Beyond the Screen: Evaluating Deepfake Detectors under Moiré Pattern Effects

Razaib Tariq Minji Heo Simon S. Woo*
Sungkyunkwan University, South Korea
{razaibtariq,minji.h0224,swoo}@g.skku.edu

Shahroz Tariq
CSIRO's Data61, Australia
shahroz.tariq@data61.csiro.au

Abstract

The detection of deepfakes is crucial for mitigating the societal impact of falsified video content. Despite the development of various algorithms for this purpose, challenges arise for detectors in real-world scenarios, especially when users capture deepfake content from screens and upload it online or when detectors operate on external devices like smartphones, requiring the capture of potential deepfakes through the camera for evaluation. A significant challenge in these scenarios is the presence of Moiré patterns, which degrade image quality and complicate conventional classification methods, notably deep neural networks (DNNs). However, the impact of Moiré patterns on the effectiveness of deepfake detection systems has not been adequately explored. This study aims to investigate how capturing deepfake videos via digital screen cameras affects the accuracy of detection mechanisms. We introduced the Moiré patterns by capturing the display of a monitor using a smartphone camera and conducted empirical evaluations using four widely recognized datasets: CelebDF, DFD, DFDC, and FF++. We compare the performance of twelve SOTA detectors on deepfake videos captured under the influence of Moiré patterns. Our findings reveal a performance decrease of up to 33.1 and 31.3 percentage points for image- and video-based detectors. Therefore, highlighting the challenges posed by Moiré patterns and other naturally induced artifacts is critical for improving the effectiveness of real-world deepfake detection efforts. To facilitate further research, we will release the Moiré pattern impact version of CelebDF, DFD, DFDC, and FF++ datasets with this paper. Our code is available here: https://github.com/Razaib-Tariq/deepmoire

1. Introduction

In the digital age, the phenomenon of deepfakes—a portmanteau of deep learning and fakes—has emerged as a double-edged sword. On one side, it represents the pinnacle of digital creativity, enabling filmmakers, artists, and content creators to push the boundaries of imagination [25]. On the other, it poses a formidable challenge to the very fabric of truth, offering tools for creating highly convincing digital forgeries [4]. These manipulated images and videos, often indistinguishable from genuine content to the untrained eye, have the potential to distort reality, manipulate perceptions, and undermine trust in digital communication [44].

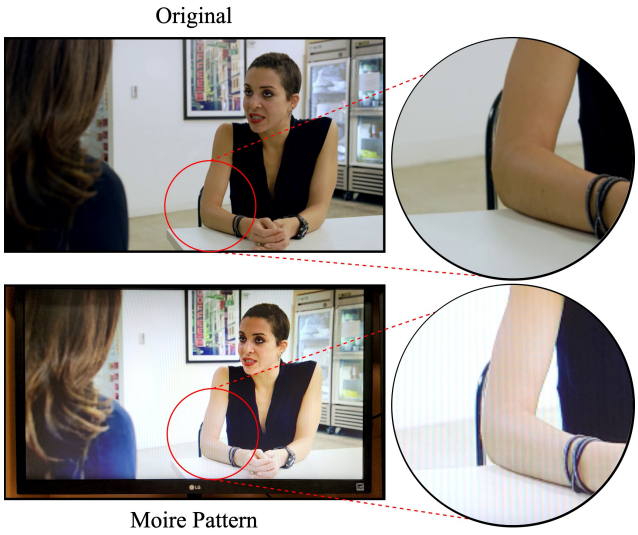

Figure 1. **Original vs. Moiré Pattern:** Comparison between an original frame (without Moiré pattern) and a camera-captured frame (with Moiré pattern).

This burgeoning capability to distort reality has catalyzed an arms race in digital forensics, spurring researchers and technologists to develop methods capable of distinguishing authentic content from manipulated ones [46]. The field of deepfake detection is a testament to this effort, evolving rapidly from rudimentary analysis of visual anomalies [58] to sophisticated machine learning models that scrutinize content at a granular level [13]. Yet, as detection methods become more advanced, so too do the techniques for creating deepfakes [37, 55], perpetuating a cycle of technological escalation.

*Corresponding author

Undoubtedly, the urgency of developing robust detection tools cannot be overstated [26, 30, 54, 56]. And, there is a need for detection methodologies that are not only effective in controlled environments but adaptable to the varied and imperfect conditions of real-world usage [54]. Amidst this, the challenge of Moiré patterns introduces a nuanced complexity to the detection of deepfakes, particularly highlighting the gap between detection capabilities in controlled settings (via benchmark datasets) and practical application challenges. Moiré patterns, often visually intricate and striking artifacts that arise from the camera's capture of digital displays, exemplify one of the unforeseen hurdles in applying effective deepfake detection in everyday scenarios.

Consider a scenario increasingly common in today's digital landscape: an individual encounters a video on a computer screen that raises suspicion of being a deepfake. To verify its authenticity, they turn to a deepfake detection application on their smartphone, capturing the video through the phone's camera for analysis. This act of capturing, seemingly straightforward, unwittingly introduces Moiré patterns into the image or video (see Fig. 1). These patterns, a byproduct of the mismatch between the screen's pixel grid and the camera's sensor, distort the digital content in ways that are often overlooked by human observers but can significantly impede the performance of detection algorithms. The Moiré patterns mask the subtle cues that many deepfake detectors rely on, potentially leading to misidentification or false confidence in the content's authenticity.

This real-world challenge underscores a critical vulnerability in our collective defense against digital disinformation: the efficacy of deepfake detection tools is contingent not only on their algorithmic sophistication but also on their resilience to the imperfect conditions under which they are deployed. The introduction of Moiré patterns through the simple act of capturing a screen with a camera exemplifies the kind of practical challenge that can undermine the integrity of digital media verification processes, highlighting the need for detection techniques that are robust against a variety of real-world complications.

This study aims to understand this critical gap. By conducting a comprehensive evaluation of the impact of Moiré patterns on a wide array of detection algorithms across multiple datasets, we seek to highlight the vulnerabilities of current methodologies and contribute to the development of more resilient detection techniques. The main contributions of our work are as follows:

- **Comprehensive Analysis of Moiré Pattern Impact:** We conducted an extensive empirical study using four deepfake datasets and twelve deepfake detectors to assess how Moiré patterns, introduced by camera-captured deepfake videos displayed on digital screens, affect the performance of deepfake detectors. This is crucial for understanding real-world application challenges.
- **Moiré Pattern Datasets:** For our evaluation, we developed the first Moiré Pattern-impacted version of the deepfake datasets for FF++ [45], CelebDF [36], DFD [24] and DFDC [9], which we will release with this work.
- **Identification of Vulnerabilities and Recommendations:** Through comparative analysis, we highlighted the specific vulnerabilities of state-of-the-art deepfake detectors to Moiré patterns. Also, we made recommendations for future research directions, including the need for more diverse training datasets that include more real-world scenarios such as Moiré-impacted videos and other artifacts.

The paper is structured as follows: We examine related work on deepfake detection and the challenges associated with the Moiré pattern in Section 2. Following this, Section 3 outlines the motivating scenarios, evaluation methodology, and experimental settings. In Section 4, we present our evaluation results along with key insights. Section 5 offers a comprehensive discussion of our findings. Finally, Section 6 concludes the paper.

2. Related Work

This section will examine the various techniques used to identify deepfakes, as well as the challenges caused by the Moiré pattern.

DEEPFAKE DETECTION. The pursuit of reliable deepfake detection has led to the development of sophisticated methodologies, evolving in complexity and specificity to counter the advancing technology of digital manipulation. This section elaborates on two primary branches of detection techniques.

Image-based Detector: At the heart of image-based deepfake detection lies the meticulous analysis of still images, seeking out the subtle fingerprints left behind by manipulation algorithms. Employing machine learning models, notably convolutional neural networks (CNNs), these detectors are trained on extensive datasets composed of both genuine and altered images [17, 18, 20–22, 28, 29, 31–33, 50–52]. Throughout the training process, the model becomes adept at discerning between authentic and manipulated content, extracting critical features such as discrepancies in facial landmarks [41], unnatural texture patterns [34], and anomalies in color distributions [1] or statistical measures [57] that are characteristic of deepfake manipulations.

Upon evaluating a new image, the detector applies its learned expertise to extract these defining features, calculating a probability score that reflects the likelihood of the image being a product of deepfake technology. This score represents the culmination of the detector's analysis, with thresholding techniques often employed to translate this probability into a definitive classification—images surpassing a certain likelihood threshold are flagged as deepfakes, while those below are considered authentic.

Video-based Detectors: Building upon the foundational principles of image-based detection, video-based deepfake detectors introduce an additional layer of analysis by incorporating the temporal dimension inherent to video content. These detectors undertake a dual approach: conducting frame-by-frame evaluations akin to image-based detection for spatial analysis and extending their scrutiny to the temporal dynamics and continuity between frames.

Frame-level analysis in video detectors mirrors the techniques used by their image-based counterparts, with an emphasis on identifying manipulation within each frame. However, video-based detectors distinguish themselves by also analyzing the motion and consistency of objects or facial features over time, searching for discrepancies that might indicate manipulation [53]. This temporal analysis leverages techniques such as motion tracking [35], frame-to-frame comparison [10], and the detection of unnatural changes or discontinuities [16], which are often telltale signs of deepfake interventions. The detector aggregates the results from individual frames and temporal analysis to decide the authenticity of the entire video.

Moreover, some video detectors enrich their analysis by examining the video's audio track, identifying anomalies or inconsistencies that might suggest tampering [12]. This comprehensive approach also includes audio-visual synchronization checks [38], assessing the coherence between audio and image to further bolster the authenticity assessment [39]. Overall, both image and video deepfake detectors aim to identify manipulation artifacts or inconsistencies that distinguish deepfake content from authentic content, with video detectors extending this analysis to temporal dynamics and audio-visual coherence. We use both image and video-based detectors in our evaluations.

Challenges Posed by Moiré Patterns. While considerable progress has been made in detecting deepfakes, less attention has been paid to the impact of Moiré patterns and other naturally induced artifacts on the detection accuracy of deepfake detectors. These patterns, often introduced when digital content is re-captured through another device, can significantly alter the visual information detectors rely on [42]. A few studies have begun to address similar challenges, such as artifact detection in digital photography and video compression artifacts [2], but the specific issue of Moiré patterns in the context of deepfake detection remains underexplored.

Addressing Moiré Patterns: Recent efforts to mitigate the effects of Moiré patterns have primarily focused on image preprocessing techniques, such as digital filtering and artifact reduction algorithms [48]. However, these solutions often require a delicate balance between artifact removal and preserving the fidelity of the underlying content, a challenge that becomes even more complex in the dynamic context of video [43]. The adaptation of these techniques specifically for deepfake detection, where preserving subtle details is crucial as detectors rely on many artifacts in deepfake videos to detect them, represents a novel area of research with significant implications for the field.

Figure 2. **Moiré Pattern-impacted Dataset Generation:** The setup is to display videos from various datasets on a computer screen. These videos are captured using a smartphone camera, leading to the recording of the authentic Moiré pattern that appears on the computer screen.

3. Methodology

In this section, we cover the motivating scenario, evaluation settings, including evaluation datasets, the Moiré pattern introduction process, and detectors used in our evaluation.

3.1. Motivating Scenario

As discussed earlier, in the evolving landscape of digital media, the proliferation of deepfake technology has necessitated the development of sophisticated detection mechanisms. These mechanisms are designed to discern the authenticity of digital content, a task of paramount importance in the era of information warfare and digital disinformation [60]. However, the practical application of deepfake detection technologies often encounters unforeseen challenges that diminish their effectiveness. An ideal example of such a challenge is the introduction of Moiré patterns during the digital content verification process.

Scenario Overview. Imagine an individual who stumbles upon a video on a digital platform or social media that, due to its content or context, raises suspicion of being a deepfake. Seeking to verify its authenticity, the individual opts to use a deepfake detection application available on their smartphone. To facilitate the analysis, the individual captures the video directly from a computer screen using the smartphone's camera. This act of capturing, a seemingly innocuous step, inadvertently introduces Moiré patterns into the captured content.

TECHNICAL IMPLICATIONS. Moiré patterns arise from the interference between the pixel grid of the digital screen and the camera's sensor array, manifesting as visually complex patterns that can significantly alter the appearance of the captured content [5]. For deepfake detection algorithms, particularly those leveraging spatial information for pattern recognition and abnormality detection, Moiré patterns introduce noise and artifacts that can obscure the subtle manipulations indicative of deepfakes. This not only leads to potential false negatives—where deepfakes are mistakenly identified as authentic content—but also false positives, undermining the reliability and trustworthiness of the detection mechanism.

RESEARCH SIGNIFICANCE. This scenario underscores a critical gap in the current state of deepfake detection methodologies—the discrepancy between their accuracy under ideal conditions on benchmark datasets and their practical effectiveness in real-world applications. The presence of Moiré patterns exemplifies the kinds of environmental and operational variables that can significantly impact detection performance. Addressing this challenge requires a nuanced understanding of both the nature of Moiré patterns and their interaction with deepfake detection algorithms.

OBJECTIVE. Motivated by this scenario, our study aims to systematically evaluate the impact of Moiré patterns on the performance of various deepfake detectors across multiple datasets. By simulating the process of capturing digital content through a smartphone camera, we introduce real-world conditions into our evaluation, offering insights into the robustness of detection algorithms against such artifacts. Our goal is to highlight the necessity for detection methodologies that are not only theoretically sound, but also practically resilient, capable of adapting to the imperfect inputs that typify everyday digital content verification efforts.

SCOPE. It is crucial to note that screen captures on the same devices do not introduce Moiré patterns. Hence, we do not consider this scenario as it is expected that typical deepfake detectors will function adequately under these conditions. Instead, we focus specifically on scenarios, where potential deepfake media is displayed on one device (such as a computer screen or television) and deepfake detectors are utilized on another device (such as a smartphone). We maintain a narrow scope for the motivating scenario to facilitate thorough evaluation. However, we will discuss the implications of this setting in Section 5.

3.2. Evaluation Settings

To accurately evaluate the effect of Moiré patterns on deepfake detection performance, we devised a meticulous experimental setup. This setup incorporates four deepfake datasets and twelve deepfake detection methods.

DATASET SELECTION AND PREPARATION. We selected four popular deepfake benchmarking dataset for our experiments: (i) FaceForensics++ (FF++) [45], (ii) Celebrity Deepfake (CelebDF) [36], (iii) Deepfake Detection (DFD) [24] and (iv) Deepfake Detection Challenge (DFDC) [9]. For each dataset, videos were systematically processed to generate two variants: the original, unaltered videos and their counterparts with Moiré patterns. This dual-version setup facilitates a direct comparison of detector performances under standard and compromised conditions, providing a clear lens through which the impact of Moiré patterns on detection accuracy can be assessed. Note that we utilized the uncompressed version of all datasets, ensuring that no additional compression artifacts were introduced, thereby isolating the evaluation to Moiré patterns. However, it is important to acknowledge that employing compressed versions could significantly amplify the impact of performance degradation.

Table 1. **Dataset Distribution:** We generated a total of 536 Moiré pattern (MP) videos using the original (OG) real and deepfake videos from FF++, DFD, DFDC and CelebDF datasets.

Datasets	Total Videos	OG-Videos		MP-Videos	
		Real	Fake	Real	Fake
DF (FF++)	100	25	25	25	25
F2F (FF++)	100	25	25	25	25
FS (FF++)	100	25	25	25	25
NT (FF++)	100	25	25	25	25
DFDC	328	82	82	82	82
DFD	108	28	28	28	28
CelebDF	232	58	58	58	58

DEVELOPMENT OF MOIRÉ PATTERN DATASET. To authentically introduce Moiré patterns, we designed a dataset generation procedure as shown in Fig. 2. This involved the playback of selected videos from our datasets on a high-resolution digital screen, followed by their recapture using a smartphone camera. This process is designed to mimic a realistic scenario in which an individual uses a handheld device to capture digital content displayed on another screen.

Experimental Setup: We employed a high-resolution (1080p) digital screen to exhibit the selected deepfake videos. These videos were subsequently captured using a high-definition smartphone camera (Samsung S22 Plus), positioned to simulate a natural recording scenario typical of an average user. Through this configuration, we amassed a corpus of 536 videos featuring the Moire pattern, each lasting approximately 10 seconds. However, the additional processing time required for file opening and saving resulted in a 30-second duration for recording each video. To maintain the device's operational integrity, we conducted multiple 30-minute recording sessions.

Controlled Variables: To ensure consistency across recordings, we meticulously controlled several variables. These included the distance between the camera and the

screen, the angle of capture, and ambient lighting conditions. Additionally, we standardized the camera settings, such as aperture, shutter speed, and ISO, to minimize potential variations that could affect the visibility of Moiré patterns. Notably, we deliberately fixed the camera in a consistent position to capture only the front-facing angle, thereby reducing variability between the captured video and the original footage. It is important to emphasize that the only alteration in the captured video is the introduction of the Moiré pattern. We intentionally deferred addressing variations such as different smartphone cameras, screen types, capture angles, motion blur, and other factors for more extensive exploration in future studies.

VIDEO SELECTION. We curated a subset of videos from FF++, CelebDF, DFD, and DFDC datasets. To ensure unbiased selection, we employed a Python script for random sampling across these datasets. Details regarding the number of videos used in our experimentation are outlined in Table 1. Maintaining a balanced representation, we adhered to a 1:1 ratio of real to fake videos. Each original video and its corresponding moire pattern variant were matched with their respective deepfake counterparts. Thus, for every original video, our dataset comprises four distinct categories: the original real video, its deepfake counterpart, the real video featuring a moire pattern, and the corresponding moire-patterned deepfake video, respectively.

SELECTION OF DEEPFAKE DETECTORS. In recognizing the diverse landscape of deepfake detection technologies, we selected a comprehensive array of detectors for evaluation. This selection spans both image-based and video-based models, incorporating a variety of underlying algorithms to ensure a broad assessment of current detection capabilities.

Image-based Detectors: These models focus on analyzing individual video frames for signs of manipulation. Their selection was based on demonstrated efficacy in identifying deepfake artifacts within still images, reflecting a wide range of detection strategies.

1. **Self-Blended Images (SBIs)** [47] serve as synthetic training data, aiding robust classifier training by mimicking common deepfake manipulation artifacts.
2. **Multi-attentional Deepfake Detection (MAT)** [61] employs a fine-grained classification approach through a multi-attentional framework.
3. **Rossler et al. (XNet)** [45] adapt XceptionNet by replacing the final layer and training for enhanced performance.
4. **ForgeryNet (FN)** [15] adapt XceptionNet and modifies for applications such as image and video classification, spatial and temporal localization, and enhancing facial manipulation detection in real-world scenarios.
5. **Capsule-forensics (CF)** [40] excels in capturing hierarchical relationships and spatial hierarchies in data, particularly beneficial for detecting forged media content. It integrates dynamic routing algorithms and introduces random noise during training to enhance robustness against various forgery attacks.
6. **The ID-unaware Deepfake Detection Model (CADDM)** [11] incorporates an Artifact Detection Module (ADM) and utilizes the Multi-scale Facial Swap (MFS) method during training. ADM identifies artifact areas in images, reducing reliance on global identity features, while MFS generates synthetic images with annotated artifact areas for training enhancement.
7. **Coccomini et al. (CCViT)** [7] combine convolutional and transformer architectures for deepfake detection, leveraging EfficientNet B0 as a pre-trained convolutional network. This approach introduces a simple yet efficient voting scheme for inference, aggregating scores from multiple faces in videos to determine the presence of manipulation.
8. **Attention-based Deepfake detection Distiller (ADD)** [27] utilize frequency attention and multi-view attention distillation techniques within a Knowledge Distillation framework to enhance the detection of highly compressed deepfake images.

Video-based Detectors: Leveraging temporal information across video frames, these detectors are designed to uncover inconsistencies or artifacts indicative of video manipulation. Their inclusion allows us to assess the added value of temporal analysis in improving detection accuracy, especially in the presence of Moiré patterns.

1. **Altfreezing** [59] introduces a training strategy suitable for data with temporal elements. It involves partitioning the model's weights into spatial and temporal groups and alternately freezing one group's weights during training.
2. **A fully temporal convolution network (FTCN)** [62] emphasizes temporal cues in video face forgery detection. Unlike conventional methods, it prioritizes learning temporal features by constraining spatial-related processing. This focus enables it to effectively capture short-term flickering and inconsistencies in manipulated face videos.
3. **LRNet** [49] integrates precise geometric features with temporal modeling. It preprocesses face images to extract geometric information, refines landmarks using the Lucas-Kanade algorithm and a Kalman filter, and incorporates them into feature sequences for classification by a two-stream Recurrent Neural Network (RNN). We employed LRNet with blazeface (BF) and retinaface (RF) configurations.
4. **LipForensics** [14] entails pre-training a convolutional neural network (CNN) on lipreading tasks using real videos to extract rich representations sensitive to anomalous mouth dynamics. Subsequently, the temporal network is fine-tuned on forged data while maintaining the

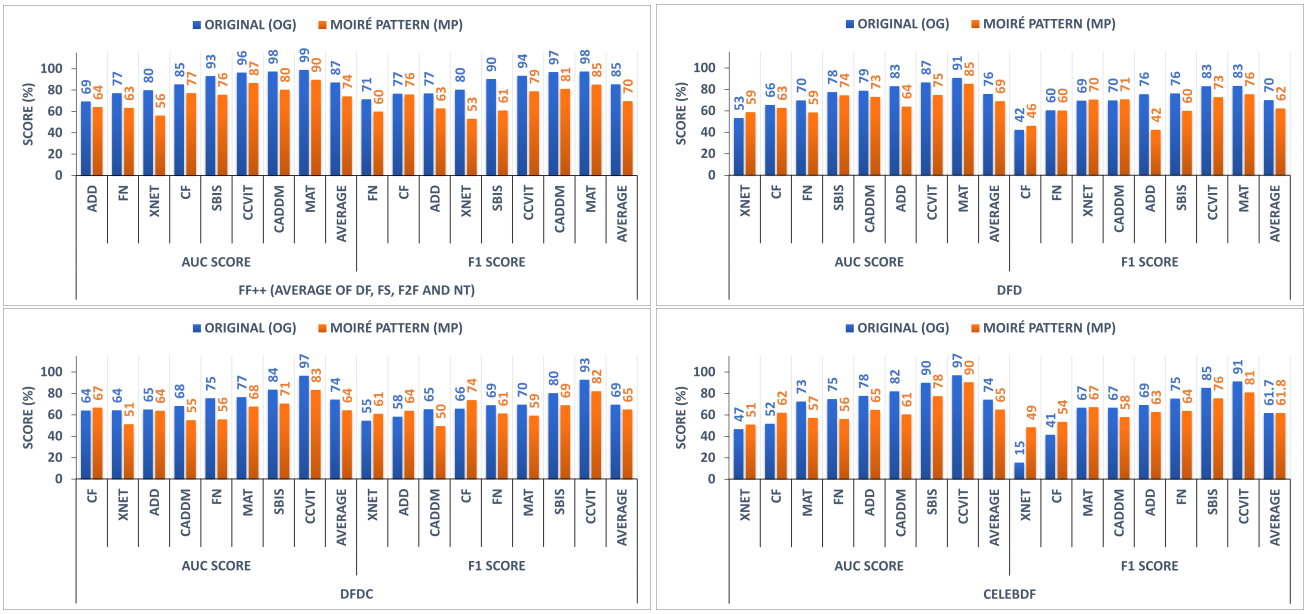

Figure 3. **Image-based Deepfake Detectors:** Detection performance in terms of AUC and F1-score of eight deepfake detectors on four popular deepfake datasets. The x-axis is ordered based on the lowest to highest (left to right) in ascending order of the performance of detectors on the original (OG) for each metric and dataset. The average value has been placed to the far right of each score. We observed a performance reduction of 8.4 percentage points on average and up to 33.1 percentage points in the worst case.

fixed feature extractor. This approach effectively targets inconsistencies in high-level semantic mouth movements, resulting in superior generalization.

Note that to avoid any bias, we exclude Altfreezing and FTCN from our evaluations on the FF++ dataset, as these methods utilize the entire FF++ dataset as the training set.

Preprocessing for evaluation. The original videos and Moiré pattern videos were standardized to a duration of 10 seconds for experimental consistency. Face extraction for image detection was performed using the dlib library [23]. Video-based detection utilized the selected video files directly as inputs. In LRnet [49], experiments were conducted based on two scenarios outlined in the original paper's code: employing BlazeFace [3] and RetinaFace [8]. For LipForensics [14], cropped images focusing on the mouth area from each video were utilized.

Performance Assessment Metrics. To rigorously evaluate the influence of Moiré patterns on deepfake detection performance, we utilized a range of established metrics, encompassing accuracy, area under the ROC curve (AUC), precision, recall, and F1-score. These metrics were chosen to afford a holistic assessment of each detector's efficacy, furnishing insights into their performance across diverse scenarios. Due to space constraints, we present AUC and F1 scores solely for the eight image-based detectors, while results for all metrics are provided for video-based detectors. Additionally, for the curious reader, we intend to include these results in tabular format within our code repository.

4. Results

In this section, we will present and analyze important observations and findings derived from the obtained results.

Image-based Detector Performance. Figure 3 shows the performance of image-based detectors, with Original (OG) results depicted in blue and Moiré Pattern (MP) results in orange. Across eight state-of-the-art deepfake detectors evaluated in this experiment, performance typically ranged from low 80s to high 90s on various metrics for the OG datasets. However, we observed an average 8.4 percentage point decrease in performance on Moiré-Impacted datasets, escalating to 33.1 percentage points in the most extreme cases. For example, on the FF++ dataset, MAT, the top performer on OG data, experienced a drop in performance from 99.0% (AUC) and 97.5% (F1) to 89.8% (AUC) and 85.3% (F1), respectively. Similarly, for DFD, DFDC, and CelebDF datasets, the F1-score decreased from 83.3% (MAT), 92.9% (CCViT), and 91.2% (MAT) to 75.6% (MAT), 82.1% (CCViT), and 81.1% (MAT), respectively. Overall, our evaluation highlights the resilience of state-of-the-art image-based detectors against deepfake manipulation in original datasets, albeit with notable performance degradation in the presence of Moiré patterns. Understanding these performance variations is crucial for advancing the robustness of detection systems in real-world scenarios, where not just Moiré patterns but other naturally induced

Table 2. **Video-based Deepfake Detectors:** Detection performance of four detectors on four deepfake dataset. Here, OG and MP represent the performance on original and moire pattern datasets, whereas Diff. represents the performance difference.

Dataset	Method	Acc			AUC			Precision			Recall			F1-Score		
		OG	*MP*	*Diff.*	*OG*	*MP*	*Diff.*	*OG*	*MP*	*Diff.*	*OG*	*MP*	*Diff.*	*OG*	*MP*	*Diff.*
DF (FF++)	*LRNet+BF*	78.0	60.4	17.6↓	87.2	64.4	22.8↓	80.0	65.2	14.8↓	80.0	62.5	17.5↓	80.0	63.8	16.2↓
	LRNet+RF	76.0	64.0	12.0↓	83.1	75.5	7.6↓	67.6	65.7	1.9↓	100.0	92.0	8.0↓	80.6	76.7	4.0↓
	LipForensics	98.0	84.0	14.0↓	100.0	95.8	4.2↓	100.0	89.3	10.7↓	100.0	100.0	0.0	100.0	94.3	5.7↓
F2F (FF++)	*LRNet+BF*	52.0	55.1	3.1↑	51.3	51.3	0.0	60.0	53.6	6.4↓	12.0	62.5	50.5↑	20.0	57.7	37.7↑
	LRNet+RF	56.0	54.0	2.0↓	56.2	60.2	4.1↑	57.1	64.7	7.6↑	64.0	44.0	20.0↓	60.4	52.4	8.0↓
	LipForensics	100.0	68.0	32.0↓	100.0	95.4	4.6↓	100.0	95.8	4.2↓	100.0	92.0	8.0↓	100.0	93.9	6.1↓
FS (FF++)	*LRNet+BF*	54.0	52.1	1.9↓	62.3	65.0	2.7↑	71.4	78.6	7.1↑	40.0	45.8	5.8↑	51.3	57.9	6.6↑
	LRNet+RF	66.0	52.0	14.0↓	69.3	46.8	22.5↓	65.4	51.2	14.2↓	68.0	84.0	16.0↑	66.7	63.6	3.0↓
	LipForensics	100.0	96.0	4.0↓	100.0	99.0	1.0↓	100.0	100.0	0.0	100.0	92.0	8.0↓	100.0	95.8	4.2↓
NT (FF++)	*LRNet+BF*	42.0	45.8	3.8↑	47.2	48.4	1.2↑	75.0	58.3	16.7↓	12.0	29.2	17.2↑	20.7	38.9	18.2↑
	LRNet+RF	50.0	50.0	0.0	49.4	50.2	0.7↑	51.5	77.8	26.3↑	68.0	28.0	40.0↓	58.6	41.2	17.4↓
	LipForensics	94.0	68.0	26.0↓	98.9	86.4	12.5↓	100.0	86.4	13.6↓	92.0	76.0	16.0↓	95.8	80.8	15.0↓
DFD	*AltFreezing*	72.2	53.7	18.5↓	83.8	63.7	20.1↓	68.6	65.4	3.2↓	92.3	65.4	26.9↓	78.7	65.4	13.3↓
	FTCN	79.6	51.9	27.8↓	90.4	59.1	31.3↓	84.6	53.9	30.8↓	86.3	60.9	25.4↓	85.4	57.1	28.3↓
	LRNet+BF	52.9	57.4	4.5↑	55.4	57.8	2.5↑	55.2	59.3	4.1↑	64.0	61.5	2.5↓	59.3	60.4	1.1↑
	LRNet+RF	64.8	57.4	7.4↓	60.1	58.1	2.0↓	76.9	53.5	23.4↓	38.5	88.5	50.0↑	51.3	66.7	15.4↑
	LipForensics	59.3	53.7	5.6↓	76.7	62.1	14.6↓	70.4	68.4	2.0↓	73.1	50.0	23.1↓	71.7	57.8	13.9↓
DFDC	*AltFreezing*	78.0	51.2	26.8↓	83.7	71.5	12.2↓	84.2	74.3	9.8↓	74.1	65.2	8.9↓	78.7	69.4	9.4↓
	FTCN	70.4	50.8	19.6↓	78.0	54.8	23.2↓	62.1	33.8	28.3↓	69.7	43.8	25.9↓	65.6	38.1	27.6↓
	LRNet+BF	61.7	67.7	6.1↑	60.1	73.3	13.3↑	62.2	71.4	9.2↑	77.7	62.5	15.2↓	69.1	66.7	2.4↓
	LRNet+RF	61.0	50.0	11.0↓	59.9	55.0	4.9↓	60.5	61.6	1.2↑	71.0	59.9	11.2↓	64.7	58.9	5.8↓
	LipForensics	71.8	59.9	11.9↓	83.8	71.0	12.8↓	79.5	65.9	13.6↓	80.5	80.0	0.5↓	80.0	72.3	7.7↓
CelebDF	*AltFreezing*	56.9	50.9	6.0↓	82.3	60.1	22.2↓	88.1	57.0	31.1↓	63.8	77.6	13.8↑	74.0	65.7	8.3↓
	FTCN	54.3	49.1	5.2↓	70.0	46.1	23.8↓	75.9	98.3	22.4↑	72.7	67.5	5.3↓	74.3	80.0	5.7↑
	LRNet+BF	50.0	48.2	1.9↓	51.9	48.8	3.1↓	56.5	60.7	4.2↑	22.4	30.9	8.5↑	32.1	41.0	8.9↑
	LRNet+RF	50.0	50.0	0.0	49.0	53.2	4.2↑	50.5	54.6	4.0↑	86.2	41.4	44.8↓	63.7	47.1	16.6↓
	LipForensics	50.9	52.6	1.7↑	78.2	69.9	8.2↓	68.4	67.7	0.7↓	89.7	72.4	17.3↓	77.6	70.0	7.6↓
Average (ALL)	*AltFreezing*	69.0	51.9	17.1↓	83.3	65.1	18.1↓	80.3	65.6	14.7↓	76.7	69.4	7.3↓	77.1	66.8	10.3↓
	FTCN	68.1	50.6	17.5↓	79.5	53.4	26.1↓	74.2	62.0	12.2↓	76.2	57.4	18.9↓	75.1	58.4	16.7↓
	LRNet+BF	55.8	55.2	0.6↓	59.3	58.4	0.9↓	65.8	63.9	1.9↓	44.0	50.7	6.7↑	47.5	55.2	7.7↑
	LRNet+RF	60.5	53.9	6.6↓	61.0	57.0	4.0↓	61.4	61.3	0.1↓	70.8	62.5	8.3↓	63.7	58.1	5.6↓
	LipForensics	82.0	68.9	13.1↓	91.1	82.8	8.3↓	88.3	81.9	6.4↓	90.8	80.3	10.4↓	89.3	80.7	8.7↓

artifacts could be present.

VIDEO-BASED DETECTOR PERFORMANCE. Table 2 presents the results of video-based detectors across five performance metrics: Accuracy, AUC, Precision, Recall, and F1-score. For each metric, we provide results for OG datasets, MP datasets, and the performance difference (Diff.), with arrows indicating performance drops or increases. Across all metrics, we observed a general trend of performance decrease, averaging 9.1 percentage points overall and reaching 44.8 percentage points in the most extreme cases. In the case of the FF++ dataset, LipForensics, the top performer in terms of AUC, experienced an average performance drop of 8.3 percentage points and up to 14.6 percentage points in the worst case. It is noteworthy that in the FF++ case, none of the other metrics conclusively determined a winner, as all detectors generally performed significantly lower on them. In contrast to FF++, for DFD, DFDC, and CelebDF datasets, performance results remained consistent across all metrics. For instance, FTCN emerged as the best performer on DFD, but experienced a 31.3 percentage point drop in AUC score. On DFDC and CelebDF datasets, AltFreezing saw decreases of 12.2 and 22.2 percentage points in AUC scores, respectively. In summary, our findings underscore the challenges faced by video-based deepfake detectors, with a consistent trend of performance decline observed across various metrics when subjected to Moiré patterns. These insights emphasize the need for further research and development efforts to enhance the efficacy of not just spatial detectors (i.e., image detectors), but also spatiotemporal detectors against naturally induced artifacts in real-world settings.

IMPACT OF METRIC CHOICE ON PERFORMANCE NUMBERS. Our observations yield three significant insights: (i) AUC scores tend to surpass accuracy and F1-score substantially (refer to Fig. 3 and Table 2). Given the absence of class imbalance in our dataset, we attribute this discrepancy in performance values to threshold dependency and the calculation methods of accuracy and AUC scores. Specifically, accuracy is contingent on the chosen classification threshold, measuring performance at a specific threshold, which is 0.5, whereas AUC evaluates the model's discrimination ability across all possible thresholds. (ii) Notably, the reliance on finding the optimal threshold is notably higher within the FF++ dataset for the evaluated de-

tectors compared to other datasets. This observation is underscored by LipForensics' average AUC score of 99.7% on OG and 94.2% on MP within the FF++ dataset, while accuracy and F1-score hover around 50% (refer to Table 2). (iii) The introduction of Moiré patterns resulted in varied performance impacts across metrics, with no discernible trend observed (refer to Table 2). Future investigations should delve into these trends to provide insights. Overall, these insights underscore the importance of carefully considering metric choices to accurately assess the effectiveness of deepfake detection algorithms across diverse datasets.

Cases where Performance Increased. In Table 2, several instances are highlighted where performance improved for Moiré-impacted datasets. However, it is noteworthy that in most cases, the original performance was notably low, hovering around or below 50%. Consequently, even with improvements, the performance often remained within the 50-60 range, which may not convey significant meaning. Moreover, in some cases, while precision increased, recall decreased, resulting in either a slight decrease or improvement in the F1-score. There were two rare cases of notable improvement: LRNet+BF for DFDC, which we observed an increase in AUC from 60.1% to 73.3%, and FTCN for CelebDF, which witnessed an improvement in F1-score from 74.3% to 80.0%. Upon further investigation, we found that while LRNet+BF exhibited an increase in AUC, its F1-score decreased from 69.1% to 66.7%. Similarly, for FTCN, while the F1-score improved, the AUC score decreased from 70.0% to 46.1%. These findings underscore the need for future research to delve deeper into the impact of different metrics, as recommended by previous studies [30].

5. Discussion and Future works

Limited Scenario Scope. Our study focuses on a specific method of generating Moiré patterns—capturing a digital screen with a smartphone camera. While this approach effectively highlights the vulnerability of detectors to such patterns, it does not encompass all potential sources of these artifacts in digital content. For example, Moiré patterns can also arise from the interaction of various digital compression algorithms. Expanding the scope to include these scenarios could yield a more comprehensive understanding of the challenge in future investigations.

Variability in Moiré Patterns. Our current scenario does not consider the variability in Moiré patterns stemming from different screen-camera combinations, lighting conditions, and capture angles. These factors can significantly influence the appearance and intensity of Moiré patterns, potentially impacting detection algorithms in nuanced ways not fully explored in our study. Addressing these variables and including recent deepfake datasets [6, 19] will be a focal point of our future work.

User Behavior and Practicality. An underlying assumption in our study is that individuals would opt for or have the capability to utilize a smartphone app for capturing and analyzing suspect content displayed on another screen. While this scenario is pivotal, there may exist alternative methods of content verification, such as capturing screenshots on the same device, which pose minimal threat to detection performance. Nonetheless, exploring various user behaviors that could lead to the introduction of Moiré patterns and consequent reduction in detection performance was imperative.

Focus on Moiré Patterns Over Other Artifacts. Although Moiré patterns pose a significant and realistic challenge, focusing exclusively on them might overlook other artifacts or issues introduced during the capture process, including motion blur, focus inconsistencies, or digital compression artifacts, which could also influence detection performance. Therefore, future research endeavors should explore these avenues as well.

Filtering with Image Processing. Employing image processing techniques to remove Moiré patterns before inputting images into detectors presents a potential solution. However, our preliminary analysis indicates that filtering Moiré patterns also eliminates certain features utilized by deepfake detectors to discern between authentic and manipulated content. Nonetheless, further exploration of this approach is warranted in future investigations.

6. Conclusion

Our work highlights the significant impact of Moiré patterns on deepfake detection. Our findings demonstrate a reduction in the performance of both image- and video-based deepfake detectors by 33.1% and 31.3%, respectively, when confronted with Moiré patterns. Moreover, we release the Moiré patterns version of four popular benchmark datasets, facilitating researchers and developers in refining the performance of their deepfake detectors. Moving forward, we plan to extend this research by curating a more comprehensive real-world Moiré pattern dataset encompassing various smartphone cameras, screen types, capture angles, motion blur, artifact types, and other pertinent factors.

Acknowledgements. This work was partly supported by Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean government MSIT: (No. 2022-0-01199, Graduate School of Convergence Security at Sungkyunkwan University), (No. 2022-0-00688, AI Platform to Fully Adapt and Reflect Privacy-Policy Changes), (No. 2019-0-00421, AI Graduate School Support Program at Sungkyunkwan University), and (No. RS-2023-00230337, Advanced and Proactive AI Platform Research and Development Against Malicious Deepfakes).

References

[1] Mohammed Thajeel Abdullah and Nada Hussein M Ali. Deepfake detection improvement for images based on a proposed method for local binary pattern of the multiple-channel color space. *International Journal of Intelligent Engineering & Systems*, 16(3), 2023. 2

[2] Eldho Abraham. Moiré pattern detection using wavelet decomposition and convolutional neural network. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1275–1279, 2018. 3

[3] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019. 6

[4] Matt Brown and David Klepper. Fake images made to show trump with black supporters highlight concerns around ai and elections. https://apnews.com/article/deepfake-trump-ai-biden-tiktok-72194f59823037391b3888a1720ba7c2, 2024. 1

[5] Yushi Cheng, Xiaoyu Ji, Lixu Wang, Qi Pang, Yi-Chao Chen, and Wenyuan Xu. mID: Tracing screen photos via Moiré patterns. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2969–2986. USENIX Association, 2021. 4

[6] Beomsang Cho, Binh M Le, Jiwon Kim, Simon Woo, Shahroz Tariq, Alsharif Abuadbba, and Kristen Moore. Towards understanding of deepfake videos in the wild. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4530–4537, 2023. 8

[7] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing*, pages 219–229. Springer, 2022. 5

[8] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 6

[9] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton-Ferrer. The deepfake detection challenge (DFDC) preview dataset. *CoRR*, abs/1910.08854, 2019. 2, 4

[10] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020. 3

[11] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3994–4004, 2023. 5

[12] Yewei Gu, Xianfeng Zhao, Chen Gong, and Xiaowei Yi. Deepfake video detection using audio-visual consistency. In *Digital Forensics and Watermarking*, pages 168–180, Cham, 2021. Springer International Publishing. 3

[13] Gourav Gupta, Kiran Raja, Manish Gupta, Tony Jan, Scott Thompson Whiteside, and Mukesh Prasad. A comprehensive review of deepfake detection using advanced machine learning and fusion methods. *Electronics*, 13(1), 2024. 1

[14] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021. 5, 6

[15] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4360–4369, 2021. 5

[16] Brian Hosler, Davide Salvi, Anthony Murray, Fabio Antonacci, Paolo Bestagini, Stefano Tubaro, and Matthew C. Stamm. Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1013–1022, 2021. 3

[17] Hyeonseong Jeon, Youngoh Bang, Junyaup Kim, and Simon S Woo. T-gd: Transferable gan-generated images detection framework. *arXiv preprint arXiv:2008.04115*, 2020. 2

[18] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon S Woo. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In *Proceedings of the 1st Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection*, pages 7–15, 2021. 2

[19] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 8

[20] Jeongho Kim, Shahroz Tariq, and Simon S Woo. Ptd: Privacy-preserving human face processing framework using tensor decomposition. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pages 1296–1303, 2022. 2

[21] Minha Kim, Shahroz Tariq, and Simon S Woo. Cored: Generalizing fake media detection with continual representation using distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 337–346, 2021.

[22] Minha Kim, Shahroz Tariq, and Simon S Woo. Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1001–1012, 2021. 2

[23] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 6

[24] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *CoRR*, abs/1812.08685, 2018. 2, 4

[25] Mark Layton. Exclusive: From generative imagery to 'star wars' deepfakes, what ai's rise means for

tv. https://tbivision.com/2023/10/12/from-generative-imagery-to-star-wars-deepfakes-what-the-rise-of-ai-means-for-the-tv-industry/, 2023. 1

[26] Binh Le, Shahroz Tariq, Alsharif Abuadbba, Kristen Moore, and Simon Woo. Why do facial deepfake detectors fail? In *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*, page 24–28, New York, NY, USA, 2023. Association for Computing Machinery. 2

[27] Binh M Le and Simon Woo. Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 122–130, 2022. 5

[28] Binh M Le and Simon S Woo. Exploring the asynchronous of the frequency spectra of gan-generated facial images. *arXiv preprint arXiv:2112.08050*, 2021. 2

[29] Binh M Le and Simon S Woo. Quality-agnostic deepfake detection with intra-model collaborative learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22378–22389, 2023. 2

[30] Binh M Le, Jiwon Kim, Shahroz Tariq, Kristen Moore, Alsharif Abuadbba, and Simon S Woo. Sok: Facial deepfake detectors. *arXiv preprint arXiv:2401.04364*, 2024. 2, 8

[31] Sangyup Lee, Shahroz Tariq, Junyaup Kim, and Simon S Woo. Tar: Generalized forensic framework to detect deepfakes using weakly supervised learning. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, pages 351–366. Springer, 2021. 2

[32] Sangyup Lee, Shahroz Tariq, Youjin Shin, and Simon S Woo. Detecting handcrafted facial image manipulations and gan-generated facial images using shallow-fakefacenet. *Applied Soft Computing*, 105:107256, 2021.

[33] Sangyup Lee, Jaeju An, and Simon S Woo. Bznet: unsupervised multi-scale branch zooming network for detecting low-quality deepfake videos. In *Proceedings of the ACM Web Conference 2022*, pages 3500–3510, 2022. 2

[34] Gen Li, Xianfeng Zhao, and Yun Cao. Forensic symmetry for deepfakes. *IEEE Transactions on Information Forensics and Security*, 18:1095–1110, 2023. 2

[35] Meng Li, Beibei Liu, Yongjian Hu, and Yufei Wang. Exposing deepfake videos by tracking eye movements. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5184–5189, 2021. 3

[36] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020. 2, 4

[37] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Comput. Surv.*, 54(1), 2021. 1

[38] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 2823–2832, New York, NY, USA, 2020. Association for Computing Machinery. 3

[39] Rami Mubarak, Tariq Alsboui, Omar Alshaikh, Isa Inuwa-Dutse, Saad Khan, and Simon Parkinson. A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access*, 11:144497–144529, 2023. 3

[40] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2307–2311. IEEE, 2019. 5

[41] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6111–6121, 2022. 2

[42] Dantong Niu, Ruohao Guo, and Yisen Wang. Morié attack (ma): A new potential risk of screen photos. In *Advances in Neural Information Processing Systems*, pages 26117–26129. Curran Associates, Inc., 2021. 3

[43] Keyurkumar Patel, Hu Han, Anil. K. Jain, and Greg Ott. Live face video vs. spoof face video: Use of moiré patterns to detect replay video attacks. In *2015 International Conference on Biometrics (ICB)*, pages 98–105, 2015. 3

[44] Alexa CorseFollow Robert McMillanFollow and Dustin VolzFollow. New era of ai deepfakes complicates 2024 elections. https://www.wsj.com/tech/ai/new-era-of-ai-deepfakes-complicates-2024-elections-aa529b9e, 2024. 1

[45] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 2, 4, 5

[46] Audrey Schomer. Sora ai videos easily confused with real footage in survey test. https://variety.com/vip/sora-ai-video-confusion-human-test-survey-1235933647/, 2024. 1

[47] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022. 5

[48] Yujing Sun, Yizhou Yu, and Wenping Wang. Moiré photo restoration using multiresolution convolutional neural networks. *CoRR*, abs/1805.02996, 2018. 3

[49] Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3609–3618, 2021. 5, 6

[50] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, pages 81–87. ACM, 2018. 2

[51] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. Gan is a friend or foe?: a framework to detect various fake face images. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1296–1303. ACM, 2019.

[52] Shahroz Tariq, Sangyup Lee, and Simon S Woo. A convolutional lstm based residual network for deepfake video detection. *arXiv preprint arXiv:2009.07480*, 2020. 2

[53] Shahroz Tariq, Sangyup Lee, and Simon Woo. One detector to rule them all: Towards a general deepfake attack detection framework. In *Proceedings of the web conference 2021*, pages 3625–3637, 2021. 3

[54] Shahroz Tariq, Sowon Jeon, and Simon S Woo. Am i a real or fake celebrity? evaluating face recognition and verification apis under deepfake impersonation attack. In *Proceedings of the ACM Web Conference 2022*, pages 512–523, 2022. 2

[55] Shahroz Tariq, Alsharif Abuadbba, and Kristen Moore. Deepfake in the metaverse: Security implications for virtual gaming, meetings, and offices. In *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*, page 16–19, New York, NY, USA, 2023. Association for Computing Machinery. 1

[56] Shahroz Tariq, Sowon Jeon, and Simon S Woo. Evaluating trustworthiness and racial bias in face recognition apis using deepfakes. *Computer*, 56(5):51–61, 2023. 2

[57] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 2

[58] Luisa Verdoliva. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020. 1

[59] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4129–4138, 2023. 5

[60] Christopher Whyte. Deepfake news: Ai-enabled disinformation as a multi-level public policy challenge. *Journal of Cyber Policy*, 5(2):199–217, 2020. 3

[61] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 5

[62] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15044–15054, 2021. 5