processing, rather than organic video-level degradations such as bit-rate variations, choices of codec, etc. We hope to see more organic degradations incorporated in the future generation of the dataset, i.e., DeeperForensics-2.0.

– *Do not contain common sense fake* Most of the fakeness lies in the image texture. However, the fake datasets usually do not contain common sense fake such as three-eye human, one-horned human, etc. These fakes are obvious to humans but DeepFake detection methods may lack the common sense to judge properly.
– *Lacking a platform for demonstrating the different fake datasets* There lacks a platform for demonstrating the different fake datasets. On such a platform, we can directly see the different image style of the various fake datasets. The platform can also provide the information of the fake dataset such as (published year, image resolution, generation method, degraded or not, download link, the best detection method on each dataset, etc.).
– *Lacking sub-categorized DeepFake detection datasets with respect to gender, age, ethnicity* DeepFake detection benchmark datasets, like many other face recognition datasets, have data biases. For example, in many cases, the majority faces are from Caucasian males, and many of the internet-crawled datasets have the celebrity biases. With unbalanced datasets with respect to gender, age, and ethnicity to train the model, the learned DeepFake detector can become data biased as well. It is worthwhile to push for a more balanced DeepFake detection benchmark. We have seen some recent attempts to build DeepFake detection dataset based on one ethnicity group such as (Kwon et al. 2021). More development in this direction is needed.
– *Lacking multi-face DeepFake detection datasets* For most of the existing DeepFake detection benchmark datasets, single face is DeepFake manipulated in the image or video, and when multiple faces are present, oftentimes, only one of the faces is DeepFake manipulated (usually the one with the largest detection bounding box size). There is a need to push for DeepFake detection benchmarks that involve multiple faces or with unknown number of DeepFake manipulated faces in the crowd. This effort will not only pose a new dimension of the challenge for DeepFake detectors, considering that the manipulation, if happens, may be hidden in the crowd and with unknown number. Also, this will foster new research into the DeepFake detection method where cues can now be drawn beyond individual faces and from the peers in the images or videos. It is good to see that one of the latest benchmarks (Le et al. 2021) is created towards that goal, and we hope to see more.

7.2 Detection of DeepFakes

We have investigated and tabulated more than 117 papers published through the peer-review process or posted on arXiv on the topic of DeepFake detection. We have observed some interesting findings and challenges, after reviewing the papers, which could inspire future work in defending DeepFakes more effectively.

– *Lacking public AI-synthesized image datasets* Almost all the existing studies build their own image dataset with various GANs to evaluate the effectiveness of their method in defending still image DeepFakes. They do not have a consensus on which forgery image datasets need to be used in evaluation. These studies claim that they have achieved competitive results in detecting various GAN-synthesized images built on their own. However, the quality of these generated fake images is still unknown, i.e., if there are any obvious artifacts that exist in the image. A public GAN-synthesized fake image dataset needs to be developed by the community for advancing this challenging research field.
– *Lacking competitive baselines in comparison* In evaluating the performance of their proposed methods, existing studies prefer to employ some simple baselines (e.g., simple DNN-based methods, naive methods by leveraging perceptible artifacts) rather than the SOTA work to demonstrate that they have beaten the prior studies. We hope that future studies could compare their work with some competitive baselines which are highlighted in Tables 7 to demonstrate the advances of their work.
– *Generalization abilities of DeepFake detectors* Tackling the unknown DeepFakes is one of the key challenges in fighting against DeepFakes. In recent years, a series of studies are working towards this goal to develop more generalized methods. Unfortunately, these works are merely evaluated on simple DeepFake video datasets, like FaceForensics++. We hope that future work can focus more on challenging datasets.

- *Robustness of DeepFake detectors* In the real-world, DeepFakes can easily suffer from various degradations, such as image/video compression, added Gaussian noises, blurring, low-light (Juefei-Xu and Savvides 2015), low-resolution (Abiantun et al. 2019), etc. Existing studies proposed various robust methods to tackle this simple degradation. However, more than 90% of methods leverage DNNs as their backbone to determine real and fake in the final classifier. The DNNs are vulnerable to adversarial noise attacks with imperceptible additive noises, which is demonstrated by prior works. Unfortunately, we observe that all the existing studies failed in evaluating their robustness against adversarial noise attacks. In addition, the SOTA detectors may fall short when faces are under occlusions such as facial masks (Zhu et al. 2022) where only the eye region is visible (Juefei-Xu and Savvides 2016; Juefei-Xu et al. 2015), heavy makeups, heavy facial hairs, etc.
- *Capabilities of DeepFake detectors* Improving the generalization capabilities to tackle the emerging unknown DeepFakes, enhancing the robustness against various DeepFake degradations including simple transformations and adversarial attacks, and explaining why the detector works are the three key factors in developing a practical DeepFake detector which could be deployed in the wild. In reviewing the recent papers with regard to detecting DeepFakes, we find that less than ten papers have evaluated the capabilities of their method from all three perspectives.
- *A Comprehensive evaluation metrics* The performance of DeepFake detectors is highly determined by the quality of DeepFakes. The low-quality DeepFakes (e.g., DeepFake-TIMIT, FaceForensic++) with observable artifacts could be easily identified by almost all the DeepFake detectors with high confidence, while, the challenging high-quality DeepFakes (e.g., Celeb-DF, DFDC) which could fool our eyes can be hardly determined by detectors. The existing studies report their experimental results by merely considering the detection accuracy and false alarms, which ignore the relation with the quality of DeepFakes, especially from the self-built DeepFake datasets. We hope that more comprehensive experimental results by considering the quality of DeepFakes should be considered in future work. Thus, new metrics for measuring the quality of DeepFakes need to be proposed by researchers.
- *A platform for evaluation* In Tables 7, we can find that the existing DeepFake detectors can easily achieve more than 90% detection accuracy in fighting the common DeepFakes. However, in a DeepFake Detection Challenge (DFDC) built by Facebook, the final competition results show that the winner can only give less than 70% accuracy in detecting DeepFakes. Another DeepFake detection challenge, called DeeperForensics Challenge 2020 (Jiang et al. 2021a), is hosted on DeeperForensics-1.0 dataset which is a real-world face forgery detection dataset. However, only 25 teams made valid submissions, and only one method adopted for generating DeepFakes in DeeperForensics-1.0. The results cannot represent the SOTA performance in DeepFake detection. Thus, the DeepFake is still a real threat to the community and academia needs to develop more practical detection methods. Obviously, the reported experimental results in academic papers can not reflect the true performance of their methods. A platform, incorporating the challenging DeepFake datasets and competitive baselines, is not ready for evaluating the true performance of existing DeepFake detectors and the future DeepFake detectors. FaceForensic++ provides a simple platform with low-quality and simple CNNs as baselines, which might fall short when tackling the ever-progressing DeepFakes.

7.3 Evasion of DeepFake Detection

We have discussed three kinds of methods for evading DeepFake detection in Sect. 6, which mainly aims at misleading the DeepFake detectors or removing artifacts introduced by DeepFake generations. In the near future, we hope that the evading methods would evade new DeepFake detectors by developing more advanced adversarial attacks that consider natural degradation in the real world and deeply removing the fake traces in both images and videos. More specifically, the following directions should be noted:

- *Misleading DeepFake detection via natural degradation* Existing adversarial attack-based evasion methods mainly rely on the additive adversarial perturbations that do not exist in the real world and might be detected by recent works on detecting adversarial examples (Pang et al. 2018; Zheng and Hong 2018). Moreover, the state-of-the-art defense methods are also able to invalidate the adversarial attacks, thus making the evasion methods less effective. A possible solution for this problem is to design natural degradation-based adversarial attacks, e.g., motion blur, light variation, shadow synthetic, etc., allowing generating realistic examples while misleading the DeepFake detection. For example, Guo et al. (2020a) realize an adversarial blur attack that can generate realistic-blurred images and mislead the state-of-the-art deep neural networks (Guo et al. 2021). Similar works are proposed for natural degradations like weather elements (Zhai et al. 2022, 2020; Gao et al. 2021b), exposure (Gao et al. 2022, 2020; Cheng et al. 2020), lighting (Gao et al. 2021c; Tian et al. 2021b, a; Sun et al. 2022), shadow (Fu et al. 2021a, b), defocus blur (Huang et al. 2021b), etc. In the future, we can employ these attacks to evade the DeepFake detection with the natural adversarial examples that can be hardly defended through the state-of-the-art defense methods designed for additive adversarial perturbations.

- *Faking the physiological signal in fake videos* The state-of-the-art DeepFake detector starts using physiological signal, e.g., heart rate extracted from the video, as an effective fake indicator (Qi et al. 2020), because even the advanced GAN methods can hardly preserve the heart rate signal that usually presents as fine-grain color variation among frames. To evading such new detectors, we should develop a novel evading method allowing the processed fake videos to also contain the normal heart rate single. This actually requires us to learn how to add sequential color variations into the frames in a fake video, letting the rhythm detection methods obtain normal heart rate.
- *Joint perception and appearance fake trace removal* Existing fake trace removal-based evading methods mainly focus on how to remove the known appearance artifacts, e.g., spectral distributions in the frequency domain, introduced by DeepFake generations while ignoring their influence on the perception, e.g., deep representations of fake faces, which seems to be the essential factor for effective DeepFake detection. Hence, in the near future, a more advanced fake trace removal could be explored by jointly removing the artifacts and perceptions of fake traces.

7.4 Epilogue and the Next Chapter

Now that we have discussed the existing challenges and opportunities for future studies, it is a good segue into some final thoughts regarding DeepFakes.

Based on the discussions throughout this survey paper, we can see that at the moment, the work on DeepFake detection heavily relies on curated datasets with the latest DeepFake generation methods incorporated that show the highest level of realisticity when the dataset is created. We would like to emphasize the significance of the continued progression of such datasets. Unlike the acclaimed ImageNet (Deng et al. 2009) classification tasks, whose image classification difficulty will remain relatively unchanged throughout the years, the DeepFake detection tasks are becoming increasingly more difficult year over year since the DeepFake generation method can produce increasingly more realistic DeepFakes. In this sense, it is imperative to hold periodic ImageNet-style contests and/or produce updated DeepFake detection datasets to keep track of the latest DeepFake generation methods and encourage competition among various research groups in order to advance the effort of countering malicious DeepFakes. A very fitting example would be the latest DeeperForensics Challenge 2020 on Real-World Face Forgery Detection (Jiang et al. 2021a).

Needless to say, the various DeepFake datasets, produced by the DeepFake generators, are valuable assets for developing next-generation DeepFake detectors. Over the years, we have seen that the datasets have grown tremendously in sizes, quality, diversity, and levels of challenging scenarios. How will the datasets evolve in the next five, ten years is unknown at this point, but we envision that the DeepFake datasets may evolve into a dichotomy following similar trends as other computer vision datasets. On one hand, there will be convergence of many dataset sources into a few very large-scale standardized evaluation datasets for the DeepFake community, similar to the scales of the ImageNet dataset or the COCO object detection dataset (Lin et al. 2014). These large-scale datasets will be less frequently updated and will most likely be served as the go-to benchmark and tools for developing and evaluating DeepFake algorithms. Next-generation large-scale general-purpose DeepFake foundational models can be developed on these large-scale datasets. In computer vision and natural language processing, foundational models (Bommasani et al. 2021) are those models trained on broad data at scale (usually in multi-modality such as vision and language) and are adaptable to a wide range of downstream tasks. Examples of vision and language foundational models include Florence (Yuan et al. 2021), CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), Wu Dao (Beijing Academy of Artificial Intelligence 2021), etc. We envision that similar general-purpose DeepFake detectors that are able to deal with the majority of DeepFake types will emerge. On the other hand, proprietary datasets that are smaller in scale that are more flexible and more frequently updated are also likely to emerge. These datasets, on the contrary, are best catered towards developing and evaluating *ad-hoc* DeepFake algorithms for particular types of DeepFake generators that are newly emerged, or for particular long-tail scenarios, and more importantly, for finetuning the aforementioned DeepFake foundational models with particular downstream DeepFake-related tasks. With the two types of datasets discussed above, i.e., large-scale general-purpose datasets vs. smaller-scale proprietary datasets, if one DeepFake dataset is out of date, it can still be beneficial to the community by being incorporated into the first type of the datasets because it is deemed very valuable for maintaining the large scale and diversity of the datasets. Meanwhile, newly emerged datasets from the latest DeepFake generators will carry the weight for pushing the development of next-generation DeepFake detectors, and when they become obsolete, they will be replaced by newer ones, and they will still find their way back to contribute to the first type of general-purpose datasets.

There have not been many studies on the intersection of DeepFake generation and adversarial attack. The current research landscape is still largely fragmented w.r.t.the two domains. Most common adversarial attacks create image-level pixel perturbations to alter the classification output through access to the white-box model parameters. However, such perturbation is not limited to image-level representations. When we decouple the face representation into various attribute/semantic latent sub-representations such as identity, expression, gender, ethnicity, etc., from the adversarial attack point of view, we can see clearly that the identity swapped DeepFake is merely an adversarial perturbation on the identity sub-representation, and similarly for expression swapped DeepFakes. From this viewing angle, how the DeepFake narrative can fit into the adversarial robustness studies is worth looking forward to.

Incorporating other multi-media modalities such as voice and sound will help to counter malicious DeepFakes. In this survey, although we have focused on the image and video modalities, it is quite intuitive to monitor the realism of the DNN generated fake voice and sound (Wang et al. 2020c) in a DeepFake video. Being able to detect multi-modal fake traces, such as traces from standalone visual cues, standalone acoustic cues, as well as the interactions between them, such as synchronization, the combined effort in detecting DeepFakes will most likely be boosted by practicing Liebig's barrel theory.

One issue surrounding the battle between DeepFake generation methods and DeepFake detection methods is that the detection method is usually lagging behind. This is generally true for adversaries and defenders in any "battle" scenarios such as the development of Covid-19 vaccination happening after the Covid-19 outbreak because the knowledge of the virus is required for the vaccination development. The same applies here to DeepFake problems. The currently developed DeepFake detection methods are still somewhat myopic in the sense that they can most confidently tackle DeepFakes generated by existing methods, but will be outcompeted by DeepFakes generated by future generation techniques. How to make DeepFake detection methods forward-looking and remain effective, with or without small tweaks, through iterations of DeepFake generation methods is an open challenge. The earlier-mentioned DeepFake foundational models can potentially provide a more friendly training paradigm for the continual evolving of the DeepFake algorithms on both sides of the battleground. Meanwhile, there are some proactive measures that the defenders (such as social media platforms where DeepFakes are most likely disseminated) can take in order to become more effective in fighting malicious DeepFakes, for example, through the responsible disclosure of generative models using GAN fingerprinting (Yu et al. 2020a), or embedding an invisible tag into the original clean image uploaded by the user which can remain retrievable after the DeepFake generation process so that at a later time when the DeepFake version is re-uploaded by the bad actor, the platform is able to retrieve the tag and block the dissemination (Wang et al. 2020b).

From an algorithmic point of view, two of the major research hot topics in the machine learning community at the moment are self-supervised learning and transformer for vision and language problems. Self-supervised learning, free of labels, enables continuous life-long learning on an infinite and smoothly changing data stream (Sun et al. 2020b). What would self-supervised learning in the domain of DeepFake detection be like? Transformer language models are now being operationalized to computer vision domain with the latest ImageGPT (Chen et al. 2020), DALL-E (OpenAI 2021), and subsequent CLIP (Radford et al. 2021) all from OpenAI, as well as Google's ALIGN (Jia et al. 2021). They, together with other latest generative approaches such as vector quantized variational autoencoder (VQ-VAE) (Razavi et al. 2019), VQGAN + autoregressive transformer (Esser et al. 2021), etc., will definitely supplement and enhance the DeepFake generation techniques. We have seen the latest vision transformers (ViT) (Dosovitskiy et al. 2020) already equipped with the state-of-the-art image generation capabilities such as the TransGAN (Jiang et al. 2021b) and Paint Transformer (Liu et al. 2021c). Although there hasn't been a ViT dedicated for DeepFake generation or detection yet, the unprecedentedly fast growing pace and increasing ubiquity in ViTs (Khan et al. 2021) to tackle various computer vision problems will eventually turn the landscape of DeepFake generation and detection into a ViT-based one. With both sides of the DeepFake battleground now equipped with the newest technological expertise, the clash between the two parties will, for sure, spark flaming research in the foreseeable future.

The humanity may reach a stage where DeepFakes have become so genuinely looking that they are beyond human and machine's capability to distinguish from the real ones, a *"DeepFake singularity"*, if you will. If this day is inevitable, be it a utopia or a dystopia, perhaps a more interesting era is upon us. Are we brave enough to embrace it?

8 Conclusion

In this survey, we have provided a comprehensive overview and detailed analysis of the research work on the topic of DeepFake generation, DeepFake detection as well as evasion of DeepFake detection, with more than 318 research papers carefully surveyed. We have presented the taxonomy of various DeepFake generation methods and the categorization of various DeepFake detection methods along with highlights of the technical evolution of the methods, and more importantly, we have showcased the battleground between the two parties with detailed interactions between the adversaries (DeepFake generation methods) and the defenders (DeepFake detection methods). The battleground allows fresh perspective into the latest landscape of the DeepFake research and can provide valuable analysis towards the research challenges and opportunities as well as research trends and directions in the field of DeepFake generation and detection. We hope that this survey paper can help empower and fast-track researchers and practitioners in this field to identify the most pressing research topics and attract more researchers to contribute to this emerging and rapidly growing field.

Acknowledgements This research was partly supported by the National Key Research and Development Program of China Under Grant No. 2021YFB3100700, the Fellowship of China National Postdoctoral Program for Innovative Talents Under No. BX2021229, the Natural Science Foundation of Hubei Province Under No. 2021CFB089, the Fundamental Research Funds for the Central Universities Under No. 2042021kf1030, the Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness under No. HNTS2022004, the National Natural Science Foundation of China (NSFC) under No. 61876134. The work was also supported by the National Research Foundation, Singapore under its the AI Singapore Programme (AISG2-RP-2020-019), the National Research Foundation, Prime Ministers Office, Singapore under its National Cybersecurity R&D Program (No. NRF2018NCR-NCR005-0001), NRF Investigatorship NRFI06-2020-0001, the National Research Foundation through its National Satellite of Excellence in Trustworthy Software Systems (NSOE-TSS) project under the National Cybersecurity R&D (NCR) Grant (No. NRF2018NCR-NSOE003-0001). We gratefully acknowledge the support of NVIDIA AI Tech Center (NVAITC) to our research.

References

115th Congress. (2018). S.3805—Malicious Deep Fake Prohibition Act of 2018. https://www.congress.gov/bill/115th-congress/senate-bill/3805

116th Congress. (2019a). H.R.3230—Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019. https://www.congress.gov/bill/116th-congress/house-bill/3230/

116th Congress. (2019b). S.2065—Deepfake Report Act of 2019. https://www.congress.gov/bill/116th-congress/senate-bill/2065

Abiantun, R., Juefei-Xu, F., Prabhu, U., & Savvides, M. (2019). SSR2: Sparse signal recovery for single-image super-resolution on faces with extreme low resolutions. *Pattern Recognition*, *90*, 308–324.

Adobe. (2021a). Adobe audition. Retrieved August 1, 2021, from https://www.adobe.com/products/audition.html (online).

Adobe. (2021b). Adobe lightroom. Retrieved August 1, 2021, from https://www.adobe.com/products/photoshop-lightroom.html (online).

Adobe. (2021c). Adobe photoshop. Retrieved August 1, 2021, from https://www.adobe.com/products/photoshop.html (online).

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). Mesonet: A compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–7). IEEE.

Afifi, M., Brubaker, M. A., & Brown, M. S. (2021). Histogan: controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7941–7950).

Agarwal, S., & Farid, H. (2021). Detecting deep-fake videos from aural and oral dynamics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 981–989).

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019). Protecting world leaders against deep fakes. In *CVPR workshops* (pp. 38–45).

Agarwal, S., Farid, H., El-Gaaly, T., & Lim, S. N. (2020a). Detecting deep-fake videos from appearance and behavior. In *2020 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–6). IEEE.

Agarwal, S., Farid, H., Fried, O., & Agrawala, M. (2020b). Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 660–661).

Amerini, I., Galteri, L., Caldelli, R., & Del Bimbo, A. (2019). Deepfake video detection through optical flow based CNN. In *Proceedings of the IEEE international conference on computer vision workshops*.

Aneja, S., Midoglu, C., Dang-Nguyen, D. T., Riegler, M. A., Halvorsen, P., Niessner, M., Adsumilli, B., & Bregler, C. (2021). Mmsys' 21 grand challenge on detecting cheapfakes. arXiv preprint arXiv:2107.05297

Antares Audio Technologies. (2021). Auto-tune. Retrieved August 1, 2021, from http://www.antarestech.com/product/auto-tune-pro/ (online).

Arjovsky M., Chintala S., & Bottou L. (2017). Wasserstein. arXiv:1701.07875

Bai, Y., Guo, Y., Wei, J., Lu, L., Wang, R., & Wang, Y. (2020). Fake generated painting detection via frequency analysis. In *2020 IEEE international conference on image processing (ICIP)* (pp.1256–1260). IEEE.

Barni, M., Kallas, K., Nowroozi, E., & Tondi, B. (2020). CNN detection of gan-generated face images based on cross-band co-occurrences analysis. In *2020 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–6). IEEE.

Beijing Academy of Artificial Intelligence. (2021). Wu Dao 2.0. https://gpt3demo.com/apps/wu-dao-20

Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., & Munos, R. (2017). The Cramer distance as a solution to biased Wasserstein gradients. arXiv preprint arXiv:1705.10743

Berthelot, D., Schumm, T., & Metz, L. (2017). Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717

Bilibili. (2010). A Chinese video sharing website. https://www.bilibili.com/

Bińkowski, M., Sutherland, D. J., Arbel, M., & Gretton, A. (2018). Demystifying MMD gans. arXiv preprint arXiv:1801.01401

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., & Brunskill, E., (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258

Bondi, L., Cannas, E. D., Bestagini, P., & Tubaro, S. (2020). Training strategies and data augmentations in CNN-based deepfake video detection. In *2020 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–6). IEEE.

Bonettini, N., Bestagini, P., Milani, S., & Tubaro, S. (2021a). On the use of benford's law to detect gan-generated images. In *2020 25th international conference on pattern recognition (ICPR)* (pp. 5495–5502). IEEE.

Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., & Tubaro, S. (2021b). Video face manipulation detection through ensemble of CNNs. In *2020 25th international conference on pattern recognition (ICPR)* (pp. 5012–5019). IEEE.

Bonomi, M., Pasquini, C., & Boato, G. (2020) Dynamic texture analysis for detecting fake faces in video sequences. arXiv preprint arXiv:2007.15271

Brock, A., Donahue, J., & Simonyan, K. (2018) Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096

Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017) Adversarial patch. arXiv preprint arXiv:1712.09665

BuzzFeed. (2018). How to spot a deepfake like the Barack Obama-Jordan Peele Video. https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed

California (2019) California assembly bill no. 730. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730

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

Carlini, N., & Farid, H. (2020). Evading deepfake-image detectors with white-and black-box attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 658–659).

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (SP)* (pp. 39–57). IEEE.

Chai, L., Bau, D., Lim, S. N., & Isola, P. (2020). What makes fake images detectable? Understanding properties that generalize. In *European conference on computer vision* (pp. 103–120). Springer.

Chandrasegaran, K., Tran, N. T., & Cheung, N. M. (2021). A closer look at Fourier spectrum discrepancies for CNN-generated images detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7200–7209).

Chen, C., Chen, Q., Xu, J., & Koltun, V. (2018a). Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3291–3300).

Chen, H., Hu, G., Lei, Z., Chen, Y., Robertson, N. M., & Li, S. Z. (2019). Attention-based two-stream convolutional networks for face spoofing detection. *IEEE Transactions on Information Forensics and Security, 15*, 578–593.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020). Generative pretraining from pixels. In *International conference on machine learning* (pp. 1691–1703). PMLR.

Chen, Q., & Koltun, V. (2017) Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 1511–1520).

Chen, R. T., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018b). Neural ordinary differential equations. In *Advances in neural information processing systems* (pp. 6571–6583).

Chen, X., Xu, C., Yang, X., Song, L., & Tao, D. (2018). Gated-gan: Adversarial gated networks for multi-collection style transfer. *IEEE Transactions on Image Processing, 28*(2), 546–560.

Chen, Y., Fan, H., Xu, B., Yan, Z., Kalantidis, Y., Rohrbach, M., Yan, S., & Feng, J. (2019b). Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE international conference on computer vision* (pp. 3435–3444).

Chen, Y. C., Xu. X., Tian, Z., & Jia, J. (2019c). Homomorphic latent space interpolation for unpaired image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2408–2416).

Chen, Z., & Yang, H. (2021). Attentive semantic exploring for manipulated face detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1985–1989). IEEE

Chen, Z., & Yang, H. (2021). Attentive semantic exploring for manipulated face detection. *ICASSP 2021–2021 IEEE international conference on acoustics* (pp. 1985–1989). IEEE: Speech and Signal Processing (ICASSP).

Chen, Z., Tondi, B., Li, X., Ni, R., Zhao, Y., & Barni, M. (2019). Secure detection of image manipulation by means of random feature selection. *IEEE Transactions on Information Forensics and Security, 14*(9), 2454–2469.

Cheng, Y., Juefei-Xu, F., Guo, Q., Fu, H., Xie, X., Lin, S. W., Lin, W., & Liu, Y. (2020) Adversarial exposure attack on diabetic retinopathy imagery. arXiv preprint arXiv:2009.09231

Cho, W., Choi, S., Park, D. K., Shin, I., & Choo, J. (2019). Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 10639–10647).

Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8789–8797).

Choi, Y., Uh, Y., Yoo, J., & Ha, J. W. (2020). Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8188–8197).

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).

Chugh, K., Gupta, P., Dhall, A., & Subramanian, R. (2020) Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 439–447).

Ciftci, U., Demir, I., & Yin, L. (2020a). Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2020.3009287

Ciftci, U.A., Demir, I., & Yin, L. (2020b). How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals. arXiv preprint arXiv:2008.11363

CNN. (2020). 'Deepfake' Queen delivers alternative Christmas speech, in warning about misinformation. https://www.cnn.com/2020/12/25/uk/deepfake-queen-speech-christmas-intl-gbr

Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., & Verdoliva, L. (2018). Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510

Cozzolino, D., Rössler, A., Thies, J., Nießner, M., & Verdoliva, L. (2020) Id-reveal: Identity-aware deepfake video detection. arXiv preprint arXiv:2012.02512

Dai, T., Cai, J., Zhang, Y., Xia, S. T., & Zhang, L. (2019). Second-order attention network for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 11065–11074).

Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5781–5790).

Dang, L. M., Hassan, S. I., Im, S., & Moon, H. (2019). Face image manipulation detection based on a convolutional neural network. *Expert Systems with Applications, 129*, 156–168.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Ding, H., Sricharan, K., Chellappa, R. (2018). Exprgan: Facial expression editing with controllable expression intensity. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32).

Ding, X., Raziei, Z., Larson, E. C., Olinick, E. V., Krueger, P., & Hahsler, M. (2020). Swapped face detection using deep learning and subjective assessment. *EURASIP Journal on Information Security, 2020*, 1–12.

Dogonadze, N., Obernosterer, J., & Hou, J. (2020) Deep face forgery detection. arXiv preprint arXiv:2004.11804

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019) The deepfake detection challenge (DFDC) preview dataset. arXiv preprint arXiv:1910.08854

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The deepfake detection challenge dataset. arXiv preprint arXiv:2006.07397

Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Chen, D., Wen, F., & Guo, B. (2020). Identity-driven deepfake detection. arXiv preprint arXiv:2012.03930

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

Du, M., Pentyala, S., Li, Y., & Hu, X. (2019). Towards generalizable forgery detection with locality-aware autoencoder. arXiv preprint arXiv:1909.05999

Dufour, N., & Gully, A. (2019). Contributing data to deepfake detection research. Google AI Blog. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

Durall, R., Keuper, M., Pfreundt, F. J., & Keuper, J. (2019). Unmasking deepfakes with simple features. arXiv preprint arXiv:1911.00686

Durall, R., Keuper, M., & Keuper, J. (2020). Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7890–7899).

Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12873–12883).

FaceApp. (2021). FaceApp. https://faceapp.com/app

Facebrity. (2021). Facebrity, Apple App Store. https://apps.apple.com/us/app/facebrity-face-swap-morph-app/id1449734851

FaceSwap. (2016). FaceSwap. https://github.com/deepfakes/faceswap

Feng, D., Lu, X., & Lin, X. (2020). Deep detection for face manipulation. In *International conference on neural information processing* (pp. 316–323). Springer.

Fernandes, S., Raj, S., Ortiz, E., Vintila, I., Salter, M., Urosevic, G., & Jha, S. (2019). Predicting heart rate variations of deepfake videos using neural ode. In *Proceedings of the IEEE international conference on computer vision workshops*.

Fernando, T., Fookes, C., Denman, S., & Sridharan, S .(2019). Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks. arXiv preprint arXiv:1911.07844

Fraga-Lamas, P., & Fernández-Caramés, T. M. (2020). Fake news, disinformation, and deepfakes: Leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality. *IT Professional, 22*(2), 53–59.

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning* (pp. 3247–3258). PMLR.

Friesen, E., & Ekman, P. (1978). Facial action coding system: A technique for the measurement of facial movement. *Palo Alto, 3*(2), 5.

Fu, L., Guo, Q., Juefei-Xu, F., Yu, H., Feng, W., Liu, Y., & Wang, S. (2021a). Benchmarking shadow removal for facial landmark detection and beyond. arXiv preprint arXiv:2111.13790

Fu, L., Zhou, C., Guo, Q., Juefei-Xu, F., Yu, H., Feng, W., Liu, Y., & Wang, S. (2021b). Auto-exposure fusion for single-image shadow removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10571–10580).

Gandhi, A., & Jain, S. (2020) Adversarial perturbations fool deepfake detectors. In *2020 International joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Ganiyusufoglu, I., Ngô, L. M., Savov, N., Karaoglu, S., Gevers, T. (2020). Spatio-temporal features for generalized detection of deepfake videos. arXiv preprint arXiv:2010.11844

Gao, G., Huang, H., Fu, C., Li, Z., & He, R. (2021a). Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3404–3413).

Gao, R., Guo, Q., Juefei-Xu, F., Yu, H., Ren, X., Feng, W., & Wang, S. (2020). Making images undiscoverable from co-saliency detection. arXiv preprint arXiv:2009.09258

Gao, R., Guo, Q., Juefei-Xu, F., Yu, H., & Feng, W. (2021b). AdvHaze: Adversarial Haze Attack. arXiv preprint arXiv:2104.13673

Gao, R., Guo, Q., Zhang, Q., Juefei-Xu, F., Yu, H., & Feng, W. (2021c). Adversarial Relighting against Face Recognition. arXiv preprint arXiv:2108.07920

Gao, R., Guo, Q., Juefei-Xu, F., Yu, H., Fu, H., Feng, W., Liu, Y., & Wang, S. (2022). Can You Spot the Chameleon? Adversarially Camouflaging Images from Co-Salient Object Detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE.

Gao, Y., Wei, F., Bao, J., Gu, S., Chen, D., Wen, F., & Lian, Z. (2021d). High-fidelity and arbitrary face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16115–16124).

Goebel, M., Nataraj, L., Nanjundaswamy, T., Mohammed, T. M., Chandrasekaran S, Manjunath B (2020) Detection, attribution and localization of gan generated images. arXiv preprint arXiv:2007.10466

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014a). Generative adversarial networks. arXiv preprint arXiv:1406.2661

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014b). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572

Gu, S., Bao, J., Yang, H., Chen, D., Wen, F.,& Yuan, L. (2019) Mask-guided portrait editing with conditional gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp 3436–3445).

Guarnera, L., Giudice, O., & Battiato, S. (2020a). Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 666–667).

Guarnera, L., Giudice, O., & Battiato, S. (2020). Fighting deepfake by exposing the convolutional traces on images. *IEEE Access, 8*, 165085–165098.

Guarnera, L., Giudice, O., Nastasi, C., & Battiato, S. (2020c). Preliminary forensics analysis of deepfake images. In *2020 AEIT international annual conference (AEIT)* (pp. 1–6). IEEE.

Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1–6). IEEE.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of Wasserstein gans. *Advances in neural information processing systems, 30*, 5767–5777.

Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Wang, J., Yu, B., Feng, W., & Liu, Y. (2020). Watch out! motion is blurring the vision of your deep neural networks. *Advances in Neural Information Processing Systems, 33*, 975–985.

Guo, Q., Cheng, Z., Juefei-Xu, F., Ma, L., Xie, X., Liu, Y., & Zhao, J. (2021). Learning to Adversarially Blur Visual Object Tracking. In *Proceedings of the IEEE international conference on computer vision (ICCV)*. IEEE.

Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision* (pp. 87–102). Springer.

Guo, Y., Chen, J., Wang, J., Chen, Q., Cao, J., Deng, Z., Xu, Y., & Tan, M. (2020b). Closed-loop matters: Dual regression networks for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5407–5416).

Guo, Z., Yang, G., Chen, J., & Sun. X. (2020c). Fake face detection via adaptive residuals extraction network. arXiv preprint arXiv:2005.04945

Gupta, P., Chugh, K., Dhall, A., Subramanian, R. (2020). The eyes know it: Fakeet-an eye-tracking database to understand deepfake perception. In *Proceedings of the 2020 international conference on multimodal interaction* (pp. 519–527).

Ha, S., Kersner, M., Kim, B., Seo, S., & Kim, D. (2020). Marionette: Few-shot face reenactment preserving identity of unseen targets. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*, 10893–10900.

Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021). Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5039–5049).

Hasan, H. R., & Salah, K. (2019). Combating deepfake videos using blockchain and smart contracts. *IEEE Access, 7*, 41596–41606.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

He, P., Li, H., & Wang, H. (2019a). Detection of fake images via the ensemble of deep representations from multi color spaces. In *2019 IEEE international conference on image processing (ICIP)* (pp. 2299–2303). IEEE.

He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., & Liu, Z. (2021) Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4360–4369).

He, Z., Zuo, W., Kan, M., Shan, S., & Chen, X. (2019). Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing, 28*(11), 5464–5478.

Hernandez-Ortega, J., Tolosana, R., Fierrez, J., & Morales, A. (2020). Deepfakeson-phys: Deepfakes detection based on heart rate estimation. arXiv preprint arXiv:2010.00400

Hsu, C. C., Lee, C. Y., & Zhuang, Y. X. (2018). Learning to detect fake face images in the wild. In *2018 international symposium on computer, consumer and control (IS3C)* (pp. 388–391). IEEE.

Hsu, C. C., Zhuang, Y. X., & Lee, C. Y. (2020). Deep fake image detection based on pairwise learning. *Applied Sciences, 10*(1), 370.

Hu, S., Li, Y., & Lyu, S. (2021). Exposing gan-generated faces using inconsistent corneal specular highlights. In *ICASSP 2021–2021 IEEE international conference on acoustics* (pp. 2500–2504). IEEE: Speech and Signal Processing (ICASSP).

Huang, H., Li, Z., He, R., Sun, Z., & Tan, T. (2018). Introvae: Introspective variational autoencoders for photographic image synthesis. arXiv preprint arXiv:1807.06358

Huang, Y., Juefei-Xu, F., Guo, Q., Xie, X., Ma, L., Miao, W., Liu, Y., & Pu, G. (2020a). Fakeretouch: Evading deepfakes detection via the guidance of deliberate noise. arXiv preprint arXiv:2009.09213

Huang, Y., Juefei-Xu, F., Wang, R., Guo, Q., Ma, L., Xie, X., Li, J., Miao, W., Liu, Y., & Pu, G. (2020b). Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction. In *Proceedings of the 28th ACM international conference on multimedia* (pp 1217–1226).

Huang, Y., Juefei-Xu, F., Gou, Q., Liu, Y., & Pu, G. (2022). Fakelocator: Robust localization of GAN-based face manipulations. *IEEE Transactions on Information Forensics and Security*.

Huang, Y., Juefei-Xu, F., Guo, Q., Ma, L., Xie, X., Miao, W., Liu, Y., & Pu, G. (2021a). Dodging DeepFake detection via implicit spatial-domain notch filtering. arXiv preprint arXiv:2009.09213

Huang, Y., Juefei-Xu, F., Guo, Q., Miao, W., Liu, Y., & Pu, G. (2021b). AdvBokeh: Learning to adversarially defocus Blur. arXiv preprint arXiv:2111.12971

Hulzebosch, N., Ibrahimi, S., & Worring, M. (2020). Detecting cnn-generated facial images in real-world scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 642–643).

Hyun, S., Kim, J., & Heo, J. P. (2021). Self-supervised video gans: Learning for appearance consistency and motion coherency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10826–10835).

Iizuka, S., Simo-Serra, E., & Ishikawa, H. (2017). Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG), 36*(4), 1–14.

IMDb. (2021). Avatar. Retrieved August 1, 2021, from https://www.imdb.com/title/tt0499549/ (online).

Jain, A. K., Flynn, P., & Ross, A. A. (2007). *Handbook of biometrics*. Springer.

Jeon, H., Bang, Y., Kim, J., & Woo, S. S. (2020a). T-gd: Transferable gan-generated images detection framework. arXiv preprint arXiv:2008.04115

Jeon, H., Bang, Y., Woo, S.S. (2020b). Fdftnet: Facing off fake images using fake detection fine-tuning network. In: IFIP international conference on ICT systems security and privacy protection (pp. 416–430). Springer.

Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918

Jiang, L., Dai, B., Wu, W., & Loy, C. C. (2020a). Focal frequency loss for generative models. arXiv preprint arXiv:2012.12821

Jiang, L., Li, R., Wu, W., Qian, C., Loy, C. C. (2020b) Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 2886–2895). IEEE.

Jiang, L., Guo, Z., Wu, W., Liu, Z., Liu, Z., Loy, C. C., Yang, S., Xiong, Y., Xia, W., Chen, B., & Zhuang, P. (2021a). Deeperforensics challenge 2020 on real-world face forgery detection: Methods and results. arXiv preprint arXiv:2102.09471

Jiang, Y., Chang, S., Wang, Z. (2021b). Transgan: Two pure transformers can make one strong gan, and that can scale up. In *35th conference on neural information processing systems*.

Jo, Y., & Park, J. (2019). Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *Proceedings of the IEEE international conference on computer vision* (pp. 1745–1753).

Juefei-Xu, F., & Savvides, M. (2015). Pokerface: Partial order keeping and energy repressing method for extreme face illumination normalization. In *Proceedings of the IEEE 7th international conference on biometrics: theory, applications, and systems (BTAS)* (pp. 1–8). IEEE.

Juefei-Xu, F., & Savvides, M. (2016). Fastfood dictionary learning for periocular-based full face hallucination. In *Proceedings of the IEEE 7th international conference on biometrics: theory, applications, and systems (BTAS)* (pp. 1–8). IEEE.

Juefei-Xu, F., Luu, K., & Savvides, M. (2015). Spartans: Single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios. *IEEE Transactions on Image Processing (TIP), 24*(12), 4780–4795.

Jung, S., & Keuper, M. (2020). Spectral distribution aware image generation. arXiv preprint arXiv:2012.03110

Karnewar, A., & Wang. O. (2020). Msg-gan: Multi-scale gradients for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7799–7808).

Karras, T., Aila, T., Laine, S., Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4401–4410).

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

Kawa, P., & Syga, P. (2020). A note on deepfake detection with low-resources. arXiv preprint arXiv:2006.05183

Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., & Brossard, E. (2016). The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4873–4882).

Khalid, H., & Woo, S. S. (2020). Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 656–657).

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2021). Transformers in vision: A survey. arXiv preprint arXiv:2101.01169

Kim, M., Tariq, S., & Woo, S. S. (2021). Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1001–1012).

Kingma, D. P., & Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems* (pp. 10215–10224).

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. arXiv preprint arXiv:1906.02691

Koopman, M., Rodriguez, A. M., & Geradts, Z. (2018). Detection of deepfake video manipulation. In *Conference: IMVIP*.

Korshunov, P., & Marcel, S. (2018). Deepfakes: A new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685

Kukanov, I., Karttunen, J., Sillanpää, H., & Hautamäki, V. (2020). Cost sensitive optimization of deepfake detector. In *2020 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC)* (pp. 1300–1303). IEEE.

Kumar, P., Vatsa, M., & Singh, R. (2020). Detecting face2face facial reenactment in videos. In *The IEEE winter conference on applications of computer vision* (pp. 2589–2597).

Kwon, P., You, J., Nam, G., Park, S., & Chae, G. (2021). Kodf: A large-scale korean deepfake detection dataset. arXiv preprint arXiv:2103.10094

Le, T. N., Nguyen, H. H., Yamagishi, J., & Echizen, I. (2021). Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. arXiv preprint arXiv:2107.14480

Li, H., Chen, H., Li, B., & Tan, S. (2018a). Can forensic detectors identify gan generated images? In *2018 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC)* (pp. 722–727). IEEE.

Li, H., Li, B., Tan, S., & Huang, J. (2020). Identification of deep network generated images using disparities in color components. *Signal Processing, 174*, 107616.

Li, J., Shen, T., Zhang, W., Ren, H., Zeng, D., & Mei, T. (2019a). Zooming into face forensics: A pixel-level analysis. arXiv preprint arXiv:1912.05790

Li, J., Li, Z., Cao, J., Song, X., & He, R. (2021a). Faceinpainter: High fidelity face adaptation to heterogeneous domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5089–5098).

Li, J., Xie, H., Li, J., Wang, Z., & Zhang, Y. (2021b). Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6458–6467).

Li, K., Zhang, T., & Malik, J. (2019b). Diverse image synthesis from semantic layouts via conditional imle. In *Proceedings of the IEEE international conference on computer vision* (pp. 4220–4229).

Li, L., Bao, J., Yang, H., Chen, D., & Wen, F. (2020b). Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5074–5083).

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020c). Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5001–5010).

Li, T., & Lin, L. (2019). Anonymousnet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*.

Li, X., Lang, Y., Chen, Y., Mao, X., He, Y., Wang, S., Xue, H., & Lu, Q. (2020d). Sharp multiple instance learning for deepfake video detection. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1864–1872).

Li, Y., & Lyu, S. (2019). Exposing deepfake videos by detecting face warping artifacts. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 1831–1839). IEEE.

Li, Y., Chang, M. C., & Lyu. S. (2018b). In ICTU oculi: Exposing AI created fake videos by detecting eye blinking. In *2018 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–7). IEEE.

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020e). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3207–3216).

de Lima, O., Franklin, S., Basu, S., Karwoski, B., & George, A. (2020). Deepfake detection using spatiotemporal convolutional networks. arXiv preprint arXiv:2006.14749

Lin, C. H., Chang, C. C., Chen, Y. S., Juan, D. C., Wei, W., & Chen, H. T. (2019). Coco-gan: Generation by parts via conditional coordinating. In *Proceedings of the IEEE international conference on computer vision* (pp. 4512–4521).

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755). Springer.

Liu, B., Zhu, Y., Song, K., & Elgammal, A. (2021). Self-supervised sketch-to-image synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*, 2073–2081.

Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., & Yu, N. (2021b). Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 772–781).

Liu, J., Zhang, W., Tang, Y., Tang. J., & Wu, G. (2020a). Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2359–2368).

Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., & Wen, S. (2019). Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3673–3682).

Liu, S., Lin, T., He, D., Li, F., Deng, R., Li, X., Ding, E., & Wang, H. (2021c). Paint transformer: Feed forward neural painting with stroke prediction. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6598–6607).

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (pp. 3730–3738).

Liu, Z., Qi, X., & Torr, P. H. (2020b). Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF* *conference on computer vision and pattern recognition* (pp. 8060–8069).

Lu, S. A. (2018). FaceSwap-GAN. https://github.com/shaoanlu/faceswap-GAN

Luo, Y., Zhang, Y., Yan, J., & Liu, W. (2021). Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16317–16326).

Lyu, S. (2020). Deepfake detection: Current challenges and next steps. In *2020 IEEE international conference on multimedia & expo workshops (ICMEW)* (pp. 1–6). IEEE.

Mansourifar, H., & Shi, W. (2020). One-shot gan generated fake face detection. arXiv preprint arXiv:2003.12244

Mao, X., Li, Q., Xie, H., Lau, R. Y., & Wang, Z. (2016). Multi-class generative adversarial networks with the l2 loss function. arXiv preprint arXiv:1611.04076 *5*, 1057–7149.

Marra, F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2018). Detection of gan-generated fake images over social networks. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)* (pp. 384–389). IEEE.

Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2019a). Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)* (pp. 506–511). IEEE.

Marra, F., Saltori, C., Boato, G., & Verdoliva, L. (2019b). Incremental learning for the detection and classification of gan-generated images. In *2019 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–6). IEEE.

Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2020). A full-image full-resolution end-to-end-trainable CNN framework for image forgery detection. *IEEE Access, 8*, 133488–133502.

Mas Montserrat, D., Hao, H., Yarlagadda, S.K., Baireddy, S., Shao, R., Horvath, J., Bartusiak, E., Yang, J., Guera, D., Zhu, F., & Delp, E. J. (2020). Deepfakes detection with automatic face weighting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 668–669).

Masi, I., Killekar, A., Mascarenhas, R. M., Gurudatt, S. P., & AbdAlmageed, W. (2020) Two-branch recurrent network for isolating deepfakes in videos. In *European conference on computer vision* (pp. 667–684). Springer.

Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE winter applications of computer vision workshops (WACVW)* (pp. 83–92). IEEE.

Maurer, U. M. (2000). Authentication theory and hypothesis testing. *IEEE Transactions on Information Theory, 46*(4), 1350–1356.

Maximov, M., Elezi, I., Leal-Taixé, L. (2020). Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5447–5456).

McCloskey, S., & Albright, M. (2019). Detecting gan-generated imagery using saturation cues. In *2019 IEEE international conference on image processing (ICIP)* (pp. 4584–4588). IEEE.

Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T. S., & Shi, H. (2020). Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5690–5699).

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR), 54*(1), 1–41.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784

MIT, T. R. (2020). Deepfake Putin is here to warn Americans about their self-inflicted doom. https://www.technologyreview.com/2020/09/29/1009098/ai-deepfake-putin-kim-jong-un-us-election/

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). Emotions don't lie: A deepfake detection method using audio-visual affective cues. arXiv preprint arXiv:2003.06711

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957

Mo, H., Chen, B., & Luo, W. (2018). Fake faces identification via convolutional neural network. In *Proceedings of the 6th ACM workshop on information hiding and multimedia security* (pp. 43–47).

Nataraj, L., Mohammed, T. M., Manjunath, B., Chandrasekaran, S., Flenner, A., Bappy, J. H., & Roy-Chowdhury, A. K. (2019). Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging, 5*, 532–1.

Natsume, R., Yatagawa, T., & Morishima, S. (2018) Rsgan: Face swapping and editing using face and hair representation in latent spaces. In *ACM SIGGRAPH 2018 posters* (pp. 1–2).

Neekhara, P., Dolhansky, B., Bitton, J., & Ferrer, C. C. (2021). Adversarial threats to deepfake detection: A practical perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 923–932).

Neves, J. C., Tolosana, R., Vera-Rodriguez, R., Lopes, V., Proença, H., & Fierrez, J. (2020). Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 1038–1048.

Nguyen, H. H., Fang, F., Yamagishi, J., & Echizen, I. (2019a). Multi-task learning for detecting and segmenting manipulated facial images and videos. In *Proceedings of the 10th international conference on biometrics theory, applications and systems (BTAS)*.

Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019–2019 IEEE international conference on acoustics* (pp. 2307–2311). IEEE: Speech and Signal Processing (ICASSP).

Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019c). Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467

Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019d). Deep learning for deepfakes creation and detection. arXiv preprint arXiv:1909.11573

Nhu, T., Na, I., & Kim, S. (2018). Forensics face detection from gans using convolutional neural network. In *Proceeding of 2018 international symposium on information technology convergence (ISITC 2018)*.

Nirkin, Y., Keller, Y., & Hassner, T. (2019). Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7184–7193).

Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2020). Deepfake detection based on the discrepancy between the face and its context. arXiv preprint arXiv:2008.12262

Noroozi, M. (2020). Self-labeled conditional gans. arXiv preprint arXiv:2012.02162

NPR. (2020). Where are the deepfakes in this presidential election? https://www.npr.org/2020/10/01/918223033/where-are-the-deepfakes-in-this-presidential-election

OpenAI. (2021). DALL-E: Creating images from text. https://openai.com/blog/dall-e/

Osakabe, T., Tanaka, M., Kinoshita, Y., & Kiya, H. (2021). Cyclegan without checkerboard artifacts for counter-forensics of fake-image detection. In *International workshop on advanced imaging technology (IWAIT)* (Vol. 11766, p. 1176609). International Society for Optics and Photonics.

OValery. (2017). Swap-face. https://github.com/OValery16/swap-face

Pang, T., Du, C., Dong, Y., & Zhu, J. (2018). Towards robust detection of adversarial examples. In *Proceedings of the 32nd international conference on neural information processing systems* (pp. 4584–4594).

Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2337–2346).

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Proceedings of the British machine vision conference (BMVC)* (pp 41.1–41.12).

Perarnau, G., Van De Weijer, J., Raducanu, B., & Álvarez, J. M. (2016). Invertible conditional gans for image editing. arXiv preprint arXiv:1611.06355

Petrov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Jiang, J., Rp, L., Zhang, S., Wu, P., & Zhang, W. (2020). Deepfacelab: A simple, flexible and extensible face swapping framework. arXiv preprint arXiv:2005.05535

Pinscreen. (2021). AI avatars, virtual assistants, and deepfakes: A real-time look. Retrieved August 1, 2021, from https://blog.siggraph.org/2021/01/ai-avatars-virtual-assistants-and-deepfakes-a-real-time-look.html/, (online).

Pinscreen. (2021). Pinscreen AI-driven virtual avatars. http://www.pinscreen.com/

Pishori, A., Rollins, B., van Houten, N., Chatwani, N., & Uraimov, O. (2020). Detecting deepfake videos: An analysis of three techniques. arXiv preprint arXiv:2007.08517

Pu, J., Mangaokar, N., Wang, B., Reddy, C. K., & Viswanath, B. (2020). Noisescope: Detecting deepfake images in a blind setting. In *Annual computer security applications conference* (pp. 913–927).

Pu, J., Mangaokar, N., Kelly, L., Bhattacharya, P., Sundaram, K., Javed, M., Wang, B., & Viswanath, B. (2021). Deepfake videos in the wild: Analysis and detection. *Proceedings of the Web Conference, 2021*, 981–992.

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., & Moreno-Noguer, F. (2018). Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 818–833).

Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., Liu, Y., & Zhao, J. (2020). Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 4318–4327).

Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision* (pp. 86–103). Springer.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., & Krueger, G. (2021). Learning transferable visual models from natural language supervision. *Image, 2*, T2.

Razavi, A., van den Oord, A., & Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems* (pp. 14866–14876).

Reface. (2021). Reface, Apple App Store. https://apps.apple.com/us/app/reface-face-swap-videos/id1488782587

Rezende, D., & Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning* (pp. 1530–1538). PMLR.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2018). Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE international conference on computer vision* (pp. 1–11).

Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P. (2019). Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI), 3*(1), 80–87.

Sambhu, N., & Canavan, S. (2020). Detecting forged facial videos using convolutional neural network. arXiv preprint arXiv:2005.08344

Schwarcz, S., & Chellappa, R. (2021). Finding facial forgery artifacts with parts-based detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 933–942).

Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., & Samaras, D. (2017). Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5541–5550).

Sohrawardi, S. J., Chintha, A., Thai, B., Seng, S., Hickerson, A., Ptucha, R., & Wright, M. (2019). Poster: Towards robust open-world detection of deepfakes. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security* (pp. 2613–2615).

Songsri-in, K., Zafeiriou, S. (2019). Complement face forensic detection and localization with faciallandmarks. arXiv preprint arXiv:1910.05455

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., & Wang, J. (2019). High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514

Sun, L., Juefei-Xu, F., Huang, Y., Guo. Q., Zhu. J., Feng, J., Liu. Y., & Pu, G. (2022). Ala: Adversarial lightness attack via naturalness-aware regularizations. arXiv preprint arXiv:2201.06070

Sun, Q., Tewari, A., Xu, W., Fritz, M., Theobalt, C., & Schiele, B. (2018). A hybrid model for identity obfuscation by face replacement. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 553–569).

Sun, X., Wu, B., & Chen, W. (2020a). Identifying invariant texture violation for robust deepfake detection. arXiv preprint arXiv:2012.10580

Sun, Y., Wang, X., & Tang, X. (2013). Hybrid deep learning for face verification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1489–1496).

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M. (2020b). Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning* (pp. 9229–9248). PMLR.

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics (TOG), 36*(4), 1–13.

Synthesia. (2021). Synthesia software. https://www.synthesia.io/

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114). PMLR.

Tarasiou, M., & Zafeiriou, S. (2020). Extracting deep local features to detect manipulated images of human faces. In *2020 IEEE international conference on image processing (ICIP)* (pp. 1821–1825). IEEE.

Tariq, S., Lee, S., Kim, H., Shin, Y., & Woo, S. S. (2018). Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd international workshop on multimedia privacy and security* (pp. 81–87).

Tariq, S., Lee, S., & Woo, S. S. (2020). A convolutional lstm based residual network for deepfake video detection. arXiv preprint arXiv:2009.07480

Texas. (2019). Texas Senate Bill No. 751. https://capitol.texas.gov/tlodocs/86R/analysis/html/SB00751F.htm

The Verge. (2019a). China makes it a criminal offense to publish deepfakes or fake news without disclosure. https://www.theverge.com/2019/11/29/20988363/

The Verge. (2019b). Virginia's 'revenge porn' laws now officially cover deepfakes. https://www.theverge.com/2019/7/1/20677800/

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2387–2395).

Thies, J., Zollhöfer, M., & Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG), 38*(4), 1–12.

Tian, B., Guo, Q., Juefei-Xu, F., Le Chan, W., Cheng, Y., Li, X., Xie, X., & Qin, S. (2021a). Bias field poses a threat to dnn-based x-ray recognition. In *2021 IEEE international conference on multimedia and expo (ICME)* (pp. 1–6). IEEE.

Tian, B., Juefei-Xu, F., Guo, Q., Xie, X., Li, X., & Liu, Y. (2021b). AVA: Adversarial vignetting attack against visual recognition. In *Proceedings of the international joint conference on artificial intelligence (IJCAI)*.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion, 64*, 131–148.

Trinh, L., Tsang, M., Rambhatla, S., & Liu, Y. (2021). Interpretable and trustworthy deepfake detection via dynamic prototypes. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1973–1983).

Tripathy, S., Kannala, J., & Rahtu, E. (2020). Icface: Interpretable and controllable face reenactment using gans. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3385–3394).

Tripathy, S., Kannala, J., & Rahtu, E. (2021). Facegan: Facial attribute controllable reenactment gan. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1329–1338).

Tursman, E., George, M., Kamara, S., & Tompkin, J. (2020). Towards untrusted social video verification to combat deepfakes via face geometry consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 654–655).

Twitter Blog. (2019). Help us shape our approach to synthetic and manipulated media. https://blog.twitter.com/en_us/topics/company/2019/synthetic_manipulated_media_policy_feedback.html

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2018). Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9446–9454).

Van Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel recurrent neural networks. In: International Conference on Machine Learning, PMLR, pp 1747–1756

Van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., & Kavukcuoglu, K. (2016). Conditional image generation with pixelcnn decoders. arXiv preprint arXiv:1606.05328

Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 910–932.

Viazovetskyi, Y., Ivashkin, V., & Kashin, E. (2020). Stylegan2 distillation for feed-forward image manipulation. In *European conference on computer vision* (pp. 170–186). Springer.

Wang, C., & Deng, W. (2021). Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14923–14932).

Wang, G., Zhou, J., & Wu, Y. (2020a). Exposing deep-faked videos by anomalous co-motion pattern detection. arXiv preprint arXiv:2008.04848

Wang, R., Juefei-Xu, F., Guo, Q., Huang, Y., Ma, L., Liu, Y., & Wang, L. (2020b). Deeptag: Robust image tagging for deepfake provenance. arXiv preprint arXiv:2009.09869

Wang, R., Juefei-Xu, F., Huang, Y., Guo, Q., Xie, X., Ma, L., & Liu, Y. (2020c). Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1207–1216).

Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., & Liu, Y. (2020d). Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. In *International joint conference on artificial intelligence (IJCAI)*.

Wang, S. Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020e). Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (Vol. 7).

Wang, T. C., Mallya, A., & Liu, M. Y. (2021). One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10039–10049).

Wang, X., Yao, T., Ding, S., & Ma, L. (2020f). Face manipulation detection via auxiliary supervision. In *International conference on neural information processing* (pp. 313–324). Springer.

Wikipedia. (2021a). Elo rating system. Retrieved December 30, 2020, from https://en.wikipedia.org/wiki/Elo_rating_system (online).

Wikipedia. (2021b). Sankey diagram. Retrieved December 17, 2020, from https://en.wikipedia.org/wiki/Sankey_diagram (online)

Woods, W., Chen, J., & Teuscher, C. (2019). Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nature Machine Intelligence, 1*(11), 508–516.

Wu, X., Xie, Z., Gao, Y., & Xiao, Y. (2020). Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features. In *ICASSP 2020–2020 IEEE international conference on acoustics* (pp. 2952–2956). IEEE: Speech and Signal Processing (ICASSP).

Wu, Y., AbdAlmageed, W., & Natarajan, P. (2019). Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9543–9552).

Xia, W., Yang, Y., Xue, J. H., & Wu, B. (2021). Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2256–2265).

Xuan, X., Peng, B., Wang, W., & Dong, J. (2019). On the generalization of gan image forensics. In *Chinese conference on biometric recognition* (pp. 134–141). Springer.

Yang, X., Li, Y., & Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. In *ICASSP 2019–2019 IEEE international conference on acoustics* (pp. 8261–8265). IEEE: Speech and Signal Processing (ICASSP).

Yang, X., Li, Y., Qi, H., Lyu, S. (2019b). Exposing gan-synthesized faces using landmark locations. In *Proceedings of the ACM workshop on information hiding and multimedia security* (pp. 113–118).

Yao, Y., Ren, J., Xie, X., Liu, W., Liu, Y. J., & Wang, J. (2019). Attention-aware multi-stroke style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1467–1475).

Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). Learning face representation from scratch. arXiv preprint arXiv:1411.7923

Yu, C. M., Chang, C. T., & Ti, Y. W. (2019a). Detecting deepfake-forged contents with separable convolutional neural network and image segmentation. arXiv preprint arXiv:1912.12184

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

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5505–5514).

Yu, N., Davis, L. S., & Fritz, M. (2019b). Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of* *the IEEE international conference on computer vision* (pp. 7556–7566).

Yu, N., Skripniuk, V., Chen, D., Davis, L., & Fritz, M. (2020a). Responsible disclosure of generative models using scalable fingerprinting. arXiv preprint arXiv:2012.08726

Yu, Y., Ni, R., & Zhao, Y. (2020b). Mining generalized features for detecting ai-manipulated fake faces. arXiv preprint arXiv:2010.14129

Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., & Liu, C. (2021). Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432

Zao. (2021). Zao, Apple App Store. https://apps.apple.com/cn/app/zao/id1465199127

Zhai, L., Juefei-Xu, F., Guo, Q., Xie, X., Ma, L., Feng, W., Qin, S., & Liu, Y. (2020). It's raining cats or dogs? adversarial rain attack on dnn perception. arXiv preprint arXiv:2009.09205

Zhai, L., Juefei-Xu, F., Guo, Q., Xie, X., Ma, L., Feng, W., Qin, S., & Liu, Y. (2022). Adversarial rain attack and defensive deraining for dnn perception. arXiv preprint arXiv:2009.09205

Zhang, C., Zhao, Y., Huang, Y., Zeng, M., Ni, S., Budagavi, M., & Guo, X. (2021a). Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3867–3876).

Zhang, G., Kan, M., Shan, S., & Chen, X. (2018). Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 417–432).

Zhang, W., Ji, X., Chen, K., Ding, Y., & Fan, C. (2021b). Learning a facial expression embedding disentangled from identity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6759–6768).

Zhang, X., Karaman, S., & Chang, S. F. (2019). Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–6). IEEE.

Zhang, Y., Zheng, L., & Thing, V. L. (2017). Automated face swapping and its detection. In *2017 IEEE 2nd international conference on signal and image processing (ICSIP)* (pp. 15–19). IEEE.

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2185–2194).

Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., & Xia, W. (2020). Learning to recognize patch-wise consistency for deepfake detection. arXiv preprint arXiv:2012.09311

Zheng, Z., & Hong, P. (2018). Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks. In *Proceedings of the 32nd international conference on neural information processing systems* (pp. 7924–7933).

Zhou, H., Sun, Y., Wu, W., Loy, C. C., Wang, X., & Liu, Z. (2021a). Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4176–4186).

Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2017). Two-stream neural networks for tampered face detection. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 1831–1839). IEEE.

Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2018). Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1053–1061).

Zhou, T., Wang, W., Liang, Z., & Shen, J. (2021b). Face forensics in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5778–5788).

Zhu, H., Fu, C., Wu, Q., Wu, W., Qian, C., & He, R. (2020). Aot: Appearance optimal transport based identity swapping for forgery detection. *Advances in Neural Information Processing Systems, 33*, 21699–21712.

Zhu, J., Guo, Q., Juefei-Xu, F., Huang, Y., Liu, Y., & Pu, G. (2022). Masked faces with faced masks. arXiv preprint arXiv:2201.06427

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).

Zhu, X., Wang, H., Fei, H., Lei, Z., & Li, S. Z. (2021a). Face forgery detection by 3d decomposition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2929–2939).

Zhu, Y., Li, Q., Wang, J., Xu, C. Z., & Sun, Z. (2021b). One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4834–4844).

Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y. G. (2020). Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2382–2390).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.