Deepfake Detection: A Comprehensive Survey from the Reliability Perspective

TIANYI WANG, Computer Science, The University of Hong Kong, Hong Kong, Hong Kong

XIN LIAO, Hunan University, Changsha, China

KAM PUI CHOW, The University of Hong Kong, Hong Kong, Hong Kong

XIAODONG LIN, University of Guelph, Guelph, Canada

YINGLONG WANG, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

The mushroomed Deepfake synthetic materials circulated on the internet have raised a profound social impact on politicians, celebrities, and individuals worldwide. In this survey, we provide a thorough review of the existing Deepfake detection studies from the reliability perspective. We identify three reliability-oriented research challenges in the current Deepfake detection domain: transferability, interpretability, and robustness. Moreover, while solutions have been frequently addressed regarding the three challenges, the general reliability of a detection model has been barely considered, leading to the lack of reliable evidence in real-life usages and even for prosecutions on Deepfake-related cases in court. We, therefore, introduce a model reliability study metric using statistical random sampling knowledge and the publicly available benchmark datasets to review the reliability of the existing detection models on arbitrary Deepfake candidate suspects. Case studies are further executed to justify the real-life Deepfake cases including different groups of victims with the help of the reliably qualified detection models as reviewed in this survey. Reviews and experiments on the existing approaches provide informative discussions and future research directions for Deepfake detection.

CCS Concepts: • **Security and privacy → Human and societal aspects of security and privacy**; • **Computing methodologies → Computer vision**; • **Applied computing → Computer forensics**; • **General and reference → Surveys and overviews**.

Additional Key Words and Phrases: Deepfake Detection, Reliability Study, Forensic Investigation, Confidence Interval

1 Introduction

In June 2022, a mother from Pennsylvania, known as a 'Deepfake mom', is sentenced to probation for three years because of her harassment of the rivals on her daughter's cheerleader team [87]. She is originally accused of using the so-called Deepfake technology to generate and spread fake videos of her daughter's opponents depicting indelicate behaviors in March 2021. However, as admitted by the prosecutors, the justification for generating Deepfake videos is unable to be confirmed without accurate evidence and tools [62]. The term Deepfake refers to a deep learning technique raised by the Reddit user 'deepfakes' [34] in 2017 [143] that could

automatically execute face-swapping from a source person to a target one while maintaining all other contents of the target image unchanged including expression, movement, and background scene. Later, the face reenactment technology [84, 169, 183, 191], which transfers attributes of a source face to a target one while maintaining the target's facial identity, is also classified as Deepfake from a comprehensive standpoint.

Deepfake can bring benefits and convenience to people's daily lives, especially from the perspective of human entertainment. Specifically, movie lovers may swap their faces onto movie clips and perform as their favorite superheroes and superheroines [89]. On the other hand, a tainted celebrity who is no longer allowed to appear on TV shows [161–163] can be face-swapped within completed TV productions instead of reshooting or removing the episode. Moreover, Deepfake is available for bringing a deceased person digitally back to life [5] and greeting families and friends with desired conversations. Besides, e-commerce is another scenario to exert positive effects of Deepfake by trying on clothes in an online fitting room [178].

Benefiting from the publicly available source code implementations on the internet, various Deepfake mobile applications [76, 96, 110] have been released. The early product in 2018, FakeApp [33], requires a large number of input images to achieve satisfactory synthetic results. Later in 2019, the popular Chinese application, ZAO [122], can generate face-swapping outputs by simply inputting a series of selfies. Recently, a more powerful facial synthetic tool, Reface [152], is built with further functionalities such as moving and singing animations based on hyper-realistic synthetic results. Although entertaining human lives, the free-access Deepfake applications require practically no experience in the field when generating fake faces, which therefore poses crucial potential threats to society. Because of the hyper-realistic quality that is indistinguishable by human eyes [126–128], Deepfake has already been ranked as the most serious artificial intelligence crime threat since 2020 [148], and the current and potential victims include politicians, celebrities, and even every human being on earth. Besides the experimental fake Obama[1], a fake president Zelensky [120] has caused panic in Ukraine during the Russia-Ukraine war. As for celebrities, fake porn videos have frequently targeted female actresses and representative victims include Emma Watson, Natalie Portman, and Ariana Grande [88, 97]. Additionally, remember the 'Deepfake mom' as discussed at the beginning? It is one of the best proofs that anyone can become a victim of Deepfake in modern society.

To protect individuals and society from the negative impacts of misusing Deepfake, Deepfake detection approaches have been frequently designed and they mostly conduct a binary classification task to identify real and fake faces with the help of deep neural networks (DNNs). In this survey, we provide an in-depth review of the developing Deepfake detection approaches from the reliability perspective. In other words, we focus on studies and topics that devote to the ultimate success of Deepfake detection in real-life usages, and more importantly, for criminal investigations in court-case judgments. Early studies mainly focus on the in-dataset model performance such that the detection models are trained and validated the performance on the same dataset. While most recent work has achieved promising detection performance for the in-dataset test, the research challenges at the current stage for Deepfake detection can be concluded in three aspects, namely, *transferability*, *interpretability*, and *robustness*. The transferability topic refers to the progress of improving the cross-dataset ability of models when evaluated on unseen data. As the detection performance keeps advancing by various approaches, interpretability is another research goal to explain the reason that the detection model determines the falsification. Moreover, when applying well-trained and well-performed detection models for real-life scenarios, robustness is considered a main topic in dealing with various real-life conditions.

While research has been incrementally attempted regarding the three challenging topics, a reliable Deepfake detection model is expected to benefit people's daily lives with good transferability on unseen data, have convincing interpretability of the detection decision, and show robust performance against practical application scenarios and conditions in real life. However, there lacks further discussion on model reliability in existing research papers and surveys. In particular, without an authenticated scheme to nominate the detection models

---

[1]https://www.youtube.com/watch?v=cQ54GDm1eL0

as reliable evidence to assist prosecutions and judgments in court, similar failures as the 'Deepfake mom' case will happen again due to the lack of reliable detection tools to support the accusation of Deepfake even though detection performance on each benchmark dataset is reported in current work. In other words, the trustworthiness of a model-derived falsification needs to be proved before it can convince people in real-life usages and for court-case judgments. To fulfill the research gap of the model reliability study, beyond the comprehensive review of Deepfake detection, we devise a scheme to scientifically validate the reliability of the well-developed Deepfake detection models using statistical random sampling knowledge [117]. To guarantee the credibility of the reliability study, we concurrently introduce a systematic workflow of data pre-processing including image frame selection and extraction from videos and face detection and cropping, which has been barely mentioned with concrete details in past work. Thereafter, we quantitatively evaluate and record the selected state-of-the-art Deepfake detection approaches by training and testing with their reported optimal settings on the same group of pre-processed datasets in a completely fair game. Thenceforth, we validate the detection model reliability following the designed evaluation scheme. In the end, a case study is enforced to justify the results derived by the detection models on four well-known real-life synthetic videos concerning the reliable detection accuracies statistically at 90% and 95% confidence levels based on the research outcomes from the model reliability study. Interesting findings and future research topics that have been scarcely concluded in previous studies [42, 89, 121, 164, 165, 178] are analyzed and discussed. Furthermore, we believe that the proposed Deepfake detection model reliability study scheme is informative and can be adopted as evidence to assist prosecutions in court once granted approval by authentication experts or institutions following the legislation.

The rest of the paper is organized as follows. We provide a brief review of the popular synthetic techniques and publicly available benchmark datasets in Section 2. Then, we define the challenges of the Deepfake detection research and provide a thorough review of the development history of the Deepfake detection approaches in Section 3. In Section 4, we illustrate the model reliability study scheme and demonstrate the algorithm details. In Section 5, we detailedly introduce a standardized data pre-processing workflow and list the participating datasets in the experiments of this paper. In Section 6, we conduct detection performance evaluation and reliability justification using selected state-of-the-art models on the benchmark datasets and following the reliability study scheme, respectively. Section 7 exhibits Deepfake detection results of the selected models when applying to the real-life videos in a case study, and discussions along with experiment results from early sections are presented in Section 8. Section 9 concludes the remarks and highlights the potential future directions in the research domain.

2 The Evolution of Deepfake Generation

2.1 Deepfake Generation

Deepfake is initially raised in the Reddit community when the open-source implementation was first published by the user 'deepfakes' simultaneously in 2017. Early research mainly focuses on subject-specific approaches, which can only swap facial identities that the models have seen during training. The most popular framework of the existing Deepfake synthesis studies [34, 137] for the subject-specific identity swap is an autoencoder [91]. In a nutshell, the autoencoder contains a shared encoder that extracts identity-independent features from the source and target faces and two unique decoders each is in charge of reconstructing synthetic faces of a corresponding facial identity. Specifically, in the training phase, faces of the source identity are fed to the encoder for identity-independent feature extraction. The extracted context vectors are then passed through the decoder that corresponds to reconstructing the faces of the source identity. Similarly, the target face reconstruction is trained following the same workflow. When using a well-trained model to operate face-swapping, a target face after context vector extraction is fed to the decoder that reconstructs the source identity. Thenceforth, the decoder generates a look maintaining the facial expression and movement of the target face while having the

identity of the desired source face. If the face-swapping model is trained for both directions, a target face may be face-swapped onto a source face following the same workflow.

Recent studies gradually focus on subject-agnostic methods to enable face-swapping for arbitrary identities with higher resolutions and have exploited Generative Adversarial Networks (GAN) [50] for better synthesis authenticity [10, 23, 36, 49, 85, 86, 123, 129, 134, 188, 195]. In other words, they aim to consistently produce high-quality face-swapping results even on facial identities that are unseen during model training. GAN is a generator-discriminator architecture that is trained by having two components battle against each other to advance the output quality. In practice, the generator is periodically trained to fool the discriminator with synthetic faces. For instance, FaceShifter [100] and SimSwap [19] each devise particular modules to preserve facial attributes that are hard to reconstruct and maintain fidelity for arbitrary facial identities. MegaFS [196] and HifiFace [176] accomplish face-swapping at high resolutions of 512 and 1,024 for arbitrary facial identities, respectively, relying on the reconstruction ability of GAN. In the last step, the generated fake face is usually blended back to the pristine target image with tuning techniques such as blurring and smoothing [190] to reduce the visible Deepfake traces.

2.2 Deepfake Benchmark Datasets

Benchmark datasets are vital in the development history of Deepfake detection models. Dolhansky et al. [197] raised the idea to break down previous datasets into three generations based on the two-generation categorization in the early work [81]. As listed in Table 1, UADFV [186], DeepfakeTIMIT [93], and FaceForensics++ (FF++) [144] are categorized to the first generation; Deepfake Detection Dataset [57], DeepFake Detection Challenge (DFDC) Preview [38], and Celeb-DF [104] are in the second generation; DeeperForensics-1.0 (DF1.0) [81] and DeepFake Detection Challenge (DFDC) [37] are in the third generation. In summary, later generations contain general improvements over the previous ones in terms of dataset magnitude or synthetic method diversity.

Unlike the summary by Dolhansky et al. [197], the agreement from individuals appearing is not considered in this study, and we instead re-define the third generation such that the datasets are of better quality, broader diversity and magnitude, higher difficulty than the early generations, or challenging detailed discrepancies in early synthetic videos are resolved in the datasets. Specifically, besides DFDC with large manipulation diversity and dataset magnitude and DF1.0 with large magnitude and considerable difficulty by adding deliberate perturbations, we further classify the following datasets in the third generation, namely, FaceShifter [100], WildDeepfake [197], and KoDF [95].

FaceShifter [100], although synthesized based on real videos of FF++, has specifically solved the so-called facial occlusion challenge that appears in previous datasets. In other words, the synthetic results are better handled even in difficult cases where parts of the face are blocked or obscured by objects such as accessories or other body parts such as hair tippings. WildDeepfake (WDF) [197] appears to be a special one in the third generation because videos are totally collected from the internet, which matches the real-life Deepfake circumstance the best. The most recent KoDF [95] dataset is so far the largest Deepfake benchmark dataset that is publicly available with reasonable diversity and contains synthetic videos at high resolutions.

3 Reliability-Oriented Challenges of Deepfake Detection

Detection work on Deepfake has been proposed since the first occurrence of Deepfake contents. Classical forgery detection approaches [28, 35, 43, 47, 133, 135] mainly focus on the intrinsic statistics and hand-crafted traces such as eye blinking [82, 102], head pose [186], and visual artifacts [119] to analyze the spatial feature manipulation

---

[2]FaceSwap, Deepfakes, Face2Face, and NeuralTextures.

[3]DF-128, DF-256, MM/NN, NTH, FSGAN, StyleGan, refinement, and audio swaps.

[4]FaceSwap, DeepFaceLab, FSGAN, FOMM, Audio-driven.

Table 1. Information of the existing Deepfake datasets categorized into three generations based on quality, diversity, and difficulty. Publication year, the number of real and fake sequences, and the source of real and fake materials are listed.

Dataset	Year	# Real / Fake	Real / Fake Source	Generation
UADFV [186]	2018	49 / 49	YouTube / FakeApp [33]	
DeepfakeTIMIT [93]	2018	− / 620	faceswap-GAN [111]	1st Generation
FF++ [144]	2019	1,000 / 4,000	YouTube / 4 methods[2]	
DFD [57]	2019	363 / 3,068	consenting actors / unknown methods	
DFDC Preview [38]	2019	1,131 / 4,119	crowdsourcing / 2 unknown methods	2nd Generation
Celeb-DF [104]	2019	590 / 5,639	YouTube / improved Deepfake	
DF1.0 [81]	2020	− / 10,000	FF++ real / DF-VAE	
FaceShifter [100]	2020	− / 1,000	FF++ real / GAN-based	
DFDC [37]	2020	23,654 / 104,500	crowdsourcing / 8 methods[3]	3rd Generation
WDF [197]	2020	3,805 / 3,509	video-sharing websites	
KoDF [95]	2021	62,166 / 175,776	lab-controlled / 6 manipulations[4]	

patterns. Besides, there are papers that have derived high accuracies and AUC scores by training and testing on the same dataset of a synthetic method. Several studies [60, 145] integrated CNN and Long Short-term Memory (LSTM) [68] for spatial and temporal features analyses, respectively, and accomplished detection for in-dataset evaluations on self-collected data and the FF++ dataset. Hsu, Zhang, and Lee [70] utilized GAN-generated fake samples for real-fake pairwise training using DenseNet [74]. Agarwal et al. [3] accomplished detection on face-swap Deepfake using CNNs with biometric information including facial expressions and head movements. However, although accomplished well-pleasing in-dataset detection performance on some early or self-collected datasets, they are mostly easily fooled by the hyper-realistic Deepfake contents in the current research domain because of limitations in dataset quality, dataset diversity, and method or model ability.

Later studies gradually consider Deepfake detection as a binary classification task using DNNs. As malicious Deepfake contents have started to jeopardize human society and the cases are even discussed in court, reliably trusted detection methods are eagerly desired by the public. In particular, three challenges (Fig. 1) of the current Deepfake detection research domain can be summarized regarding the reliability goal, namely, transferability, interpretability, and robustness.

3.1 Transferability

Deep learning models usually exhibit satisfactory performance on the same type of data that are seen in the training process but perform poorly on unseen data. In real life, Deepfake materials can be generated via various synthetic techniques [19, 49, 111, 115, 196] as abundant on-the-shelf easily accessible face-swapping implementations are publicly available, and a reliable Deepfake detection model is expected to perform well on unseen data to imitate real-life Deepfake cases. Therefore, guaranteeing the transferability of the detection models for cross-dataset performance is necessary and frequently discussed.

With the fast development of deep learning techniques, various methods are devised using simple Convolutional Neural Network (CNN) based models. Zhou et al. [194] fused a CNN stream with a support vector machine (SVM) [66] stream to analyze face features with the assistance of local noise residuals. Afchar et al. [2] studied the mesoscopic features of images with successive convolutional and pooling layers. Nguyen et al. [124] designed a
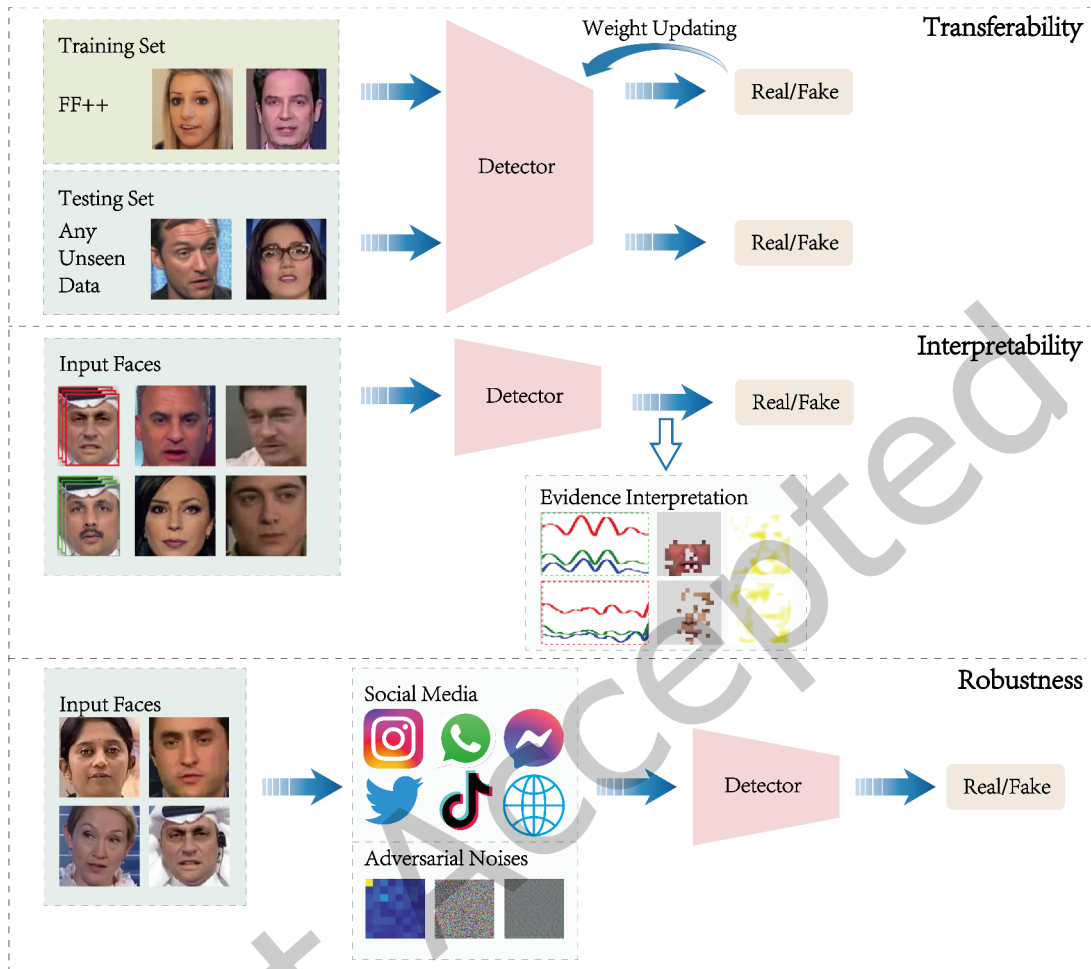
Fig. 1. Demonstrations of the three challenges from top to bottom. Transferability (top) refers to models that focus on stable detection ability on unseen benchmark datasets; interpretability (middle) refers to efforts on explaining the model detected falsification; robustness (bottom) refers to models that handle Deepfake suspects under different real-life conditions and scenarios.

multi-task learning scheme to simultaneously perform detection and localization using CNNs. DFT-MF [77] thinks that the mouth features can be important for detection and utilizes a convolutional model to detect Deepfake by verifying, analyzing, and isolating mouth and lip movements. Face X-ray [101] performs a medical x-ray on the candidate Deepfake faces by revealing whether the blending of two different source images can be decomposed.

Rather than the basic CNN architectures, well-designed and pre-trained CNN backbones are frequently exploited to improve model performance on Deepfake detection, especially for the cross-dataset performance on unseen data as more benchmark datasets have been released. An early approach [4] proposes optical flow analysis with pre-trained VGG16 [153] and ResNet50 [64] CNN backbones and achieves preliminary in-dataset test performance on the FaceForensics++ (FF++) dataset [144]. Capsule [125] employs capsule architectures [146]

with light VGG19-based network parameters but achieves similar detection performance to the traditional approaches leveraging CNNs. Li et al. [103] conducted a strengthened model, DSP-FWA, with the help of the spatial pyramid pooling [63] with ResNet50 as the backbone. This method is shown to be applicable to Deepfake materials at different resolution levels. FFD [31] leverages the popular attention mechanism by element-wise multiplication to study the feature maps and utilizes the XceptionNet CNN backbone, achieving marginally more promising performance than the work by Rossler et al. [144]. SSTNet [184] exploits Spatial, Steganalysis, and Temporal features using XceptionNet [24] and exhibits reasonable intra-cross-dataset performance on FF++. Given the assumption that Deepfake only modifies the internal part of the face, Nirkin et al. [130] adopted two streams of XceptionNet [24] for face and context (hair, ears, neck) identification and another XceptionNet to classify real and fake based on the learned discrepancies between the two. Later, Bonettini et al. [9] and Tariq et al. [158] studied the ensemble of various pre-trained convolutional models including EfficientNetB4 [157], XceptionNet [24], DenseNet [74], VGGNet [153], ResNet [64], and NASNet [198] backbones to detect Deepfake. Rossler et al. [144] employed the pre-trained well-designed XceptionNet [24] network and achieved state-of-the-art detection performance at the time on FF++. Besides, a special study introduced by Wang et al. [171] proves the transferability of the model trained on one CNN-generated dataset to the rest ten using pre-trained ResNet50 [64].

Meanwhile, frequency cues are also noticed and analyzed by researchers. While early image forgery detection work [6] focus on all high, medium, and low frequencies via Fourier transform, Frank et al. [46] were the first that raised the idea of finding frequency inconsistency between real and fake by employing low-frequency features to assist Deepfake detection using RGB information. Since low-frequency traces are mostly hidden by blurry facial features, later studies mainly analyze high-level features. F3-Net [141] exploits both low- and high-frequency features without RGB feature extraction operation. DFT [41], Two-Branch [118], SPSL [107], and MPSM-RFAM [20] accomplish promising detection results by analyzing high-frequency spectrum along with the on-the-shelf RGB features. Li et al. [99] extracted both middle- and high-frequency features and correlated frequency and RGB features for detection. By pointing out the drawbacks of using coarse-grained frequency information, Gu et al. [53] combined fine-grained frequency information with RGB features for better feature richness in a latest work. Moreover, Jeong et al. [79] designed a new training scheme with frequency-level perturbation maps added, which further enhanced the generalization ability of the detection model regarding all GAN-based generators.

Since the CNN architecture and backbones lack generalization ability and mainly focus on local features, even XceptionNet is restricted in learning the global features for further performance improvements. Therefore, developed solutions have introduced convolutional spatial attention to enlarge the local feature area and learn the corresponding relations, and the detection AUC scores have been gradually raised above 70% on average upon unseen datasets accordingly. The SRM [113] approach makes two streams of network using XceptionNet as the CNN backbone and focuses on the high-frequency information and RGB frames with a spatial attention module in each stream. The MAT model [192] proposes the CNN backbone EfficientNetB4 and borrows the convolutional attention idea to study different local patches of the input image frame. Specifically, the artifacts in shallow features are zoomed in for fine-grained enhancement in the detection performance. With the success of the transformer architecture [170] in the natural language processing (NLP) domain, different versions of vision transformer [40, 108, 136, 167, 175] have derived reasonable performance in the computer vision domain due to its ability on global image feature learning. The architectures of vision transformers (ViT) have also been employed for Deepfake detection with promising results in recent work [21, 67, 78, 172, 179].

While reaching the bottleneck specially for the cross-dataset performance even resorting to the advanced powerful but potentially time-consuming neural networks, most recent work gradually focuses on strategies to enrich diversity in training data. Such attempts have improved the detection AUC scores even up to 80% on some unseen datasets. PCL [193] is introduced with an inconsistency image generator to add synthetic

diversity and provide richly annotated training data. Sun et al. [155] proposed Dual Contrastive Learning (DCL) to study positive and negative paired data and enrich data views for Deepfake detection with better transferability. FInfer [71] inferences future image frames within a video and is trained based on the representation-prediction loss. Shiohara and Yamasaki [151] presented a novel synthetic training data called self-blended images (SBIs) by blending solely pristine images from the original training dataset, and classifiers are thus boosted to learn more generic representations. Chen et al. [17] proposed the SLADD model to further synthesize the original forgery data for model training to enrich model transferability on unseen data using pairs of pristine images and randomly selected forgery references with forgery configurations including forgery region, blending type, and mix-up blending ratio. Cao et al. [13] introduced the RECCE model to emphasize the common compact representations of genuine faces based on reconstruction-classification learning on real faces. The reconstruction learning over real images enhances the learning representations to be aware of unknown forgery patterns. Liang et al. [105] proposed an easily embeddable disentanglement framework to remove content information while maintaining artifact information for training and Deepfake detection using reconstructed data with various combinations of content and artifact features of real and fake samples. The OST [18] method improves the detection performance by preparing pseudo-training samples based on testing images to update the model weights before applying to the true samples.

3.2 Interpretability

Despite the promising ability of deep learning models, they suffer the weak interpretability problems due to the black-box characteristic [109]. In other words, it is hard to explain how and why a model comes up with a particular result. For the Deepfake detection task, while achieving promising detection performance statistically for in- and cross-dataset evaluations, the interpretability issue is still maintained to be fully resolved at the current stage. In other words, people tend to trust methods that are easily understandable via common sense rather than those with satisfactory accuracies but derived based on features that are hard to explain. Consequently, forensic evidence is critical to be probed to interpret and support the detection model performance by highlighting the reasons for classifying fake samples. In a nutshell, the interpretability challenge is to answer the following questions regarding a Deepfake suspect in order to be reliable:

- Why is the content classified as fake?
- Based on which part does the detection model determine the content as fake?

As confirmed by Baldassarre et al. [7] using a series of quantitative metrics for evaluating the interpretability of the detection models, heatmaps are generally futile and unacceptable when being adopted to explain the detected artifacts [185]. Moreover, heatmaps of the real faces are displayed in like manners, which are indistinguishable from that of the fake ones. Therefore, to answer the above questions, a group of noise-based research has been attempted to probe the Deepfake forensic traces for explanations upon the detection results. Early work mostly relies on the Photo Response Non-Uniformity (PRNU), a small factory-defects-generated noise pattern in the light-sensitive sensors of a digital camera [112]. PRNU has shown strong abilities in source anonymization [138] and source device identification [116, 147]. Unfortunately, most of the PRNU-based Deepfake detection studies [32, 92] have failed to show strong detection performance statistically. Therefore, the PRNU noise pattern can be a useful instrument for source device identification tasks, but it may not be a meaningful forensic noise tracing tool to satisfy the purpose of the Deepfake detection task with respect to the interpretability goal.

Later approaches [4, 16, 77, 101] step up to utilize CNNs for noise extraction and analyze the trace differences between real and fake. A milestone denoiser, DnCNN, proposed by Zhang et al. [189] is able to perform blind Gaussian denoising with promising performance, and has been later extended to study the camera model fingerprint for various downstream tasks including forgery detection and localization for face-swapping images based on the underlying noises [29, 30]. Recently, studies [55, 56] aim to extract the manipulation traces for

detection and interpretation. Guo et al. [58] proposed the AMTEN method to suppress image content and highlight manipulation traces as more filter iterations are applied. Wang et al. [173, 174] utilized the pre-trained denoisers for Deepfake forensic noise extraction and investigated the underlying Deepfake noise pattern consistency between face and background squares with the help of the siamese architecture [11]. Guo et al. [59] proposed a guided residual network to maintain the manipulation traces within the guided residual image and analyzed the difference between real and fake.

Besides studies [26, 140] that rely on biological signals by analyzing minuscule periodic changes through the faces visualizing distinguishable sequential signal patterns as indicators, other studies mostly attempt to explore universal and representative artifacts directly from the visual concepts of the Deepfake materials. The PRRNet [150] approach studies region-wise relation and pixel-wise relation within the candidate image for detection and can roughly locate the manipulated region using pixel-wise values. Trinh et al. [168] fed the dynamic prototype to the detection model and successfully visualized obvious artifact fluctuation via the prototype. Yang et al. [65] proposed to re-synthesis testing images by incorporating a series of visual tasks using GAN models, and they finally extracted visual cues to help perform detection. Obvious artifact differences can be observed by their Stage5 model on real and fake samples. Most recently, Dong et al. [39] introduced FST-matching to disentangle source-, target-, and artifact-relevant features from the input image and improved the detection performance by utilizing artifact-relevant features solely. The adopted features within the fake samples are visualized to explain the detection results.

3.3 Robustness

As the transferability and interpretability challenges have been frequently undertaken and reasonable or even promising results have been achieved accordingly, there lacks a stage to be fulfilled in order for the detection approaches to be useful in real-life cases. The challenge of this stage can be summarized as robustness. To be specific, the quality of the candidate Deepfake material in real life is not as ideal as the benchmark datasets in experimental conditions most of the time. Consequently, they may experience different levels of compression operation in multiple scenarios due to objectively limited conditions [114]. Moreover, post-processing strategies and artificially added perturbations can cause further challenges and even disable the well-trained and well-performed detection models. Therefore, the robustness of the detection approaches is necessary to be significantly considered when facing various real-life application scenarios under restricted and special conditions.

Multiple studies have been conducted mainly regarding two types of real-life conditions, namely, the passive condition and active condition. The passive condition refers to scenarios with objective limitations such as video compression due to network flow settings. In an early study, Kumar, Vatsa, and Singh [94] designed multiple streams of ResNet18 [64] to specifically deal with face reenactment manipulation at various compression levels. Hu et al. [72] proposed a two-stream method that specifically analyzes the features of compressed videos that are widely spread on social networks. The LRNet [156] model is developed to stay robust when detecting highly compressed or noise-corrupted videos. Cao et al. [14] solved the difficulty of detecting against compressed videos by feeding compression-insensitive embedding feature spaces that utilize raw and compressed forgeries to the detection model. Wu et al. [181, 182] analyzed noise patterns generated via online social networks before feeding image data into the detection model for training. The method has won top ranking against existing approaches especially when facing forgeries after being transmitted through social networks. Le and Woo [8] employed attention distillation in the frequency perspective and have successfully raised the detection ability on highly compressed data using ResNet50. RealForensics [61] aims to solve Deepfake contents with real-life quality by training the detection model with an auxiliary dataset containing real talking faces before utilizing the benchmark datasets. The method has significantly advanced the detection performance against multiple objective scenarios and adversarial attacks such as Gaussian noise and video compression. Besides, a recent work [177] has

constructed a new dataset containing Deepfake contents under the near-infrared condition to prevent potential future Deepfake attacks in the corresponding scenarios.

On the other hand, the active condition summarizes deliberate adversarial attacks such as distortions and perturbations. Gandhi and Jain [48] explored Lipschitz regularization [180] and Deep Image Prior (DIP) [98] to remove the artificial perturbations they generated and maintain model robustness. Yang et al. [25] simulated commonly seen data corruption techniques on the benchmark datasets to increase data diversity in model training. Operations such as resolution down-scaling, bit-rate adjustment, and artifacting have effectively boosted the detection ability against in-the-wild Deepfake contents. Hooda et al. [69] designed a Disjoint Deepfake Detection (D3) detector that improves adversarial robustness against artificial perturbations using an ensemble of models. The LTTD [54] framework is enforced to conquer the challenges brought by post-processing procedures of Deepfake generation like visual compression to models that rely on low-level image feature patterns. Lin et al. [106] proved the fact that having temporal information to participate in detection makes the detector less prone to black-box attacks.

Moreover, in the latest studies, researchers have frequently illustrated the necessity of robustness by fooling the well-trained detection models with stronger adversarial attacks [51, 75]. Jia et al. [80] even emphasized the robustness of the detection models against potential adversarial attacks by injecting frequency perturbations to fool the state-of-the-art approaches. Also, the robustness of Deepfake detectors is evaluated via the Fast Gradient Sign Method (FGSM) and the Carlini-Wagner L2-norm attack in several studies [48, 149]. Moreover, Carlini and Farid [15] employed white- and black-box attacks on a well-trained detector in five case studies. Obvious performance damping can be observed from the reported results. Faces with imperceptible visual variation even after perturbations and noises are added in have highlighted the importance of studying model robustness in the current Deepfake detection research domain.

4 Detection Model Reliability Study

4.1 Overview

In the ideal scenario, a reliable Deepfake detection model should retain promising transferability on unseen data with unknown synthetic techniques, pellucid explanation upon the falsification, and robust resistance against different real-life application conditions. As reviewed in Section 3, despite a method favorably satisfying all three challenges simultaneously is not yet accomplished, a metric that nominates the reliability of a method for real-life usages and court-case judgments is needed.

Regardless of the largely improved but still unsatisfied cross-dataset performance in the evolution of Deepfake detection, existing studies have only evaluated the model performance on each testing dataset to show the model detection abilities while the values for each evaluation metric (accuracy and AUC score) vary depending on different testing sets adopted in experiments. On the contrary, in real-life cases, people have no clue about fake content regarding its source dataset or the corresponding facial manipulation techniques, and the malicious attacker is unlikely to reveal such crucial information. Consequently, for a victim of Deepfake to defend his or her innocence or accuse the attacker [22, 88, 97, 120], simply presenting a model detection decision and listing the numerical model performance on each benchmark dataset may not be convincing and reliable. Specifically, a unique statistical claim is necessary regarding the detection performance to discuss the model trustworthiness on any arbitrary candidate suspect instead of varying on each testing dataset when adopting the detection model as forensic evidence for criminal investigation and court-case judgments. To conclude, the following questions are to be solved:

- Can a detection model assist or act as evidence in forensic investigations in court?
- How reliable is the detection model when performing as forensic evidence in real-life scenarios?
- How accurate is the detection model regarding an arbitrary falsification?

Unfortunately, to the best of our knowledge, no existing work has studied the model reliability or come up with a reliable claim for the model to play the role of forensic evidence. Therefore, in this study, inspired by the reliability study on antigen diagnostic tests [1, 27, 52] for the recent COVID-19 pandemic [132], we conduct a quantitative study by investigating the detection model reliability with a new evaluation metric with statistical techniques. Unlike the studies for antigen diagnostic tests that prefer to achieve a perfect specificity rather than sensitivity as the goal is to avoid missing any positive case [131], we wish to correctly identify both real and fake materials with no priority. In particular, we construct a population to imitate real-life Deepfake distribution and design a scientific random sampling scheme to analyze and compute the confidence intervals for the values of accuracy and AUC score metrics regarding the Deepfake detection models. As a result, numerical ranges indicating the reliable model performance at 90% and 95% confidence levels can be derived for both accuracy and AUC score.

4.2 Deepfake Population

In reality, a candidate Deepfake suspect can only have two possible categories, namely, real and fake. Admittedly, most public images and videos circulating on the internet are pristine without artificial changes. However, whenever a real-life Deepfake case is raised such that the authenticity of the candidate material needs to be justified, we could not consider all images and videos in the world as the target population [166] because most of the real ones are not likely to be disputed in the discussion of Deepfake. At the same time, the probability that the candidate material is fake does not necessarily equal to the proportion of fake ones regarding all images and videos in the world.

Despite the uncertainty of the real-life Deepfake population and distribution, in this study, we construct a sampling frame [12] with the accessible high-quality Deepfake benchmark datasets to imitate the target population of Deepfake in real life for detection model reliability analysis. Details of the participating datasets are introduced in Section 5.2.

4.3 Random Sampling

We perform random sampling [117] from the constructed sampling frame with a sample size of s for t trials. Two sampling options are considered in this study: balanced and imbalanced. For a balanced sampling setting, we maintain the condition that the same amount of real and fake samples are randomly drawn. On the contrary, an imbalanced setting allows a completely random stochastic rule with respect to real and fake samples.

For an arbitrary Deepfake detection model M after sufficient training, we randomly draw s samples from the sampling frame following the sampling option. Then the s samples are fed to model M for authenticity prediction, deriving predicted labels and prediction scores. After that, the accuracy and AUC score metrics are computed accordingly based on the ground-truth authenticities of the sampled faces. Such a sampling process is repeated for t trials and a total of t accuracies and t AUC scores are derived in the end. Take the t accuracies as an example, we first compute the mean value $\bar{x}$ and standard deviation σ following

$$\bar{x} = \frac{\Sigma_{i=1}^{t} x_i}{t}, \tag{1}$$

and

$$\sigma = \sqrt{\frac{\Sigma_{i=1}^{t} (x_i - \bar{x})^2}{t - 1}}, \tag{2}$$

where x_i refers to the accuracy value of the i-th trial.

According to the central limit theorem (CLT) [45], the distribution of sample means tends toward a normal distribution as the sample size gets larger. Therefore, the normal distribution confidence interval CI can be calculated by

$$CI = \bar{x} \pm z\frac{\sigma}{\sqrt{s}}, \tag{3}$$

where parameter z represents the z-score, an indicator of the confidence level following the instruction of the z-table [16]. When deducing statistical results for the t AUC scores, the above workflow applies identically.

Since the target population is of unknown distribution, different values of the sample size s are adopted in order to settle the confidence intervals at different confidence levels. Meanwhile, considering that an insufficient number of trials per sample size may cause bias when locating the sample mean, different values of trails for t are chosen to eliminate the potential bias. Detailed steps of the model reliability study can be summarized as Algorithm 1.

Algorithm 1 Deepfake detection model reliability study.

Input: M = well-trained detection model
Output: 90% and 95% CI for each sample size
 1: $f_r \leftarrow$ list of real samples
 2: $f_f \leftarrow$ list of fake samples
 3: $o \leftarrow$ balance sampling option
 4: $t \leftarrow$ number of trials
 5: $S \leftarrow [s_1, s_2, ..., \text{len}(S)]$
 6: shuffle$(\cdot) \leftarrow$ function to shuffle the list
 7: acc$(\cdot) \leftarrow$ function to calculate accuracy
 8: auc$(\cdot) \leftarrow$ function to calculate AUC score
 9: **for** $i \leftarrow 0$ **to** $\text{len}(S) - 1$ **do**
10:    acc_lst $\leftarrow []$
11:    auc_lst $\leftarrow []$
12:    **for** $j \leftarrow 0$ **to** t **do**
13:       **if** $o ==$ True **then**
14:          $f_r \leftarrow$ shuffle(f_r)
15:          $f_f \leftarrow$ shuffle(f_f)
16:          samples $\leftarrow f_r[0 : \frac{S[i]}{2}] + f_f[0 : \frac{S[i]}{2}]$
17:       **else**
18:          $f_a \leftarrow$ shuffle$(f_r + f_f)$
19:          samples $\leftarrow f_a[0 : S[i]]$
20:       **end if**
21:       preds, pred_scores, labels $\leftarrow M(\text{samples})$
22:       acc_lst.append(acc(preds, labels))
23:       auc_lst.append(auc(pred_scores, labels))
24:    **end for**
25:    compute $\bar{x}$ and σ for acc_lst and auc_lst
26:    compute and record 90% and 95% CI
27: **end for**

---

5 Dataset Preparation

The choice of training dataset and data pre-processing scheme can significantly affect the performance of a deep learning model. Among the evolution of Deepfake datasets, as introduced in Section 2.2, various benchmark datasets are frequently adopted for training and testing to boost the detection model performance, but the workflow of data pre-processing operations has been barely discussed in detail in existing studies. Moreover, there lacks a standard pre-processing scheme in the current domain, causing difficulty in model comparison due to non-uniform training datasets after pre-processing by different detection work. On the other hand, using heedlessly prepared datasets for sampling frame construction can lead to improbable results towards the reliability study. Therefore, in this paper, we expound a standard and systematic workflow of data preparation and pre-processing to resolve the inconsistency and benefit both veterans and new starters in the research domain, ensuring a fair game for other work to compare with the model performance as exhibited in this paper following the same settings.

5.1 Dataset Pre-processing

While a video-level detector may rely on special data processing arrangement directly upon videos, for the frame-level detectors, since the detection results are evaluated based on all selected frames, to avoid potential biases toward particular videos, it is meaningful to keep the amount of extracted faces from each video equivalent during model training and testing. Therefore, we firstly obtain c image frames using FFmpeg [44] for each candidate video with an equal frame interval between every two adjacent extracted frames following

$$P = \frac{iN}{c} \text{ for } 0 \leq i < c \text{ and } i \in \mathbb{Z}, \tag{4}$$

where N refers to the number of frames that contain faces in the video and $P = \{p_0, p_1, ..., p_{c-1}\}$ contains the sequentially ordered indices for which frames to be extracted from the video. In other words, image frames with no face detected are excluded from the sequential ordering and indexing. Besides, videos with fewer than c frames containing detected faces are also omitted. The dlib library [90] is utilized for face detection and cropping where the face detector provides coordinates of the bounding box that locates the detected face. For the sequence of frames with frame indices $P = \{p_0, p_1, ..., p_{c-1}\}$ from a video, we fix the size l of a squared bounding box b for all faces by

$$l = \max\{\max_i w_i, \max_i h_i\} \text{ for } 0 \leq i < c \text{ and } i \in \mathbb{Z}, \tag{5}$$

where w_i and h_i are the widths and heights of each bounding box. We then locate the center of each face f_i with the help of the corresponding bounding box b_i and place the fixed squared bounding box b at the centers for face cropping.

5.2 Datasets Involved and Detailed Arrangements

Following the convention of the existing Deepfake detection work and considering the qualities of available benchmark datasets, we consider five datasets in experiments in this study, namely, FF++ [144], FaceShifter [100], DFDC [37], Celeb-DF [104], and DF1.0 [81]. In detail, early datasets [186] are excluded due to low quantity and diversity. Meanwhile, although WDF [197] is similar to real-life Deepfake materials, the videos collected from the internet are manually labeled without knowing the ground-truth labels, which leads to credibility issues. KoDF [95] is the largest Deepfake dataset up to date, but its huge magnitude requires unreasonably large storage (∼4 TB) that we are unable to acquire and process[5]. All involved benchmark datasets follow the

---

[5]The 6 manipulation algorithms in KoDF are highly overlapped with the 8 manipulation algorithms in DFDC. This favorably suggests that the adoption of DFDC satisfies the demand for model diversity even without KoDF.

pre-processing scheme for face extraction as discussed in Section 5.1, and special settings are mentioned in the following subsections when necessary.

5.2.1 FaceForensics++.

FaceForensics++ (FF++) is currently the most widely adopted dataset in the existing Deepfake detection studies. The dataset contains 1,000 real videos collected from YouTube and 4,000 Deepfake videos synthesized based on the real ones. In specific, four facial manipulation techniques are each applied to the 1,000 real videos to derive the corresponding 1,000 fake ones. Among the four facial manipulation techniques, FaceSwap (FS) [115] and Deepfakes (DF) [34] are face-swapping algorithms that synthesize the faces by swapping facial identities, while Face2Face (F2F) [160] and NeuralTextures (NT) [159] perform face reenactment by modifying facial attributes such as expressions and accessories.

The FF++ dataset has provided a subject-independent official dataset split with a ratio of 720:140:140 for training, validation, and testing. Meanwhile, three dataset qualities have been released, namely, Raw, HQ (c23), and LQ (c40), where the latter two are compressed with different video compression levels following the H.264 codec. In recent Deepfake detection work, FF++ is frequently adopted as the training dataset due to its manipulation diversity and data orderliness, and the HQ (c23) version is mostly utilized because it has a similar video compression level and video quality to the real-life Deepfake contents. In this survey, whenever necessary, we adopt FF++ for model training following the official dataset split. The key image frames are also extracted and employed since the performance enhancement by the participating key image frames in the training process has been proved in the early studies [2, 104, 172]. In the training process, unless specially designed, commonly used data augmentation is performed upon the real faces to construct a balanced training dataset for real and fake. In the testing phase, the testing set is constructed following the official split without further augmentation.

5.2.2 Deepfake Detection Challenge.

Deepfake Detection Challenge (DFDC) is one of the largest public Deepfake datasets with 23,654 real videos and 104,500 fake ones. Among the fake videos, there are eight synthetic techniques [34, 73, 86, 129, 139, 187] that have been applied based on the real ones. Due to its large data quantity, we randomly pick 10 of the 50 video folders from the official dataset and randomly shuffle 100 real videos and 100 fake ones from each folder. Since most existing approaches focus on detecting Deepfake visually and most benchmark datasets are published without audio, fake videos using the pure audio swap technique are easily classified as pristine because there is no visual artifact on the faces. Meanwhile, the official DFDC dataset only provides labels for real and fake while sub-labels for specific synthetic techniques are unavailable. Therefore, in this paper, the randomly picked 1,000 fake videos are manually examined to eliminate fake videos with the pure audio swap technique to omit noises in detection and guarantee a fair experimental setting. The selected videos are then fed through the data pre-processing scheme in Section 5.1 for model evaluation.

5.2.3 Celeb-DF.

Celeb-DF is one of the most challenging benchmark Deepfake datasets that are publicly available. It contains 590 celebrity interview videos collected from YouTube and 5,639 face-swapped videos based on the real ones using an improved face-swapping algorithm with resolution enhancement, mismatched color correction [142], inaccurate face mask adjustment, and temporal flickering reducing [83] on the basic face-swapping auto-encoder architecture. A set of 518 official testing videos with high visual quality has failed most of the existing baseline models at a time because obvious visual artifacts can be barely found. We resort to the official testing set with 178 real videos and 340 fake ones for model evaluation.

5.2.4 DeeperForensics-1.0.

DeeperForensics-1.0 (DF1.0) is the first large-scale dataset that is manually added with deliberate distortions and perturbations to the clean face-swapped videos. A strengthened face-swapping algorithm, Deepfake Variational Auto-Encoder (DF-VAE), is introduced for superior synthetic performance with better reenactments on expression and pose, fewer style mismatches, and more stable temporal continuity. There are a total of 10,000 synthesized videos where 1,000 of them are face-swapped from lab-controlled source videos onto the FF++ real videos using DF-VAE and the rest 9,000 videos are derived using the 1,000 raw manipulated

videos by applying combinations of seven distortions[6] under five intensity levels. Since the HQ and LQ versions of FF++ contain the same visual content and only differ in compression levels, DF1.0 with sufficient visual quality diversity in the manipulated videos serves as a perfect substitution for the LQ version of FF++ in the experiments, providing a convincing evaluation of all quality circumstances. Produced based on FF++, the dataset has only provided the official split ratio of 7:1:2 for the fake videos, and we thus execute model evaluation with merely the 2,000 fake testing videos.

5.2.5 FaceShifter. FaceShifter refers to a subject-agnostic GAN-based face-swapping algorithm that solves the facial occlusion challenge with a novel Heuristic Error Acknowledging Refinement Network (HEAR-Net). A subset with 1,000 synthetic videos is later included in the FF++ dataset by applying the FaceShifter face-swapping model to the 1,000 real videos. Since FaceShifter and FF++ share the same set of real videos, we take only the 140 fake videos for model evaluation following the FF++ official split.

6 Detection Model Evaluation

In this section, we first adopt several state-of-the-art Deepfake detection models that are mainly designed regarding each of the three challenges as defined in Section 3 and report their detection performance on each benchmark testing set. Then the models are further discussed regarding the reliability following Algorithm 1 along with case studies on real-life Deepfake materials.

6.1 Experiment Settings

Based on the Deepfake detection developing history and the three challenges of the current Deepfake detection domain, in the experiment, we selected several representative milestone baseline models and the most recent ones that have source code publicly available for reproduction. Specifically, Xception [24], MAT [192], and RECCE [13] mainly attempt on the transferability challenge, Stage5 [189] and FSTMatching [39] focus on the interpretability topic, and MetricLearning [14] and LRNet [156] are designed for the robustness issue. Models with publicly available trained weights are directly adopted for evaluation if the model is trained on FF++ or special arrangements other than the five benchmark datasets are necessary during training. The rest models are trained on FF++ in our experiment as discussed in Section 5 and all models converge commonly. The selected models are tested on all benchmark datasets. To guarantee complete fairness, we applied optimal parameter settings as reported in the corresponding published papers during training and testing.

During model testing, we recorded the video-level Deepfake detection performance. In particular, detection results of the cropped faces of each video are averaged to a unique output for detectors that are designed for detecting individual images. Methods that can directly generate a single output for each video are fed with raw videos via the corresponding processing scheme as provided by their published source codes. The well-trained models are firstly evaluated on the FF++ testing set for the in-dataset setting, in other words, tested on the same dataset they have seen during training. Then, to further validate the model performance on unseen datasets, the cross-dataset evaluation is conducted to test the models on DFDC, Celeb-DF, DF1.0, and FaceShifter. We set $N = 10$ for training and $N = 20$ for testing regarding Eq. (4) for frame extraction during data pre-processing when applicable.

We adopted accuracy (ACC) and AUC score at the video level as the evaluation metrics. In detail, the accuracy refers to the proportion of the correctly classified data items regarding all testing data, and the AUC score represents the area under the receiver operating characteristic (ROC) curve. In other words, the AUC score demonstrates the probability that a random positive sample scores higher than a random negative sample from

---

[6]Change of color saturation, local block-wise distortion, change of color contrast, Gaussian blur, white Gaussian noise in color components, JPEG compression, and change of video constant rate factor.

the testing set, that is, the ability of the classifier to distinguish between real and fake faces. For testing sets that contain only fake samples, the AUC score is inapplicable and thus withdrawn.

6.2 Model Performance on Benchmark Testing Sets

Model performance for both in- and cross-dataset evaluations is listed in Table 2. It can be observed that all models of the transferability topic have derived reasonable detection performance on FF++ with accuracy values and AUC scores over 90% since their goal is to achieve better performance in cross-dataset experiments after maintaining promising performance on seen data. In particular, MAT [192] wins the comparison with the highest 97.40% accuracy and 99.67% AUC score. On the contrary, models that are designed specifically for interpretability or robustness purposes have exhibited relatively poor detection performance on FF++, and the potential causation is discussed together with their detection performance on other benchmark datasets in the following paragraphs.

As for cross-dataset evaluation, most models have suffered a performance damping since the testing data are unseen during training. In specific, no model has reached over 80% AUC scores on DFDC or Celeb-DF and some models even exhibit abnormal detection performance when validated on unseen fake testing sets solely (DF1.0 and FaceShifter). This may be caused by oblivious overfitting on real or fake data by the models. While models that are solving the transferability challenge have all exhibited normal and reasonable functionalities, hidden trouble can be discovered regarding the interpretability and robustness topics. Specifically, model weights of Stage5 [65] are adopted for testing because the model is trained with re-synthesized samples using exclusively GAN models under special settings, but this at the same time has led to unsatisfied results when detecting fake samples that are not synthesized using GAN architectures. FSTMatching [39] spends huge computing power on disentangling source and target artifacts for the explanation, which thus results in the failure against other models although similarly trained on FF++. MetricLearning [14] and LRNet [156] both are proposed and trained to deal with highly compressed Deepfake contents in special scenarios. Unfortunately, they suffer performance fluctuation when the compression condition varies without expectation. Furthermore, LRNet [156] executes detection based on facial landmarks solely, which is another main reason that leads to the unsatisfactory.

Besides, models are generally unstable on different testing sets. For instance, RECCE [13] wins the competition on Celeb-DF for both accuracy and AUC score, but its detection ability deteriorates immensely when facing DFDC. On the other hand, MAT [192] wins the battle on DFDC and achieves competitive performance on Celeb-DF,

Table 2. Quantitative video-level accuracy (ACC) and AUC score performance comparison on each testing set. (†: trained weights directly adopted for evaluation.)

Model	Test Dataset							
	FF++ [144]		DFDC [37]		Celeb-DF [104]		DF1.0 [81]	FaceShifter [100]
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	ACC
Xception [24]	93.92%	97.31%	65.21%	71.57%	70.27%	70.71%	56.87%	57.55%
MAT [192]	97.40%	99.67%	66.63%	74.83%	71.81%	77.16%	41.74%	18.71%
RECCE [13]	90.72%	95.26%	62.06%	66.94%	71.81%	77.90%	51.21%	56.12%
Stage5† [65]	19.97%	50.21%	51.02%	48.08%	34.36%	39.88%	0.00%	0.00%
FSTMatching [39]	81.33%	77.01%	44.88%	39.90%	38.61%	44.27%	25.68%	13.67%
MetricLearning† [14]	80.03%	77.71%	48.98%	61.89%	65.64%	60.37%	100.00%	100.00%
LRNet† [156]	55.22%	67.82%	53.41%	53.91%	51.54%	59.72%	52.19%	41.30%

Table 3. Dataset statistics of the sampling frame for model testing regarding the number of videos with cropped faces. Datasets with no real samples are marked with '–' sign.

	FF++ [144]	DFDC [37]	Celeb-DF [104]	DF1.0 [81]	FaceShifter [100]	Total
Num Real	140	1,000	178	–	–	1,318
Num Fake	560	1,000	340	2,010	140	4,050
Total	700	2,000	518	2,010	140	5,368

but an obvious overdependence on the real samples can be concluded from its poor accuracies on DF1.0 and FaceShfiter. Meanwhile, Xception has derived reasonable detection performance on each testing set even though not winning the comparison on any dataset. As a result, no model appears to be the overall winner according to Table 2 and it is hard to determine which model to use when facing an arbitrary candidate Deepfake suspect in real-life cases.

Moreover, in most cases, a well-trained model usually achieves a higher AUC score than the accuracy on each testing set. The reason is that the threshold to classify real and fake with softmax or sigmoid function applied is always fixed at 0.5 for the accuracy evaluation upon the output scores within the range of [0, 1] where 0 refers to real and 1 represents fake, while the actual threshold for the optimal model performance is usually located differently regarding 0.5. Therefore, despite a classifier with a threshold value set to 0.5 does not perform well, the model may still distinguish between real and fake with a relatively high AUC score. However, it is also worth noting that although a high AUC score may reveal the model's ability to separate real and fake samples, the threshold may vary depending on different testing sets and different models. Hence, in order to stably determine real or fake, finding a fixed threshold to consistently satisfy the detection goal on arbitrary images and videos may help boost the overall model detection ability in the research domain.

6.3 Model Reliability Experiment Results

The model reliability evaluation is conducted on a sampling frame with 5,368 videos composed of the testing set of each benchmark dataset, and the detailed data quantity is listed in Table 3. Following the workflow of Algorithm 1, we proposed reliability analyses on the well-trained Deepfake detection approaches. We computed the 90% and 95% confidence intervals in experiments. Specifically, two values of t, 500 and 3,000, are chosen to ensure a sufficient number of trials when locating the sample means. Various sample sizes $s \in \{10, 100, 500, 1,000, 1,500, 2,000, 2,500\}$ are selected to find the settled confidence intervals. Besides, both sampling options, balanced and imbalanced sampling, are executed in experiments. Detailed results are listed in Tables 4 to 7 following the Cartesian product settings of balancing options $O = \{True, False\}$, the number of trials $T = \{500, 3,000\}$, and evaluation metrics $E = \{ACC, AUC\}$.

It can be easily observed that for all models in the tables, the mean values $\bar{x}$ gradually settle as sample sizes become larger and the standard deviation (Std.) values σ consistently decrease concurrently. Similarly, the 90% and 95% confidence intervals are progressively straitened and settled around the mean values. Moreover, statistical results with 3,000 sampling trials converge faster and are more stable than those with 500 trials as sample size increases for all experiments, and both trial numbers lead to similar final values once settled.

Regarding the balanced sampling setting, the results generally match the model detection performance in Table 2. On the other hand, since the constructed sampling frame is imbalanced for real and fake faces, an imbalanced sampling option may lead to more fake data than the real ones in the sample set. Consequently, for models that have shown relatively poor performance when evaluated on fake testing sets (DF1.0 and FaceShifter) solely, the mean values and confidence intervals of the accuracy are located at lower levels. Meanwhile, models

Table 4. Accuracy statistics (%) with balanced sampling for 500 and 3,000 trials.

Model Names	Num Trials	Statistics	Sample Sizes						
			10	100	500	1,000	1,500	2,000	2,500
Xception [24]	500	90% CI	57.89−79.55	65.67−72.21	67.40−70.24	67.94−69.78	68.22−69.64	68.33−69.39	68.42−69.24
		95% CI	55.81−81.63	65.04−72.84	67.13−70.51	67.76−69.96	68.08−69.78	68.23−69.49	68.34−69.31
		Mean	68.72	68.94	68.82	68.86	68.93	68.86	68.83
		Std.	14.70	4.44	1.92	1.25	0.97	0.72	0.56
	3,000	90% CI	58.90−78.54	65.87−71.98	67.55−70.16	67.98−69.70	68.25−69.50	68.37−69.36	68.50−69.26
		95% CI	57.02−80.42	65.29−72.57	67.31−70.41	67.82−69.87	68.13−69.62	68.28−69.45	68.42−69.34
		Mean	68.72	68.93	68.86	68.84	68.88	68.86	68.88
		Std.	14.62	4.55	1.94	1.28	0.93	0.73	0.57
MAT [192]	500	90% CI	59.72−77.44	66.23−71.87	67.59−70.06	68.18−69.72	68.30−69.52	68.40−69.38	68.52−69.31
		95% CI	58.02−79.14	65.69−72.41	67.35−70.30	68.03−69.87	68.19−69.64	68.31−69.47	68.45−69.39
		Mean	68.58	69.05	68.82	68.95	68.91	68.89	68.92
		Std.	13.17	4.19	1.83	1.15	0.91	0.73	0.59
	3,000	90% CI	59.45−77.74	66.10−71.84	67.67−70.15	68.04−69.71	68.28−69.50	68.38−69.37	68.49−69.29
		95% CI	57.70−79.49	65.55−72.39	67.43−70.39	67.88−69.87	68.16−69.62	68.29−69.47	68.41−69.37
		Mean	68.60	68.97	68.91	68.87	68.89	68.88	68.89
		Std.	13.61	4.27	1.85	1.24	0.91	0.74	0.60
RECCE [13]	500	90% CI	51.46−71.14	57.49−64.22	59.38−62.06	59.95−61.81	60.16−61.43	60.30−61.26	60.42−61.17
		95% CI	49.56−73.04	56.84−64.86	59.13−62.32	59.77−61.99	60.03−61.55	60.20−61.35	60.35−61.24
		Mean	61.30	60.85	60.72	60.88	60.79	60.78	60.79
		Std.	14.63	5.00	1.99	1.38	0.94	0.72	0.56
	3,000	90% CI	49.56−72.51	57.19−64.21	59.28−62.27	59.78−61.76	60.08−61.52	60.25−61.35	60.37−61.22
		95% CI	47.36−74.71	56.52−64.88	58.99−62.55	59.59−61.95	59.94−61.66	60.15−61.46	60.29−61.30
		Mean	61.04	60.70	60.77	60.77	60.80	60.80	60.80
		Std.	15.59	4.77	2.03	1.35	0.98	0.75	0.58
Stage5 [65]	500	90% CI	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00
		95% CI	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00
		Mean	50.00	50.00	50.00	50.00	50.00	50.00	50.00
		Std.	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	3,000	90% CI	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00
		95% CI	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00
		Mean	50.00	50.00	50.00	50.00	50.00	50.00	50.00
		Std.	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FSTMatching [39]	500	90% CI	44.61−59.71	49.85−54.76	51.42−53.37	51.68−53.01	51.91−52.87	51.96−52.78	52.06−52.69
		95% CI	43.16−61.16	49.37−55.23	51.24−53.56	51.55−53.14	51.82−52.96	51.88−52.86	52.00−52.75
		Mean	52.16	52.30	52.40	52.34	52.39	52.37	52.37
		Std.	12.12	3.95	1.56	1.07	0.77	0.66	0.50
	3,000	90% CI	44.62−59.82	50.15−54.80	51.42−53.42	51.73−53.07	51.85−52.84	51.98−52.78	52.06−52.70
		95% CI	43.16−61.28	49.71−55.24	51.23−53.61	51.60−53.19	51.76−52.94	51.91−52.85	51.99−52.76
		Mean	52.22	52.48	52.42	52.40	52.35	52.38	52.38
		Std.	12.23	3.73	1.60	1.07	0.80	0.64	0.52
MetricLearning [14]	500	90% CI	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00
		95% CI	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00
		Mean	50.00	50.00	50.00	50.00	50.00	50.00	50.00
		Std.	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	3,000	90% CI	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00
		95% CI	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00	50.00−50.00
		Mean	50.00	50.00	50.00	50.00	50.00	50.00	50.00
		Std.	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LRNet [156]	500	90% CI	41.70−62.46	48.22−55.13	50.53−53.24	51.09−52.97	51.31−52.72	51.42−52.51	51.62−52.37
		95% CI	39.70−64.46	47.56−55.79	50.27−53.51	50.91−53.15	51.18−52.86	51.32−52.61	51.55−52.44
		Mean	52.08	51.67	51.89	52.03	52.02	51.96	52.00
		Std.	15.43	5.13	2.02	1.40	1.05	0.81	0.55
	3,000	90% CI	41.70−61.81	49.05−55.25	50.74−53.29	51.15−52.80	51.38−52.64	51.51−52.49	51.62−52.36
		95% CI	39.77−63.73	48.46−55.84	50.49−53.54	50.99−52.96	51.26−52.77	51.41−52.58	51.55−52.43
		Mean	51.75	52.15	52.01	51.98	52.01	52.00	51.99
		Std.	16.17	4.98	2.05	1.33	1.01	0.79	0.59

Table 5. Accuracy statistics (%) with imbalanced sampling for 500 and 3,000 trials.

Model Names	Num Trials	Statistics	Sample Sizes						
			10	100	500	1,000	1,500	2,000	2,500
Xception [24]	500	90% CI	55.90−75.50	62.70−69.19	64.66−67.62	65.08−67.19	65.48−67.06	65.55−66.91	65.58−66.89
		95% CI	54.02−77.38	62.07−69.82	64.37−67.90	64.88−67.39	65.32−67.21	65.42−67.05	65.45−67.02
		Mean	65.70	65.94	66.14	66.13	66.27	66.23	66.23
		Std.	14.56	4.83	2.20	1.57	1.18	1.01	0.97
	3,000	90% CI	55.36−77.38	62.75−69.79	64.74−67.81	65.06−67.25	65.32−67.06	65.44−66.99	65.51−66.94
		95% CI	53.24−79.50	62.08−70.46	64.44−68.10	64.85−67.46	65.15−67.23	65.29−67.13	65.37−67.07
		Mean	66.37	66.27	66.27	66.15	66.19	66.21	66.22
		Std.	14.97	4.78	2.09	1.49	1.19	1.05	0.97
MAT [192]	500	90% CI	50.11−70.45	57.75−64.32	59.17−61.98	59.61−61.72	59.82−61.55	59.93−61.38	59.93−61.28
		95% CI	48.15−72.41	57.12−64.95	58.90−62.24	59.41−61.92	59.65−61.72	59.79−61.52	59.80−61.41
		Mean	60.28	61.03	60.57	60.66	60.68	60.66	60.61
		Std.	15.12	4.88	2.08	1.57	1.29	1.08	1.00
	3,000	90% CI	49.79−70.58	57.22−63.89	59.21−62.13	59.57−61.64	59.80−61.52	59.90−61.39	59.97−61.26
		95% CI	47.80−72.57	56.59−64.53	58.94−62.40	59.37−61.83	59.64−61.69	59.76−61.53	59.85−61.38
		Mean	60.18	60.56	60.67	60.60	60.66	60.64	60.61
		Std.	15.47	4.96	2.17	1.54	1.28	1.11	0.96
RECCE [13]	500	90% CI	53.35−73.89	59.48−66.01	61.53−64.47	62.15−64.17	62.37−64.00	62.30−63.70	62.43−63.76
		95% CI	51.37−75.87	58.86−66.64	61.25−64.75	61.96−64.36	62.21−64.16	62.16−63.84	62.30−63.88
		Mean	63.62	62.75	63.00	63.16	63.18	63.00	63.09
		Std.	15.27	4.85	2.18	1.50	1.21	1.04	0.98
	3,000	90% CI	52.85−73.05	59.88−66.32	61.64−64.53	62.01−64.07	62.25−63.89	62.30−63.77	62.36−63.66
		95% CI	50.91−74.99	59.26−66.94	61.36−64.81	61.81−64.27	62.09−64.04	62.16−63.91	62.24−63.79
		Mean	62.95	63.10	63.09	63.04	63.07	63.04	63.01
		Std.	15.04	4.80	2.15	1.54	1.22	1.09	0.97
Stage5 [65]	500	90% CI	16.38−33.50	22.20−27.30	23.73−26.06	24.04−25.65	24.20−25.34	24.33−25.28	24.37−25.15
		95% CI	14.73−35.15	21.71−27.79	23.50−26.28	23.88−25.80	24.10−25.45	24.24−25.37	24.30−25.23
		Mean	24.94	24.75	24.89	24.84	24.77	24.80	24.76
		Std.	13.75	4.09	1.87	1.29	0.91	0.76	0.63
	3,000	90% CI	16.38−33.20	22.13−27.53	23.61−25.96	24.02−25.54	24.21−25.37	24.31−25.25	24.40−25.17
		95% CI	14.77−34.82	21.62−28.05	23.39−26.18	23.88−25.69	24.10−25.49	24.22−25.34	24.33−25.25
		Mean	24.79	24.83	24.79	24.78	24.79	24.78	24.79
		Std.	13.52	4.34	1.89	1.22	0.93	0.76	0.62
FSTMatching [39]	500	90% CI	31.38−51.22	37.87−44.02	39.65−42.21	40.20−41.92	40.43−41.76	40.56−41.64	40.67−41.54
		95% CI	29.47−53.13	37.28−44.61	39.40−42.45	40.03−42.08	40.31−41.88	40.45−41.74	40.58−41.62
		Mean	41.30	40.95	40.93	41.06	41.09	41.10	41.10
		Std.	15.93	4.94	2.06	1.38	1.06	0.87	0.70
	3,000	90% CI	31.47−50.66	38.02−44.20	39.83−42.47	40.19−41.94	40.42−41.74	40.56−41.64	40.65−41.53
		95% CI	29.63−52.50	37.43−44.79	39.58−42.72	40.02−42.11	40.29−41.87	40.45−41.75	40.57−41.62
		Mean	41.07	41.11	41.15	41.06	41.08	41.10	41.09
		Std.	15.43	4.97	2.12	1.41	1.06	0.87	0.71
MetricLearning [14]	500	90% CI	66.17−83.91	72.88−77.98	74.09−76.40	74.42−76.06	74.65−75.78	74.69−75.67	74.86−75.59
		95% CI	64.47−85.61	72.39−78.47	73.87−76.62	74.27−76.22	74.54−75.89	74.60−75.77	74.79−75.66
		Mean	75.04	75.43	75.25	75.24	75.22	75.18	75.22
		Std.	14.23	4.09	1.85	1.31	0.91	0.79	0.59
	3,000	90% CI	66.53−83.60	72.81−78.12	74.05−76.36	74.46−75.97	74.66−75.83	74.75−75.67	74.85−75.62
		95% CI	64.89−85.24	72.30−78.63	73.83−76.58	74.32−76.12	74.54−75.95	74.66−75.76	74.77−75.69
		Mean	75.07	75.46	75.20	75.22	75.25	75.21	75.23
		Std.	13.73	4.28	1.86	1.22	0.95	0.74	0.62
LRNet [156]	500	90% CI	44.08−62.36	49.99−55.83	51.62−54.32	51.89−53.62	52.08−53.41	52.14−53.19	52.19−53.07
		95% CI	42.32−64.12	49.43−56.40	51.36−54.58	51.72−53.79	51.96−53.53	52.04−53.29	52.10−53.16
		Mean	53.22	52.91	52.97	52.75	52.74	52.66	52.63
		Std.	14.68	4.69	2.17	1.39	1.06	0.84	0.71
	3,000	90% CI	42.91−62.33	49.72−55.84	51.33−53.99	51.79−53.58	51.99−53.36	52.13−53.22	52.23−53.14
		95% CI	41.05−64.19	49.13−56.43	51.08−54.25	51.62−53.76	51.86−53.49	52.03−53.33	52.15−53.23
		Mean	52.62	52.78	52.66	52.69	52.68	52.68	52.69
		Std.	15.61	4.92	2.14	1.44	1.10	0.88	0.73

Table 6. AUC score statistics (%) with balanced sampling for 500 and 3,000 trials.

Model Names	Num Trials	Statistics	Sample Sizes						
			10	100	500	1,000	1,500	2,000	2,500
Xception [24]	500	90% CI	68.57–84.16	74.21–78.76	75.67–77.53	76.06–77.32	76.18–77.13	76.29–77.04	76.38–76.92
		95% CI	67.08–85.66	73.77–79.20	75.49–77.71	75.94–77.44	76.09–77.22	76.22–77.11	76.33–76.98
		Mean	76.37	76.49	76.60	76.69	76.65	76.66	76.65
		Std.	15.69	4.58	1.87	1.27	0.95	0.76	0.55
	3,000	90% CI	69.14–84.68	74.50–79.02	75.69–77.64	76.02–77.29	76.23–77.14	76.30–77.02	76.36–76.93
		95% CI	67.65–86.17	74.07–79.45	75.51–77.82	75.90–77.42	76.14–77.23	76.23–77.09	76.30–76.98
		Mean	76.91	76.76	76.66	76.66	76.69	76.66	76.64
		Std.	15.67	4.55	1.96	1.28	0.92	0.73	0.57
MAT [192]	500	90% CI	66.83–83.07	73.00–77.79	74.79–76.75	74.88–76.21	75.13–76.16	75.31–76.12	75.31–75.93
		95% CI	65.27–84.63	72.54–78.25	74.60–76.94	74.75–76.34	75.03–76.26	75.23–76.20	75.25–75.99
		Mean	74.95	75.39	75.77	75.54	75.64	75.71	75.62
		Std.	16.34	4.81	1.97	1.34	1.05	0.82	0.62
	3,000	90% CI	67.60–83.56	73.37–78.00	74.68–76.61	74.94–76.27	75.13–76.12	75.24–76.03	75.33–75.96
		95% CI	66.07–85.09	72.93–78.45	74.50–76.79	74.82–76.40	75.03–76.21	75.17–76.10	75.27–76.02
		Mean	75.58	75.69	75.65	75.61	75.62	75.63	75.64
		Std.	16.09	4.67	1.94	1.34	1.00	0.79	0.64
RECCE [13]	500	90% CI	57.27–75.43	63.82–69.22	65.88–68.13	66.02–67.53	66.18–67.32	66.39–67.27	66.40–67.14
		95% CI	55.52–77.18	63.30–69.74	65.66–68.35	65.87–67.67	66.07–67.43	66.31–67.35	66.33–67.21
		Mean	66.35	66.52	67.00	66.77	66.75	66.83	66.77
		Std.	18.28	5.43	2.26	1.52	1.14	0.88	0.75
	3,000	90% CI	57.80–75.48	64.02–69.42	65.65–67.90	66.05–67.59	66.25–67.41	66.35–67.26	66.44–67.19
		95% CI	56.11–77.17	63.50–69.94	65.43–68.12	65.90–67.73	66.14–67.52	66.26–67.35	66.37–67.26
		Mean	66.64	66.72	66.77	66.82	66.83	66.80	66.82
		Std.	17.81	5.45	2.27	1.55	1.17	0.92	0.75
Stage5 [65]	500	90% CI	42.69–57.45	48.82–52.75	49.70–51.45	49.80–50.95	50.07–50.95	50.17–50.88	50.20–50.76
		95% CI	41.27–58.86	48.44–53.12	49.53–51.62	49.68–51.07	49.99–51.03	50.10–50.95	50.15–50.81
		Mean	50.07	50.78	50.57	50.38	50.51	50.53	50.48
		Std.	20.02	5.33	2.37	1.57	1.19	0.97	0.76
	3,000	90% CI	43.18–57.79	48.23–52.59	49.59–51.37	49.81–51.03	50.06–50.96	50.13–50.83	50.22–50.75
		95% CI	41.78–59.20	47.81–53.01	49.42–51.55	49.70–51.14	49.97–51.04	50.06–50.90	50.16–50.80
		Mean	50.49	50.41	50.48	50.42	50.51	50.48	50.48
		Std.	19.36	5.78	2.36	1.61	1.19	0.94	0.71
FSTMatching [39]	500	90% CI	47.67–66.44	54.04–59.66	55.80–58.10	56.09–57.72	56.39–57.53	56.45–57.40	56.53–57.28
		95% CI	45.87–68.24	53.49–60.20	55.58–58.32	55.93–57.88	56.29–57.64	56.36–57.49	56.46–57.35
		Mean	57.06	56.85	56.95	56.91	56.96	56.92	56.90
		Std.	18.88	5.66	2.31	1.64	1.15	0.96	0.76
	3,000	90% CI	47.06–66.20	54.10–59.72	55.84–58.24	56.16–57.78	56.27–57.49	56.45–57.37	56.56–57.31
		95% CI	45.23–68.04	53.56–60.26	55.61–58.47	56.00–57.93	56.16–57.61	56.36–57.45	56.49–57.38
		Mean	56.63	56.91	57.04	56.97	56.88	56.91	56.93
		Std.	19.29	5.66	2.42	1.63	1.23	0.92	0.75
MetricLearning [14]	500	90% CI	64.48–76.77	67.55–71.48	68.57–70.18	69.01–70.08	68.99–69.81	69.22–69.78	69.28–69.69
		95% CI	63.29–77.95	67.18–71.85	68.41–70.33	68.90–70.19	68.91–69.89	69.17–69.83	69.24–69.73
		Mean	70.62	69.51	69.37	69.55	69.40	69.50	69.49
		Std.	16.26	5.19	2.13	1.42	1.09	0.73	0.54
	3,000	90% CI	60.11–78.39	66.82–72.23	68.35–70.58	68.78–70.23	68.91–69.96	69.10–69.90	69.20–69.78
		95% CI	58.36–80.14	66.30–72.74	68.14–70.79	68.64–70.37	68.80–70.07	69.03–69.98	69.14–69.83
		Mean	69.25	69.52	69.47	69.50	69.44	69.50	69.49
		Std.	17.56	5.20	2.14	1.40	1.02	0.76	0.56
LRNet [156]	500	90% CI	46.90–64.71	51.93–57.71	53.66–56.02	54.26–55.81	54.44–55.61	54.55–55.43	54.69–55.32
		95% CI	45.19–66.43	51.37–58.27	53.43–56.25	54.11–55.96	54.32–55.73	54.46–55.52	54.63–55.38
		Mean	55.81	54.82	54.84	55.04	55.03	54.99	55.01
		Std.	17.93	5.83	2.38	1.57	1.19	0.89	0.64
	3,000	90% CI	45.24–63.92	52.49–58.15	53.84–56.20	54.25–55.76	54.47–55.60	54.57–55.46	54.68–55.33
		95% CI	43.45–65.71	51.95–58.69	53.62–56.43	54.10–55.90	54.36–55.71	54.48–55.55	54.62–55.40
		Mean	54.58	55.32	55.02	55.00	55.03	55.02	55.01
		Std.	18.82	5.70	2.37	1.52	1.14	0.90	0.66

Table 7. AUC score statistics (%) with imbalanced sampling for 500 and 3,000 trials.

| Model Names	Num Trials	Statistics	Sample Sizes						
			10	100	500	1,000	1,500	2,000	2,500
Xception [24]	500	90% CI	–	74.64–78.75	75.80–77.60	75.98–77.21	76.12–77.19	76.17–77.06	76.31–77.15
		95% CI	–	74.25–79.15	75.62–77.77	75.86–77.33	76.02–77.29	76.09–77.15	76.23–77.23
		Mean	–	76.70	76.70	76.60	76.65	76.62	76.73
		Std.	–	5.43	2.38	1.62	1.41	1.18	1.11
	3,000	90% CI	–	74.78–78.80	75.79–77.56	76.02–77.29	76.13–77.17	76.20–77.11	76.24–77.05
		95% CI	–	74.39–79.19	75.61–77.73	75.89–77.41	76.03–77.27	76.11–77.20	76.17–77.13
		Mean	–	76.79	76.67	76.65	76.65	76.66	76.65
		Std.	–	5.33	2.35	1.69	1.38	1.21	1.07
MAT [192]	500	90% CI	–	74.03–78.10	74.64–76.44	75.09–76.35	75.09–76.11	75.24–76.14	75.23–76.02
		95% CI	–	73.64–78.49	74.47–76.62	74.97–76.47	74.99–76.21	75.15–76.23	75.16–76.09
		Mean	–	76.07	75.54	75.72	75.60	75.69	75.63
		Std.	–	5.24	2.32	1.62	1.32	1.16	1.01
	3,000	90% CI	–	73.65–77.52	74.71–76.41	75.07–76.28	75.16–76.16	75.25–76.10	75.25–76.03
		95% CI	–	73.28–77.89	74.55–76.58	74.96–76.40	75.07–76.26	75.17–76.18	75.18–76.10
		Mean	–	75.58	75.56	75.68	75.66	75.68	75.64
		Std.	–	5.13	2.26	1.60	1.32	1.13	1.03
RECCE [13]	500	90% CI	–	64.69–69.07	65.67–67.67	66.32–67.58	66.27–67.36	66.35–67.31	66.37–67.26
		95% CI	–	64.27–69.50	65.48–67.86	66.20–67.70	66.17–67.46	66.26–67.40	66.29–67.34
		Mean	–	66.88	66.67	66.95	66.81	66.83	66.82
		Std.	–	5.80	2.64	1.66	1.44	1.27	1.17
	3,000	90% CI	–	64.67–68.97	65.85–67.78	66.17–67.56	66.26–67.38	66.34–67.28	66.38–67.25
		95% CI	–	64.26–69.38	65.66–67.96	66.04–67.69	66.16–67.48	66.25–67.37	66.30–67.33
		Mean	–	66.82	66.81	66.87	66.82	66.81	66.81
		Std.	–	5.69	2.56	1.84	1.47	1.25	1.15
Stage5 [65]	500	90% CI	–	48.58–53.40	49.39–51.49	49.74–51.14	50.00–51.09	50.08–50.98	50.13–50.83
		95% CI	–	48.12–53.86	49.19–51.69	49.61–51.27	49.89–51.19	50.00–51.07	50.07–50.90
		Mean	–	50.99	50.44	50.44	50.54	50.53	50.48
		Std.	–	6.37	2.78	1.85	1.44	1.18	0.92
	3,000	90% CI	–	48.10–52.96	49.49–51.58	49.78–51.18	49.97–51.02	50.06–50.91	50.11–50.81
		95% CI	–	47.64–53.42	49.29–51.78	49.65–51.31	49.87–51.12	49.98–50.99	50.04–50.88
		Mean	–	50.53	50.53	50.48	50.49	50.49	50.46
		Std.	–	6.44	2.77	1.85	1.39	1.13	0.93
FSTMatching [39]	500	90% CI	–	54.79–59.43	55.72–57.70	56.33–57.62	56.45–57.46	56.53–57.33	56.64–57.30
		95% CI	–	54.34–59.87	55.53–57.89	56.20–57.74	56.35–57.56	56.45–57.40	56.57–57.37
		Mean	–	57.11	56.71	56.97	56.96	56.93	56.97
		Std.	–	6.14	2.62	1.71	1.34	1.06	0.88
	3,000	90% CI	–	54.53–59.18	56.08–58.08	56.21–57.55	56.42–57.44	56.52–57.36	56.60–57.27
		95% CI	–	54.08–59.63	55.89–58.27	56.09–57.68	56.33–57.53	56.44–57.44	56.54–57.34
		Mean	–	56.86	57.08	56.88	56.93	56.94	56.94
		Std.	–	6.16	2.64	1.77	1.34	1.11	0.89
MetricLearning [14]	500	90% CI	–	66.94–72.04	68.33–70.44	68.75–70.16	69.06–70.11	69.06–69.91	69.02–69.76
		95% CI	–	66.45–72.53	68.13–70.64	68.62–70.29	68.96–70.21	68.98–69.99	68.95–69.83
		Mean	–	69.49	69.38	69.46	69.59	69.48	69.39
		Std.	–	6.74	2.78	1.85	1.38	1.12	0.97
	3,000	90% CI	–	67.07–72.09	68.45–70.49	68.72–70.12	68.93–70.01	69.07–69.91	69.16–69.87
		95% CI	–	66.59–72.58	68.25–70.69	68.58–70.25	68.83–70.12	68.99–69.99	69.09–69.94
		Mean	–	69.58	69.47	69.42	69.47	69.49	69.51
		Std.	–	6.65	2.71	1.86	1.43	1.12	0.94
LRNet [156]	500	90% CI	–	52.42–57.52	54.03–56.28	54.33–55.77	54.56–55.71	54.48–55.37	54.59–55.42
		95% CI	–	51.93–58.01	53.81–56.50	54.19–55.91	54.45–55.81	54.40–55.46	54.51–55.50
		Mean	–	54.97	55.15	55.05	55.13	54.93	55.01
		Std.	–	6.56	2.90	1.85	1.47	1.14	1.06
	3,000	90% CI	–	52.53–57.66	53.86–56.03	54.32–55.78	54.43–55.56	54.56–55.46	54.64–55.39
		95% CI	–	52.04–58.16	53.65–56.24	54.17–55.92	54.32–55.67	54.47–55.54	54.57–55.46
		Mean	–	55.10	54.95	55.05	55.00	55.01	55.01
		Std.	–	6.80	2.88	1.95	1.50	1.19	0.99
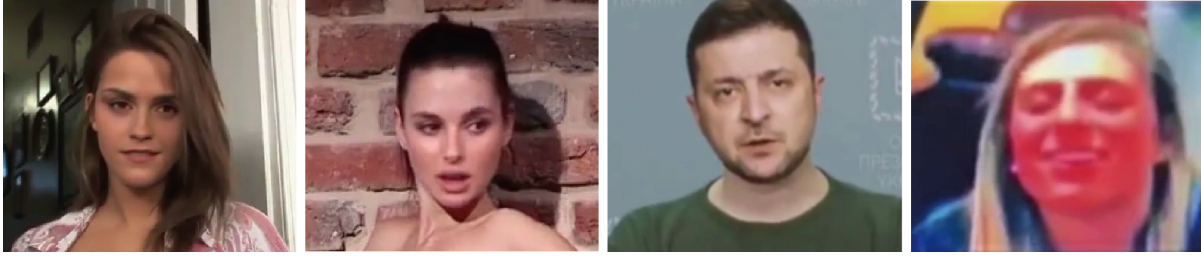
Fig. 2. Screenshots of the four real-life Deepfake videos for Deepfake detection in the case study. The videos are hyper-realistic with different resolutions and no obvious artifacts can be observed visually by human eyes.

that have achieved promising performance on merely the fake testing sets have led to higher accuracy values for mean and confidence intervals. The AUC score metric, as mentioned in Section 6.1, is impervious regarding the imbalanced dataset. Therefore, mean values and confidence intervals under balanced and imbalanced sampling options are generally identical.

With a closer look at the tables, the leading approaches, Xception [24] and MAT [192], both have achieved mean accuracies above 68% and confidence intervals around 69% with respect to the balanced sampling option. All models regarding the interpretability and robustness topics have derived accuracies and confidence intervals around 50%. Stage5 [65] and MetricLearning [14] convey results with all values being 50% since the former recognizes all candidate samples as real and the latter classifies all as fake when checking the predicted labels accordingly. While Stage5 [65] may only be interpretable when facing GAN-based synthetic contents depending on its model design, MetricLearning [14] relies on a fixed threshold of 5 upon the output value without softmax or sigmoid activation. Since a perfect threshold value may vary depending on the testing dataset, the fixed threshold may be the main fact that causes the mistake. By looking at the AUC scores in Table 6 and Table 7, MetricLeaning [14] actually has displayed a reasonable ability in distinguishing real and fake with a mean value of 69.49% AUC score. As for FSTMatching [39] and LRNet [156], the experimental results are generally consistent with the ones in Table 2.

As for experiments under the imbalanced sampling setting, besides the foreseeably high and low performance by MetricLearning [14] and Stage5 [65] as discussed, Xception [24] wins with the highest mean accuracy of 66.23% due to its stable performance in detecting both real and fake faces. Regarding the AUC scores, ignoring the imperceptible differences and taking a look at Table 6, 76.64% is derived by Xception because of its ability to separate real and fake samples at a certain threshold even though performing relatively unsatisfactory with the threshold value of 0.5 regarding the softmax output scores. Besides, MAT [192] is the only other model that has achieved above 70% AUC score. In the remaining approaches, RECCE has reached above 65% AUC scores, while the other two methods have performed relatively unsatisfactory in comparison.

It is also worth noting that despite the imbalanced sampling setting does not affect the final values of mean and confidence intervals, it may cause the AUC score incomputable for a tiny sample size. In particular, as shown in Table 7, there is a high possibility to randomly draw 10 samples that belong to the same category, which then leads to an incomputable AUC score since the sample set lacks data from the other category. Meanwhile, it is unlikely to randomly draw 100 or even more samples that are of the same category even though there is a possibility theoretically.

7 Real-life Case Study

In this section, we made use of the experiment results of the model reliability study. Experiments are conducted to analyze the reliabilities of the existing Deepfake detection approaches when applied to real-life cases. Specifically, four famous Deepfake cases that have jeopardized individuals and society from 2018 to 2022 are considered in this case study (Fig. 2). In 2018, when the technique of Deepfake had just been released shortly, the well-known actress Emma Watson who performed in the Harry Potter movie series was face-swapped onto porn videos [88]. In the same year, another famous actress Natalie Portman encountered a similar fake scandal because of Deepfake [97]. The porn videos were widely spread at the time and had gravely influenced their reputations because the term 'Deepfake' was unfamiliar to the public and people were easily tricked and believed the videos to be genuine upon their first appearances. Later in March 2021, a Bucks County mom was accused of creating Deepfake videos of the underage girls on her daughter's cheerleader team and threatening them to quit the team [22]. The videos exhibited the girls that were naked, drinking alcohol, or vaping, and are accused to be fake. Nevertheless, two months later in May, the prosecutors admitted that they could not prove the fake-video claims without reliable tools and evidence [62]. This is a representative case that confirms that anyone can become a victim of Deepfake nowadays. One of the most famous and most recent Deepfake events, the fake Zelensky video, had caused a short panic in the country during the Russia-Ukraine war [120]. In that fake video, a synthetic president Zelensky was telling Ukrainians to put down their weapons and give up resistance.

7.1 Detection Results

We obtained the available video clips of the four Deepfake cases from the internet and performed Deepfake detection using each of the well-trained models. Specifically, for video clips with sufficient numbers of image frames that contain faces, we randomly extracted 100 frames for face cropping when using frame-level detectors. As for the cheerleader case, as the sensitive contents are omitted, we were only able to acquire a total of 75 faces from the news clip. For frame-level detectors, the numbers of faces classified as real or fake by each model for each video are exhibited in Table 8, and except MetricLearning [14] that uses a fixed threshold, the softmax scores for each video are averaged to obtain a single score that indicates the model determined authenticity. In particular, except for MetricLearning, the fake scores lie in the range of [0, 1] where 0 refers to real and 1 refers to fake. For video-level detectors, the ultimate results are straightforwardly demonstrated in the table. As a result, given the fact that all four videos are known to be fake, about half of the selected models have made the correct classifications regarding both the number of fake faces and the average softmax score. Besides, we provided reliably quantified 95% confidence intervals regarding the models' detection accuracy for reference when utilizing the results.

Despite achieving high statistical values regarding the accuracy and AUC score metrics in early experiments, Xception has failed to classify all four fake videos such that most faces are classified as real and all average softmax scores are below 0.43. Besides, the fake Emma Watson and Natalie Portman videos have also tricked the MAT model such that roughly two-thirds of the faces are classified as real and the final scores are below 0.4. Stage5 and FSTMatching, which proved to be relatively unsatisfactory in early experiments and discussions, have both failed to detect all four fake videos. MetricLearning and LRNet, although performing poorly on lab-controlled datasets, have shown robust detection ability especially because the videos circulated and downloaded from the internet have suffered incrementally heavy compressions. The RECCE model, as a result, turns out to be the winner with the highest fake scores and the number of correctly classified faces when facing real-life Deepfake suspects.

7.2 Deepfake Intelligence

Considering real-life usages, accessible Deepfake detection models such as the ones being discussed in this survey can be gathered to an online platform and to provide real-time detection services. To go even further, similar to

Table 8. Deepfake detection results on real-life cases by the well-trained models. 95% confidence intervals following the balanced sampling option for accuracies are listed along the models. (†: threshold fixed as 5 where greater values refers to fake; ‡: video-level detector with no intermediate frame-level result.)

| Model | Real-life Videos | | | | | | | | 95% CI Balanced Sampling ACC (%) |
| | Watson (2018) | | Portman (2018) | | Cheerleader (2021) | | Zelensky (2022) | | |
	# Real / Fake	Fake Score	# Real / Fake	Fake Score	# Real / Fake	Fake Score	# Real / Fake	Fake Score	
Xception [24]	78 / 22	0.290 (Real)	59 / 41	0.426 (Real)	68 / 7	0.401 (Real)	69 / 31	0.328 (Real)	68.42–69.34
MAT [192]	66 / 34	0.365 (Real)	67 / 33	0.378 (Real)	3 / 72	0.951 (Fake)	34 / 66	0.630 (Fake)	68.41–69.37
RECCE [13]	10 / 90	0.779 (Fake)	8 / 92	0.808 (Fake)	0 / 75	0.902 (Fake)	15 / 85	0.723 (Fake)	60.29–61.30
Stage5 [65]	100 / 0	0.000 (Real)	100 / 0	0.000 (Real)	75 / 0	0.000 (Real)	100 / 0	0.000 (Real)	50.00–50.00
FSTMatching [39]	94 / 6	0.259 (Real)	89 / 11	0.261 (Real)	75 / 0	0.113 (Real)	95 / 5	0.245 (Real)	51.99–52.76
MetricLearning† [14]	35 / 65	11.698 (Fake)	27 / 73	11.996 (Fake)	23 / 52	11.928 (Fake)	27 / 73	11.951 (Fake)	50.00–50.00
LRNet‡ [156]	–	0.655 (Fake)	–	0.667 (Fake)	–	0.667 (Fake)	–	0.538 (Fake)	51.55–52.43

VirusTotal [154], a famous virus and malware intelligence platform that reports malicious threat intelligence, a fake video intelligence web portal can be established regarding the Deepfake detection research domain by endlessly integrating the detection models and real-life Deepfake intelligence beyond detection.

8 Discussion

Our study provides a scientific workflow to prove the reliability of the detection models when applied to real-life cases. Unlike research that simply enumerates detection performance on each benchmark dataset, the reliability study scheme derives statistical claims regarding the detectors on arbitrary candidate Deepfake suspects with the help of confidence intervals. Particularly, the interval values are reliable based on sufficient trials from a sampling frame that ideally imitates real-world Deepfake distribution according to CLT. Considering that the prosecutors are unable to prove the fake-video claim due to the challenge brought to the video evidence authentication standard by Deepfake, the experiment results in Section 4 have solved the problem favorably. Specifically, the model reliability study scheme can be qualified by expert witnesses for the validity of the detection models based on the expert's testimony following corresponding rules, and thus, the reliable statistical metrics regarding the detection performance may assist the video evidence for criminal investigation cases. The accuracies are picked to assist the claims while the AUC scores are more helpful at the research level.

As a result, a reliable justification can be claimed based on values in Table 4 and Table 5 with the help of the confidence intervals and mean values once a sampling option is chosen. For example, suppose the RECCE model is used to help justify the authenticity of the cheerleader video of the 'Deepfake mom' case following the balanced sampling results, a claim can be made such that the video is fake with accuracies lying in the range of [60.37%, 61.22%] and [60.29%, 61.30%] with 90% and 95% confidence levels, respectively. In other words, we are 90% and 95% confident to declare that the video is fake with accuracies in the range of [60.37%, 61.22%] and [60.29%, 61.30%], respectively. If the imbalanced sampling option is trusted, a similar claim can be concluded in the range of [62.36%, 63.66%] and [62.24%, 63.79%] accuracies for the 90% and 95% confidence levels, respectively.

In real life, since the authenticity of the candidate suspect is normally unknown, based on the experiment results in this study, the dominant MAT model is likely to be adopted for Deepfake detection and the following conclusion can be provided following the balanced sampling setting: we are 95% confident to classify the video as real (or fake) with an accuracy between 68.41% and 69.37%. If the imbalanced sampling setting is trusted, the winning Xception model can be employed to offer the justification such that the video is real (or fake) with an accuracy between 65.37% and 67.07% at the 95% confidence level.

Meanwhile, several findings can be concluded based on the results in Table 2 and Tables 4 to 8. Firstly, trade-offs are objectively unavoidable when attempting each of the three challenges. Specifically, regarding Table 2, models with promising transferability on lab-controlled benchmark datasets might lack interpretability on their performance and robustness in sophisticated real-life scenarios. While successfully explaining the detection decision with pellucid evidence and common sense or smoothly resolving challenges in specific real-life conditions with robustness as reported in the published papers, there usually remains limited computational power and model ability to enrich the transferability and detection accuracy on unseen data. Secondly, although the detection performance in Sections 6.2 and 6.3 is unremarkable compared to other approaches, the RECCE model appears to derive the best detection results on real-life Deepfake videos given the fact that we are aware that they are all fake, while on the contrary, the winning MAT and Xception models from early experiments both have failed in classifying multiple real-life fake videos. In other words, based on the experiment results, at the current research stage, a detection model showing promising performance on the benchmark datasets does not necessarily perform well on real-life Deepfake materials. This may be caused by the potential adversarial attacks such that the facial manipulation technique of the fake materials can easily fool models that are mainly based on certain feature extraction perspectives or techniques, and the detectors, therefore, need to be improved to better cooperate with the reliability study scheme. Lastly, for models that have a large gap between accuracy and AUC score values, it may be meaningful to locate a classification threshold other than 0.5 in order to achieve satisfactory detection accuracy since their high AUC scores have highlighted the ability to separate real and fake data.

9 Conclusion

This paper provides a thorough survey of reliability-oriented Deepfake detection approaches by defining the three challenges of Deepfake detection research: transferability, interpretability, and robustness. While the early methods mainly solve puzzles on seen data, improvements by persistent attempts have gradually shown promising up-to-date detection performance on unseen benchmark datasets. Considering the lack of usage for the well-trained detection models to benefit real life and even specifically for criminal investigation, this paper conducts a comprehensive survey regarding model reliability by introducing an unprecedented model reliability study scheme that bridges the research gap and provides a reliable path for applying on-the-shelf detection models to assist prosecutions in court for Deepfake related cases. A barely discussed standardized data preparation workflow is simultaneously designed and presented for the reference of both starters and veterans in the domain. The reliable accuracies of the detection models derived by random sampling at the 90% and 95% confidence levels are informative and may be adopted as or to assist forensic evidence in court for Deepfake-related cases under the qualification of expert testimony.

Based on the informative findings in validating the detection models, potential future research trends are worth discussing. Although a Deepfake detection model can be verified the reliability for real-life usages via the presented reliability study scheme in this paper, an ideal model that simultaneously solves transferability, interpretability, and robustness challenges is not yet accomplished in the current research domain. Consequently, obvious trade-offs have been observed when resolving each of the three challenges. Therefore, considering that the ground-truth attacks and labels of current and future Deepfake contents will not be visible to victims in Deepfake-related cases, researchers may continuously advance the detection model performance regarding each of the three challenges, but more importantly, a model that satisfies all three goals at the same time is urgently desired. At the same time, based on the model reliability study scheme that is first put forward in this study, subsequent improvements and discussions can also be conducted to achieve the general reliability goal progressively. For instance, tracing original sources of synthetic contents and recovering synthetic operation sequences are worth exploiting in future work to further enhance the reliability of a falsification. Moreover, although videos in this study are either real or fake as a whole, as the real-world Deepfake materials become complex, videos with only

partial image frames being fake can lead to further potential risks, and the corresponding benchmark datasets and detectors are also desired.

10 Acknowledgements

References

[1] U.S Food & Drug Administration. 2022. COVID-19 Antigen Home Test Package Insert for Healthcare Providers. https://www.fda.gov/media/152698/download. Accessed: 2022-09-09.

[2] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (2018), 1–7.

[3] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. 2020. Detecting Deep-Fake Videos from Appearance and Behavior. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. 1–6. https://doi.org/10.1109/WIFS49906.2020.9360904

[4] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. 2019. Deepfake Video Detection through Optical Flow Based CNN. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 1205–1207.

[5] Rima Sabina Aouf. 2019. Museum creates deepfake Salvador Dalí to greet visitors. https://www.dezeen.com/2019/05/24/salvador-dali-deepfake-dali-musuem-florida/. Accessed: 2022-09-08.

[6] Yong Bai, Yuanfang Guo, Jinjie Wei, Lin Lu, Rui Wang, and Yunhong Wang. 2020. Fake Generated Painting Detection Via Frequency Analysis. In *2020 IEEE International Conference on Image Processing (ICIP)*. 1256–1260. https://doi.org/10.1109/ICIP40778.2020.9190892

[7] Federico Baldassarre, Quentin Debard, Gonzalo Fiz Pontiveros, and Tri Kurniawan Wijaya. 2022. Quantitative Metrics for Evaluating Explanations of Video DeepFake Detectors. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press. https://bmvc2022.mpi-inf.mpg.de/0972.pdf

[8] Le Minh Binh and Simon Woo. 2022. ADD: Frequency Attention and Multi-View Based Knowledge Distillation to Detect Low-Quality Compressed Deepfake Images. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 1 (Jun. 2022), 122–130. https://doi.org/10.1609/aaai.v36i1.19886

[9] Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. 2021. Video Face Manipulation Detection Through Ensemble of CNNs. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 5012–5019. https://doi.org/10.1109/ICPR48806.2021.9412711

[10] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*. https://openreview.net/forum?id=B1xsqj09Fm

[11] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature Verification Using a "Siamese" Time Delay Neural Network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems (NIPS'93)*. 737–744.

[12] R.S. Brown. 2010. Sampling. In *International Encyclopedia of Education (Third Edition)* (third edition ed.), Penelope Peterson, Eva Baker, and Barry McGaw (Eds.). Elsevier, Oxford, 142–146. https://doi.org/10.1016/B978-0-08-044894-7.00294-3

[13] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. 2022. End-to-End Reconstruction-Classification Learning for Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4113–4122.

[14] Shenhao Cao, Qin Zou, Xiuqing Mao, Dengpan Ye, and Zhongyuan Wang. 2021. Metric Learning for Anti-Compression Facial Forgery Detection. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM 2021)*. 1929–1937.

[15] Nicholas Carlini and Hany Farid. 2020. Evading Deepfake-Image Detectors With White- and Black-Box Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[16] Jiaxin Chen, Xin Liao, Wei Wang, Zhenxing Qian, Zheng Qin, and Yaonan Wang. 2022. SNIS: A Signal Noise Separation-based Network for Post-processed Image Forgery Detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2022), 1–1. https://doi.org/10.1109/TCSVT.2022.3204753

[17] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. 2022. Self-supervised Learning of Adversarial Examples: Towards Good Generalizations for DeepFake Detections. In *CVPR*.

[18] Liang Chen, Yong Zhang, Yibing Song, Jue Wang, and Lingqiao Liu. 2022. OST: Improving Generalization of DeepFake Detection via One-Shot Test-Time Training. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.).

[19] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In *MM '20: The 28th ACM International Conference on Multimedia*.

[20] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. 2021. Local Relation Learning for Face Forgery Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 2 (May 2021), 1081–1088. https://ojs.aaai.org/index.php/AAAI/article/view/16193

[21] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. 2023. Voice-Face Homogeneity Tells Deepfake. *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, 3, Article 76 (nov 2023), 22 pages. https://doi.org/10.1145/3625231

[22] Rudy Chinchilla. 2021. Mom Made Deepfake Nudes of Daughter's Cheer Teammates to Harass Them: Police. https://www.nbcphiladelphia.com/news/local/mom-made-deepfake-nudes-of-daughters-cheer-teammates-to-harass-them-police/2740906/. Accessed: 2022-09-08.

[23] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8789–8797. https://doi.org/10.1109/CVPR.2018.00916

[24] Francois Chollet. 2017. Xception: Deep Learning With Depthwise Separable Convolutions. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*. 1251–1258.

[25] Yang Chuming, Daniel Wu, and Ken Hong. 2021. Practical Deepfake Detection: Vulnerabilities in Global Contexts. In *Responsible AI (RAT) - ICLR 2021 workshop*.

[26] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. 2020. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. https://doi.org/10.1109/TPAMI.2020.3009287

[27] EU Health Security Committee. 2022. EU Common list of COVID-19 antigen tests. https://health.ec.europa.eu/system/files/2022-07/covid-19_eu-common-list-antigen-tests_en.pdf. Accessed: 2022-09-09.

[28] Davide Cozzolino, Diego Gragnaniello, and Luisa Verdoliva. 2014. Image forgery localization through the fusion of camera-based, feature-based and pixel-based techniques. In *2014 IEEE International Conference on Image Processing (ICIP)*. 5302–5306. https://doi.org/10.1109/ICIP.2014.7026073

[29] D. Cozzolino and L. Verdoliva. 2020. Noiseprint: A CNN-Based Camera Model Fingerprint. *IEEE Transactions on Information Forensics and Security* 15 (2020), 144–159. https://doi.org/10.1109/TIFS.2019.2916364

[30] Davide Cozzolino Giovanni Poggi Luisa Verdoliva. 2019. Extracting camera-based fingerprints for video forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[31] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. 2020. On the Detection of Digital Face Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[32] Catherine de Weever and S. Wilczek. 2020. Deepfake detection through PRNU and logistic regression analyses.

[33] deepfakes. 2018. FakeApp. https://www.malavida.com/en/soft/fakeapp/. Accessed: 2022-09-08.

[34] deepfakes. 2019. DeepFakes. https://github.com/deepfakes/. Accessed: 2022-09-08.

[35] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[36] Hui Ding, Kumar Sricharan, and Rama Chellappa. 2018. ExprGAN: Facial Expression Editing with Controllable Expression Intensity. In *AAAI*.

[37] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The DeepFake Detection Challenge (DFDC) Dataset. https://doi.org/10.48550/ARXIV.2006.07397

[38] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. 2019. The Deepfake Detection Challenge (DFDC) Preview Dataset. https://doi.org/10.48550/ARXIV.1910.08854

[39] Shichao Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. 2022. Explaining Deepfake Detection by Analysing Image Matching. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 18–35.

[40] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[41] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. 2020. Unmasking DeepFakes with simple Features. https://doi.org/10.48550/ARXIV.1911.00686

[42] Hany Farid. 2022. Creating, Using, Misusing, and Detecting Deep Fakes. *Journal of Online Trust and Safety* 1, 4 (Sep. 2022). https://doi.org/10.54501/jots.v1i4.56

[43] Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. 2012. Image Forgery Localization via Fine-Grained Analysis of CFA Artifacts. *IEEE Transactions on Information Forensics and Security* 7 (2012), 1566–1577.

[44] FFmpeg. 2021. FFmpeg. https://www.ffmpeg.org/. Accessed: 2021-08-29.

[45] Hans Fischer. 2011. *The Central Limit Theorem from Laplace to Cauchy: Changes in Stochastic Objectives and in Analytical Methods*. Springer New York, New York, NY, 17–74. https://doi.org/10.1007/978-0-387-87857-7_2

[46] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Article 304, 12 pages.

[47] Jessica Fridrich and Jan Kodovsky. 2012. Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security* 7, 3 (2012), 868–882. https://doi.org/10.1109/TIFS.2012.2190402

[48] Apurva Gandhi and Shomik Jain. 2020. Adversarial Perturbations Fool Deepfake Detectors. In *2020 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207034

[49] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. 2021. Information Bottleneck Disentanglement for Identity Swapping. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3403–3412. https://doi.org/10.1109/CVPR46437.2021.00341

[50] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *The 27th Neural Information Processing Systems Advances*. 2672–2680.

[51] Dou Goodman, Hao Xin, Wang Yang, Wu Yuesheng, Xiong Junfeng, and Zhang Huan. 2020. Advbox: a toolbox to generate adversarial examples that fool neural networks. arXiv:2001.05574 [cs.LG]

[52] Hong Kong Special Administrative Region Government. 2022. Rapid Antigen Test (RAT) for COVID-19. https://www.coronavirus.gov.hk/pdf/RapAgTest_FAQ_ENG.pdf. Accessed: 2022-09-09.

[53] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. 2022. Exploiting Fine-Grained Face Forgery Clues via Progressive Enhancement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 1 (Jun. 2022), 735–743. https://doi.org/10.1609/aaai.v36i1.19954

[54] Jiazhi Guan, Hang Zhou, Zhibin Hong, Errui Ding, Jingdong Wang, Chengbin Quan, and Youjian Zhao. 2022. Delving into Sequential Patches for Deepfake Detection. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.).

[55] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. 2020. DeepFake Detection by Analyzing Convolutional Traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[56] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. 2020. Fighting Deepfake by Exposing the Convolutional Traces on Images. *IEEE Access* 8 (2020), 165085–165098. https://doi.org/10.1109/ACCESS.2020.3023037

[57] Nick Dufourand Andrew Gully. 2019. Contributing Data to Deepfake Detection Research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. Accessed: 2022-09-08.

[58] Zhiqing Guo, Gaobo Yang, Jiyou Chen, and Xingming Sun. 2021. Fake face detection via adaptive manipulation traces extraction network. *Computer Vision and Image Understanding* 204 (2021), 103170. https://doi.org/10.1016/j.cviu.2021.103170

[59] Zhiqing Guo, Gaobo Yang, Jiyou Chen, and Xingming Sun. 2022. Exposing Deepfake Face Forgeries with Guided Residuals. https://doi.org/10.48550/ARXIV.2205.00753

[60] David Güera and Edward J. Delp. 2018. Deepfake Video Detection Using Recurrent Neural Networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 1–6. https://doi.org/10.1109/AVSS.2018.8639163

[61] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. 2022. Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14930–14942. https://doi.org/10.1109/CVPR52688.2022.01453

[62] Drew Harwell. 2021. Remember the 'deepfake cheerleader mom'? Prosecutors now admit they can't prove fake-video claims. https://www.washingtonpost.com/technology/2021/05/14/deepfake-cheer-mom-claims-dropped/. Accessed: 2022-09-08.

[63] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 9 (2015), 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824

[64] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[65] Yang He, Ning Yu, Margret Keuper, and Mario Fritz. 2021. Beyond the Spectrum: Detecting Deepfakes via Re-synthesis. In *30th International Joint Conference on Artificial Intelligence (IJCAI)*.

[66] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications* 13, 4 (1998), 18–28. https://doi.org/10.1109/5254.708428

[67] Young-Jin Heo, Woon-Ha Yeo, and Byung-Gyu Kim. 2022. DeepFake detection algorithm based on improved vision transformer. *Applied Intelligence* (2022). https://doi.org/10.1007/s10489-022-03867-9

[68] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9 (12 1997), 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

[69] Ashish Hooda, Neal Mangaokar, Ryan Feng, Kassem Fawaz, Somesh Jha, and Atul Prakash. 2022. Towards Adversarially Robust Deepfake Detection: An Ensemble Approach. https://doi.org/10.48550/ARXIV.2202.05687

[70] Chih-Chung Hsu, Yi-Xiu Zhuang, and Chia-Yen Lee. 2020. Deep Fake Image Detection Based on Pairwise Learning. *Applied Sciences* 10, 1 (2020). https://doi.org/10.3390/app10010370

[71] Juan Hu, Xin Liao, Jinwen Liang, Wenbo Zhou, and Zheng Qin. 2022. FInfer: Frame Inference-Based Deepfake Detection for High-Visual-Quality Videos. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 1 (Jun. 2022), 951–959. https://doi.org/10.1609/aaai.v36i1.19978

[72] Juan Hu, Xin Liao, Wei Wang, and Zheng Qin. 2022. Detecting Compressed Deepfake Videos in Social Networks Using Frame-Temporality Two-Stream Convolutional Network. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 3 (2022), 1089–1102. https://doi.org/10.1109/TCSVT.2021.3074259

[73] Dong Huang and Fernando De La Torre. 2012. Facial Action Transfer with Personalized Bilinear Regression. In *Computer Vision – ECCV 2012*, Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 144–158.

[74] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[75] Shehzeen Hussain, Paarth Neekhara, Brian Dolhansky, Joanna Bitton, Cristian Canton Ferrer, Julian McAuley, and Farinaz Koushanfar. 2022. Exposing Vulnerabilities of Deepfake Detection Systems with Robust Attacks. *Digital Threats* 3, 3, Article 30 (feb 2022), 23 pages. https://doi.org/10.1145/3464307

[76] Wombo Studios Inc. 2021. Wombo: Make your selfies sing. https://play.google.com/store/apps/details?id=com.womboai.wombo&hl=en&gl=US. Accessed: 2022-09-08.

[77] Mousa Tayseer Jafar, Mohammad Ababneh, Mohammad Al-Zoube, and Ammar Elhassan. 2020. Forensics and Analysis of Deepfake Videos. In *2020 11th International Conference on Information and Communication Systems (ICICS)*. 053–058. https://doi.org/10.1109/ICICS49469.2020.239493

[78] Hyeonseong Jeon, Youngoh Bang, and Simon S. Woo. 2020. FDFtNet: Facing Off Fake Images Using Fake Detection Fine-Tuning Network. In *ICT Systems Security and Privacy Protection*, Marko Hölbl, Kai Rannenberg, and Tatjana Welzer (Eds.). Springer International Publishing, Cham, 416–430.

[79] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. 2022. FrePGAN: Robust Deepfake Detection Using Frequency-Level Perturbations. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 1 (Jun. 2022), 1060–1068. https://doi.org/10.1609/aaai.v36i1.19990

[80] Shuai Jia, Chao Ma, Taiping Yao, Bangjie Yin, Shouhong Ding, and Xiaokang Yang. 2022. Exploring Frequency Adversarial Attacks for Face Forgery Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 4093–4102. https://doi.org/10.1109/CVPR52688.2022.00407

[81] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In *CVPR*. 2889–2898.

[82] Tackhyun Jung, Sangwon Kim, and Keecheon Kim. 2020. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access* 8 (2020), 83144–83154. https://doi.org/10.1109/ACCESS.2020.2988660

[83] R. E. Kalman. 1960. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering* 82, 1 (03 1960), 35–45. https://doi.org/10.1115/1.3662552

[84] Wonjun Kang, Geonsu Lee, Hyung Il Koo, and Nam Ik Cho. 2022. One-Shot Face Reenactment on Megapixels. https://doi.org/10.48550/ARXIV.2205.13368

[85] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Hk99zCeAb

[86] T. Karras, S. Laine, and T. Aila. 2021. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 12 (dec 2021), 4217–4228. https://doi.org/10.1109/TPAMI.2020.2970919

[87] Katie Katro. 2022. Bucks County mother gets probation in harassment case involving daughter's cheerleading rivals. https://6abc.com/raffaela-spone-bucks-county-pa-cheerleaders-harassment-case-victory-vipers-squad/11939419/. Accessed: 2022-09-08.

[88] Leo Kelion. 2018. Deepfake porn videos deleted from internet by Gfycat. https://www.bbc.com/news/technology-42905185. Accessed: 2022-09-08.

[89] Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. 2020. Deepfakes: Trick or treat? *Business Horizons* 63, 2 (2020), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006 ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING.

[90] Davis King. 2021. dlib 19.22.1. https://pypi.org/project/dlib/. Accessed: 2021-08-29.

[91] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).

[92] Marissa Koopman, Andrea Macarulla Rodriguez, and Zeno Geradts. 2018. Detection of Deepfake Video Manipulation. In *Proceedings of the 20th Irish Machine Vision and Image Processing conference*. 133–136.

[93] Pavel Korshunov and Sébastien Marcel. 2018. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. *CoRR* abs/1812.08685 (2018). arXiv:1812.08685 http://arxiv.org/abs/1812.08685

[94] Prabhat Kumar, Mayank Vatsa, and Richa Singh. 2020. Detecting Face2Face Facial Reenactment in Videos. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2578–2586. https://doi.org/10.1109/WACV45572.2020.9093628

[95] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. 2021. KoDF: A Large-Scale Korean DeepFake Detection Dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 10744–10753.

[96] Laan Labs. 2017. Face Swap Live. https://play.google.com/store/apps/details?id=com.laan.labs.faceswaplive&hl=en&gl=US. Accessed: 2022-09-28.

[97] Dave Lee. 2018. Deepfakes porn has serious consequences. https://www.bbc.com/news/technology-42912529. Accessed: 2022-09-08.

[98] Victor Lempitsky, Andrea Vedaldi, and Dmitry Ulyanov. 2018. Deep Image Prior. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9446–9454. https://doi.org/10.1109/CVPR.2018.00984

[99] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. 2021. Frequency-Aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 6458–6467. https://doi.org/10.1109/CVPR46437.2021.00639

[100] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2019. FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping. *arXiv preprint arXiv:1912.13457* (2019).

[101] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face X-Ray for More General Face Forgery Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5000–5009. https://doi.org/10.1109/CVPR42600.2020.00505

[102] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. 2018. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. 1–7. https://doi.org/10.1109/WIFS.2018.8630787

[103] Yuezun Li and Siwei Lyu. 2019. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[104] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3204–3213. https://doi.org/10.1109/CVPR42600.2020.00327

[105] Jiahao Liang, Huafeng Shi, and Weihong Deng. 2022. Exploring Disentangled Content Information for Face Forgery Detection. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 128–145.

[106] Dongdong Lin, Benedetta Tondi, Bin Li, and Mauro Barni. 2022. Exploiting temporal information to prevent the transferability of adversarial examples against deep fake detectors. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*. 1–8. https://doi.org/10.1109/IJCB54206.2022.10007959

[107] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. 2021. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 772–781. https://doi.org/10.1109/CVPR46437.2021.00083

[108] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[109] Octavio Loyola-González. 2019. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. *IEEE Access* 7 (2019), 154096–154113. https://doi.org/10.1109/ACCESS.2019.2949286

[110] FaceApp Technology Ltd. 2017. FaceApp: Face Editor. https://play.google.com/store/apps/details?id=io.faceapp&hl=en&gl=US. Accessed: 2022-09-28.

[111] Shao-An Lu. 2018. faceswap-GAN. https://github.com/shaoanlu/faceswap-GAN. Accessed: 2022-09-08.

[112] J. Lukas, J. Fridrich, and M. Goljan. 2006. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security* 1, 2 (2006), 205–214. https://doi.org/10.1109/TIFS.2006.873602

[113] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. 2021. Generalizing Face Forgery Detection With High-Frequency Features. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*. 16317–16326.

[114] Federico Marcon, Cecilia Pasquini, and Giulia Boato. 2021. Detection of Manipulated Face Videos over Social Networks: A Large-Scale Study. *Journal of Imaging* 7, 10 (2021). https://doi.org/10.3390/jimaging7100193

[115] Marek MarekKowalski. 2019. FaceSwap. https://github.com/MarekKowalski/FaceSwap. Accessed: 2021-08-29.

[116] Francesco Marra, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. 2017. Blind PRNU-Based Image Clustering for Source Identification. *IEEE Transactions on Information Forensics and Security* 12, 9 (2017), 2197–2211. https://doi.org/10.1109/TIFS.2017.2701335

[117] Luca Martino, David Luengo, and Joaquín Míguez. 2018. *Direct Methods*. Springer International Publishing, Cham, 27–63. https://doi.org/10.1007/978-3-319-72634-2_2

[118] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. 2020. Two-Branch Recurrent Network for Isolating Deepfakes in Videos. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 667–684.

[119] Falko Matern, Christian Riess, and Marc Stamminger. 2019. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. 83–92. https://doi.org/10.1109/WACVW.2019.00020

[120] Joshua Rhett Miller. 2022. Deepfake video of Zelensky telling Ukrainians to surrender removed from social platforms. https://nypost.com/2022/03/17/deepfake-video-shows-volodymyr-zelensky-telling-ukrainians-to-surrender/. Accessed: 2022-09-08.

[121] Yisroel Mirsky and Wenke Lee. 2021. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* 54, 1, Article 7 (jan 2021), 41 pages. https://doi.org/10.1145/3425780

[122] Hello Group Inc. (Momo). 2019. ZAO. https://apps.apple.com/cn/app/id1465199127. Accessed: 2022-09-08.

[123] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. 2018. RSGAN: Face Swapping and Editing Using Face and Hair Representation in Latent Spaces. In *ACM SIGGRAPH 2018 Posters* (Vancouver, British Columbia, Canada) *(SIGGRAPH '18)*. Association for Computing Machinery, New York, NY, USA, Article 69, 2 pages. https://doi.org/10.1145/3230744.3230818

[124] Huy Hoang Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. 2019. Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos. *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)* (2019), 1–8.

[125] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Use of a Capsule Network to Detect Fake Images and Videos. arXiv:1910.12467 [cs.CV]

[126] Sophie J. Nightingale, Shruti Agarwal, Erik Härkönen, Jaakko Lehtinen, and Hany Farid. 2021. Synthetic faces: how perceptually convincing are they? *Journal of Vision* 21, 9 (2021), 2015. https://doi.org/10.1167/jov.21.9.2015

[127] Sophie J. Nightingale and Hany Farid. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences* 119, 8 (2022), e2120481119. https://doi.org/10.1073/pnas.2120481119

[128] Sophie J. Nightingale and Hany Farid. 2022. Synthetic Faces Are More Trustworthy Than Real Faces. *Proceedings of the National Academy of Sciences* 22, 14 (2022), 3068. https://doi.org/10.1167/jov.22.14.3068

[129] Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*. 7184–7193.

[130] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. 2022. DeepFake Detection Based on Discrepancies Between Faces and Their Context. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 10 (2022), 6111–6121. https://doi.org/10.1109/TPAMI.2021.3093446

[131] World Health Organization. 2022. Use of SARS-CoV-2 antigen-detection rapid diagnostic tests for COVID-19 self-testing. https://apps.who.int/iris/bitstream/handle/10665/352350/WHO-2019-nCoV-Ag-RDTs-Self-testing-2022.1-eng.pdf?sequence=1. Accessed: 2022-09-09.

[132] Andreas Osterman, Maximilian Iglhaut, Andreas Lehner, Patricia Späth, Marcel Stern, Hanna Autenrieth, Maximilian Muenchhoff, Alexander Graf, Stefan Krebs, Helmut Blum, Armin Baiker, Natascha Grzimek-Koschewa, Ulrike Protzer, Lars Kaderali, Hanna-Mari Baldauf, and Oliver T. Keppler. 2021. Comparison of four commercial, automated antigen tests to detect SARS-CoV-2 variants of concern. *Medical Microbiology and Immunology* 210, 5 (01 Dec 2021), 263–275. https://doi.org/10.1007/s00430-021-00719-0

[133] Xunyu Pan, Xing Zhang, and Siwei Lyu. 2012. Exposing image splicing with inconsistent local noise variances. In *2012 IEEE International Conference on Computational Photography (ICCP)*. 1–10. https://doi.org/10.1109/ICCPhot.2012.6215223

[134] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic Image Synthesis With Spatially-Adaptive Normalization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2332–2341. https://doi.org/10.1109/CVPR.2019.00244

[135] Bo Peng, Wei Wang, Jing Dong, and Tieniu Tan. 2017. Optimized 3D Lighting Environment Estimation for Image Forgery Detection. *IEEE Transactions on Information Forensics and Security* 12 (2017), 479–494.

[136] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye. 2021. Conformer: Local Features Coupling Global Representations for Visual Recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 357–366. https://doi.org/10.1109/ICCV48922.2021.00042

[137] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. 2021. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. arXiv:2005.05535 [cs.CV]

[138] Francesco Picetti, Sara Mandelli, Paolo Bestagini, Vincenzo Lipari, and Stefano Tubaro. 2020. DIPPAS: A Deep Image Prior PRNU Anonymization Scheme. arXiv:2012.03581 [cs.MM]

[139] Adam Polyak, Lior Wolf, and Yaniv Taigman. 2019. TTS Skins: Speaker Conversion via ASR. https://doi.org/10.48550/ARXIV.1904.08983

[140] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. 2020. DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 4318–4327. https://doi.org/10.1145/3394171.3413707

[141] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 86–103.

[142] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. 2001. Color transfer between images. *IEEE Computer Graphics and Applications* 21, 5 (2001), 34–41. https://doi.org/10.1109/38.946629

[143] Learn & revise. 2019. Deepfakes: What are they and why would I make one? https://www.bbc.co.uk/bitesize/articles/zfkwcqt. Accessed: 2022-09-08.

[144] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 1–11.

[145] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and P. Natarajan. 2019. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. In *CVPR Workshops*.

[146] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic Routing between Capsules. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 3859–3869.

[147] Shota Saito, Yoichi Tomioka, and Hitoshi Kitazawa. 2017. A Theoretical Framework for Estimating False Acceptance Rate of PRNU-Based Camera Identification. *IEEE Transactions on Information Forensics and Security* 12, 9 (2017), 2026–2035. https://doi.org/10.1109/TIFS.2017.2692683

[148] ScienceDaily. 2020. 'Deepfakes' ranked as most serious AI crime threat. https://www.sciencedaily.com/releases/2020/08/200804085908.htm. Accessed: 2021-05-01.

[149] Shaikh Akib Shahriyar and Matthew Wright. 2022. Evaluating Robustness of Sequence-Based Deepfake Detector Models by Adversarial Perturbation. In *Proceedings of the 1st Workshop on Security Implications of Deepfakes and Cheapfakes (WDC '22)*. Association for Computing Machinery, New York, NY, USA, 13–18. https://doi.org/10.1145/3494109.3527194

[150] Zhihua Shang, Hongtao Xie, Zhengjun Zha, Lingyun Yu, Yan Li, and Yongdong Zhang. 2021. PRRNet: Pixel-Region relation network for face forgery detection. *Pattern Recognition* 116 (2021), 107950. https://doi.org/10.1016/j.patcog.2021.107950

[151] Kaede Shiohara and Toshihiko Yamasaki. 2022. Detecting Deepfakes with Self-Blended Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18720–18729.

[152] Dima Shvets. 2022. Reface. https://hey.reface.ai/. Accessed: 2022-09-08.

[153] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun (Eds.).

[154] Hispasec Sistemas. 2004. VirusTotal. https://www.virustotal.com. Accessed: 2023-02-28.

[155] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. 2022. Dual Contrastive Learning for General Face Forgery Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 2 (Jun. 2022), 2316–2324. https://doi.org/10.1609/aaai.v36i2.20130

[156] Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia. 2021. 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 (CVPR)*. 3609–3618.

[157] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 6105–6114.

[158] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S. Woo. 2018. Detecting Both Machine and Human Created Fake Face Images In The Wild. In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security* (Toronto, Canada) *(MPS '18)*. Association for Computing Machinery, New York, NY, USA, 81–87. https://doi.org/10.1145/3267357.3267367

[159] Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred Neural Rendering: Image Synthesis Using Neural Textures. *ACM Trans. Graph.* 38, 4, Article 66 (July 2019), 12 pages.

[160] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. 2016. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*. 2387–2395.

[161] Global Times. 2021. Chinese social media platforms delete actor's accounts of for hurting the nation after controversial photos of Yasukuni Shrine. https://www.globaltimes.cn/page/202108/1231473.shtml. Accessed: 2022-09-27.

[162] Global Times. 2021. Chinese surrogacy scandal actress Zheng Shuang fined $46 million for tax evasion, shows banned. https://www.globaltimes.cn/page/202108/1232636.shtml. Accessed: 2022-09-27.

[163] Global Times. 2021. Works of scandals-hit actress Zhao Wei removed from platforms, following ban on actor Zhang Zhehan for visiting Yasukuni Shrine. https://www.globaltimes.cn/page/202108/1232631.shtml. Accessed: 2022-09-27.

[164] Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez. 2021. DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance. In *Pattern Recognition. ICPR International Workshops and Challenges*, Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani (Eds.). Springer International Publishing, Cham, 442–456.

[165] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion* 64 (2020), 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

[166] David Tonucci. 2005. 44 - New and Emerging Testing Technology for Efficacy and Safety E valuation of Personal Care Delivery Systems. In *Delivery System Handbook for Personal Care and Cosmetic Products*, Meyer R. Rosen (Ed.). William Andrew Publishing, Norwich, NY, 911–929. https://doi.org/10.1016/B978-081551504-3.50049-3

[167] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 10347–10357.

[168] Loc Trinh, Michael Tsang, Sirisha Rambhatla, and Yan Liu. 2021. Interpretable and Trustworthy Deepfake Detection via Dynamic Prototypes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1973–1983.

[169] Soumya Tripathy, Juho Kannala, and Esa Rahtu. 2019. ICface: Interpretable and Controllable Face Reenactment Using GANs. *arXiv preprint arXiv:1904.01909* (2019).

[170] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

[171] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. 2020. CNN-Generated Images Are Surprisingly Easy to Spot… for Now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8692–8701. https://doi.org/10.1109/CVPR42600.2020.00872

[172] Tianyi Wang, Harry Cheng, Kam Pui Chow, and Liqiang Nie. 2023. Deep Convolutional Pooling Transformer for Deepfake Detection. *ACM Transactions on Multimedia Computing, Communications, and Applications* 19, 6.

[173] Tianyi Wang and Kam Pui Chow. 2023. Noise Based Deepfake Detection via Multi-Head Relative-Interaction. *Proceedings of the AAAI Conference on Artificial Intelligence* (2023).

[174] Tianyi Wang, Ming Liu, Wei Cao, and Kam Pui Chow. 2022. Deepfake noise investigation and detection. *Forensic Science International: Digital Investigation* 42 (2022), 301395. https://doi.org/10.1016/j.fsidi.2022.301395 Proceedings of the Twenty-Second Annual DFRWS USA.

[175] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 568–578.

[176] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2021. HifiFace: 3D Shape and Semantic Prior Guided High Fidelity Face Swapping. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 1136–1142. https://doi.org/10.24963/ijcai.2021/157

[177] Yukai Wang, Chunlei Peng, Decheng Liu, Nannan Wang, and Xinbo Gao. 2022. ForgeryNIR: Deep Face Forgery and Detection in Near-Infrared Scenario. *IEEE Transactions on Information Forensics and Security* 17 (2022), 500–515. https://doi.org/10.1109/TIFS.2022.3146766

[178] Mika Westerlund. 2019. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review* 9 (11 2019), 40–53. https://doi.org/10.22215/timreview/1282

[179] Deressa Wodajo and Solomon Atnafu. 2021. Deepfake Video Detection Using Convolutional Vision Transformer. https://arxiv.org/abs/2102.11126

[180] Walt Woods, Jack Chen, and Christof Teuscher. 2019. Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nature Machine Intelligence* 1, 11 (01 Nov 2019), 508–516. https://doi.org/10.1038/s42256-019-0104-6

[181] Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. 2022. Robust Image Forgery Detection over Online Social Network Shared Images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13430–13439. https://doi.org/10.1109/CVPR52688.2022.01308

[182] Haiwei Wu, Jiantao Zhou, Jinyu Tian, Jun Liu, and Yu Qiao. 2022. Robust Image Forgery Detection Against Transmission Over Online Social Networks. *IEEE Transactions on Information Forensics and Security* 17 (2022), 443–456. https://doi.org/10.1109/TIFS.2022.3144878

[183] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. 2018. ReenactGAN: Learning to Reenact Faces via Boundary Transfer. In *ECCV*.

[184] Xi Wu, Zhen Xie, YuTao Gao, and Yu Xiao. 2020. SSTNet: Detecting Manipulated Faces Through Spatial, Steganalysis and Temporal Features. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2952–2956. https://doi.org/10.1109/ICASSP40776.2020.9053969

[185] Ying Xu, Kiran Raja, and Marius Pedersen. 2022. Supervised Contrastive Learning for Generalizable and Explainable DeepFakes Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 379–389.

[186] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing Deep Fakes Using Inconsistent Head Poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8261–8265. https://doi.org/10.1109/ICASSP.2019.8683164

[187] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9458–9467. https://doi.org/10.1109/ICCV.2019.00955

[188] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. 2021. FACIAL: Synthesizing Dynamic Talking Face with Implicit Attribute Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*

(ICCV). 3867–3876.

[189] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. 2017. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing* 26, 7 (2017), 3142–3155.

[190] Shanghang Zhang, Xiaohui Shen, Zhe Lin, Radomír Mech, João P. Costeira, and Jose M.F. Moura. 2018. Learning to Understand Image Blur. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 6586–6595. https://doi.org/10.1109/CVPR.2018.00689

[191] Yunxuan Zhang, Siwei Zhang, Yue He, Cheng Li, Chen Change Loy, and Ziwei Liu. 2019. One-shot Face Reenactment. In *British Machine Vision Conference (BMVC).*

[192] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. 2021. Multi-Attentional Deepfake Detection. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR).* 2185–2194.

[193] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia. 2021. Learning Self-Consistency for Deepfake Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV).* IEEE Computer Society, Los Alamitos, CA, USA, 15003–15013. https://doi.org/10.1109/ICCV48922.2021.01475

[194] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. 2017. Two-Stream Neural Networks for Tampered Face Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* 1831–1839. https://doi.org/10.1109/CVPRW.2017.229

[195] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV).* 2242–2251. https://doi.org/10.1109/ICCV.2017.244

[196] Yuhao Zhu, Qi Li, Jian Wang, Chengzhong Xu, and Zhenan Sun. 2021. One Shot Face Swapping on Megapixels. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).* 4834–4844.

[197] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. 2020. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. 2382–2390. https://doi.org/10.1145/3394171.3413769

[198] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. 2018. Learning Transferable Architectures for Scalable Image Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8697–8710. https://doi.org/10.1109/CVPR.2018.00907