SoK: Facial Deepfake Detectors

Binh M. Le*, Jiwon Kim*, Shahroz Tariq†, Kristen Moore†, Alsharif Abuadbba†, and Simon S. Woo*
*Sungkyunkwan University, South Korea, {bmle,merwl0,swoo}@g.skku.edu
†CSIRO's Data61, Australia, {shahroz.tariq, kristen.moore, sharif.abuadbba}@data61.csiro.au

***Abstract*—Deepfakes have rapidly emerged as a profound and serious threat to society, primarily due to their ease of creation and dissemination. This situation has triggered an accelerated development of deepfake detection technologies. However, many existing detectors rely heavily on lab-generated datasets for validation, which may not effectively prepare them for novel, emerging, and real-world deepfake techniques. In this paper, we conduct an extensive and comprehensive review and analysis of the latest state-of-the-art deepfake detectors, evaluating them against several critical criteria. These criteria facilitate the categorization of these detectors into 4 high-level groups and 13 fine-grained sub-groups, all aligned with a unified standard conceptual framework. This classification and framework offer deep and practical insights into the factors that affect detector efficacy. We assess the generalizability of 16 leading detectors across various standard attack scenarios, including black-box, white-box, and gray-box settings. Our systematized analysis and experimentation lay the groundwork for a deeper understanding of deepfake detectors and their generalizability, paving the way for future research focused on creating detectors adept at countering various attack scenarios. Additionally, this work offers insights for developing more proactive defenses against deepfakes.**

1. Introduction

Deep learning-based approaches to manipulating human faces, notably deepfakes, have recently gained significant attention [1]–[5] due to their potential misuse in malicious endeavors, including pornography, the propagation of fake news, and privacy violation. The advanced realism of deepfakes, largely due to advances in Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), presents a significant challenge to humans and classifiers attempting to differentiate them from authentic content. Furthermore, the accessibility of open source, user-friendly deepfake generation tools [6]–[8] has facilitated an unprecedented ease of disseminating manipulated media. Apart from their entertainment value, deepfakes have profound implications that pose threats to various social aspects [9], [10].

From a cybersecurity standpoint and amidst the rapid advancements in deepfake technology, Facial Liveness Verification (FLV) is currently encountering unprecedented challenges [11]. In fact, FLV, a widely used tool for identity authentication across various security-sensitive domains, is often offered as a Platform-as-a-Service (PaaS) by major cloud vendors. Consequently, any vulnerabilities present in the APIs offered by FLV PaaS providers can be inherited by downstream applications (apps) that use them, posing a threat to millions of end-users. Notably, broken authentication in APIs, ranked second in the "OWASP API Security Top 10" [12], was highlighted when tax scammers used deepfake techniques to infiltrate a government FLV system, resulting in fraudulent tax invoices worth $76.2 million [11].

To mitigate the potential harm of deepfake, a myriad of researchers are investing efforts to improve deepfake detection methods and fortify existing detection systems [1], [2], [13]. These methods adopt a variety of approaches including spatial analysis [1], [14], [15], frequency analysis [16]–[18], temporal analysis [4], and the identification of underlying artifacts or fingerprints [19]. Simultaneously, researchers are striving to build detectors robust enough to withstand various forms of corruption, such as noise [20], [21], compression [1], [18], and, most critically, to identify previously unseen deepfakes in the wild [22]–[24]. Indeed, the limitations of available training datasets make deepfake detectors especially prone to performance degradation when encountering unseen deepfakes, potentially resulting in performance worse than random guesses [22], [25]. Therefore, improving the detector's robustness against novel attacks is paramount.

While recent studies claimed generalizability of their model against various types of deepfakes [21], [24], they mainly focus on standard academic datasets [26], [27]. Meanwhile, there is limited understanding of deepfake detectors, creation tools, and datasets, particularly regarding their characteristics, functionalities, and performance across different scenarios. This results in a considerable gap between reported efficacy and actual performance, which can only be properly assessed through extensive and systematic tests against popular deepfake tools and deepfakes found in the wild. We find that recent attempts to systematically categorize deepfake generation and detection methods fall short in providing comprehensive evaluations [28], [29] or a detailed classification of deepfake creation tools and advanced detectors [30].

To the best of our knowledge, our work represents the first attempt to systematically scrutinize the comprehensiveness of existing studies on deepfake detection, seeking to answer the following research questions:

RESEARCH QUESTIONS

RQ1: What are the influential factors in facial deepfake detection?
RQ2: How do leading detectors perform when evaluated for generalizability?
RQ3: How do the identified influential factors impact the performance of the detectors when evaluated for generalizability?

To answer **RQ1**, we explore well-known detectors that were recently published in top-tier venues and analyze them according to key determining factors. Our contributions to addressing this research question are as follows:

(1) Comprehensive Literature Review. We systematically investigated the existing literature from 2019 to 2023 and curated a selection of 51 deepfake detectors from top-tier (CORE A*) venues in AI and security areas.

(2) Factor Identification. Within this curated selection, we identified and isolated 25 crucial factors commonly used to build deepfake detectors, considering aspects such as deepfake types, each artifact type, input data representation methods, network architectures, training style, and more.

(3) Framework Development. We developed a comprehensive conceptual framework with clear and reflective indicators, mapping each detector to the specific factors that we identified above. In this framework, we provide valuable insights into the varying degrees of importance of different factors and offer potential explanations for such disparities.

In another pioneering effort to resolve **R2** and **R3**, we evaluate the generalizability of leading detectors through a security lens by evaluating them in *black-box*, *gray-box*, and *white-box* settings. This approach aims to overcome the limitations of previous surveys that primarily focused on gray-box settings and rarely explored black-box settings.

While **R2** could potentially be addressed through validation in gray-box scenarios, the lack of transparency in gray-box settings in previous studies has hindered a thorough examination of **R3**. Therefore, we designed and conducted white-box experiments, providing a unique exploration of detectors under the conditions where the deepfake generator is known and both the source and target are controlled. This unique approach enables us to comprehensively evaluate the influential factors that have been rigorously identified from **R1**, where such an evaluation would be challenging to conduct in black-box or gray-box settings. Our contributions include the following key steps:

(1) Stringent Selection Criteria. We set rigorous selection criteria to identify 16 (out of the curated selection of 51) detectors renowned for their accessibility, ensuring both reuse and reproducibility for future evaluations.

(2) Gray-Box Evaluation. To assess the generalizability of these detectors, we first subjected them to gray-box dataset settings, employing diverse benchmark datasets where we only know partial information and do not have full control over the victim/driver.

(3) White-Box Evaluation. We evaluate the selected 16 detectors against the controlled and systematic settings of a white-box dataset. Namely, we created a new white-box evaluation dataset where we exercised control over the deepfake generator, source, and destination videos.

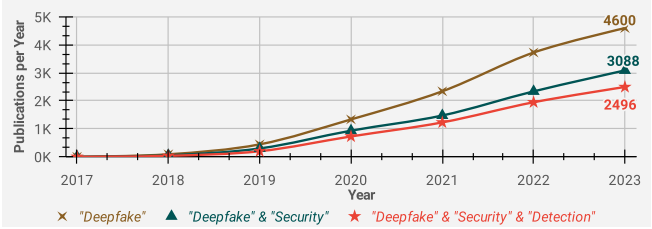


Figure 1. Publications (per year) for deepfake-related keywords. *"Security"* and *"Detection"* are common topics in the most deepfake publications.

(4) Black-Box Evaluation. To assess the generalizability of these detectors, we also subjected them to black-box dataset settings, employing the diversity and heterogeneity of in-the-wild deepfake. For that purpose, we used the recent dataset from [31], containing real-world deepfakes of 2,000 samples from 4 different online platforms to ensure comprehensive black-box testing.

Prior Deepfake Surveys *vs.* This SoK. As illustrated in Fig. 1, the number of publications related to three major keywords: Deepfake, Security, and Detection, has exponentially increased over the years. Consequently, there is a need to consolidate and systematize existing literature into a single study. While numerous studies have explored deepfake detection and generation, *a critical gap exists in research that systematically summarizes deepfake detectors under various influential factors and assesses their impact on well-known detectors using diverse protocol settings.* In Table 1, we categorize and summarize prior survey studies across various criteria. The initial work by Verdoliva [32] focused primarily on deepfake detection but was too brief to provide in-depth insights into deepfake detectors and lacked detailed information on up-to-date detection methods. Later, more thorough studies [28], [29], [33], [34] provided comprehensive summaries covering various aspects of deepfake applications, threats, generation, and detection. Recent evaluation studies by Yan et al. [30] and Khan & Nguyen [35] have garnered more attention, yet they still lacked diversity in evaluation protocols, only considering gray-box settings.

To the best of our knowledge, we are the first to present a thorough overview of the varying and dynamic deepfake detection landscape and to comprehensively evaluate it. Our paper distinguishes itself from previous surveys with the following unique features:

(i) Timeliness. We collect and analyze the latest SoTA deepfake detectors.

(ii) Detailedness. This is the first paper to introduce an end-to-end conceptual framework that can generalize for most recent deepfake detectors, establishing a new standard for comparison and analysis.

(iii) In-depth Evaluation. Beyond systematizing detectors, we are the first to comprehensively evaluate the most recent SoTA detectors under three fundamental security settings: gray-box, white-box, and black-box. This enables the characterization of influential factors in detectors, and supports future studies.

Table 1. COMPARISON OF CONTRIBUTION BETWEEN OUR SURVEY AND RELEVANT SURVEYS. † INDICATES THAT THE STUDIES DO NOT CONDUCT EXPERIMENTS BY THEMSELVES BUT SOLELY REPORT NUMBERS.

Prior surveys	Year	Years Covered (*Detectors*)	Conceptual Framework	Detectors Analysis	Their Own Evaluation	Evaluation Dataset		(Cross) Evaluation Strategy			Notes
						Same	*Cross*	*Gray-box*	*White-box*	*Black-box*	
Verdoliva [32]	2020	2005 – 2020	✗	Brief	✗	✓†	✗	✗	✗	✗	Summarization
Tolosana et al. [33]	2020	2018 – 2020	✗	Thorough	✗	✓†	✗	✗	✗	✗	Summarization
Mirsky and Lee [28]	2021	2017 – 2020	✗	Thorough	✗	✓†	✗	✗	✗	✗	Summarization
Juefei-Xu et al. [34]	2022	2016 – 2021	✗	Thorough	✗	✓†	✗	✗	✗	✗	Summarization
Rana et al. [36]	2022	2018 – 2020	✗	Brief	✗	✗	✗	✗	✗	✗	Summarization of Detectors
Nguyen et al. [37]	2022	2018 – 2021	✗	Brief	✗	✗	✗	✗	✗	✗	Summarization
Malik et al. [38]	2022	2018 – 2021	✗	Brief	✗	✗	✗	✗	✗	✗	Summarization
Seow et al. [29]	2022	2018 – 2021	✗	Thorough	✗	✗	✗	✗	✗	✗	Summarization
Naitali et al. [39]	2023	2022 – 2023	✗	Brief	✗	✗	✗	✗	✗	✗	Summarization
Yan et al. [30]	2023	2018 – 2023	✗	Brief	✓ (15)	✓	✓	✓	✗	✗	Evaluation of Published Detectors
Khan & Nguyen [35]	2023	-	✗	Brief	✓ (8)	✓	✓	✓	✗	✗	Evaluation of General NN Models
Ours	2023	2019 – 2023	✓	Thorough	✓ (16)	✓	✓	✓	✓	✓	Summarization & Evaluation of Published Detectors

Table 2. STRUCTURE OF THIS SYSTEMATIZATION OF KNOWLEDGE.

SoK	Description	Section
Introduce	Background on deepfake detection and generation	2
Introduce	Methodology for answering the research questions	3
Systematize	Taxonomy of deepfake detection and conceptual framework	4
Evaluate	Detector selection criteria for evaluation	5.1
Evaluate	Evaluation strategies and datasets for evaluations	5.2, 5.3
Evaluate	Pre-training sources and metrics for evaluations	5.4
Evaluate	Evaluation of results and analysis of influential factors	6
Discuss	Limitations, current challenges, and future directions	7
Discuss	Research gaps and concluding thoughts	7

Roadmap. This paper is structured as follows, as detailed in Table 2: Section 2 revisits the foundational concepts of deepfake generation and detection, laying the groundwork for subsequent discussions. In Sec. 3, we outline our methodology for addressing the primary research questions. Section 4 provides preliminary observations on the 51 detectors analyzed in this study, and introduces a conceptual framework for categorizing these detectors from various critical perspectives. The standard deepfake detectors, datasets, tools, and are systematized and examined in Sec. 5, setting the stage for the validation process described in Sec. 6. We discuss the limitations and challenges of our study and conclude with final thoughts in Sec. 7. Our evaluation code is on this link: https://anonymous.4open.science/r/DeepfakeSoK. Once the paper is accepted, we will organize the repository and make it public on GitHub.

2. Background and Related Work

Deepfake Generation. Since Ian Goodfellow *et al.* introduced Generative Adversarial Networks (GANs) [40], significant advances have been made in realistic image synthesis, particularly for human faces [41], [42]. GANs employ a generator ($\mathcal{G}$) to create images and a discriminator ($\mathcal{D}$) to distinguish between real and generated data. Additionally, AutoEncoders (AE), pioneered by Yann LeCun *et al.* [43] and revitalized by variational auto-encoders (VAEs) [44], compress data into compact forms, utilized in deepfake technology to alter facial features like expressions and styles.

With the increasing prevalence of deepfakes, or AI-synthesized videos, in the digital world [9], [45], and the academic response through initiatives like the Deepfake Detection Challenge (DFDC) [46], FaceForensics++ (FF++) [26], Celebrity Deepfake (CelebDF) [27] and Audio-Video Deepfake (FakeAVCeleb) [47], [48], they can be broadly segmented into three types: face swaps (or replacements), reenactments, and synthesis. For the reader's convenience, throughout this paper, the terms reference (source or driver) and target (destination or victim) identities are $\mathcal{R}$ and $\mathcal{T}$, respectively. $\mathcal{V}_{\mathcal{R}}$ signifies the reference video (perhaps sourced from the Internet), while $\mathcal{V}_{\mathcal{T}}$ refers to images or videos of the targeted individual. The deepfakes made from $\mathcal{V}_{\mathcal{R}}$ and $\mathcal{V}_{\mathcal{T}}$ are symbolized by $\mathcal{V}_{\mathcal{D}}$.

Faceswap. Faceswap methods, such as FaceSwap [7], DeepFakes [49], Faceshifter [50], and FSGAN [51], merge facial features from a target face ($\mathcal{V}_{\mathcal{T}}$) into a recipient video ($\mathcal{V}_{\mathcal{R}}$), creating a new video ($\mathcal{V}_{\mathcal{D}}$) where the target's face replaces the recipient's, while maintaining the original body and background. A notable example is superimposing a celebrity's face, like Scarlett Johansson's ($\mathcal{T}$), onto another person in a video ($\mathcal{V}_{\mathcal{R}}$) [52], achieved using tools like DeepFaceLab [6], Dfaker [53], and SimSwap [54].

Reenactment. The reenactment process combines $\mathcal{V}_{\mathcal{T}}$'s facial features with $\mathcal{V}_{\mathcal{R}}$'s expressions and movements to create $\mathcal{V}_{\mathcal{D}}$, using techniques like Talking Head (TH) [55], First-Order Motion Model (FOM) [8], Face2Face [56], and Neural Textures [57]. This method has animated public figures, such as in altered speeches of Donald Trump [58] and Richard Nixon [59].

Synthesis. Synthesis deepfakes vary in method, with Diffusion or GAN recently surpassing facial blends in popularity. Focused on image synthesis ($\mathcal{I}$), these techniques blend identities, such as $\mathcal{R}^1$ and $\mathcal{R}^2$, to create a new image $\mathcal{I}_{\mathcal{D}}$. For instance, Diffusion can blend Donald Trump ($\mathcal{I}^1_{\mathcal{R}}$) and Joe Biden ($\mathcal{I}^2_{\mathcal{R}}$) to create a synthesized identity [60].

Deepfake Detection. Image forgery detection has been extensively researched, as illustrated in the comprehensive survey [61]. In this work, we focus on methods specifically designed to identify deepfakes involving human faces. The objective of deepfake detection is to isolate and differentiate between representations of counterfeit images or videos and their genuine counterparts. Recently proposed methods generally adopt either supervised [15], [23], [26], [62]–[69] or self-supervised approaches [24], [70]. A significant advantage of self-supervised methods is their ability to leverage the virtually unlimited facial datasets freely available online,thereby potentially reducing the model's bias about

race or environment. However, they often require hypothesizing the artifact patterns found in synthesized faces. Conversely, supervised approaches typically employ deep learning models to discern the distinctions between real and fake images, overseeing the model's performance through predefined metrics.

Moreover, different approaches have their own methods of handling input, which can vary significantly. Some utilize spatial input through color images [26], [71], [72], while others employ frequency inputs, leveraging techniques such as DCT or Fourier transform [16], [73]. Yet another approach involves spatiotemporal input through sequences of frames [4], [15], [74]. Some methods focus specifically on particular artifacts, such as mouth movements [21], or the gradient artifacts present in synthesized images [19]. Furthermore, the field has seen a wide variety of detector categories arising from diverse training schemes and network architectures, including knowledge distillation frameworks [18], Siamese networks [75], [76], transformers [70], [77], and 3D ResNet configurations [2], [4], [78]. *Nevertheless, a systematic approach is notably absent for aggregating, comparing, and contrasting the myriad of frontier deepfake detection methods introduced in recent years and subjecting them to comprehensive evaluation using unified criteria.*

3. Methodology

This section details our methodology for obtaining answers to our main research questions.

Methodology for RQ1. To address the first research question, we conducted a systematization of the existing literature on deepfake detectors. To do this, we followed a stringent process with the objective to identify the key influential factors in facial deepfake detection. The process began with the meticulous collection of research papers related to deepfake detection. We describe the entire process, encompassing the inclusion and exclusion criteria for selecting these papers, in Sec. 4.1. This process produced 51 relevant papers. Subsequently, each of these 51 articles were systematically reviewed, considering factors such as dataset utilization, methodology focus, pre-processing methods, model architecture, and model evaluation criteria (see Sec. 4.2). After decomposing the methodologies of individual deepfake detectors, we collated their information to construct a conceptual framework providing a holistic view of the detection process. This conceptual framework facilitates the systematization of the existing literature, mapping each method to its place in the conceptual framework. We further categorize the detectors using this conceptual framework based on the focus of their methodology into 4 high-level groups and 13 fine-grained sub-groups, providing more clarity to our systematization. The complete process is expounded on in Sec. 4. Finally, we present the influential factors derived from this exercise in Sec. 6.2.

Methodology for RQ2. To address the second research question, we systematically evaluated the performance of deepfake detectors on various deepfake datasets to ensure a rigorous and equitable comparison. We started this process by meticulously selecting a subset of detectors from those previously identified in *RQ1*, employing stringent inclusion and exclusion criteria, as detailed in Sec. 5.1. This careful curation yielded 16 detectors, forming the basis for our subsequent generalization experiments. To facilitate this evaluation, the subsequent step involved the selection of appropriate evaluation strategies and datasets for experimentation. We selected, collected, and generated deepfake datasets encompassing three distinct evaluation settings: gray-box, white-box, and black-box (see Sec. 5.2 and 5.3). Finally, the results of our detector generalizability experiments are presented and analyzed in Sec. 6.1.

Methodology for RQ3. To answer the third research question, in Sec. 6.2, we take the generalizability results from gray-box, white-box, and black-box settings and analyze them using the lens of the conceptual framework and the influential factors identified in *RQ1* in Sec 6.1. Based on this analysis, we provide our insights and recommendations in Sec. 6.2 as well as discuss the gaps in the literature and provide directions for future research in Sec. 7

4. Systematization of Deepfake Detectors

Overview. This section provides detailed information on our systemized selection criteria for the deepfake detectors that we used in this study. Moreover, we analyze the selected detectors to obtain insights to systematically identify and analyze all the influential factors of deepfake detectors, elucidating their impact on accurate deepfake video detection in different scenarios. This section is dedicated to addressing ***RQ1: What are the influential factors in facial deepfake detection?*** To answer it, we develop a novel conceptual framework based on our extensive analysis of selected detectors. It provides a high-level perspective and a visual representation of key concepts and their interrelationships.

4.1. Paper Selection Criteria

First, we describe our paper collection process, including the inclusion and exclusion criteria.

Paper Collection Process. We focused on recent developments in deepfake detection in the last four years from 2019 to 2023, a period marked by significant growth in the field following the introduction of the FaceForensics++ benchmark [26]. Utilizing the Google Scholar search query `"deepfake detection"` for this timeframe period, we identified 4,220 relevant publications.

Inclusion and Exclusion Criteria. We exclude papers that are not specifically related to deepfake detection and that do not propose a deepfake detector. Furthermore, to ensure their credibility, we exclude papers that have not undergone a rigorous peer review process. We achieve this by selecting articles published in CORE A* venues[1]. This process yielded 51 relevant papers.

1. Note: We make exceptions to this criterion in certain cases, retaining some papers from lower-tier venues (i.e., CORE A and B venues) due to their exceptional popularity, as evidenced by citation scores or superior performance on widely recognized deepfake benchmarking datasets.

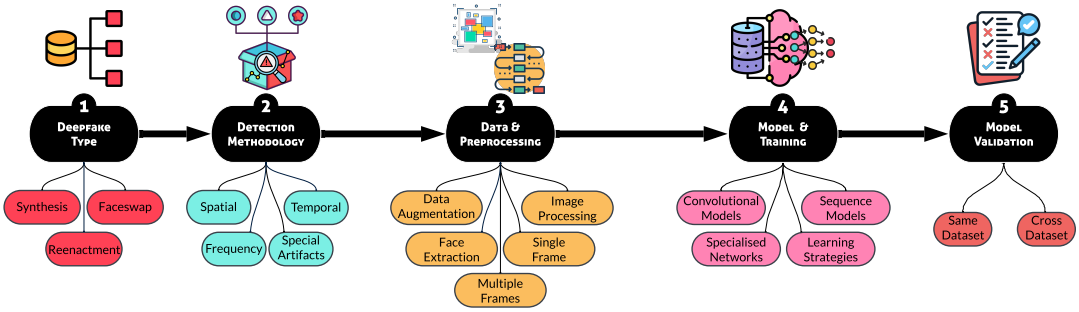


Figure 2. Our Five-Step Conceptual Framework: All detection methods adhere to this framework: Stage #1 (Deepfake Type), #2 (Detection Methodology), #3 (Data & Preprocessing), #4 (Model & Training), and #5 (Model Validation).

4.2. Analysis of Detectors

This section outlines our methodology for analyzing the 51 selected deepfake detectors. Our examination commenced by identifying the primary focus of these detectors, revealing a prevalent emphasis on detecting both faceswap and reenactment-based deepfakes. Interestingly, only 5 detectors utilize synthesis of synthetic images [16], [19], [75], [79], [80].

Moving on to artifact and pattern analysis, we observed that the majority of detectors concentrate on spatial features, either independently or in conjunction with temporal or frequency domain features. Exceptions include two approaches [16], [80], where each exclusively focuses on the frequency domain, and three methods [2], [75], [81], which consider special artifacts such as Voice Sync and Noise Traces.

Subsequently, we thoroughly investigated the preprocessing methods utilized by these detectors for data preparation. This exploration covered a range of techniques, including image processing, data augmentation, and face extraction. We identified diverse approaches in input data representation among the reviewed literature, notably the use of both single-frame and multi-frame sequences. These methods were observed to be employed with nearly equal frequency across our pool of literature.

Next, our investigation extended to the architectural characteristics of the classification models used by the detectors. Deep neural networks (DNNs) form the core of most approaches, including convolutional models such as VGG [82] and ResNet [83], sequence models such as BiLSTM [84] and Vision Transformer [85], in addition to specialized networks such as graph learning [86] or capsule networks [14]. These DNNs were deployed in standalone configurations or in combination with each other, employing various learning strategies such as knowledge distillation [87], Siamese networks [88], etc. This analysis helped us understand the landscape of model architectural choices employed in detectors and later helped us in our categorization.

Finally, our investigation of validation methodologies revealed two predominant approaches: training and testing on the same dataset, and training on one dataset whilst testing on another to demonstrate model generalisability. This analytical endeavor yielded two key outcomes: (i) it elucidates the typical procedural steps followed by deepfake detectors, and (ii) it delineates the specific activities encompassed within these steps. This information serves as the foundation for consolidating the entire process into a conceptual framework, which we present in the next section.

4.3. Conceptual Framework

Our meticulous review of the 51 selected papers on deepfake detection helped us identify a common 5-step pipeline frequently used to develop deepfake detection methods. This 5-step process serves as the foundation of our conceptual framework, as shown in Fig. 2, which features 25 influential factors that we identified for RQ1. Our conceptual framework is delineated below:

❶ **Deeepfake Type.** The first step of our framework involves identifying the specific type(s) of facial deepfake attacks that the detector will target. Figure 2 outlines the three categories of deepfakes considered in our framework, namely faceswap (or replacement), face reenactment, and face synthesis, which were defined in Sec. 2. Recent literature on deepfake detectors primarily focus on faceswap and reenactment, as evidenced by [28], [89].

❷ **Detection Methodology.** The second step involves detailing the detection methodology employed by the deepfake detector. These methodologies can be broadly classified into four main categories as presented in Fig. 2: *Spatial* artifact, *Temporal* artifact, *Frequency* artifact, and *Special* artifact-based detectors, each focusing on specific aspects of deepfake identification. Spatial artifact-based detectors analyze individual images or video frames and focus on intra-frame visual anomalies or discrepancies, which could manifest as irregularities in texture, color, lighting, misalignments, or inconsistent blending between different segments of the image. Temporal artifact detectors aim to identify inter-frame inconsistencies across multiple video frames over a specified duration.

On the other hand, frequency artifact-based detectors differ from visual methods by operating within the frequency domain. Indeed, deepfake manipulation often alters the rate at which the pixel values are changing, creating a distinctive frequency 'signature' affecting the image's spectral char-

acteristics, serving as discriminative cues for these detectors. In addition, *special artifacts* concentrate on identifying unique, specific manipulation signatures or anomalies that are characteristic of deepfake generation methods. Examples of this include models that aim to detect anomalies in synchronization features, i.e., audio-visual features that aim to convey the temporal alignment between vision and sound (i.e., between lip movement and voice) [2]. Our comprehensive categorization is presented in columns 1 and 2 of Table 3, where we derive more fine-grained level subgroups.

❸ **Data & Preprocessing.** The third step in our framework concerns preparing and transforming the dataset(s) used for detector training. We categorize data preprocessing into three key areas as shown in Fig. 2: 1) Data Augmentation, 2) Image Processing, and 3) Face Extraction. Additionally, we classify data representation into two categories: Single-frame and mult-frame. First, *Data Augmentation* plays the pivotal role of synthesizing training data, employing techniques such as Suspicious Forgeries Erasing [90], Self-Blended Images (SBI) [24], as well as Temporal Repeat and Dropout [4]. Collectively, these methods strengthen the detector's ability to identify subtle anomalies indicative of deepfakes. Also, *Image Preprocessing* techniques collectively contribute to the effective preparation and transformation of datasets, including methods such as 3D Dense Face Alignment (3DDFA) [91] to enable accurate feature extraction, and others such as Face Alignment [92] and RetinaFace with 4 key points [93] to ensure the standardization of facial features across images. Lastly, *Face Extraction* techniques involve accurately identifying and isolating human faces in a video or image, using popular tools such as Dlib [94] and Multi-task Cascaded CNNs (MTCNN) [95].

❹ **Model & Training.** The fourth step in our framework encompasses different aspects of the model and training settings, including the detector model's architecture. Our framework classifies the structure of the model into three broad categories: convolutional models, sequence models, and specialized networks (see Fig. 2). Models in the *Convolutional Models* category leverage common Convolutional Neural Networks (CNNs) such as ResNet, VGG, or XceptionNet, which discern authentic images from manipulated ones by identifying subtle inconsistencies and anomalies in pixel patterns and textures. Besides, *Sequence Models* utilize Recurrent Neural Network (RNN) or Transformer-based model architectures, including BiLSTM, Vision Transformer, or Transformer Encoder, to analyze sequential inconsistencies. Inputs are sequences of frames of video data for spatiotemporal models, which identify deepfakes by tracking the continuity and flow of sequential frames. Alternatively, it may be for spatial detectors, which divide a single frame into multiple patches and then provide these patches as a sequence as input to the detector. *Specialized Networks* models differ from those in the convolutional models' category by integrating novel architectures such as U-Net [96] or Capsule Network [97], to capture more nuanced and complex indicators of deepfakes. Finally, Step 4 also includes *Learning Strategies* for training, such as meta-learning [98], Graph Information Interaction layers [86], Dual Cross-Modal Attention [99], and Siamese learning [76].

❺ **Model Validation.** Our fifth and final step of the framework addresses the critical task of validating the trained deepfake detection model. Based on our literature study, this validation process can be broadly categorized into two distinct approaches: the same dataset and cross-dataset validation. *Same dataset* validation approaches entail assessing the model's performance on the test set of the same dataset(s) from which the training data were taken (e.g., both training and test sets taken from FF++). *Cross dataset*, in contrast, involves testing the detector on a dataset different from the one from which the training dataset was taken (e.g., training data taken from FF++ and test data from CelebDF). This evaluation method is vital for assessing the model's generalizability and robustness.

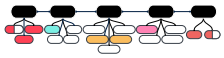
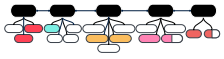
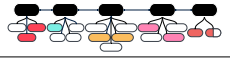
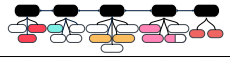
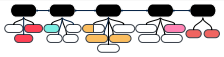
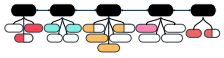
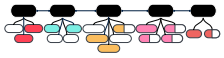
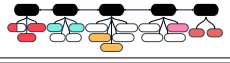
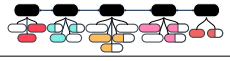
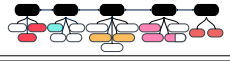
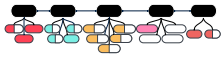
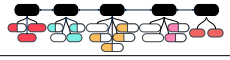
4.4. Detector Taxonomy

In this section, we take the 51 detectors identified through our systematic selection process in Sec. 4.1, and map each detector into a unified taxonomy based on our conceptual framework. We present our taxonomy in Table 3, which naturally segments the 51 detectors into 4 high-level groups based on their Detector Methodology. Then, Table 3 refines each group into sub-groups, based on the shared conceptual framework representation among the papers in that sub-group, shown in the "C.F. REPRESENTATION" column (column six). We visually depict each framework using color-coded nodes. A fully colored node indicates that every paper in the framework belongs to the specified category. In contrast, a white node means none of the papers in that framework fit the category. A half-colored node signifies a mixed scenario: some papers in the framework belong to the category, while others do not. Overall, this resulted in 13 distinct conceptual framework representations, which depict the different clusters of detector methodology and evaluation in leading research publications since 2019. We outline each of the representations of the 13 conceptual framework and how they fit into the taxonomy below.

Spatial artifact models. We find 22 spatial artifact-based detectors are all capable of performing both faceswap and reenactment attacks. All the spatial approaches are single-frame models that utilize face extraction techniques. Of the 22 detectors, 13 of them (roughly 59%) feature the ConvNet model architecture. And, the spatial artifact models can be sub-grouped into 5 distinct conceptual framework representations based primarily on their differences in Step 4 model and training. Conceptual framework #1 contains purely convolutional models that do not utilize sequence models or any additional specialized networks or learning strategies.

Conceptual Framework #2 stands out in our study for its use of specialized networks such as EfficientNets [121].Similarly, conceptual Frameworks #3 and #4, similar to Framework #1, rely on convolutional models. However, Framework #3 incorporates additional special learning strategies, enhancing its detection capabilities. In contrast, Framework

Table 3. **SYSTEMATIZATION OF DEEPFAKE DETECTORS.** THE "C.F. REPRESENTATION" COLUMN SHOWS EACH CONCEPTUAL FRAMEWORK REPRESENTATION. WHITE NODES MEAN NO PAPERS FIT THE CATEGORY, HALF-COLORED NODES MEAN SOME PAPERS BELONG TO THE CATEGORY, WHILE FULLY COLORED NODES MEAN ALL PAPERS BELONG TO THE CATEGORY (SEE APPX. TABLE 5 FOR DETAILS ON DETECTORS). THE "FF++ SCORE" COLUMN SHOWS THE DETECTOR'S PERFORMANCE ON THE FF++ DATASET. † INDICATES THE DETECTORS USED FOR EVALUATIONS.

FOCUS OF METHODOLOGY		DETECTOR NAME	VENUE	YEAR	UNIQUE FEATURE (NOVELTY)	C.F. REPRESENTATION	FF++ SCORE
SPATIAL ARTIFACTS	**1. *ConvNet Models***	Cap.Forensics † [14]	ICASSP	'19	*Capsule Network*		96.6 (AUC)
		XceptionNet † [26]	ICCV	'19	*Depthwise Convolutions*		99.26 (ACC)
		Face X-ray [71]	CVPR	'20	*Face X-ray Clues*		98.52 (AUC)
		FFD [100]	CVPR	'20	*Unified Methodology*		-
		RECCE [72]	CVPR	'22	*Bipartite Graphs*		99.32 (AUC)
		CORE [101]	CVPRW	'22	*Consistency Loss*		99.94 (AUC)
		IID [102]	CVPR	'23	*Face Implicit Identities*		99.32 (AUC)
		MCX-API † [103]	WACVW	'23	*Multiple Color Spaces*		99.68 (AUC)
	2. *Specialized Networks*	EffB4Att † [76]	ICPR	'20	*Siamese Training*		94.44 (AUC)
		LTW [104]	AAAI	'21	*Intra-class Compact Loss*		99.17 (AUC)
		MAT † [23]	CVPR	'21	*Multi-attention losses*		99.27 (AUC)
		DCL [105]	AAAI	'22	*Intra-instance CL*		99.30 (AUC)
		SBIs † [24]	CVPR	'22	*Self-blend Image*		99.64 (AUC)
	3. *ConvNet Models with Learning Strategies*	MLAC [106]	ACMMM	'21	*Adversarial Learning*		88.29 (AUC)
		FRDM [99]	CVPR	'21	*High Frequency Pattern*		-
		OST [98]	NeurIPS	'22	*Meta-learning*		98.20 (AUC)
	4. *ConvNet with Specialized Networks*	CADDM † [107]	CVPR	'23	*Identity Representation*		99.70 (AUC)
		QAD [1]	ICCV	'23	*Collaborative Learning*		95.60 (AUC)
	5. *Sequence Models*	ICT † [70]	CVPR	'22	*Facial & Other Inconsistency*		98.56 (AUC)
		UIA-ViT [108]	ECCV	'22	*Unsupervised Inconsistency*		99.33 (AUC)
		AUNet [13]	CVPR	'23	*Action Units*		99.89 (AUC)
SPATIOTEMPORAL ARTIFACTS	**6. *ConvNet Models***	ADDNet-3d [109]	ACMMM	'20	*Facial Attentive Mask*		86.69 (ACC)
		DeepRhythm [110]	ACMMM	'20	*Anomaly Heartbeat Rhythm*		98.50 (ACC)
		S-IML-T [111]	ACMMM	'20	*Multi-instance Learning*		98.39 (ACC)
		TD-3DCNN [112]	IJCAI	'21	*Time Discrepancy Modeling*		72.22 (AUC)
		DIA [113]	IJCAI	'21	*Global-Local frame learning*		98.80 (AUC)
		DIL [114]	AAAI	'22	*Local Dynamic Sync*		98.93 (ACC)
		FInfer [74]	AAAI	'22	*Faces Predictive Learning*		95.67 (AUC)
		HCIL [115]	ECCV	'22	*Contrastive Learning*		99.01 (ACC)
		AltFreezing † [4]	CVPR	'23	*Alternate Modules Freezing*		98.60 (AUC)
	7. *ConvNet with Specialized Networks & Learning Strategies*	STIL [116]	ACMMM	'21	*SpatioTemporal Inconsistency*		98.57 (ACC)
		LipForensics † [21]	CVPR	'21	*Reading Mouth Movements*		97.10 (AUC)
		FTCN † [78]	ICCV	'21	*Temporal Transformer*		-
	8. *Sequence Models*	CCViT † [77]	ICIAP	'21	*Combine ViT and CNN*		80.00 (ACC)
		CLRNet † [15]	WWW	'21	*Spatial-temporal Modules*		99.35 (F1)
		LTTD [117]	NeurIPS	'22	*Unsupervised Learning*		97.72 (AUC)
FREQUENCY ARTIFACTS	**9. *ConvNet Models***	F3-Net [16]	ECCV	'20	*Frequency Learning*		98.10 (AUC)
		FDFL [118]	CVPR	'21	*Single-center Loss*		97.13 (AUC)
		ADD † [18]	AAAI	'22	*Knowledge Distillation*		-
		LRL [73]	AAAI	'23	*Spatial & Frequency Learning*		99.46 (AUC)
	10. *ConvNet with Sequence Model & Learning Strategies*	TRN [119]	ECCV	'20	*SpatioTemporal Frequency*		99.12 (AUC)
		SPSL [80]	CVPR	'21	*Phase Spectrum Learning*		95.32 (AUC)
	11. *Specialized Network & Learning Strategies*	CD-Net [17]	ECCV	'22	*Intra-Sync with Frequency*		98.5 (AUC)
		SFDG [86]	CVPR	'23	*Collaborative Learning*		95.98 (AUC)
SPECIAL ARTIFACTS	**12. *ConvNet Models***	RFM [90]	CVPR	'21	*Region Tracking*		99.97 (AUC)
		FD2Net [120]	CVPR	'21	*Facial Features Modeling*		99.68 (AUC)
		SOLA [81]	CVPR	'22	*2nd Order Anomaly*		98.10 (AUC)
		AVAD [2]	CVPR	'23	*Audio-video Anomaly*		-
		LGrad † [19]	CVPR	'23	*Grad Pattern Learning*		66.70 (ACC)
	13. *Sequence Model with Learning Strategies*	LRNet † [79]	CVPR	'21	*Temporal Landmark Learning*		99.90 (AUC)
		NoiseDF [75]	AAAI	'23	*Noise Pattern Learning*		93.99 (AUC)

#4 integrates further specialized networks, diversifying its approach. Lastly, Framework #5 is unique in its reliance on purely sequential models, foregoing the use of special learning strategies or specialized networks, thus presenting a more streamlined approach to deepfake detection.

Spatiotemporal artifact models. There were 17 spatiotemporal artifact-based detectors, all of which use multiple frames in Step 3. We remark that due to the visual nature of facial deepfakes, all spatiotemporal models focus on spatial artifacts in addition to temporal, which is why there are no temporal-only detectors. All the spatiotemporal models additionally utilize face extraction in Step 3. Among the spatiotemporal detectors, there are three distinct conceptual framework representations. The first one, Framework #6, features purely convolutional models that do not utilize sequence models or any learning strategies. And, Framework #7 includes convolutional models that can additionally be categorized as specialized networks, and/or sequence models, and may also utilize learning strategies. Framework #8 is comprised of spatiotemporal sequence networks.

Frequency artifact models. There were 8 frequency-based detectors, all of which can detect both faceswap

and reenactment attacks. The frequency artifact models can be divided into 3 distinct conceptual frameworks. The first one, conceptual Framework #9 consists of the purely convolutional frequency-based models, whereas conceptual Framework #10 contains the ConvNet models that are also sequence models and/or utilize special learning strategies. Lastly, conceptual Framework #11 consists of specialized networks with learning strategies.

Special artifact models. We identified 7 models specifically designed for detecting special artifacts. All of these are capable of identifying both faceswap and reenactment attacks. These special artifact models are categorized into two distinct frameworks. Conceptual Framework #12 encompasses purely convolutional models, focusing on spatial artifact detection. Meanwhile, Framework #13 comprises sequence models, which are distinct in their inclusion of temporal artifacts within their learning algorithms. This division highlights the varied approaches in detection, addressing different types of deepfake manipulations.

Preliminary observations. Our systematization offers insights into the key factors influencing facial deepfake detection. Notably, ConvNets dominate the field, as evidenced by their prevalence in the majority of detectors analyzed across all 4 methodologies. This trend persists in the latest research, with ConvNets featuring prominently in papers published in 2023. Additionally, spatial and spatiotemporal approaches remain a significant focus in our literature review.

As shown in column 7 in Table 3, each detector's performance on the FF++ is self-reported, using a variety of metrics such as accuracy, AUC, and F1 score. This diversity of standards hinders the direct comparison of detector efficacy. A further challenge in comparing models arises from the inconsistency in training datasets. Not all models were trained on the FF++ dataset, resulting in a mix of cross-dataset and same-dataset evaluations in our analysis. The aforementioned discrepancy in evaluation contexts clearly show the necessity of our comprehensive and systematic evaluation of deepfake detectors, as outlined in Sec. 5. Our approach aims to highlight influential factors more effectively, providing a clearer understanding of the deepfake detection landscape.

5. Evaluation Settings

5.1. Detectors for Evaluation

To address ***RQ2*** (*How do the current leading detectors perform when evaluated for generalizability?*), we rigorously evaluated the performance of various deepfake detectors across various datasets, ensuring a meticulous and equitable comparison. To this end, we selected a subset of detectors from the 51 identified in *RQ1* (see Table 3) by employing the following inclusion criteria:

Generalization Claims. The detector papers should provide evidence of their methods demonstrating generalization in one or more unseen datasets. This criterion allows us to focus solely on detectors explicitly developed to exhibit generalizability across different types of deepfakes.

Open Source with Model Weights. Replicating the original training environment for each detector is challenging. To maintain fairness, we only consider open-source detectors with publicly available pre-trained model weights. Following these criteria, we identified 16 SoTA detectors, which are marked with † in Table 3. A common trend observed in all detectors that claim generalizability is that the FF++ dataset was used to train the detector, and the detector's generalizability was assessed on second-generation deepfake datasets. One of our evaluation strategies will also adopt similar settings. In the forthcoming sections, we will discuss our evaluation strategies and the specific deepfake datasets used for these assessments.

5.2. Evaluation Strategies

Our evaluation strategy is driven by the question: ***How do the current leading detectors perform when evaluated for generalizability?*** To address this, we have devised three distinct evaluation strategies that emulate varied levels of information and control regarding source videos, destination videos, and the utilized deepfake generation methods. These strategies mirror evaluation scenarios reminiscent of gray-box, white-box, and black-box settings.

(i) Gray-box Generalizability Evaluation. Our objective is to conduct a thorough evaluation using datasets where we possess partial knowledge of deepfake, yet lack control over the source, destination, or generation method. By subjecting all detectors to gray-box dataset settings, and employing various benchmark datasets (Sec. 5.3.1), we aim to simulate scenarios where detectors have limited information about the deepfake generation process. This scenario represents a middle ground between black-box and white-box evaluations, where detectors operate with partial information, reflecting common real-world scenarios where some knowledge exists but complete control is lacking.

(ii) White-box Generalizability Evaluation. A distinct aspect of our study involves evaluating detectors in controlled settings, where we possess complete control over the source and destination videos and deepfake generation process. By identifying 20 tools, our rigorous selection criteria narrowed the choices down to 7 (Sec. 5.3.2). This evaluation setup aims to mimic scenarios where we have full information and control over the deepfake generation, providing insights into the detector's performance under different conditions. This controlled setting allows us to highlight how detectors perform when all factors are known and controlled.

(iii) Black-box Generalizability Evaluation. This evaluation setting prioritizes real-world dataset assessments without specific knowledge of the deepfake generation methods or their origins. Our focus here is to measure the overall performance of the detectors in scenarios that mimic real-world encounters with deepfakes. Given that the generation methods for these deepfakes remain unknown, this renders the situation analogous to the black-box evaluation setting. We simulate a black-box scenario by assembling a comprehensive dataset from links provided by [31] comprising 2,000 samples sourced from 4 online platforms (Sec.

5.3.3). The lack of information regarding deepfake generation methods aligns with the challenges akin to real-world detection scenarios, emphasizing the need for detectors to perform effectively under such conditions.

5.3. Datasets for Evaluations

In order to facilitate the implementation of our generalization evaluation strategies, we utilize a variety of deepfake datasets. These comprise publicly available benchmark datasets, those collected by the research community from the internet, and datasets generated by our team.

5.3.1. Dataset for Gray-box Evaluation. For the implementation of the gray-box evaluation strategy, we utilize two primary datasets: DFDC [46] andCelebDF [122]. Both exemplify second-generation deepfakes, characterized by amplified visual quantity and increased detection complexity, rendering them well-suited for our assessment.

The DFDC dataset, released by Facebook, contains over 100,000 faceswap videos from 3,426 actors, diverse in gender, age, and ethnicity. Due to limited public information about its creation, it suits gray-box evaluation. The subsequent CelebDF dataset presents more sophisticated deepfakes with 590 original YouTube-sourced videos of celebrities with diverse demographics in terms of age, ethnicity, and gender, leading to 5,639 DeepFake videos. This enhances the variety of challenging samples for evaluation.

5.3.2. Dataset for White-box Evaluation. We found existing deepfake datasets inadequate for thorough white-box evaluation. FF++ offers only limited insights due to its inherent biases and homogeneity, such as static front-facing shots. It also misses key deepfake generation methods used in real-world scenarios. To address such gaps, we created a stabilized deepfake dataset specifically for our study in Sec. 6.1.2. Our dataset, generated through a controlled pipeline, isolates the deepfake generation method as the explanatory variable, allowing us to examine detector performance variations against different deepfake techniques. Initially, we identified 23 deepfake tools (see Appendix - Table 4), and subsequently selected 7 based on our defined inclusion criteria. We outline the details of dataset creation below.

Generator Selection Process. We conducted a thorough survey to explore the current landscape of deepfake generation methods. Information was gathered from each paper's official websites such as GitHub repositories. To maintain consistency of the white-box dataset, the selected generation methods should be *applicable* to our task and ensure user *freedom* in the victim and driver selection. More specified conditions are provided in Table 4 in Appendix.

Selected Generators. Following a methodical evaluation process and the elimination of methods that did not meet our criteria, we curated a set of 7 distinct methods: DeepFaceLab [6], Faceswap [7], specifically the LightWeight variant within Faceswap, DeepFaker [123], FOM-Animation [8], and FSGAN [51]. Further details on all generators are outlined in Table 4 in Appendix.

Driver Video Selection. We chose the real videos from the deepfake detection dataset (DFD) [124] as the driver and victim videos for our deepfake video generation process for the following reasons: (i) All individuals featured in these videos are paid actors who have provided explicit consent for their videos to be utilized in deepfake generation for research purposes, and (ii) The dataset encompasses a wide variety of scenarios, enhancing its diversity in this context.

Deepfake Generation Process. By rigorously following a systematic process, we produced deepfake videos for each of the 7 selected generation methods. To generate these dataset videos, we conducted the following steps: (i) Selection of two random actors from a pool of 28 actors, (ii) Matching the scenarios portrayed in the original videos to both the source and target actors, emphasizing crucial elements such as facial expressions, body posture, and non-verbal cues to augment the video quality, and (iii) Provision of these videos to the selected deepfake generation methods[2]. The real (source and target) videos utilized for deepfake generation constitute the real segment of the dataset, comprising 54 videos. Meanwhile, the deepfake segment of the dataset encompasses 28 videos for each of the 7 distinct methods, resulting in a total of 196 videos ($28 \times 7 = 196$). In aggregate, our stabilized dataset comprises 250 videos.

5.3.3. Dataset for Black-box Evaluation. For our black-box evaluation, we used the Real-World DeepFake (RWDF-23) dataset [31], which includes 2,000 deepfake videos from platforms such as Reddit, YouTube, Bilibili, and TikTok. To ensure a varied representation of deepfake creation methods, intentions, demographics and contexts, they searched the selected platforms using the following deepfake terms in four different languages: "Deepfake", "딥페이크", "深度伪造", and "дипфейк". We adopted the approach outlined in [78] to extract sub-clips from each video. Subsequently, we carefully labeled the first clip of each video, resulting in 513 genuine and 1,383 manipulated clips. We excluded 104 clips due to false positives from the face extractor, often mistaking faces in static artwork.

5.4. Pre-train Model and Metrics

Pre-training Sources. Most methods leverage pre-training on the FF++ datasets [26], either partially or entirely. Notably, the CLRNet and CCViT models undergo pre-training using the DFDC dataset, specifically prepared for the Deepfake Detection Challenge; therefore, we omitted them from DFDC results in Fig. 3. A distinct approach to pre-training is observed in LGrad, where the authors employ a novel dataset generated with ProGAN.

Inference Process. During inference, detector configurations adhere strictly to the specifications in the respective paper. For frame-based prediction methodologies (*i.e.*,

2. Note: In most generation methods, deepfakes are produced iteratively, with visual quality progressively improving. To ensure consistency, a coauthor manually reviewed the visual fidelity, terminating the process when no further improvement was observed after multiple iterations.

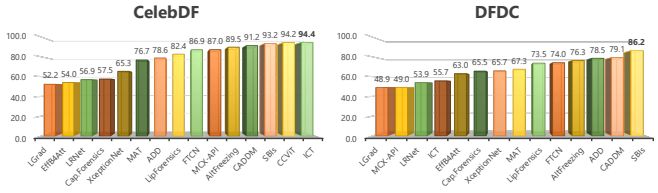


Figure 3. Gray-box results. Performance (AUC%) of selected deepfake detectors on CelebDF and DFDC datasets.

Rossler, Cap.Forensics, EffB4Att, MAT, SBIs, ICT, MCX-API, CADDM, LGrad, ADD), we aggregate frame predictions to derive video probability or scores. Conversely, for multi-frame dectors (*i.e.*, FTCN, LRNet, LipForensics, CCViT, AltFreezing, CLRNet), we selectively sample frames based on their designated temporal length for prediction.

Evaluation Metrics. In assessing deepfake detection methods, metrics like accuracy (ACC), F1 score, and Area Under the Receiver Operating Characteristic Curve (AUC) are commonly used. The diversity of metrics across studies complicates direct comparison. Due to the distinct characteristics of detectors and class imbalance in our datasets, AUC is the most relevant, evaluating detector performance across various thresholds. A higher AUC indicates better discrimination between real and manipulated videos, balancing false positives and negatives.

6. Evaluation Results

This section presents the results from our evaluations. Our primary focus here is to answer the following research questions. ***RQ2: How do leading detectors perform when evaluated for generalizability? RQ3: How do the identified influential factors may impact the performance of the detectors when evaluated for generalizability?***

6.1. Detection Results

6.1.1. Gray-box generalizability. This section presents generalizability results derived from the raw performance metrics on the CelebDF and DFDC datasets.

CelebDF. ICT achieves the highest AUC score of 94.4%, followed closely by CCViT with 94.2% and SBIs with 93.2% on the CelebDF dataset. Notably, the authors of ICT employed pre-training on the MS-Celeb-1M dataset [125], boasting an extensive collection of approximately 1 million real celebrity faces. This substantial pre-training dataset could account for ICT'sexceptional performance on the CelebDF dataset, where the emphasis is on celebrities. On the other hand, SBIs is carefully pre-trained in a self-supervised manner with only real samples, and they synthesize deepfakes during training to generate difficult samples that present subtle artifacts. Moreover, Altfreezing (89.5%), CADMM(87.0%), MCX-API (87.0%), and FTCN (86.9%) all perform in the high 80s by learning a secondary artifact such as identity leakage (CADMM) or color channels' inconsistency, whereas LGrad (52.2%), EffB4Att (54.0%), and LRNet (56.9%) are the worst performers on CelebDF. This could be explained by their reliance on fully supervised learning using cross-entropy loss on the FaceForensics++ dataset, without auxiliary robust learning modules.

DFDC. SBIs (86.2%) is the best performer, with the AUC score of CADDM (79.1%), ADD (78.5%), and Altfreezing (76.3%) ranging within 7 to 10 percentage points of SBIs. The training approach using SBIs involves blending source and target images derived from individual pristine images, thereby creating a more versatile dataset with subtle, hard-to-detect fake samples, which in turn prompts classifiers to develop more universal and resilient representations for detecting deepfakes. The worst performers were LGrad (48.9%), MCX-API (49.0%), LRNet (53.9%), and ICT (55.7%), which performed similarly to a random guess. And, LGrad is highly competent against GAN imagery; however, it is not robust against deepfake images. Meanwhile, LRNet, MCX-API, and ICT, which showed great performance on the CelebDF dataset, failed to detect deepfakes in the DFDC dataset. This may be because they focus on high-level semantics, specifically identity information, or because they may have optimized to the FF++ dataset, which in some ways has similar settings (e.g., front-facing and interview-style videos) as CelebDF.

For average performance scores, we found that ADD (78.6%), AltFreezing (82.9%), CADDM (85.2%), FTCN (80.5%), and SBIs (89.7%) were the top-5 performers in the gray-box generalizability test. This success could be attributed to their efforts in detecting subtle artifacts in videos from the DFDC dataset, which is well-known for its diversity in quality. Specifically, these models excel in various aspects: SBIs in the diversity of deepfake detection, AltFreezing and FTCN in identifying temporal inconsistencies, and ADD in recognizing low-quality deepfakes.

Insights. (1) While DFDC and CelebDF are both considered second-generation deepfake benchmark datasets, results reported in the CelebDF paper [27] suggest that CelebDF poses a greater challenge than DFDC, as detectors were reported to perform lower on average on the CelebDF (56.9%) compared to DFDC (64.7%). Interestingly, our study reveals the opposite for a subset of 10 detectors that were characterized by increased diversity and recency, exhibited lower performance on DFDC than CelebDF. This trend holds true except for EffB4Att, Cap.Forensics, and XceptionNet. However, it is worth noting that their performance on both datasets remains relatively low, ranging from the mid-50s to mid-60s, rendering them not competitive. And, XceptionNet and Cap.Forensics, both old models having experimented only on FF++, appear to be more vulnerable to the diverse manipulation methodologies employed in CelebDF. Additionally, EffB4Att, which is pre-trained on FF++, exhibits relatively lower cross-validation results on DFDC, with a score of 63.0% (sourced from the FD2Net paper [120]). Despite the low performance, since the model itself was designed for participation in the DFDC challenge, there is a high likelihood that it has undergone tuning specifically for the DFDC dataset. This is a plausible explanation for the

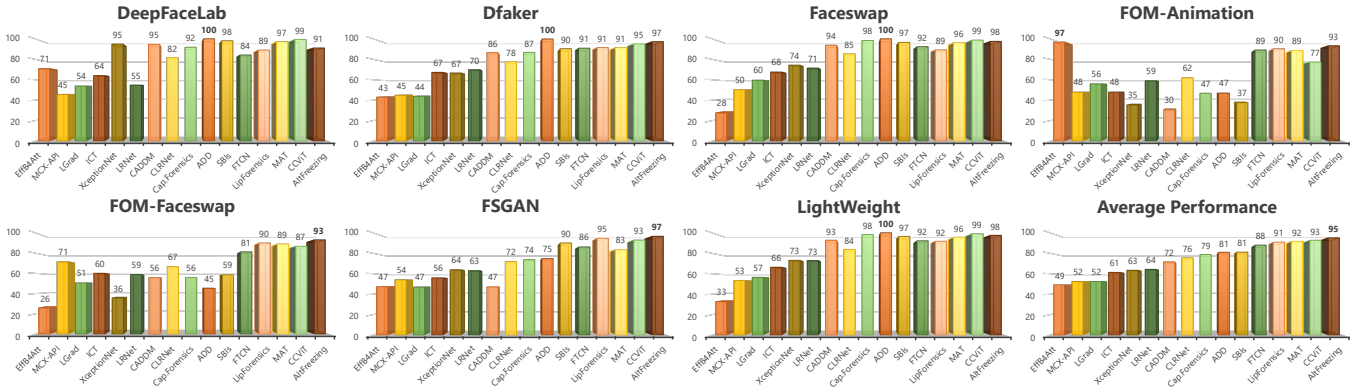


Figure 4. White-box results (Stabilized Set). Performance (AUC%) of selected deepfake detectors on our stabilized datasets .

observed trends in the performance of these three models.

(2) ICT is the best performer on CelebDF (94.4%) and one of the worst performers on DFDC (55.7%). This sharp contrast can potentially be attributed to overfitting, as ICT was trained on celebrity faces, a subset that significantly overlaps with the CelebDF dataset. A similar pattern is observed in MCX-API (87.0% to 49.0%). This is a performance difference of 38.0% points for MCX-API and 38.7% points for ICT. It is worth noting that MCX-API heavily relies on detecting inconsistencies in the color space, which may degrade in the lower-quality deepfakes found in the DFDC dataset, contributing to its reduced efficacy.

(3) ADD is very consistent across both datasets (78.6% and 78.5%) with a difference of only 0.1% points. It seems that the ADD exhibits this consistency because it has been trained on highly compressed and low-quality images, allowing it to generalize well to similar characteristics in other datasets. The main reason is that the model learns general features rather than being dependent on specific datasets, and as a result, it can achieve good performance across different datasets. In contrast, the difference between the average performance of the detectors in CelebDF (77.3%) and DFDC (66.9%) is 10.4%. Recently, contemporary models have primarily used CelebDF as an evaluation dataset to assess robustness and generalization. Therefore, in current research, there appears to be such a performance gap between the two datasets, indicating a contrary trend. This resulting trend originally came from the latest SoTA methods published after the year 2021, such as MAT (2021), FTCN (2021), LipForensics (2021), SBIs (2022), ICT (2022), MCX-API (2023), CADDM (2023), and AltFreezing (2023).

6.1.2. White-box generalizability. In this experiment, we aim to further explore the generalizability of deepfake detectors by utilizing a stabilized dataset, which was generated using 7 selected deepfake generation tools, all employing the same source and driver faces. This enables us to expose underlying artifacts that are not readily identifiable in gray-box settings, including specific generation methods. Notably, the DFDC, which is utilized in our gray-box experiments, employs various faceswap methods. However, the exact nature of these methods remains unknown. This introduces a significant degree of uncertainty, particularly when addressing research question **R3**, which seeks the effectiveness of deepfake detection under varying conditions.

We noted that AltFreezing demonstrates the highest average performance at 95%, notably excelling in detecting deepfakes generated by Dfaker, Faceswap, FSGAN, and LightWeight compared to the other three methods. In particular, AltFreezing is designed as a spatiotemporal model with a specific focus on enhancing the generalization capability of the forgery detection model. Uniquely, it alternately freezes spatial and temporal weights. Moreover, its performance does not notably decrease, when confronted with geometric manipulation methods such as FOM (93%) or primarily used deepfake methods such as DeepFaceLab (91%). Further investigation revealed that the AltFreezing model used in this experiment was pre-trained on FaceForensics++ with various augmentations, ensuring the model's generalization to our unseen deepfake creation methods.

Moving beyond AltFreezing's performance, CCViT emerges as a notably consistent performer, achieving an average accuracy of 93%. This method utilizes temporal inconsistency and learns to generalize detectors through augmentations or voting, similar to AltFreezing. Additionally, MAT maintains a relatively high average performance at 92%.And, MAT serves as a fine-grained classifier, combining global/semantic features with local/textural features for the first time. Similar to what Altfreezing has experienced, MAT excels on certain datasets, also faltering with unseen data generated FOM methods (89%) and FSGAN (83%). It is not surprising, considering that first-order-motion techniques were rarely employed in previous deepfake datasets.

On the other hand, LGrad (44%), MCX-API (52%), and EffB4Att (47%) are the lowest average performers in our study. This low performance may be attributed to the specific focus of each model. For instance, LGrad specializes in discerning fully synthesized fake images generated by techniques such as GANs or Diffusion models, rather than faceswap or reenactments, which are more prevalent in our white-box test setting. Besides, MCX-API and EffB4Att also suffer from low generalization issues. They both em-

ploy contrastive loss during their learning processes. This approach, while effective in certain contexts, may lead to an overemphasis on specific artifacts in the training data. Consequently, the models may overfit to these characteristics, resulting in a poor generalized decision boundary that fails to accurately detect a broader range of deepfakes.

6.1.3. Black-box generalizability. In this experiment, we delve further into the deepfake detector generalizability, subjecting it to scrutiny on a corpus of 2,000 in-the-wild deepfake videos. Our investigation yields the following key findings: (1) None of the detectors managed to surpass the 70% accuracy threshold, underscoring the persistent challenge of effectively detecting in-the-wild deepfakes. (2) MAT (68.8%), CCViT (66.4%) and ICT (63.4%) emerge as the top performers in black-box testing, coming from various methodological backgrounds. A common influential factor among them is the integration of the multihead attention mechanism, suggesting its critical role in enhancing detection efficacy. (3) Notably, three out of the top four performers in this dataset—MAT, ICT, and CADDM—originate from spatial artifact-focused methods. While they are different in their methodological categorization, a closer examination reveals shared influential factors: the utilization of the EfficientNet model in CADDM and MAT, and the incorporation of multi-head attention in MAT and ICT. Interestingly, when EfficientNet is used in a Siamese network setting (EffB4Att), the results prove less promising, registering an accuracy of 52.4%. (4) On the contrary, two of the least effective performers, XceptionNet and Cap.Forensics, stem from spatial artifact-focused methods that employ ConvNet models, without additional supervised learning objectives or mining steps. This shared characteristic further explores the potential factors that contribute to their suboptimal performance in this experimental context.

Interestingly, ADD, the frequency-based artifact method, renowned for its efficacy on white-box and DFDC datasets, achieves a modest 49.5% AUC on black-box deepfakes. A notable drawback is that neither the teacher nor student networks are explicitly designed for generalization to unseen deepfakes, possibly leading to diminished accuracy in addressing in-the-wild scenarios.

Additionally, specialized methods targeting special artifacts, like LRNet with facial landmarks and LGrad with gradient noises, exhibited low performance at 50.0% and 48.5%, respectively. This is attributed to their limited consideration of spatial information in input deepfake videos. Notably, these methods fail to capture certain features, such as unconventional blending boundaries or unnatural temporal movements.

6.2. Impact of Influential Factors

The results obtained from our comprehensive evaluation utilizing gray-box, white-box, and black-box strategies demonstrate the impact of the identified influential factors within our conceptual framework. To elucidate this impact, we provide specific illustrations without loss of generality:

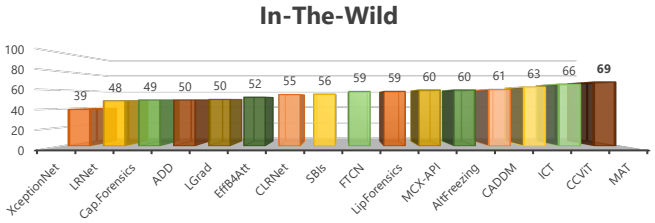


Figure 5. Black-box results (in-the-wild). Performance (AUC%) of selected deepfake detectors on the in-the-wild dataset.

(a) Gray-box evaluation: The performance of the ICT excelled in CelebDF but exhibited suboptimal results in DFDC. This discrepancy can be attributed to the model's training on the MS-Celeb-1M dataset, evident in the overlap between the training (MS-Celeb-1M) and validation (CelebDF) datasets. The influence of conceptual Framework #5 (Model validation) is evident, emphasizing the importance of distinct training and evaluation datasets. Notably, ICT performed the poorest on DFDC, a completely new dataset, highlighting the challenges associated with generalization.

(b) Black-box evaluation: Multiple detectors from diverse groups, they are spatial artifacts (MAT & ICT) and spatiotemporal artifacts (CCViT), demonstrated superior results. A common thread among these successful models was the incorporation of a multihead attention mechanism, emphasizing the influence of conceptual Framework #4 (Model & Training). Additionally, the consistent utilization of spatial artifacts played a pivotal role in enhancing detection performance across various detectors and groups, emphasizing conceptual Framework #2 (Detection Methodology).

(c) White-box evaluation: LGrad exhibited a significant drop in AUC. This decline can be attributed to LGrad's training on synthetically generated fake images using GAN/Diffusion techniques, in contrast to the faceswap and reenactment methods employed in our white-box setup. This outcome demonstrates the usefulness of conceptual Framework #1 (Deepfake Type), emphasizing the critical role of targeted deepfake types. Moreover, our framework can capture the nuanced interplay of the identified influential factors within different evaluation strategies, shedding light on the perplexity of model performance in various scenarios.

7. Discussion

Reproducing SOTA is Challenging. Examining more than 50 deepfake detection models published in leading venues from 2019 to 2023 reveals a worrying trend. Only 15 (30%) of these models publicly released their pre-trained models. This lack of transparency, evident in the remaining 70%, hampers reproducibility and limits understanding of their actual limitations, thereby obstructing effective comparative analysis. This accessibility issue slows down the evaluation of different methodologies, potentially hindering progress in deepfake detection. Promoting the release of pre-trained models is vital for enhancing comparative studies, acceler-

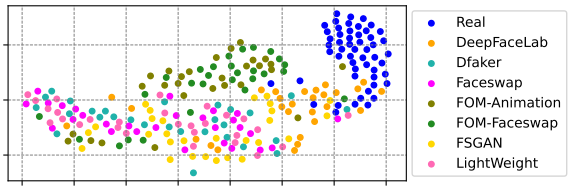


Figure 6. t-SNE visualization of 7 stabilized datasets (and *Real* part) with AltFreezing method. Each dot corresponds to one video representation.

ating advancements, and ensuring the robustness of these methodologies in real-world applications.

Real-World Deepfake Detection is still an Open Issue. It is evident from our results that no single detector consistently excels across all categories within our proposed three-tiered evaluation framework (black, gray, and white boxes). Although many detectors claim generalizability based on gray-box evaluations, our findings suggest they are proficient mainly in specific scenarios. Detectors tailored for certain deepfake types, like faceswap or reenactment, often falter when identifying other synthetic variants. Moreover, the difficulty of training on one dataset and testing on another poses a significant challenge, potentially invalidating the generalizability claims of these detectors in the broader context of deepfake detection. To visually illuminate these distinctions, we employ dimensionality reduction via t-distributed stochastic neighbor embedding (t-SNE) to illustrate the divergent characteristics of samples from seven datasets, as perceived by the model (See Fig. 6). Our findings demonstrate the need for a more comprehensive evaluation of generalizability, advocating for a thorough examination through our proposed evaluation strategy.

Synthesis Deepfake Type is Overlooked. Among the 51 detectors we examined, more than 96% focus mainly on reenactment or faceswap detection. Notably, the newly introduced diffusion models are synthesis-based, and limited research has been conducted on devising effective detection mechanisms for them. Although some initial efforts have been made to identify fully synthetic images generated by diffusion models, this avenue is still in its infancy and requires substantial attention [126]. The prospect of developing meta-detectors capable of distinguishing between faceswap, reenactment, and synthesis could serve as an initial triaging step in the detection process.

Influential Factors. While we meticulously identify the influential factors that researchers consider in developing their detectors and outline the impact of many as use cases, quantifying the individual influence of each factor on bottom-line efficacy across all settings proves challenging. This difficulty arises from the inherently predictive nature of AI-based models and the complexities associated with retraining each detector with diverse combinations of influential factors, owing to insufficient details on their construction. Addressing this challenge remains an open avenue for future exploration. Nevertheless, the comprehensive identification of these factors holds significant value, offering the potential to enhance the qualification of methods and detectors by leveraging a full understanding of these critical elements.

Ethical Considerations. We emphasize the careful and ethical creation of deepfake datasets, the utilization of existing datasets, and the incorporation of deepfake creation and detection tools from the research community. Our approach has been thoroughly reviewed and approved by the ethics review boards of our respective organizations, reflecting our unwavering commitment to maintaining high ethical standards in our research endeavors.

Future Directions. We propose four strategic directions to combat the proliferation of deep-fakes effectively.

ⓐ **Open Detectors and Three-Level Evaluations.** We advocate for researchers to release their developed detector models and subject them to a thorough evaluation using gray-box, white-box, and black-box assessments. This approach ensures the validation of generalizability, promoting transparency and reliability in deepfake detection.

ⓑ **Multimodal Detection.** Future research in deepfake detection should transcend reliance on a single data source. Instead, we encourage the exploration of multimodal models that integrate various cues, including audio, language models, visual elements, and metadata analysis. This comprehensive approach harnesses the synergistic effect of combining multiple data types, thereby enhancing detection accuracy and robustness.

ⓒ **Proactive Rather Than Reactive.** An essential research direction involves moving beyond reactive deepfake detection strategies. Instead, we propose the development of unique fingerprinting techniques for deepfake media, enabling the proactive tracing of their origin. This approach facilitates the identification and tracking of deepfake sources, allowing for their proactive removal and significantly mitigating the spread of misinformation through deepfakes.

ⓓ **Holistic Approach.** Effectively addressing the deepfake challenge requires a multi-faceted approach. This includes the integration of advanced detection technologies, provenance tracking methods, comprehensive public education to raise awareness, and robust government policies to regulate usage. By synthesizing these diverse strategies, we can establish a resilient and comprehensive defense against the manipulation and misuse of deepfake technology.

Concluding Thoughts. We believe deepfakes are becoming a more serious and real threat, as deepfakes will continue to grow and emerge in scale with more complexity and sophistication. Therefore, there is an immediate need to examine various deepfake detection tools and understand their limitations by thoroughly analyzing and evaluating them. Our work aims to address this gap by providing a comprehensive framework and evaluation of the current deepfake detection research. We find that there is a considerable gap between the claimed efficacy and the actual performance of the deepfake detectors. Future research should focus on developing more generalized detection methods, which can be based on gray, black, and white-box settings. Also, more proactive defenses, such as deepfake generation suppression and watermarking sources, should be researched to defend against ever-evolving deepfakes. Also, we hope that our

proposed framework and evaluation methodology can be readily extendable to handle new emerging deepfakes.

References

[1] B. M. Le and S. Woo, "Quality-agnostic deepfake detection with intra-model collaborative learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[2] C. Feng, Z. Chen, and A. Owens, "Self-supervised video forensics by audio-visual anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[3] S. Tariq, S. Jeon, and S. Woo, "Am i a real or fake celebrity? measuring commercial face recognition web apis under deepfake impersonation attack," *arXiv preprint arXiv:2103.00847*, 2021.

[4] Z. Wang, J. Bao, W. Zhou, W. Wang, and H. Li, "Altfreezing for more general video face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[5] S. Tariq, S. Jeon, and S. S. Woo, "Evaluating trustworthiness and racial bias in face recognition apis using deepfakes," *Computer*, vol. 56, no. 5, pp. 51–61, 2023.

[6] DeepFaceLab GitHub Community, "Deepfacelab," https://github.com/iperov/DeepFaceLab, 2023, accessed: 2023-01-01.

[7] FaceSwap GitHub Community, "Faceswap," https://github.com/MarekKowalski/FaceSwap, 2016, accessed: 2021-01-01.

[8] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Conference on Neural Information Processing Systems*, 2019.

[9] Federal Bureau of Investigation (FBI), "Deepfakes and stolen pii utilized to apply for remote work positions," https://www.ic3.gov/Media/Y2022/PSA220628, 2022, accessed: 2022-07-01.

[10] S. Tariq, A. Abuadbba, and K. Moore, "Deepfake in the metaverse: Security implications for virtual gaming, meetings, and offices," in *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*, ser. WDC '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 16–19. [Online]. Available: https://doi.org/10.1145/3595353.3595880

[11] C. Li, L. Wang, S. Ji, X. Zhang, Z. Xi, S. Guo, and T. Wang, "Seeing is living? rethinking the security of facial liveness verification in the deepfake era," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022.

[12] The Open Worldwide Application Security Project (OWASP), "Broken authentication," https://owasp.org/API-Security/editions/2023/en/0xa2-broken-authentication/, 2023.

[13] W. Bai, Y. Liu, Z. Zhang, B. Li, and W. Hu, "Aunet: Learning relations between action units for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[14] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[15] S. Tariq, S. Lee, and S. Woo, "One detector to rule them all: Towards a general deepfake attack detection framework," in *Proceedings of the web conference 2021*, 2021.

[16] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *European conference on computer vision*. Springer, 2020.

[17] L. Song, Z. Fang, X. Li, X. Dong, Z. Jin, Y. Chen, and S. Lyu, "Adaptive face forgery detection in cross domain," in *European Conference on Computer Vision*. Springer, 2022.

[18] B. M. Le and S. S. Woo, "Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images," in *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, 2022.

[19] C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei, "Learning on gradients: Generalized artifacts representation for gan-generated images detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[20] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[21] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "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*, 2021.

[22] J. Pu, N. Mangaokar, L. Kelly, P. Bhattacharya, K. Sundaram, M. Javed, B. Wang, and B. Viswanath, "Deepfake videos in the wild: Analysis and detection," in *Proceedings of the Web Conference 2021*, 2021.

[23] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[24] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[25] B. Le, S. Tariq, A. Abuadbba, K. Moore, and S. Woo, "Why do facial deepfake detectors fail?" in *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*, ser. WDC '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 24–28. [Online]. Available: https://doi.org/10.1145/3595353.3595882

[26] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[27] L. Yuezun, Y. Xin, S. Pu, Q. Honggang, and L. Siwei, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, 2020.

[28] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys (CSUR)*, 2021.

[29] J. W. Seow, M. K. Lim, R. C. Phan, and J. K. Liu, "A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities," *Neurocomputing*, vol. 513, pp. 351–371, 2022.

[30] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "Deepfakebench: A comprehensive benchmark of deepfake detection," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[31] B. Cho, B. M. Le, J. Kim, S. Woo, S. Tariq, A. Abuadbba, and K. Moore, "Towards understanding of deepfake videos in the wild," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 4530–4537.

[32] L. Verdoliva, "Media forensics and deepfakes: an overview," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[33] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, 2020.

[34] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering malicious deepfakes: Survey, battleground, and horizon," *International journal of computer vision*, 2022.

[35] S. A. Khan and D. T. Dang Nguyen, "Deepfake detection: A comparative analysis," *arXiv preprint arXiv:2308.03471*, 2023.

[36] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE access*, 2022.

[37] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, 2022.

[38] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "Deepfake detection for human face images and videos: A survey," *Ieee Access*, 2022.

[39] A. Naitali, M. Ridouani, F. Salahdine, and N. Kaabouch, "Deepfake attacks: Generation, detection, datasets, challenges, and research directions," *Computers*, 2023.

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, 2020.

[41] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[42] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[43] Y. LeCun, "Phd thesis: Modeles connexionnistes de l'apprentissage (connectionist learning models)," 1987.

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

[45] D. O'Sullivan and J. Passantino, "'verified' twitter accounts share fake image of 'explosion' near pentagon, causing confusion," https://edition.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/index.html, 2023, accessed: 2023-05-31.

[46] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[47] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "Fakeavceleb: A novel audio-video multimodal deepfake dataset," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[48] H. Khalid, M. Kim, S. Tariq, and S. S. Woo, "Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors," in *Proceedings of the 1st Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection*, 2021, pp. 7–15.

[49] DeepFakes GitHub Community, "Deepfakes," https://github.com/deepfakes/faceswap, 2017, accessed: 2021-01-01.

[50] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *arXiv preprint arXiv:1912.13457*, 2019.

[51] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[52] Druuzil Tech & Games - Youtube Channel, "Scarlett johansson in "moulin rouge!" - deepfake," https://www.youtube.com/watch?v=E5VdGFjbf8E, 2023, accessed: 2023-05-01.

[53] DFaker GitHub Community, "Dfaker," https://github.com/dfaker/df, 2017.

[54] R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020. [Online]. Available: https://doi.org/10.1145%2F3394171.3413630

[55] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[56] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[57] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Transactions on Graphics (TOG)*, 2019.

[58] Media Education Lab, "Belgium climate politics - trump deep fake," https://www.youtube.com/watch?v=8o0iOm-2sLw, 2021, accessed: 2023-05-01.

[59] MIT's Center for Advanced Virtuality, "In event of moon disaster," https://moondisaster.org/, 2020, accessed: 2023-05-01.

[60] Andrew, Stable Diffusion Art, "Fine-tune your ai images with these simple prompting techniques," https://stable-diffusion-art.com/fine-tune-your-ai-images-with-these-simple-prompting-techniques/, 2022, accessed: 2023-05-01.

[61] L. Zheng, Y. Zhang, and V. L. Thing, "A survey on image tampering and its detection in real-world photos," *Journal of Visual Communication and Image Representation*, 2019.

[62] M. Kim, S. Tariq, and S. S. Woo, "Cored: Generalizing fake media detection with continual representation using distillation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 337–346.

[63] S. Tariq, S. Lee, and S. S. Woo, "A convolutional lstm based residual network for deepfake video detection," *arXiv preprint arXiv:2009.07480*, 2020.

[64] S. Lee, S. Tariq, J. Kim, and S. S. Woo, "Tar: Generalized forensic framework to detect deepfakes using weakly supervised learning," in *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer, 2021, pp. 351–366.

[65] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo, "Gan is a friend or foe?: a framework to detect various fake face images," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. ACM, 2019, pp. 1296–1303.

[66] J. Kim, S. Tariq, and S. S. Woo, "Ptd: Privacy-preserving human face processing framework using tensor decomposition," in *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 2022, pp. 1296–1303. [Online]. Available: https://doi.org/10.1145/3477314.3507036

[67] S. Lee, S. Tariq, Y. Shin, and S. S. Woo, "Detecting handcrafted facial image manipulations and gan-generated facial images using shallow-fakefacenet," *Applied Soft Computing*, vol. 105, p. 107256, 2021.

[68] M. Kim, S. Tariq, and S. S. Woo, "Fretal: Generalizing deepfake detection using knowledge distillation and representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1001–1012.

[69] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo, "Detecting both machine and human created fake face images in the wild," in *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*. ACM, 2018, pp. 81–87.

[70] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo, "Protecting celebrities from deepfake with identity consistency transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[71] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[72] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[73] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, "Local relation learning for face forgery detection," in *Proceedings of the AAAI conference on artificial intelligence*, 2021.

[74] J. Hu, X. Liao, J. Liang, W. Zhou, and Z. Qin, "Finfer: Frame inference-based deepfake detection for high-visual-quality videos," in *Proceedings of the AAAI conference on artificial intelligence*, 2022.

[75] T. Wang and K. P. Chow, "Noise based deepfake detection via multi-head relative-interaction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[76] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of cnns," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021.

[77] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection," in *International conference on image analysis and processing*. Springer, 2022.

[78] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[79] Z. Sun, Y. Han, Z. Hua, N. Ruan, and W. Jia, "Improving the efficiency and robustness of deepfakes detection through precise geometric features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[80] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[81] J. Fei, Y. Dai, P. Yu, T. Shen, Z. Xia, and J. Weng, "Learning second order local anomaly for general face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[82] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[84] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, 1997.

[85] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[86] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng, "Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[87] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[88] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," *Advances in neural information processing systems*, 1993.

[89] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied intelligence*, 2023.

[90] C. Wang and W. Deng, "Representative forgery mining for fake face detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[91] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[92] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017.

[93] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[94] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, 2009.

[95] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, 2016.

[96] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[97] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *Advances in neural information processing systems*, 2017.

[98] L. Chen, Y. Zhang, Y. Song, J. Wang, and L. Liu, "Ost: Improving generalization of deepfake detection via one-shot test-time training," *Advances in Neural Information Processing Systems*, 2022.

[99] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[100] J. Stehouwer, H. Dang, F. Liu, X. Liu, and A. Jain, "On the detection of digital face manipulation," *arXiv*, 2019.

[101] Y. Ni, D. Meng, C. Yu, C. Quan, D. Ren, and Y. Zhao, "Core: Consistent representation learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[102] B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang, and D. Ye, "Implicit identity driven deepfake face swapping detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[103] Y. Xu, K. Raja, L. Verdoliva, and M. Pedersen, "Learning pairwise interaction for generalizable deepfake detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

[104] K. Sun, H. Liu, Q. Ye, Y. Gao, J. Liu, L. Shao, and R. Ji, "Domain general face forgery detection by learning to weight," in *Proceedings of the AAAI conference on artificial intelligence*, 2021.

[105] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji, "Dual contrastive learning for general face forgery detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[106] S. Cao, Q. Zou, X. Mao, D. Ye, and Z. Wang, "Metric learning for anti-compression facial forgery detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[107] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[108] W. Zhuang, Q. Chu, Z. Tan, Q. Liu, H. Yuan, C. Miao, Z. Luo, and N. Yu, "Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection," in *European Conference on Computer Vision*. Springer, 2022.

[109] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020.

[110] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, and J. Zhao, "Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms," in *Proceedings of the 28th ACM international conference on multimedia*, 2020.

[111] X. Li, Y. Lang, Y. Chen, X. Mao, Y. He, S. Wang, H. Xue, and Q. Lu, "Sharp multiple instance learning for deepfake video detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020.

[112] D. Zhang, C. Li, F. Lin, D. Zeng, and S. Ge, "Detecting deepfake videos with temporal dropout 3dcnn." in *IJCAI*, 2021.

[113] Z. Hu, H. Xie, Y. Wang, J. Li, Z. Wang, and Y. Zhang, "Dynamic inconsistency-aware deepfake video detection." in *IJCAI*, 2021, pp. 736–742.

[114] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, and L. Ma, "Delving into the local: Dynamic inconsistency learning for deepfake video detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[115] Z. Gu, T. Yao, Y. Chen, S. Ding, and L. Ma, "Hierarchical contrastive inconsistency learning for deepfake video detection," in *European Conference on Computer Vision*. Springer, 2022.

[116] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, F. Huang, and L. Ma, "Spatiotemporal inconsistency learning for deepfake video detection," in *Proceedings of the 29th ACM international conference on multimedia*, 2021.

[117] J. Guan, H. Zhou, Z. Hong, E. Ding, J. Wang, C. Quan, and Y. Zhao, "Delving into sequential patches for deepfake detection," *Advances in Neural Information Processing Systems*, 2022.

[118] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "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*, 2021.

[119] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020.

[120] X. Zhu, H. Wang, H. Fei, Z. Lei, and S. Z. Li, "Face forgery detection by 3d decomposition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[121] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019.

[122] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[123] DeepFaker Application, "Deepfaker," https://deepfaker.app/, 2021, accessed: 2023-01-01.

[124] N. Dufour and A. Gully, "Contributing data to deepfake detection research," https://blog.research.google/2019/09/contributing-data-to-deepfake-detection.html, 2019, accessed: 2023-01-01.

[125] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016.

[126] J. Ricker, S. Damm, T. Holz, and A. Fischer, "Towards the detection of diffusion model deepfakes," *arXiv preprint arXiv:2210.14571*, 2022.

[127] FacePlay Application, "Faceplay app," https://www.faceplay.cc/, 2021, accessed: 2023-01-01.

[128] DeepFakesWeb Site, "Deepfakesweb," https://deepfakesweb.com/, 2021, accessed: 2023-01-01.

[129] DeepFaceLive, "Deepfacelive," https://drive.google.com/file/d/1KS37b2IBuljJuZiJsgnWuzs7Y5OfkOyI/view/, 2023, accessed: 2023-01-01.

[130] FaceApp, "Faceapp," https://www.faceapp.com/, 2021, accessed: 2023-01-01.

[131] Reface Application, "Reface," https://reface.app/, 2021, accessed: 2023-01-01.

[132] S. A. Lu, "faceswap-gan," https://github.com/shaoanlu/faceswap-GAN, 2023, accessed: 2023-01-01.

[133] Revive Application, "Revive," https://play.google.com/store/apps/details?id=revive.app&hl=en&gl=US/, 2021, accessed: 2023-01-01.

[134] Fakeit Application, "Fakeit," https://vk.com/fakeit/, 2021, accessed: 2023-01-01.

[135] DeepFaker Bot Site, "Deepfakerbot," https://t.me/DeepFakerBot/, 2021, accessed: 2023-01-01.

[136] Revel AI BV, "Revelai," http://revel.ai/, 2021, accessed: 2023-01-01.

[137] R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, oct 2020. [Online]. Available: https://doi.org/10.1145%2F3394171.3413630

[138] LicoLico Application, "Licolico," http://licolico.cn/home/, 2021, accessed: 2023-01-01.

[139] DeepfakeStudio Application, "Deepfakestudio," https://play.google.com/store/apps/details?id=com.deepworkings.dfstudio&hl=en&gl=US&pli=1/, 2021, accessed: 2023-01-01.

[140] Deepcake.io Site, "Deepcakeio," http://deepcake.io/, 2021, accessed: 2023-01-01.

Appendix A. Deepfake Generation Tools

Table 4. **DEEPFAKE GENERATION TOOLS.** OUR SELECTED GENERATORS ARE HIGHLIGHTED IN GREEN. IN THE TABLE, THE 'BEING SERVICED' COLUMN INDICATES WHETHER THE PROGRAM IS OUTDATED, MARKED WITH THE LAST UPDATE YEAR INFORMATION BESIDE IT. NOTABLY, THE LIGHTWEIGHT MODEL, AVAILABLE AS AN ALTERNATIVE IN THE FACESWAP SOURCE, IS INCLUDED AS ANOTHER OPTION. † INDICATES THAT THE PROGRAM IS OFFICIALLY CLOSED.

Program	Open Source	Star No.	Fork No.	Being Seviced	Freedom of Victim&Driver	Score (max:6)
FacePlay [127]	✗	-	-	✓(2023)	✓ ✗	2
DeepFakesWeb [128]	✗	-	-	✓	✓ ✓	3
DeepFaceLab [6]	✓	42k	9.4k	✓(2022)	✓ ✓	6
DeepFaceLive [129]	✓	17.1k	2.5k	✓(2023)	✗ ✓	5
FaceApp [130]	✗	-	-	✓(2023)	✓ ✗	2
Reface [131]	✗	-	-	✓(2023)	✓ ✗	2
Dfaker [53]	✓	461	151	✓(2020)	✓ ✓	6
Faceswap [7]	✓	46.7k	12.6k	✓(2023)	✓ ✓	6
LightWeight [7]	✓	46.7k	12.6k	✓(2023)	✓ ✓	6
deepfakes's faceswap [49]	✓	3k	1k	✗(2018)	✓ ✓	5
Faceswap-GAN [132]	✓	3.3k	840	✗(2019)	✓ ✓	5
FOM-Animation [8]	✓	13.7k	3.1k	✓(2023)	✓ ✓	6
FOM-Faceswap [8]	✓	13.7k	3.1k	✓(2023)	✓ ✓	6
FSGAN [51]	✓	702	143	✓(2023)	✓ ✓	6
DeepFaker [123]	✗	-	-	✓(2023)	✓ ✗	2
Revive [133]	✗	-	-	✓(2023)	✓ ✗	2
Fakeit [134]	✗	-	-	✗	✗ ✗	0
DeepFaker Bot [135]	✗	-	-	✗	✓ ✓	2
Revel.ai [136]	✗	-	-	✓(2023)	✓ ✓	3
SimSwap [137]	✓	2	703	✓(2023)	✓ ✓	5
licolico [138]	✗	-	-	✗†	✓ ✗	1
Deepfake Studio [139]	✗	-	-	✓(2023)	✓ ✓	3
Deepcake.io [140]	✗	-	-	✗†	✗ ✗	0

Appendix B. Further Details on Detectors

To structurally overview the overall published deepfake detectors, we introduce a new categorization methodology in Table 5. All of the appropriate deepfake detection methods provide some mutual series of processes and components to build their deep learning network. Based on this knowledge, a deepfake detector could be easily broken down into several key components: *Deepfakes for Training, Artifacts, Data Pre-processing, Model Training, and Model Validation*.

Table 5. **FURTHER DETAILS ON DETECTORS.** F2F, NT, AND FOM STANDS FOR FACIAL REENACTMENT METHODS FACE2FACE AND NEURAL TEXTURES. FS, DF, FSGAN, AND FASH STAND FOR FACE SWAP METHODS FACESWAP, DEEPFAKE, FACESWAPGAN, AND FACESHIFTER.

Paper Name	Deepfakes for Training	Artifacts	Data Pre-processing	Model Training	Model Validation
Cap.Forensics	F2F, DF	Spatial	Single Frame	VGG	F2F, DF
XceptionNet	NT, F2F, FS, DF	Spatial	dlib,Single Frame	XceptionNet	NT, F2F, FS, DF
Face X-ray	NT, F2F, FS, DF	Spatial	Single Frame	HRNet	NT, F2F, FS, DF, GAN
FFD	NT, F2F, FS, DF, GAN	Spatial	InsightFace, Single Frame	XceptionNet, VGG	NT, F2F, FS, DF, GAN
RECCE	NT, F2F, FS, DF	Spatial	RetinaFace, Single Frame	XceptionNet	NT, F2F, FS, DF, GAN
CORE	NT, F2F, FS, DF, GAN	Spatial	MTCNN, Single Frame	XceptionNet	NT, F2F, FS, DF, GAN
IID	NT, F2F, FS, DF	Spatial	RetinaFace, Single Frame	ResNet	NT, F2F, FS, DF, FaSh, GAN
MCX-API	NT, F2F, FOM, FS, DF	Spatial	MTCNN, Single Frame	XceptionNet	NT, F2F, FOM, FS, DF, FSGAN, GAN
EffB4Att	NT, F2F, FS, DF, GAN	Spatial	BlazeFace, Single Frame	EfficientNet, Siamese	NT, F2F, FS, DF, GAN
LTW	NT, F2F, FS, DF	Spatial	MTCNN, Single Frame	EfficientNet	NT, F2F, FS, DF + GAN
MAT	NT, F2F, FS, DF	Spatial	RetinaFace, Single Frame	EfficientNet	NT, F2F, FS, DF + GAN
DCL	NT, F2F, FS, DF	Spatial	DSFD, Single Frame	EfficientNet	NT, F2F, FS, DF + GAN
SBIs	FS	Spatial	dlib, RetinaFace, Single Face, Single Frame	EfficientNet	NT, F2F, FS, DF, FSGAN, GAN
MLAC	NT, F2F, FS, DF	Spatial	dlib, Single Frame	XceptionNet, GAN learning	NT, F2F, FS, DF
FRDM	NT, F2F, FS, DF	Spatial	dlib, Single Frame	XceptionNet, Dual Cross Modal Attention	NT, F2F, FS, DF, GAN,VAE
OST	NT, F2F, FS, DF	Spatial	dlib, Single Frame	XceptionNet, Meta Training	NT, F2F, FS, DF, GAN,VAE
CADDM	NT, F2F, FS, DF, FaSh	Spatial	MTCNN, Single Frame	ResNet, EfficientNet	NT, F2F, FS, DF, FaSh, GAN
QAD	NT, F2F, FS, DF, FaSh	Spatial	dlib, Single Frame	ResNet, EfficientNet, Collaborative learning	NT, F2F, FS, DF, FaSh, GAN
ICT	FS	Spatial	RetinaFace, Self-blend on real image, Single Frame	Vision Transformer	NT, F2F, FS, DF, GAN,VAE
UIA-ViT	NT, F2F, FS, DF	Spatial	dlib, Single Frame	Vision Transformer	NT, F2F, FS, DF, GAN
AUNet	NT, F2F, FS, DF	Spatial	dlib, RetinaFace, Single Frame	Vision Transformer	NT, F2F, FS, DF, FSGAN, GAN
ADDNet-3d	FS, DF, GAN	Spatial, Temporal	MTCNN, Multiple Frames	Convolutional layers	FS, DF, GAN
DeepRhythm	NT, F2F, FS, DF, GAN	Spatial, Temporal	dlib, MTCNN, Multiple Frames	ResNet	NT, F2F, FS, DF, GAN
S-IML-T	NT, F2F, FS, DF, GAN	Spatial, Temporal	dlib, MTCNN, Multiple Frames	XceptionNet	NT, F2F, FS, DF, GAN
TD-3DCNN	NT, F2F, FS, DF	Spatial, Temporal	MobileNet, Multiple Frames	3D Inception	NT, F2F, FS, DF, GAN
DIA	NT, F2F, FS, DF, GAN	Spatial, Temporal	RetinaFace, 4 keypoints, Multiple Frames	ResNet	NT, F2F, FS, DF, GAN
DIL	NT, F2F, FS, DF, GAN	Spatial, Temporal	dlib, MTCNN, Multiple Frames	ResNet	NT, F2F, FS, DF, GAN
FInfer	NT, F2F, FS, DF	Spatial, Temporal	dlib, Multiple Frames	Convolutional layers	NT, F2F, FS, DF, GAN,VAE
HCIL	NT, F2F, FS, DF	Spatial, Temporal	dlib, MTCNN, Multiple Frames, Multiple Frames	ResNet	NT, F2F, FS, DF, GAN
AltFreezing	NT, F2F, FS, DF	Spatial, Temporal	Temporal drop, Temporal repeat, Self-blend on real, Multiple Frames	3D ResNet	NT, F2F, FS, DF, FaSh, VAE
STIL	NT, F2F, FS, DF, GAN	Spatial, Temporal	MTCNN, Multiple Frames	ResNet, Spatial-Temporal Inconsistency	NT, F2F, FS, DF, GAN
LipForensics	NT, F2F, FS, DF, FaSh	Spatial, Temporal	RetinaFace, Face Alignment, Cropping Mouths, Multiple Frames	ResNet Temporal CNN	NT, F2F, FS, DF, FaSh, GAN, VAE
FTCN	NT, F2F, FS, DF	Spatial, Temporal	InsightFace, Face Alignment, Multiple Frames	3D ResNet, Transformer Encoder	NT, F2F, FS, DF, FaSh, VAE
CCViT	FS, DF, GAN	Spatial	MTCNN, Single Frame	EfficientNet, Vision Transformer	NT, F2F, FS, DF, FaSh, GAN
CLRNet	NT, F2F, FS, DF, GAN	Spatial, Temporal	MTCNN, Multiple Frames	3D ResNet	NT, F2F, FS, DF, GAN
LTTD	NT, F2F, FS, DF	Spatial, Temporal	MTCNN, Multiple Frames	Vision Transformer	NT, F2F, FS, DF, FaSh, GAN,VAE
F3-Net	NT, F2F, FS, DF	Frequency	F2F (RGB Tracking), Single Frame	XceptionNet	NT, F2F, FS, DF
FDFL	NT, F2F, FS, DF	Spatial, Frequency	RetinaFace, DCT transform, Single Frame	XceptionNet	NT, F2F, FS, DF
ADD	NT, F2F, FS, DF, FaSh	Spatial, Frequency	dlib, Single Frame	ResNet, Knowledge Distillation	NT, F2F, FS, DF, FaSh
LRL	NT, F2F, FS, DF	Spatial, Frequency	Single Frame	Convolutional layers	NT, F2F, FS, DF + GAN
TRN	NT, F2F, FS, DF	Spatial, Temporal, Frequency	dlib, Multiple Frames	DenseNet, BiLSTM	NT, F2F, FS, DF + GAN
SPSL	NT, F2F, FS, DF	Frequency	IDCT Transform, Single Frame	XceptionNet	NT, F2F, FS, DF
CD-Net	NT, F2F, FS, DF	Spatial, Temporal, Frequency	DCT and IDCT Transform, Multiple Frames	SlowFast	NT, F2F, FS, DF, GAN,VAE
SFDG	NT, F2F, FS, DF	Spatial, Frequency	dlib, Single Frame	Information Interaction layers, Graph CNN, U-Net, EfficientNet	NT, F2F, FS, DF, GAN,VAE
RFM	NT, F2F, FS, DF, GAN	Spatial, Forgery Attention Map	Suspicious Forgeries Erasing, Single Frame	XceptionNet	NT, F2F, FS, DF, GAN
FD2Net	NT, F2F, FS, DF	Spatial, 3D	3DDFA, Single Frame	XceptionNet	NT, F2F, FS, DF, GAN
SOLA	NT, F2F, FS, DF	Spatial, Frequency, Noise Traces	RetinaFace, ASRM, Single Frame	ResNet	NT, F2F, FS, DF, FaSh
AVAD	Real Video	Spatial, Temporal, Voice Sync	S3FD, Face Alignment, Multiple Frames	3D ResNet, VGG, Transformer Encoder	FOM, FS, FSGAN, GAN
LGrad	GAN	Gradient	Pre-trained StyleGAN, Single Frame	ResNet	NT, F2F, FS, DF, GAN
LRNet	NT, F2F, FS, DF	Temporal, Landmarks	dlib, Openface, Multiple Frames	LRNet	NT, F2F, FS, DF, VAE
NoiseDF	NT, F2F, FS, DF	Noise Traces	RIDNet, Single Frame	Siamese	NT, F2F, FS, DF, GAN,VAE