Through the Lens: Benchmarking Deepfake Detectors Against Moiré-Induced Distortions

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

Deepfake detection remains a pressing challenge, particularly in real-world settings where smartphone-captured media from digital screens often introduces Moiré artifacts that can distort detection outcomes. This study systematically evaluates state-of-the-art (SOTA) deepfake detectors on Moiré-affected videos, an issue that has received little attention. We collected a dataset of 12,832 videos, spanning 35.64 hours, from the Celeb-DF, DFD, DFDC, UADFV, and FF++ datasets, capturing footage under diverse real-world conditions, including varying screens, smartphones, lighting setups, and camera angles. To further examine the influence of Moiré patterns on deepfake detection, we conducted additional experiments using our DeepMoiréFake, referred to as (DMF) dataset and two synthetic Moiré generation techniques. Across 15 top-performing detectors, our results show that Moiré artifacts degrade performance by as much as 25.4%, while synthetically generated Moiré patterns lead to a 21.4% drop in accuracy. Surprisingly, demoiréing methods, intended as a mitigation approach, instead worsened the problem, reducing accuracy by up to 17.2%. These findings underscore the urgent need for detection models that can robustly handle Moiré distortions alongside other real-world challenges, such as compression, sharpening, and blurring. By introducing the DMF dataset, we aim to drive future research toward closing the gap between controlled experiments and practical deepfake detection.

1 Introduction

The rise of deepfake technology has transformed how digital media can be manipulated, presenting a growing threat across the internet and social networking platforms. Deepfakes, which are artificially generated or altered videos that convincingly imitate real individuals, pose significant risks to privacy, security, and the spread of misinformation. The increasing ease with which deepfakes can be created exacerbates this issue [1], as their realism often deceives the general public and sophisticated detection algorithms. Advances in deepfake generation techniques, such as those using Generative Adversarial Networks (GANs) [2] and other deep learning models [3, 4], including diffusion models [5], have made detection an extremely challenging task [6–9] in real-world scenarios on the Internet. While efforts have been made to develop robust detection systems [10–21], such algorithms are predominantly evaluated in controlled environments using benchmark datasets. However, real-world scenarios introduce various challenges, including environmental factors and media-sharing distortions, which can significantly impact detection accuracy [22]. One of the most prominent challenges arises when deepfake content is viewed on screens and recorded using smartphone cameras. Although naive screen capture is available, Digital Rights Management (DRM) on many platforms often disables it. Accordingly, we focus on the prevalent smartphone screen-recapture scenario that users adopt for

*Equal contribution.
†Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

casual sharing or unauthorized reproduction. In practice, the same deepfake can exhibit drastically different visual characteristics when viewed directly on a screen versus when captured by a camera, adding an extra layer of complexity for detection systems (See Figure 1). This common real-world scenario introduces visual artifacts known as Moiré patterns, which occur due to the interference between the pixel grid of the display and the camera sensor [23]. These Moiré patterns, often undetected by the human eye, severely disrupt deepfake detection algorithms, highlighting a critical gap between controlled environment performance and practical, real-world conditions.

In this paper, we investigate the impact of Moiré patterns and compression on deepfake detection systems across three scenarios: (i) Authentic Moiré patterns, (ii) Synthetic Moiré patterns, and (iii) Compression Attacks. Authentic Moiré patterns are introduced when users record content displayed on the screen with smartphones, degrading detection accuracy by distorting key visual features [24]. For instance, a deepfake video of President Putin was shown nationwide on television declaring martial law, which was captured with a smartphone, clearly showing signs of a Moiré pattern [25, 26] shared on X (formerly Twitter). Another example of a smartphone-captured deepfake on social media is creating false narratives by a broadcaster announcing President Macron rescheduling a visit due to an assassination attempt [27]. Synthetic Moiré patterns, on the other hand, are deliberately generated either through pixel-level manipulation or by capturing screen-displayed content with controlled distortions, obscuring the artifacts that deepfake detectors rely on. Finally, Compression Attacks simulate real-world video uploads to social networking sites (SNS), where compression artifacts combine with Moiré patterns to impair detection systems further.

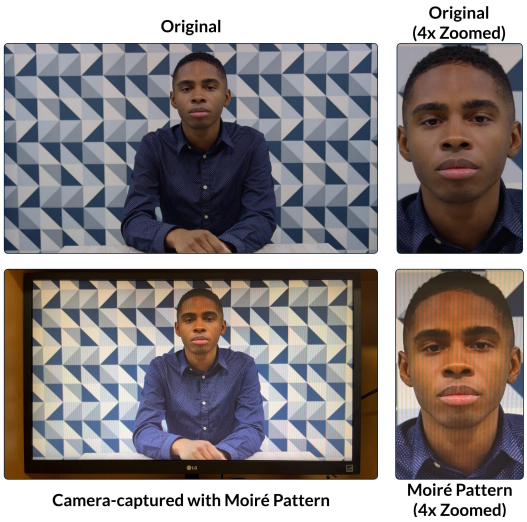

Figure 1: **Original vs. Moiré pattern**

We address these challenges by introducing the DMF dataset, the first deepfake dataset to incorporate Moiré patterns into public deepfake datasets. It includes diverse videos captured from four screens under two lighting conditions with two smartphones, providing a realistic benchmark for evaluating the resilience of state-of-the-art deepfake detectors. Table 1 and Table 2 detail the dataset's variations, and video-capturing specifications, reflecting practical challenges in deepfake detection. Unlike our previous studies [33] and [34], this work adds details not covered before. Specifically, we present a broader analysis of deepfake detectors, examine distortions beyond Moiré patterns, and evaluate compression effects and mitigation strategies in depth. These differences are discussed in Section 2, and the various angles used in our analysis are outlined in Appendix P. To assess the effectiveness of our dataset, we performed an extensive evaluation using 15 different deepfake generation methods. Additionally, we examine the impact of defense methods such as demoiréing techniques on the performance of these detection algorithms as a potential mitigation strategy. We summarize our main contributions as follows:

Table 1: **DeepMoiréFake Details:** We selected a subset of videos from five famous deepfake datasets and manually captured them under various conditions, which resulted in a total playback time of 35.64 hours containing Moiré patterns across 12,832 videos (802×4 (screens)×2 (phones)×2 (lightning conditions).

Name	Real Videos *(People)*	Fake Videos *(People)*	Videos from Dataset	Duration per Video *(secs)*	Total Videos Captured	Captured Duration *(hours)*
FF++ [28]	200	200	400	10	6400	17.78
DFD [29]	28	28	56	10	896	2.49
DFDC [30]	66	66	132	10	2112	5.87
CelebDF [31]	58	58	116	10	1856	5.16
UADFV [32]	49	49	98	10	1568	4.36
Total	**401**	**401**	**802**	-	12832	**35.64**

1. **Moiré Pattern Attacks, Scenarios, and Datasets:** We propose Authentic Moiré patterns and Synthetic Moiré patterns. Constructively, we developed the first Moiré Pattern-impacted deepfake datasets to evaluate both real-world cases. They are captured with four different computer screens

using two different smartphone cameras under two different lighting conditions on videos from FaceForensics++ (FF++) [28], CelebDF [31], the DeepFake Detection (DFD) [29], the DeepFake Detection Challenge (DFDC) [30] and UADFV [32] dataset. DMF is released publicly under DOI-based restricted terms and conditions to support further research on Moiré-induced challenges in deepfake detection[3], and our evaluation codes are publicly available here[4].

2. **Extensive Moiré Pattern Evaluation and Benchmarking:** We conducted an extensive empirical study using our DMF dataset and 15 detectors to determine how Moiré patterns from camera-captured deepfake videos on digital screens affect detector performance. This helps in understanding real-world application challenges and vulnerabilities with current detection methods.
3. **Mitigation and Defense Approach with Demoiréing.** To remove the Moiré pattern from DMF videos, we propose the state-of-the-art defense methods and apply them using five image and two video demoiréing methods, evaluate these demoiréd videos using the identical 15 deepfake detectors, and present the effectiveness and implications of defense methods.

Our real-world evaluation revealed that the presence of Moiré patterns caused an average performance decline of 10.7% in deepfake detectors, with reductions reaching up to 25.4% in extreme cases. Additionally, implementing demoiréing as a defense further decreased detection accuracy, with an average decline of 6.1% and up to 17.2% in severe cases. These findings highlight the need for further research to understand better the interaction between demoiréing techniques and deepfake detection algorithms.

Table 2: Specifications and variations in the video-capturing setup.

NAME (VARIATIONS)	DETAILS
CAMERA ANGLES (4)	Center, 45° left, 45° right and Handheld
LIGHTNING CONDITIONS (2)	On and Off
SCREENS (4) (60 (Hz)	LG (LED), BenQ (LED), Samsung (QHD-IPS), and Lenovo (UHD-IPS)
PHONES (2)	iPhone 13 Samsung S22 Plus
SCREEN RESOLUTION (2)	1980x1080, 3840x2160
CAPTURE RESOLUTION (1)	1980x1080
FRAME RATE (1)	30 fps
VIDEO CAPTURE APPS (2)	OBS Studio, DroidCam (iOS) IP Webcam (Android)

2 Related Work

DEEPFAKE GENERATION. Deepfake video generation leverages advanced deep learning techniques such as variational autoencoders (VAEs) [35], generative adversarial networks (GANs) [2], and diffusion models [36] to produce highly realistic manipulated videos. Common deepfake manipulations include face swapping, face reenactment, face attribute editing, and face synthesis [37, 38]. *Face swapping* replaces a target face with a source face while preserving attributes such as skin color, expressions, and the surrounding environment [3]. *Face reenactment* transfers expressions and movements from a source face to a target, retaining the target's appearance and identity. This technique uses facial motion capture and deep learning to modify the target's movements based on a driving image, video, or pose [39–41]. *Face attribute* editing alters specific facial features, such as age, expressions, or skin tone, using generative models among GANs and VAEs. It can focus on single attributes or edit multiple attributes simultaneously [42, 43]. Finally, *face synthesis* employs GANs to create hyper-realistic human faces that do not exist. While it has applications in gaming and fashion, it also poses risks, such as enabling fake identities on social networks to spread misinformation [44, 45].

DEEPFAKE DETECTION. As deepfake generation technology advances, effective detection methods become increasingly critical to prevent misuse. Deepfake detection relies on deep learning models that identify subtle artifacts often

Table 3: A comparison of publicly available Deepfake datasets.

DATASET	REAL VIDEOS	FAKE VIDEOS	TOTAL VIDEOS	ENCODING ARTIFACTS	ACQUISITION ARTIFACTS
UADFV [32]	49	49	98	✗	✗
DeepfakeTIMIT [46]	640	320	960	✓ (Compress.)	✗
FF++ [28]	1,000	4,000	5,000	✓ (Compress.)	✗
CelebDF [31]	590	5,639	6,229	✗	✗
DFD [29]	363	3,000	3,363	✗	✗
DeeperForensics [47]	50,000	10,000	60,000	✗	✗
DFDC [48]	23,654	104,500	128,154	✓ (Compress.)	✓ (Lighting)
KoDF [49]	62,166	175,776	237,942	✗	✗
FakeAVCeleb [50]	500	19,500	20,000	✗	✗
Ours	**401**	**401**	**802**	✗	✓ **(Moiré)**

[3] https://doi.org/10.7910/DVN/XYOSYW
[4] https://github.com/Razaib-Tariq/DeepMoireFake

imperceptible to the human eye. Techniques include convolutional neural networks [51–56], temporal analysis [57], frequency domain analysis [19], and attention mechanisms using transformers [58–60]. Detection methods are typically developed and evaluated using datasets such as FaceForensics++ (FF++) [28], CelebDF [31], UADFV [32], and FakeAVCeleb [50]. A comparative analysis of existing datasets is presented in Table 3.

Our work evaluates deepfake detection systems under three distinct real-world scenarios. Captured Moiré Pattern Attack (CMPA) simulates authentic Moiré patterns generated when users record deepfake videos from screens, causing distortions that obscure critical visual features and degrade detection accuracy. Synthetic Moiré Pattern Attacks (SMPA) investigate the effects of artificial Moiré patterns using methods such as SMPA-MA [61] and SMPA-SPS [62]. Finally, Compression Attack (CA) explores how video compression artifacts from SNS uploads interact with Moiré patterns, further degrading deepfake detection performance.

Moiré Patterns in Deepfake Detection. While deepfake detection has seen significant advancements, the impact of Moiré patterns on detection performance remains an underexplored challenge. Prior studies, including our previous works [33] and [34], have investigated deepfake detection under various conditions; however, these efforts were limited in scope. The former employed a restricted set of detection methods, offering only preliminary insights on a constrained dataset, whereas the latter expanded the evaluation to more detectors but lacked real-world scenarios where Moiré patterns could be actively exploited. Furthermore, these studies did not systematically assess critical factors such as image distortion, compression effects, and mitigation strategies, essential for improving robustness in practical applications.

In contrast, our study extensively explores these missing aspects and introduces the novel DMF dataset to fill these gaps. The dataset comprises videos captured from four screens under two lighting conditions using two smartphones, enabling robust and practical evaluations. By capturing real-world interference patterns under controlled variations, the DMF dataset enables robust and practical evaluations, providing a more comprehensive benchmark for deepfake detection. Moreover, we analyze both authentic and synthetic Moiré patterns, extending prior research [33] and [34]. Beyond this, we systematically investigate mitigation strategies, including demoiréing, denoising, and deblurring, revealing significant trade-offs where removing Moiré patterns may inadvertently reduce detection accuracy. Our empirical analysis spans 15 deepfake detectors and rigorously evaluates the interplay between Moiré patterns and compression artifacts, providing a robust real-world assessment.

3 Dataset Collection and Generation

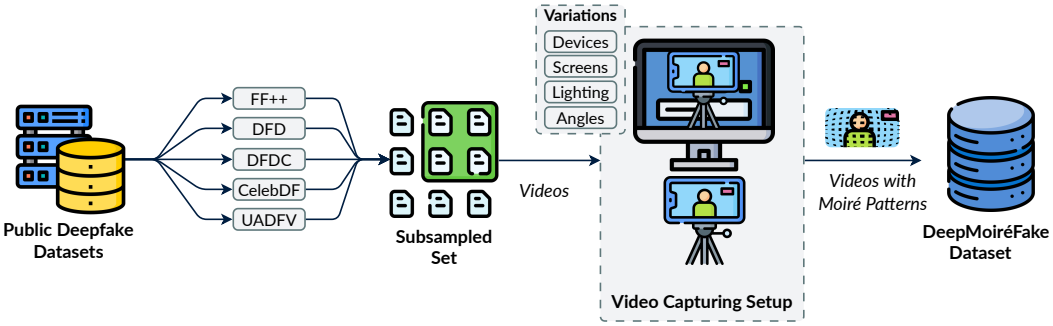

Figure 2: Our manual Moiré pattern collection pipeline.

Selection and Collection. We selected five public deepfake datasets, UADFV [32], FaceForensics++ [28], DFD [29], DFDC [30], and CelebDF [31] as a representative set, covering a variety of settings detailed in Table 3. The overall process of generating our DeepMoiréFake dataset from these source datasets is illustrated in Figure 2, which shows the pipeline of subsampling videos and capturing them under various screen and device conditions to induce Moiré patterns. UADFV [32] dataset was created using the FakeApp tool and contains 49 real YouTube videos, each paired with a corresponding deepfake video. Videos are approximately 11 seconds long, with a resolution of 294×500 pixels. FaceForensics++ (FF++) [28] dataset includes 1,000 real YouTube videos and 1,000 deepfake videos generated using four techniques: Deepfake [4], Face2Face [40], Faceswap [3], and

NeuralTexture [39]. It provides 4,000 manipulated videos in three quality levels uncompressed (raw), lightly compressed (C23), and highly compressed (C40), enabling evaluations across varying compression levels. DFD [29] dataset by Google/Jigsaw consists of 3,068 deepfake videos generated from 363 original videos of 28 consented individuals representing diverse genders, ages, and ethnicities. DFDC [30] dataset is the largest public faceswap dataset, featuring over 100,000 videos from 3,426 paid actors. Most videos are in 1080p resolution and include a mix of deepfake, GAN-based, and non-learned techniques, with an average of 14.4 videos per individual. CelebDF [31] dataset contains 590 real and 5,639 deepfake videos, sourced from over two million frames of YouTube interviews with 59 celebrities of diverse genders, ages, and ethnicities.

DATASET SUBSET SELECTION. We initially selected a subset of videos from five public deepfake datasets for dataset generation, ensuring a balanced representation across gender and ethnicity. Firstly, from the FaceForensics++ (FF++) dataset, we randomly chose 50 real and 50 fake videos from each of the four sub-datasets, totaling 400 videos. For the DFD, DFDC[5], CelebDF, and UADFV datasets, we selected one real and one fake video for each unique individual, resulting in 56, 132, 116, and 98 videos, respectively (see Table 1), resulting in a carefully selected 802 videos from these datasets. Our main objective was to ensure a diverse representation from these deepfake datasets while maintaining a manageable number of manual videos for manual handling when creating Moiré pattern videos for our DMF dataset.

GENERATION OF DEEPMOIRÉFAKE (DMF) DATASET. The DMF dataset addresses the limitations of existing deepfake datasets by replicating real-world conditions, where videos are captured on monitor screens using mobile devices. This approach introduces distortions, such as Moiré patterns and screen-specific characteristics, which are captured through mobile device cameras, thereby facilitating the evaluation of deepfake detection methods in practical scenarios. The dataset includes real and deepfake content recordings displayed on four monitors using two smartphone cameras (iPhone 13 and Samsung S22 Plus). Each smartphone was positioned on a stand 35 cm from the screen at various angles. Videos were captured under varied lighting conditions, with specifications detailed in Table 2. This carefully designed setup ensures a diverse range of screen types, lighting conditions, and device configurations, making the dataset valuable for advancing deepfake detection in real-world applications. During this stage, we ensured label accuracy via an automated process followed by manual verification and enforced consistency across screen camera setups. We also categorized each sample by device type, display screen, viewing angle, and lighting environment. These attributes are recorded in the dataset metadata to support reproducibility and downstream analysis.

DEEPFAKE DETECTION. To evaluate deepfake detection performance across different datasets, we utilized 15 deepfake detectors. Specifically, we employed 10 image-based detectors, including SelfBlended [63], Rossler [28], ForgeryNet [64], Capsule-Forensics (Capsule) [65], MAT [66], CADDM [67], CCViT [58], and ADD [19] to assess detection results on the original dataset, Moiré pattern dataset, and demoiréd dataset. For the Rossler [28], we used pre-trained weights from three variations of the FaceForensics++ dataset: raw, C23, and C40, referred to as Rossler, Rossler C23, and Rossler C40, respectively. In our deepfake video detection experiments, we employed 5 detectors: AltFreezing [68], FTCN [69], LRNet [70], and LipForensics [71]. For LRNet, we used the BlazeFace (LRNet BF) and RetinaFace (LRNet RF) variants.

4 Experimental Scenarios and Settings

Authentic Moiré Patterns or Captured Moiré Pattern Attack (CMPA). To simulate and evaluate a user-generated distortion, we propose a ***Captured Moiré Pattern Attack (CMPA)***. Moiré patterns, caused by interference between the pixel grids of the camera and monitor display [72], are more intense with more significant resolution mismatches and are particularly pronounced on older technologies such as LCD, LED, and IPS displays compared to OLEDs. This effect diminishes as the distance between the camera and monitor increases [73]. These distortions degrade deepfake detection accuracy by introducing artifacts that interfere with identifying critical visual features, especially over multiple frames. Deepfakes were generated using a range of monitors with varying resolutions, such as 1080p LED, 1440p QHD IPS, and 4K UHD IPS displays, to simulate real-world

[5]For DFDC, we selected videos from the preview version, containing 5000 videos of 66 unique individuals.

conditions. Future datasets should include OLED displays and 8K monitors to improve the robustness of detection algorithms against these evolving challenges [74].

SYNTHETIC MOIRÉ PATTERN ATTACKS (SMPA). To evaluate the impact of interference patterns on deep learning models, we propose ***Synthetic Moiré Pattern Attacks (SMPA)***, which replicate noise artifacts commonly observed in screen recordings (see Figure 3). These patterns degrade model performance by introducing complex distortions that are difficult to detect and eliminate. The SMPA approach we used incorporates two methods: (1) **SMPA-MA**, which simulates real-world capture conditions by applying scaling, resampling, and random rotations to input images, as proposed by [61], and (2) **SMPA-SPS**, which modulates parameters such as skew, contrast, and deviation while introducing non-linear distortions such as sine waves to replicate complex Moiré patterns [62]. By mimicking real-world Moiré artifacts, the SMPA demonstrates an efficient and effective adversarial approach, emphasizing the importance of designing robust detection algorithms to counter such attacks.

COMPRESSION ATTACK WITH CMPA AND SMPA. Uploading videos to Social Networking Sites (SNS) often introduces compression and quality degradation, adding new artifacts to the content. To replicate real-world scenarios, we propose the ***Compression Attack (CA)*** on Moiré Patterns, which combines Moiré distortions with compression artifacts to simulate the impact of social media uploads. This approach leverages the widely used H.264 compression algorithm, adopted by platforms such as TikTok and YouTube [75], and standard techniques from FaceForensics++ [28]. By mirroring these real-world compression methods, we provide a realistic evaluation of how compression affects deepfake detection performance. We generated two compressed versions of our dataset, C23 and C40, to simulate the quality degradation caused by SNS uploads [76, 77]. Compression reduces high-frequency information, introducing noise that interacts with existing Moiré patterns to create more complex distortions. These combined artifacts significantly degrade deepfake detection systems' performance, with increasing compression noise leading to further reductions in detection accuracy. This attack reveals a critical vulnerability in current detection systems, which often overlook the combined effects of Moiré patterns and compression artifacts. By exploiting compression, attackers can obscure signs of manipulation, enabling altered videos to evade detection on SNS platforms.

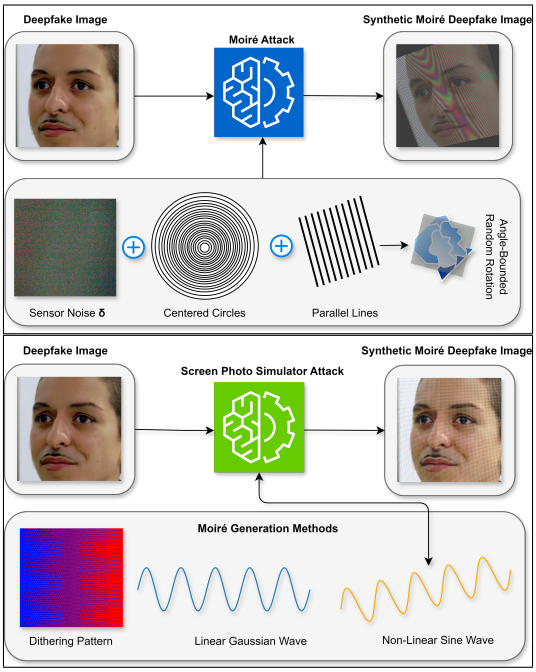

Figure 3: Synthetic Moiré Generation

PREPROCESSING AND PRETRAINED WEIGHTS. Our dataset comprised videos from various sources, including CelebDF [31], DFD [29], DFDC [30], and FaceForensics++ [28]. Each dataset was preprocessed according to the specific requirements of the respective deepfake detectors. For the demoiré experiments, an additional preprocessing step was applied to remove the Moiré pattern using state-of-the-art demoiréing methods [78–81]. We used the pretrained weights for each detector during the evaluation. We selected the top five performers on CMPA from the image detectors Rossler C23, MAT, CADDM, SelfBlended, and CCViT for SMPA, CA, demoiréing, fine-tuning, and retraining experiments. However, SelfBlended and CCViT did not exhibit any performance improvement during training, remaining at a static accuracy of 50%. As a result, we excluded them from further analysis and focused on Rossler C23, MAT, and CADDM.

MOIRÉ REMOVAL METHODS. We employed state-of-the-art demoiréing methods such as DMCNN [78], MBCNN [79], and ESDNet [80] (under two settings) alongside DDA [81], which is tailored for mobile devices, to remove the Moiré pattern from DMF videos: Firstly, (i) **DMCNN** utilizes groups of convolutional layers for downsampling and deconvolutional layers to restore resolution. The final output image is produced by summing feature maps from all branches; Secondly,

(ii) **MBCNN** features a learnable bandpass filter (LBF) for effective Moiré texture removal and employs a two-step tone mapping strategy for color restoration. This includes global tone mapping to correct color shifts and local fine-tuning for per-pixel accuracy. Thirdly, (iii) **ESDNet** integrates a Semantic-Aligned Scale-Aware Module (SAM) to handle scale variations of Moiré patterns. It enhances model effectiveness by extracting and dynamically fusing multi-scale features within the same semantic level, maintaining a lightweight network structure. and Lastly, (iv) **DDA** is optimized for real-time deployment on mobile devices. This method employs a parameter-shared supernet paradigm, ensuring resource efficiency without adding an extra parameter burden. It was selected because our data collection was performed using mobile devices. The demoiréing experiments were conducted on image and video-based techniques, which are provided in the Appendix Table 11 and Table 12.

METRICS. We evaluated the performance of deepfake detectors in our experiments using accuracy, AUC score, precision, recall, and F1-score. The main text reports the results based on the AUC score. For the CA, fine-tuning, and retraining settings, we report the best Accuracy. The results for the other metrics are available in the Appendix H.

5 Results

CMPA – PERFORMANCE UNDER VARIOUS PLAYBACK SCREEN SETTINGS.

In Table 4, the most significant visual Moiré artifacts were observed when videos were captured from the LG and BenQ screens, both of which use backlit LED technology with low pixel density and traditional RGB stripe subpixel layouts. These structural characteristics tend to amplify aliasing effects, particularly when captured through camera sensors, resulting in severe Moiré distortions. Correspondingly, the most substantial performance degradation in detection was also recorded for these two screens. This indicates that certain display technologies might amplify Moiré artifacts more than others. The variations in pixel arrangements, refresh rates, and anti-aliasing techniques across different screens likely contribute to the severity of these distortions. CCViT [58] demonstrated the best detection performance across all screen environments, with an average AUC of 84.5%. Meanwhile, Capsule [65] and LRNet showed robustness in this with the different capturing devices scenarios, and performance dropping by only 2-3 percentage points, for instance, on average, Rossler C23 [28] performance dropped to 16.1%, whereas Capsule experienced only a 2.2% drop. The performance from Capsule and LRNet is significantly low (around the mid-60s), making them impractical in the real world. Overall, we observed a similar trend in performance results across different screen configurations. In addition, we include performance results on videos captured at ±45° viewing angles in the Appendix Table 15, further examining how angled perspectives affect detection robustness under Moiré interference.

Table 4: Performance on different playback screens.

	DETECTORS (*Type and Name*)	ORIGINAL PERFORMANCE	Videos captured from screens			
			LG	*BenQ*	*Lenovo*	*Samsung*
VIDEO	LRNet BF	*61.7*	54.9	55.3	55.9	53.2
	LRNet RF	*62.2*	58.8	60.5	58.7	58.8
	FTCN	*90.2*	65.9	65.3	70.6	68.9
	LipForensics	*90.6*	80.3	80.8	**84.4**	79.8
	AltFreezing	***92.5***	**80.4**	**81.3**	83.7	**82.9**
IMAGE	Rossler	*67.7*	56.2	54.5	59.4	56.9
	ADD	*69.7*	65.4	64.3	66.3	63.4
	Capsule	*71.3*	71.2	69.6	69.0	66.6
	ForgeryNet	*76.9*	61.5	61.8	66.5	63.6
	Rossler C40	*77.0*	67.7	66.9	67.3	67.8
	Rossler C23	*86.5*	68.6	67.4	74.5	70.9
	MAT	*87.0*	72.4	74.9	80.1	76.6
	CADDM	*87.1*	71.3	71.8	80.9	79.5
	SelfBlended	*88.8*	73.7	75.5	80.9	76.4
	CCViT	***95.0***	**81.9**	**83.7**	**86.4**	**86.0**
Avg. Performance loss (Moiré vs. Original)			-11.6	-11.4	-8.0	-10.2

CMPA – PERFORMANCE WITH DIFFERENT CAPTURING DEVICES. In Figure 4, we illustrate the performance of detectors with original and Moiré pattern captured videos using iPhone and Samsung devices, showing a significant performance drop, highlighting the impact of Moiré artifacts on deepfake detection. The detection performance on videos captured using the Samsung S22 Plus was slightly worse on average than that captured with the iPhone. CCViT [58] achieved the best performance across all scenarios, with 95% on the original, 85% on iPhone-captured, and 83% on Samsung-captured images. The worst performance was observed with the LRNet models and Rossler model [28], where Rossler scored 68% on the original, 58% on iPhone-captured, and 55% on Samsung-captured images, suggesting that the severity of Moiré interference may vary across different smartphone camera sensors and image processing pipelines. Overall, all detectors have a

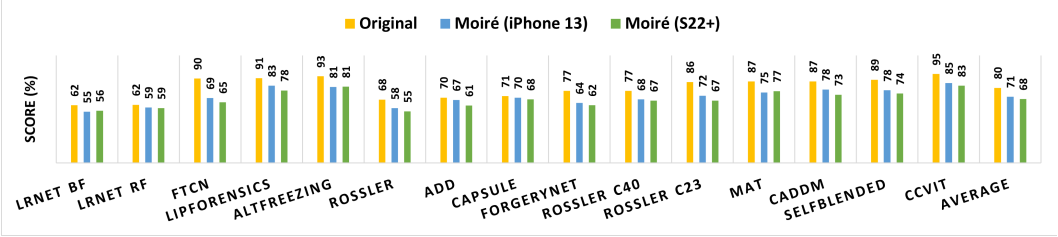

Figure 4: **Different Capturing Devices:** AUC performance of detectors dropped by 9.5 and 12.0 percentage points on average for videos with Moiré patterns captured by iPhone 13 and Samsung S22 Plus, with a maximum drop of 25.4 percentage points in the worst case.

significant drop in performance, regardless of the capturing device used. This consistent degradation raises concerns about the generalizability of deepfake detectors in real-world scenarios where Moiré artifacts are commonly introduced during video playback or screen recording.

CMPA – Performance Under Different Lighting Conditions. Table 5 shows the performance of the (top-3) detectors when videos are captured by the camera on different screens that are exposed to different lighting conditions. In this scenario, we observed a very minimal performance change (around 1%), which shows that the impact of the Moiré pattern remained the same irrespective of the lighting conditions (see Appendix Figure 14 for results of all detectors).

Table 5: Lighting Conditions

Detectors	On	Off
LipForensics	81.3	78.7
AltFreezing	82.1	80.6
CCViT	84.5	83.5

SMPA – Synthetic Moiré Pattern Attacks Results. We examined two types of Synthetic Moiré Attacks. One is SMPA-MA, and the other is SMPA-SPS. Each Synthetic Moiré Attack framework is shown in (Figure 3). By using a subset of one variation of the camera-captured videos. MAT shows the most severe performance degradation by synthetic Moiré attack among three detector models, with a performance drop of 21.4 percentage points (see Table 6). Unlike MAT, which shows performance degradation after the Synthetic Moiré Attack, Rossler and CADDM show improved performance after SMPA-MA.

Table 6: Comparison of deepfake detector performance in the presence and absence of Moiré Attacks.

Detectors	Without Attack	Moiré Attack		
		CMPA	*SMPA-MA*	*SMPA-SPS*
Rossler C23	**78.1**	**81.9**	83.1	75.4
MAT	76.8	68.8	55.4	61.8
CADDM	73.0	73.1	**86.8**	**80.7**

Table 7: CA baseline results under C23 and C40, evaluated with each detector's pretrained weights.

Detectors	C23				C40			
	OG	*CMPA*	*SMPA-MA*	*SMPA-SPS*	*OG*	*CMPA*	*SMPA-MA*	*SMPA-SPS*
Rosseler C23	**98.4**	**96.5**	**87.7**	80.5	**87.5**	**99.3**	83.2	**98.7**
MAT	86.7	66.1	55.4	56.5	75.3	66.5	52.2	60.8
CADDM	97.7	96.4	86.7	**90.3**	80.1	99.0	**84.8**	96.8

Table 8: Performance of fine-tuned and retrained models on C23 and C40 compression attacks.

	Detectors	Fine-tune				Retrain			
		OG	*CMPA*	*SMPA-MA*	*SMPA-SPS*	*OG*	*CMPA*	*SMPA-MA*	*SMPA-SPS*
C23	Rossler C23	98.0	**96.5**	88.6	91.0	97.8	96.1	87.8	91.5
	MAT	99.2	91.8	94.8	**98.5**	99.3	92.1	**95.8**	**97.5**
	CADDM	**99.8**	96.3	**95.0**	92.0	**99.4**	**96.2**	90.0	91.8
C40	Rossler C23	82.5	**99.6**	85.7	97.0	86.7	99.5	85.7	95.7
	MAT	**98.0**	90.6	**94.4**	97.9	**99.1**	84.2	**94.4**	98.1
	CADDM	90.9	99.3	90.8	**99.2**	96.0	**99.7**	90.71	**99.3**

Compression Attacks (CA). In Table 7 we observe that for the CA baseline, methods such as Rossler C23, MAT, and CADDM show distinct accuracy ranges, with Rossler C23 and CADDM achieving around 80.1–99.0% and MAT lagging behind at 55.2–86.7%. Following fine-tuning and retraining, however, the overall trend is upward, as detailed in Table 8. Most notably, the MAT model's accuracy surged to 99.2% in the best case, effectively closing the gap and becoming competitive with

the other methods. This indicates that fine-tuning or retraining models on specific datasets or with targeted adjustments can enhance their ability to adapt to Moiré patterns and compression artifacts, ultimately improving detection accuracy. This improvement suggests that models benefit from being updated to handle new types of distortions or patterns, which may not have been fully accounted for in the original training process.

IMAGE DISTORTION ATTACKS. We evaluated the impact of Gaussian blurring and sharpening on deepfake detection by applying these techniques to the original datasets. Gaussian blurring, implemented with OpenCV's GaussianBlur function [82], smooths images by reducing fine details, while sharpening, using a high-pass filter via filter2D, enhances edges [83]. This systematic approach ensures consistent application, allowing direct comparison of detection performance. In Appendix Table 10, we present AUC scores before and after applying these transformations.

MITIGATION STRATEGIES – PERFORMANCE AFTER DEMOIRÉING. The top-performing deepfake detectors across all demoiréing techniques were CCViT [58], CADDM [67], and Rossler C23 [28], consistently ranking 1st, 2nd, and 3rd, respectively (see Appendix Table 11 for detailed results). CCViT achieved the highest average score of 79.2%, maintaining superior performance across original, Moiré-affected, and demoiréd images. Among demoiréing methods, ESDNet [80], trained on the FHDMi dataset, exhibited the lowest performance loss, indicating its effectiveness in mitigating Moiré-induced degradation. Conversely, DDA [81] demonstrated the highest performance loss, likely due to its optimization for mobile devices, which compromises its detection capabilities compared to other techniques. A significant finding from this experiment was that while demoiréing methods effectively removed most Moiré patterns from the images (see Figure 5), they also eliminated certain deepfake artifacts that detectors rely on for classification. As a result, performance on demoiréd images dropped more than on images with Moiré patterns. Specifically, Moiré patterns caused an average performance drop of 10.1 percentage points for detectors. In contrast, demoiréd images resulted in an average drop of 14.7 percentage points. This underscores the need for advanced mitigation strategies to address Moiré patterns without inadvertently removing critical deepfake artifacts, ensuring robust detection performance. We conducted additional experiments by processing Moiré videos using VD-Moiré [84] and FPANet [85] demoiréing methods are outlined in Table 12 and denoising and deblurring from NAFNet [86], detailed results from these experiments are provided in (see Appendix Table 13 and Table 14).

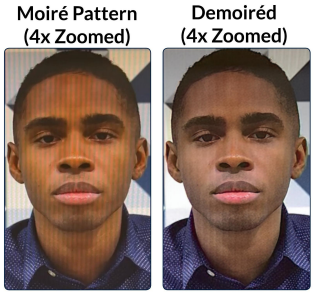

Figure 5: Moiré vs. Demoiréd

Table 9: Overview of detectors with Fine-Tune and Retrain.

DETECTORS	FINE-TUNE				RETRAIN			
	OG	*CMPA*	*SMPA-MA*	*SMPA-SPS*	*OG*	*CMPA*	*SMPA-MA*	*SMPA-SPS*
Rossler C23	77.0	80.6	94.4	81.1	87.9	84.7	**94.9**	79.5
MAT	**94.5**	**85.4**	70.3	**95.6**	**97.9**	**89.0**	71.3	**96.5**
CADDM	86.3	84.6	**94.4**	95.0	85.1	81.9	92.9	95.4

MITIGATION STRATEGIES – PERFORMANCE AFTER FINE-TUNING AND RETRAINING. For fine-tuning, we utilized pretrained weights derived from the original dataset, which were also employed to assess the model's performance on the same data. The test dataset for fine-tuning and retraining comprises original data, captured Moiré data, and synthetic Moiré data. In the case of MAT, performance after retraining exhibited an improved score (see Table 9). However, for CADDM, fine-tuning demonstrated superior performance compared to retraining.

ADDITIONAL ANALYSIS OF MOIRÉ IMPACT AND MITIGATION. We evaluate eight image detectors on original datasets (CelebDF, DFD, DFDC, FF++, and UADFV) and under the most severe Moiré distortion (LED screen) with multiple variations (light on/off, iPhone 13/Samsung S22+). We also assess the performance after demoiréing, denoising, and deblurring effects. The corresponding ROC curves are presented in (see appendix Figure 15—Figure 25), showing varying performance on image detectors and random guess prediction when impacted by Moiré patterns. Furthermore, our investigation extends to evaluating the impact of Moiré patterns on frequency analysis, Appendix Figure 26, and deepfake generative models, with results provided in (see Appendix Figure 27 and Figure 28), with non-GAN and GAN showing distinct frequency patterns.

REMARKS. These results demonstrate that just preprocessing methods (e.g., demoiréing) are insufficient to address the challenge posed by deepfake videos containing Moiré patterns or other artifacts. This highlights the need for more robust detection models capable of handling such distortions. In this context, our DMF dataset provides a valuable addition to public deepfake datasets for training these detectors.

6 Discussion

Challenges in Data Collection. Capturing Moiré patterns in real-world conditions required careful consideration of screen types, lighting variations, angles, and smartphone camera differences. Our dataset comprises 12,832 videos spanning 35.64 hours, sourced from CelebDF, DFD, DFDC, FF++, and UADFV, ensuring diverse representation. Differences in screen pixel structures influenced the intensity of Moiré artifacts. Additionally, smartphone cameras introduced variability in artifact appearance, further complicating the data collection process. These challenges highlight the complexity of generating a dataset that accurately represents Moiré-induced distortions in deepfake detection.

Limitation and Future work. While we acknowledge that real-world Moiré-inducing conditions span a wide range of factors, including variations in camera and display hardware, and dynamic motion, this work focuses on analyzing the impact of Moiré patterns on deepfake detection. Our experimental setup was intentionally designed to control these variables in a reproducible environment, enabling a focused investigation of Moiré-related effects. Broader scenarios involving diverse hardware configurations, motion artifacts, and platform-specific filters (e.g., beautification or AR effects on apps like TikTok and Instagram) remain essential directions for future work.

7 Conclusion

In this paper, we investigated the impact of Moiré patterns on deepfake detection, exposing a significant vulnerability in current methods. Our experiments showed that both Authentic and Synthetic Moiré patterns can degrade detector performance, reducing accuracy by up to 25.4%. This issue is further exacerbated by compression artifacts, where the combined effect leads to even greater performance deterioration. These findings highlight that existing models, often designed for clean, high-quality inputs, struggle with real-world artifacts introduced by screen captures and digital processing. While demoiréing techniques can mitigate these distortions, they may also inadvertently weaken detection performance. This underscores the need for more resilient deepfake detection systems capable of handling practical distortions like Moiré patterns and compression without significant accuracy loss.

SOCIAL IMPACT. Our work highlights the need for advanced deepfake detection to mitigate real-world artifacts. The dataset we share contains the real and deepfake videos captured with different mobile devices. The package also contains detailed documentation with all relevant metadata specified to users. We recommend using DMF as a training dataset to enhance detector robustness, aiding efforts to curb the spread of malicious deepfakes. To promote responsible, impactful use of the DMF dataset and to discourage misuse aimed at bypassing detectors, we provide access through a DOI-based request system. This process enhances security and ensures the dataset is used strictly for legitimate academic research.

Acknowledgement

This work was partly supported by Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean government MSIT: (RS-2022-II221199, RS-2022-II220688, RS-2019-II190421, RS-2023-00230337, RS-2024-00437849, RS-2021-II212068, and RS-2025-02263841). Also, this work was supported by the Cyber Investigation Support Technology Development Program (No.RS-2025-02304983) of the Korea Institute of Police Technology (KIPoT), funded by the Korean National Police Agency. Lastly, this work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00356293).

References

[1] Arick Wierson. Data breach, identity fraud trends reveal deepfake and generative ai threats. https://www.biometricupdate.com/202402/data-breach-identity-fraud-trends-reveal-deepfake-and-generative-ai-threats, 2024.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[3] Marek Kowalski. Faceswap - github repository. https://github.com/MarekKowalski/FaceSwap, 2016.

[4] deepfakes. faceswap. https://github.com/deepfakes/faceswap, 2024. GitHub repository.

[5] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

[6] Jia Wen Seow, Mei Kuan Lim, Raphaël CW Phan, and Joseph K Liu. A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing*, 513:351–371, 2022.

[7] Simon S Woo Hyeongjun Choi, Inho Jung. Combating dataset misalignment for robust ai-generated image detection in the real world. In *Proceedings of the 4th Workshop on Security Implications of Deepfakes and Cheapfakes*, pages 15–20, 2025.

[8] Hyeongjun Choi and Simon S Woo. Gan or dm? in-depth analysis and evaluation of ai-generated face data for generalizable deepfake detection. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pages 759–766, 2025.

[9] 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*, WDC '23, page 16–19, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702037. doi: 10.1145/3595353.3595880. URL https://doi.org/10.1145/3595353.3595880.

[10] Xin Wang, Hui Guo, Shu Hu, Ming-Ching Chang, and Siwei Lyu. Gan-generated faces detection: A survey and new perspectives. *ECAI 2023*, pages 2533–2542, 2023.

[11] Manoj Kumar, Hitesh Kumar Sharma, et al. A gan-based model of deepfake detection in social media. *Procedia Computer Science*, 218:2153–2162, 2023.

[12] Ali Javed, Khalid Mahmood Malik, et al. Faceswap deepfakes detection using novel multi-directional hexadecimal feature descriptor. In *2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pages 273–278. IEEE, 2022.

[13] Anubhav Jain, Nasir Memon, and Julian Togelius. A dataless faceswap detection approach using synthetic images. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2022.

[14] Hyeonseong Jeon, Young Oh Bang, Junyaup Kim, and Simon Woo. T-gd: Transferable gan-generated images detection framework. In *International Conference on Machine Learning*, pages 4746–4761. PMLR, 2020.

[15] 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.

[16] Muhammad Shahid Muneer and Simon S Woo. Towards safe synthetic image generation on the web: A multimodal robust nsfw defense and million scale dataset. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1209–1213, 2025.

[17] Inzamamul Alam, Muhammad Shahid Muneer, and Simon S Woo. Ugad: Universal generative ai detector utilizing frequency fingerprints. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4332–4340, 2024.

[18] Chan Park, Bohyun Moon, Minsun Jeon, Jee-weon Jung, and Simon S Woo. X3a: Efficient multimodal deepfake detection with score-level fusion. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pages 767–774, 2025.

[19] Simon Woo et al. 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*, volume 36, pages 122–130, 2022.

[20] 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*, WWW '21, page 3625–3637, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449809. URL https://doi.org/10.1145/3442381.3449809.

[21] 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*, ADGD '21, page 7–15, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386821. doi: 10.1145/3476099.3484315. URL https://doi.org/10.1145/3476099.3484315.

[22] Minsun Jeon and Simon S. Woo. Seeing through the blur: Unlocking defocus maps for deepfake detection, 2025.

[23] Bin He, Ce Wang, Boxin Shi, and Ling-Yu Duan. Mop moire patterns using mopnet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[24] Cong Yang, Zhenyu Yang, Yan Ke, Tao Chen, Marcin Grzegorzek, and John See. Doing more with moiré pattern detection in digital photos. *IEEE Transactions on Image Processing*, 32: 694–708, 2023.

[25] Joy Reid. Deepfake of purported putin declaring martial law fits disturbing pattern. https://www.msnbc.com/the-reidout/reidout-blog/putin-deepfake-russia-rcna88014.

[26] igorsushko. Putin's new year's address was computer-generated. https://x.com/igorsushko/status/1741777672647418168.

[27] Shayan86. President macron has cancelled a scheduled visit to ukraine over fears of an assassination attempt. https://x.com/Shayan86/status/1758235524957893064.

[28] 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.

[29] Nicholas Dufour, Andrew Gully, Per Karlsson, Alexey Victor Vorbyov, Thomas Leung, Jeremiah Childs, and Christoph Bregler. Deepfakes detection dataset by google & jigsaw. https://research.google/blog/contributing-data-to-deepfake-detection-research/, 2019. arXiv preprint arXiv:1901.08971; dataset release by Google & Jigsaw.

[30] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

[31] 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.

[32] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[33] Razaib Tariq, Shahroz Tariq, and Simon S. Woo. Exploring the impact of moire pattern on deepfake detectors. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 3813–3819, 2024. doi: 10.1109/ICIP51287.2024.10647902.

[34] Razaib Tariq, Minji Heo, Simon S Woo, and Shahroz Tariq. Beyond the screen: Evaluating deepfake detectors under moire pattern effects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4429–4439, 2024.

[35] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[36] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[37] Ángel Fernández Gambín, Anis Yazidi, Athanasios Vasilakos, Hårek Haugerud, and Youcef Djenouri. Deepfakes: current and future trends. *Artificial Intelligence Review*, 57(3):64, 2024.

[38] Binh M Le, Jiwon Kim, Simon S Woo, Kristen Moore, Alsharif Abuadbba, and Shahroz Tariq. Sok: Systematization and benchmarking of deepfake detectors in a unified framework. In *2025 IEEE 10th European Symposium on Security and Privacy (EuroS&P)*, pages 883–902. IEEE, 2025.

[39] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[40] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[41] Arian Beckmann, Anna Hilsmann, and Peter Eisert. Fooling state-of-the-art deepfake detection with high-quality deepfakes. In *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*, pages 175–180, 2023.

[42] Tianyi Wang, Mengxiao Huang, Harry Cheng, Bin Ma, and Yinglong Wang. Robust identity perceptual watermark against deepfake face swapping. *arXiv preprint arXiv:2311.01357*, 2023.

[43] Minji Heo and Simon S. Woo. Fakechain: Exposing shallow cues in multi-step deepfake detection, 2025.

[44] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *arXiv preprint arXiv:2212.13344*, 2022.

[45] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[46] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *CoRR*, abs/1812.08685, 2018. URL http://arxiv.org/abs/1812.08685.

[47] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2889–2898, 2020.

[48] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[49] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. *arXiv preprint arXiv:2103.10094*, 2021.

[50] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

[51] 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*, MPS '18, page 81–87, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359887. doi: 10.1145/3267357.3267367. URL https://doi.org/10.1145/3267357.3267367.

[52] 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*, SAC '19, page 1296–1303, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359337. doi: 10.1145/3297280.3297410. URL https://doi.org/10.1145/3297280.3297410.

[53] 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 (CVPR) Workshops*, pages 1001–1012, June 2021.

[54] 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*, MM '21, page 337–346, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517. doi: 10.1145/3474085.3475535. URL https://doi.org/10.1145/3474085.3475535.

[55] 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. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2021.107256. URL https://www.sciencedirect.com/science/article/pii/S1568494621001794.

[56] Sangyup Lee, Shahroz Tariq, Junyaup Kim, and Simon S. Woo. TAR: Generalized Forensic Framework to Detect Deepfakes Using Weakly Supervised Learning. In Audun Jøsang, Lynn Futcher, and Janne Hagen, editors, *ICT Systems Security and Privacy Protection*, pages 351–366, Cham, 2021. Springer International Publishing. ISBN 978-3-030-78120-0.

[57] Shahroz Tariq, Sangyup Lee, and Simon S Woo. A Convolutional LSTM based Residual Network for Deepfake Video Detection. *arXiv preprint arXiv:2009.07480*, 2020.

[58] 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.

[59] Aminollah Khormali and Jiann-Shiun Yuan. Dfdt: An end-to-end deepfake detection framework using vision transformer. *Applied Sciences*, 12(6), 2022. ISSN 2076-3417. doi: 10.3390/app12062953. URL https://www.mdpi.com/2076-3417/12/6/2953.

[60] Junke Wang, Zuxuan Wu, Jingjing Chen, and Yu-Gang Jiang. M2TR: multi-modal multi-scale transformers for deepfake detection. *CoRR*, abs/2104.09770, 2021. URL https://arxiv.org/abs/2104.09770.

[61] Dantong Niu, Ruohao Guo, and Yisen Wang. Moiré attack (MA): A new potential risk of screen photos. *CoRR*, abs/2110.10444, 2021. URL https://arxiv.org/abs/2110.10444.

[62] Sungjun Choi. Simulating moire effects seen in photos of digital device screens. https://github.com/mr3coi/screen_photo_simulator, 2024.

[63] 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.

[64] 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.

[65] 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.

[66] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. *CoRR*, abs/2103.02406, 2021. URL https://arxiv.org/abs/2103.02406.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] Jesse Schallek, Hongbin Li, R. Kardon, Young H. Kwon, M. Abràmoff, P. Soliz, and D. Ts'o. Stimulus-evoked intrinsic optical signals in the retina: spatial and temporal characteristics. *Investigative ophthalmology & visual science*, 50 10:4865–72, 2009. doi: 10.1167/iovs.08-3290.

[73] Vladimir Saveljev and Sung-Kyu Kim. Simulation and measurement of moiré patterns at finite distance. *Opt. Express*, 20(3):2163–2177, Jan 2012. doi: 10.1364/OE.20.002163. URL https://opg.optica.org/oe/abstract.cfm?URI=oe-20-3-2163.

[74] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: a comprehensive benchmark of deepfake detection. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

[75] Video Compression. Video compression algorithm by social media application. https://getstream.io/glossary/video-compression/.

[76] ian bremmer. Visual moiré artifacts in video of putin calling for martial law. https://x.com/ianbremmer/status/1665841241349668864, 2023.

[77] Vincent Flibustier. interesting technique to give credibility to a fake: filming. https://x.com/vinceflibustier/status/1758521285628383505, 2024.

[78] Yujing Sun, Yizhou Yu, and Wenping Wang. Moiré photo restoration using multiresolution convolutional neural networks. *IEEE Transactions on Image Processing*, 27(8):4160–4172, 2018.

[79] Bolun Zheng, Shanxin Yuan, Gregory Slabaugh, and Ales Leonardis. Image demoireing with learnable bandpass filters. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3636–3645, 2020.

[80] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Jiajun Shen, Jia Li, and Xiaojuan Qi. Towards efficient and scale-robust ultra-high-definition image demoiréing. In *European Conference on Computer Vision*, pages 646–662. Springer, 2022.

[81] Yuxin Zhang, Mingbao Lin, Xunchao Li, Han Liu, Guozhi Wang, Fei Chao, Shuai Ren, Yafei Wen, Xiaoxin Chen, and Rongrong Ji. Real-time image demoireing on mobile devices. *arXiv preprint arXiv:2302.02184*, 2023.

[82] OpenCV. Gaussianblur. https://docs.opencv.org/4.x/d4/d13/tutorial_py_filtering.html, 2009.

[83] OpenCV. Sharpening. https://docs.opencv.org/4.x/d2/dbd/tutorial_distance_transform.html, 2009.

[84] Peng Dai, Xin Yu, Lan Ma, Baoheng Zhang, Jia Li, Wenbo Li, Jiajun Shen, and Xiaojuan Qi. Video demoireing with relation-based temporal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[85] Gyeongrok Oh, Sungjune Kim, Heon Gu, Sang Ho Yoon, Jinkyu Kim, and Sangpil Kim. Fpanet: Frequency-based video demoireing using frame-level post alignment. *Neural Networks*, 184: 107021, 2025.

[86] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022.

Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: [Yes]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: [Yes]

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: [Yes]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: [Yes]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: [Yes]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: [Yes]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.