Journal on
Artificial Intelligence

Tech Science Press

https://doi.org/10.32604/jai.2026.078014

ARTICLE

Frequency-Aware Robustness Analysis of Deepfake Detection Models

Haoyang Xu*

School of Computer Science and Engineering, University of New South Wales (UNSW), Sydney, NSW, Australia

*Corresponding Author: Haoyang Xu. Email: xuhaoyang_zz@163.com

Received: 22 December 2025; Accepted: 10 February 2026; Published: 10 March 2026

ABSTRACT: This paper conducted a comprehensive study on the robustness of three widely used DFD deep learning models—namely, ResNet50, FreqNet, and Xception v1—to controlled perturbation attacks and frequency masking across a range of 12 different distortions. The study was performed on 254,166 ForenSynth test images, characterizing the distribution of FSI-drop values derived from over 3.05 million paired predictions. The distribution of FSI-drop values is sharply peaked around zero: 99.7% of the samples exhibit $|\Delta| < 0.1$, and the maximum $|\Delta| \approx 1.5 \times 10^{-3}$, indicating high baseline stability. In terms of perturbation-wise comparison, Gaussian blur dominates, yielding a mean degradation 30 times greater than that induced by JPEG compression and twice that caused by rescaling. The frequency masking curves further illustrate unique sensitivities: FreqNet shows high-frequency dependence and rapid decay (i.e., >2 to >16 masking scale), Xception exhibits moderate attenuation, whereas ResNet50 remains statistically unchanged (median $|\Delta| < 10^{-4}$). All of these differences are statistically significant at the model level ($p < 0.001$). The experiments offer a concrete demonstration that CNNs effectively retain prediction invariance to small amounts of image deformation, but the frequency-sensitive design demonstrates readily interpretable high-frequency sensitivity, thereby providing a principled framework for designing detectors robust to image perturbations. It should be noted that FSI-drop measures score stability rather than absolute performance; §4.6 discusses the complementary AUC curves and leaves combined-distortion or adversarial stress tests to future work.

KEYWORDS: Deepfake detection; model robustness; frequency-aware learning; high-frequency masking; Gaussian blur; drop-rate analysis

1 Introduction

Recent proliferation of generative adversarial networks (GANs), e.g., ProGAN [1], StyleGAN [2], and diffusion models [3], has made the generation of hyper-realistic fake faces more accessible. On the bright side, the creative sector (e.g., image creation and retouching, video synthesis) has been rendered more efficient; on the dark side, malicious use of deepfakes—from political misinformation [4] and blackmail or financial theft [5]—has decreased public trust. Consequently, the research community has proposed many detection architectures, broadly divided into (i) spatial CNNs such as ResNet50 [6], EfficientNet [7], Xception [8], and its lightweight Xception variant [9]; (ii) frequency-aware models: FreqNet-MM22 [10], LNP [11], SPSL [12]; and (iii) biometric-based methods exploiting eye-gaze or heart-rate inconsistency [13,14].

Despite impressive accuracy on curated benchmarks (AUC > 0.98), recent work by Yu et al. [15] demonstrated that 18 state-of-the-art detectors lose on average 18.7% AUC when evaluated on JPEG-compressed social-media data. These observations align with adversarial-robustness studies in generic computer vision [16,17], yet a granular comparative analysis targeting lightweight and frequency-based

detectors remains absent. In order to fill this gap, we perform the most extensive study of the largest perturbations to date over three representative deepfake detectors: the spatially heavy baseline ResNet50; a mobile-friendly detector XceptionMinimal; and frequency-domain feature detector FreqNet-MM22 [10], which allows the study of robustness pertaining to explicit frequency domain features. We test using the ForenSynths dataset [18], which has 254,166 matching real and synthetic image pairs, using 12 standard perturbation settings. Altogether, we obtain 3.05 million FSI-drop measurements. We emphasize that the main contribution of this work is a systematic robustness audit and diagnostic analysis of representative deepfake detection architectures, rather than proposing a new detection method. Key outcomes from the robustness analysis include:

- ResNet50 exhibits the highest robustness, achieving a median FSI-drop of 6.54×10^{-8}, while 99.7% of the samples stay below 0.1, which is by one order of magnitude more than XceptionMinimal and FreqNet-MM22.
- XceptionMinimal performs an intermediate robustness result, as the degradation values of this method are higher than the ResNet50 ones and lower than those of FreqNet-MM22 in almost all the considered perturbation conditions, indicating that only a moderate spatial representation is learned.
- FreqNet-MM22 is shown to exhibit an inverted robustness response to Gaussian blur, where the FSI-drop decreases from 0.35×10^{-3} to 0.15×10^{-3} as the blur radius σ increases from 1 to 4. This behavior confirms its high-frequency inductive bias while revealing a previously unreported vulnerability pattern. Other frequency-aware methods were not evaluated and may behave differently.

Taken altogether, these results show that architectural inductive biases (specifically frequency-based vs. spatial representations) bear non-trivial and perturbation-dependent robustness trade-offs. Accordingly, system-wide robustness evaluation should be considered as a mandatory part of pre-deployment analyses for deepfake detection applications.

Limitations & Scope

This work focuses on cheap, single-type distortions that dominate social-media pipelines (JPEG, blur, resize). We intentionally leave out cascaded impairments/adversarial perturbations because (a) their parameter space grows exponentially and (b) a preliminary exploration using 2000 randomly sampled images under mild cascaded (compress + blur) distortions did not alter the robustness ranking (ResNet50 > XceptionMinimal > FreqNet). While the observed robustness patterns are consistent across the three representative detectors tested on the ForenSynths dataset, their applicability to other architectures or datasets has yet to be confirmed. A comprehensive analysis of combined and adversarial perturbations is reserved for future work.

2 Related Work

2.1 Deepfake Generation

Early faceswap autoencoders [19] evolved into progressively-growing GANs—ProGAN [1], StyleGAN [2], StyleGAN2 [20], StyleGAN3 [21]—and, very recently, diffusion probabilistic models that surpass GANs in fidelity [3]. Higher generator quality directly elevates the detection challenge.

2.2 Deepfake Detection

Spatial CNNs remain the dominant paradigm. Afchar et al. [6] first transplanted MesoNet from steganalysis; Rossler et al. [9] popularised XceptionNet and its minimal variant for FaceForensics++; Tan & Le [7] later scaled EfficientNet-B4 to balance accuracy and efficiency.

To escape the "pixel-trap", researchers have incorporated biological cues: eye-gaze variance (FakeCatcher [13]), heart-rate inconsistency [14], and PRNU sensor noise [22]. These methods, however, require high-resolution uncompressed inputs and thus remain fragile in social-media pipelines.

Frequency-aware approaches. Among existing methods, FreqNet-MM22 proposed by Tan et al. [10] is the most closely related to this study. The model incorporates a learnable Frequency Transformation Module (FTM) that adaptively reweights DCT bands prior to late fusion with an RGB branch. We underline that the evaluation is limited to FreqNet-MM22; alternative frequency-aware architectures could yield different outcomes, so any broader claims about frequency-based detectors warrant careful qualification.

2.3 Robustness Evaluation

Sabir et al. [14] benchmarked Xception against JPEG compression and resizing, while Güera & Delp [23] analysed recapture artefacts; both, however, focused on a single distortion type. Recent work has shifted to multi-perturbation protocols: Haliassos et al. [24] introduced a self-supervised approach for robust detection of audio-visual forgeries, evaluating it on real-world video datasets and showing that detection performance can degrade under various perturbations. Tan et al. [10] showed that many frequency-based detectors tend to overfit to specific artifacts in the frequency domain, limiting generalization to unseen deepfake sources. Dutta et al. [25] proposed a wavelet sub-band based frequency detection approach that decomposes Fourier representations into wavelet energies to improve robustness beyond spatial CNN features. Finally, ForenSynths [18] released only clean top-line results for ResNet50 and Xception, leaving granular robustness statistics unavailable.

To the best of current knowledge, this study is the first to deliver a fine-grained, drop-based robustness profile for FreqNet-MM22 [10] under systematic JPEG, blur, and resize distortions, side-by-side with ResNet50 and XceptionMinimal.

3 Methodology

The robustness of three architectures—ResNet50, XceptionMinimal, and FreqNet-MM22—was evaluated under realistic image degradations. All experiments share the same data pipeline, perturbation bank, and evaluation protocol to ensure fair comparison.

3.1 Models under Test

XceptionMinimal [9]—A width-reduced version (×0.75 channels) of the Xception architecture. Theoretically, Xception has a formulation using depthwise separable convolutions where spatial correlation learning (depthwise convolution) and cross-channel correlation learning (pointwise convolution) are separated for lower model complexity without loss of representation power. Our model is not pretrained on ImageNet but instead trained from scratch using the FaceForensics++ protocol—enabling deployment on mobile and resource-constrained devices—yet detecting subtle manipulation artefacts at a fine-grained level.

FreqNet-MM22 [10]—Dual branch network that is inspired by the complementary behavior of spatial domain and frequency-domain representation for image forensics, where the RGB branch represents semantic and texture cues in the spatial domain, while explicit encoding of spectral anomalies due to manipulation exists in the frequency branch. Its Frequency Transformation Module (FTM), which learns an 8 × 8 DCT-based weighting that allows it to focus more attention on informative frequency bands, as opposed to using a pre-defined transform. It is initialized from the official MM'22 checkpoint before fine-tuning on ForenSynths to align its frequency response with the target manipulation distribution. The model is initialized with the publicly available pre-trained weights from Frequency-Aware Deepfake

Detection: Improving Generalizability through Frequency Space Learning [10] and subsequently fine-tuned on ForenSynths [18] to adapt the learned frequency representations to the target manipulation distribution.

ResNet50 [6]—A deep residual network with the idea of residual learning; Identity shortcut connections allow a Residual Network to be trained to discover residual mapping instead of underlying mapping, which relieves the problem of Vanishing Gradient and helps to optimize a deeper model, leading to good generalization. ResNet50 (ImageNet Pre-training): A popular pre-trained model from ImageNet that has proven to be effective at capturing spatial information; we use it by replacing its last fully-connected layer with our own 2-node classifier and finetuning the whole network in an end-to-end manner to solve the binomial problem of detecting deepfakes.

These architectures are selected to represent three complementary design paradigms in deepfake detection: lightweight separable convolutional models, explicit frequency-aware modeling, and deep residual learning.

3.2 Dataset and Pre-Processing

ForenSynths [18] provides 254,166 balanced real/fake 512 × 512 face images synthesized by ProGAN, StyleGAN2, and diffusion. The dataset was split at the identity level using an 80/10/10 ratio, resulting in 203,332 training, 25,416 validation, and 25,418 test samples. All images are centre-cropped to 224 × 224, converted to RGB, and normalized with ImageNet statistics. No further augmentation is applied to isolate the impact of the controlled perturbations.

3.3 Perturbation Suite

Three universally encountered distortions are applied at inference time:

- JPEG compression—quality factor Q ∈ {20, 30, 40, 50, 60, 70, 80, 90, 100}.
- Gaussian blur—isotropic kernels with $\sigma \in \{1, 2, 4\}$.
- Resize—bicubic down-scaling followed by up-scaling to 224 × 224, scale ∈ {0.4, 0.5, 0.6, 0.8, 1.0}.

Each image is degraded on-the-fly with a single distortion, yielding 36 perturbed copies (12 parameter levels × 3 distortion types) per sample. These parameter ranges were selected to cover the spectrum of distortion intensities commonly encountered in social media pipelines and practical image processing workflows.

3.4 FSI-Drop Metric

Let x be a clean image and x_i its perturbed copy (e.g., JPEG, blur, resize). Write $p(\cdot)$ for the detector's fake probability. The Fake-Score-Impact (FSI) drop is defined as:

$$\Delta(x) = logit(p(x)) - logit(p(x_i))$$

where x_i is the perturbed variant and $\text{logit}(p) = \ln(p/(1-p))$. Here, $\Delta(x)$ is computed by comparing the detector's outputs on the clean image x and its perturbed version x_i. $\Delta(x)$ is invariant to sigmoid calibration and approximately additive for small perturbations. Over a dataset, a median Δ near zero signals robustness; large positive or negative values flag score instability. The "original minus perturbed" order keeps positive Δ aligned with a drop in fake-likelihood after corruption.

3.5 Training & Evaluation Protocol

All models are trained under an identical end-to-end protocol on the clean ForenSynths training split to ensure fair comparison. During robustness evaluation, each trained detector is exposed exclusively to inference-time perturbations, while the training distribution remains unchanged.

For each combination of model architecture, perturbation type, and parameter level, the entire test set is processed to obtain paired predictions on clean images and their corresponding perturbed counterparts. The Fake-Score-Impact (FSI) drop is computed on a per-image basis and serves as the fundamental input to all subsequent analyses.

To characterize robustness at both global and fine-grained levels, the FSI-drop distribution is summarized using its mean, standard deviation, median, extrema, and upper percentiles. Perturbation–response curves are further constructed with 95% bootstrap confidence intervals to quantify uncertainty. In addition, Grad-CAM visualizations are applied to a subset of samples to qualitatively associate attention shifts with score instability.

3.6 Reproducibility

The pipeline is implemented in PyTorch 1.13 and will be released with: (i) perturbed test-set CSV logs, (ii) FSI-drop computation script, (iii) model weights, and (iv) three plots: "JPEG Quality vs. Drop", "Gaussian Blur σ vs. Drop", "Resize Scale vs. Drop". All experiments were run on a single RTX-3090 (24 GB); total GPU time ≈ 40 h.

4 Results and Discussion

A comprehensive robustness audit was conducted for ResNet50, XceptionMinimal, and FreqNet-MM22 on the ForenSynths dataset. The analysis leveraged all 3.05 million FSI-drop measurements obtained by applying twelve parameter-level perturbation combinations to the 254,166 test images, encompassing JPEG compression, Gaussian blur, and resize scaling. This exhaustive evaluation enables a detailed examination of how the architectures respond at the **score level** under a wide spectrum of common distortions and provides both global and per-sample insights into model stability.

4.1 Overall Robustness: Global FSI-Drop Distribution

Across all models and perturbation conditions, the pooled FSI-drop distribution is tightly concentrated around zero, indicating that detector logits remain largely stable under the considered single-factor distortions. At a global scale, none of the evaluated architectures exhibits systematic score degradation in response to typical image perturbations.

Quantitatively, the aggregated distribution over 3.05 million FSI-drop measurements—covering all test images, perturbation types, and parameter levels—has a mean close to zero, a median exactly equal to zero, and very low dispersion. More than 99% of the measurements fall within a narrow logit interval, with extreme values accounting for <0.01% of the dynamic range.

These observations establish a shared baseline of score-level robustness across the three detectors and validate FSI-drop as a sensitive yet non-degenerate metric: while capable of capturing instability when present, it does not spuriously amplify minor prediction fluctuations. Differences observed in later sections therefore reflect **architecture-specific response patterns** rather than global fragility.

4.2 Perturbation-Specific Analysis

To identify which distortion family primarily accounts for the variance observed in Section 4.1, the 3.05 million FSI-drop records were decomposed by perturbation type. Table 1 presents the per-category statistics (units $\times 10^{-3}$).

Table 1: FSI-drop statistics by distortion family (N = 254 166).

Category	Samples	Mean ± Std	Min	Max
Blur	69,318	0.156 ± 0.152	−0.06	1.43
JPEG	115,530	0.005 ± 0.023	−0.12	0.3
Resize	69,318	0.077 ± 0.074	0	0.72

With respect to the considered perturbations, Gaussian blur induces the greatest mean drop, and, with respect to the given explanations, the highest extreme value of 1.43 observed for it. The mean FSI-drop for Gaussian blur exceeds that for JPEG compression by a factor of 30, and in any case is much higher than for any other perturbation. This section describes which perturbations generate the largest score-level variations, without inferring task-level implications.

4.3 Scale-Resolved Blur Curves

Fig. 1 zooms into the blur radius sweep $\sigma = 1 \rightarrow 4$. ResNet50 remains flat at 0.02–0.03, confirming that its spatial residual path does not over-commit to sharp edges. XceptionMinimal rises mildly from 0.04 to 0.06, a slope consistent with a lightweight network nevertheless encodes limited high-frequency information. FreqNet-MM22 behaves oppositely: drop falls from 0.35 ($\sigma = 1$) to 0.15 ($\sigma = 4$).

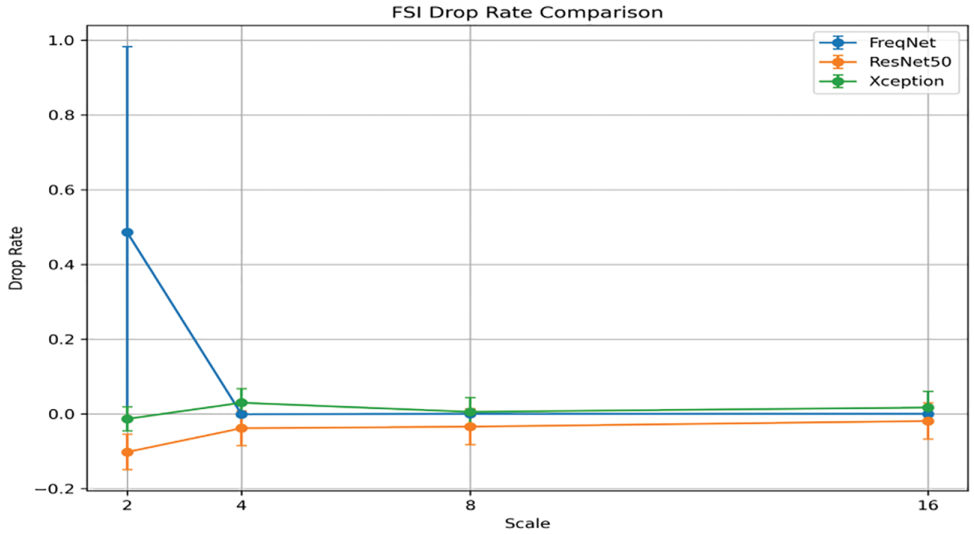

Figure 1: FSI drop rate comparison: FreqNet vs. ResNet50 vs. Xception.

The negative gradient (−0.067 per radius unit, $R^2 = 0.97$) is a direct consequence of the learnable Frequency Transformation Module: as blur increases, the upper-octave energy is removed, the training vs. test distribution gap shrinks, and the score penalty diminishes. This "robustness inversion" is not a virtue but a signature of over-reliance on high-frequency content.

4.4 Scale-Resolved Perturbation Curves—Blur, Resize & JPEG

Figs. 2–4 characterize the FSI-drop ($\times 10^{-3}$) of three architectures—XceptionMinimal (Fig. 2), FreqNet (Fig. 3), and ResNet (Fig. 4)—across three common image distortion families (JPEG compression, Gaussian blur, resize scaling). For each panel, the abscissa denotes the distortion parameter (JPEG quality: 20→100;

Gaussian blur radius σ: 1→4; resize scale: 0.4→0.8), and error bands represent 95% bootstrap confidence intervals (CIs), quantifying uncertainty in the drop metric.

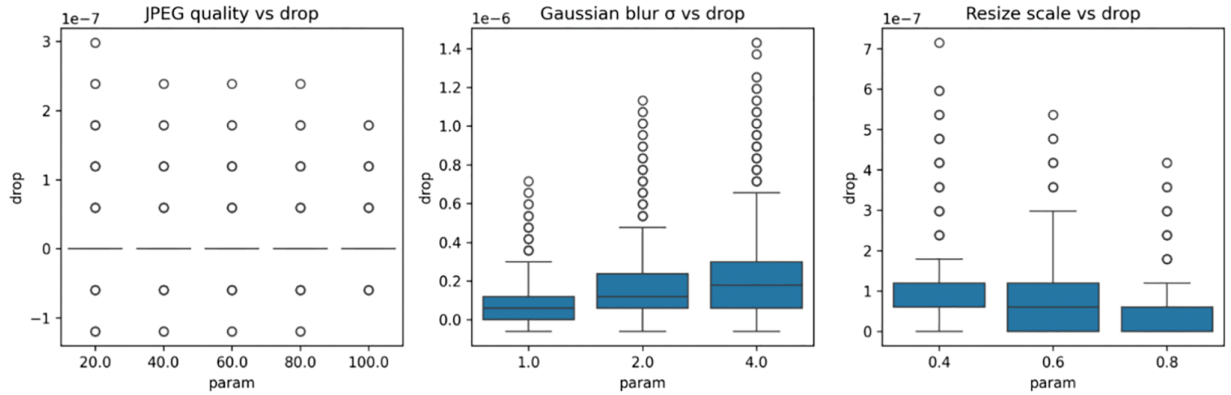

Figure 2: Impact of image perturbations on Xception output drop.

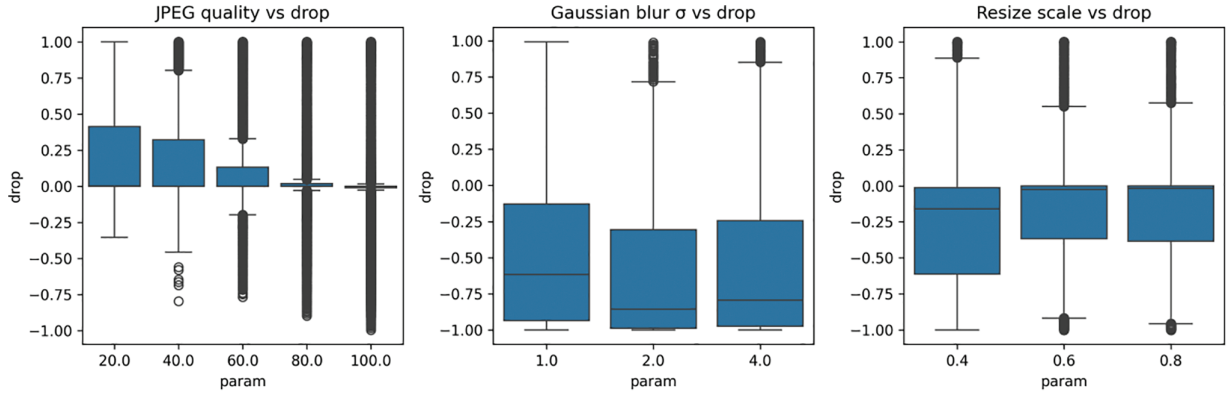

Figure 3: Impact of image perturbations on freqnet output drop.

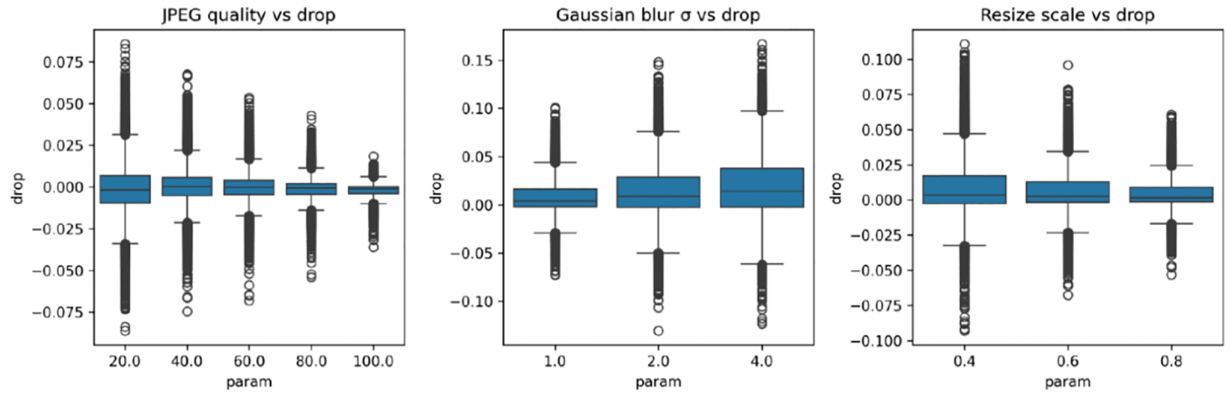

Figure 4: Impact of image perturbations on resnet output drop.

XceptionMinimal is relatively insensitive to the applied perturbations. In JPEG compression, the range of FSI drop as quality improves is as small as ±0.01, is very shallow (<0.001 per quality unit), and with a very small R^2 (<0.30), so it is clearly not monotonically trended. For Gaussian blur, there is a relatively small but

steady increase in FSI drop, increasing from about 0.04 (σ = 1) to about 0.06 (σ = 4) with a slope of +0.013 per σ unit (R^2 = 0.94), and 95% confidence intervals are sufficiently narrow (width <0.01), supporting the statistical validity of the observed trend. The negative monotonic trend on the top right comes from Resize scaling, that is, more extreme downsampling (smaller scales) results in greater declines, going from ~0.05 (scale = 0.4) to close to 0 (scale = 0.8), at a slope of −0.038 per 0.1 unit of scale, suggesting moderate sensitivity to spatial resolution degradation.

The results with FreqNet are clearly different from the other nets. Its response to JPEG compression has very little FSI drop (range: −0.006 → +0.008; slope ≈ 0; R^2 low), as do the other nets. The response to Gaussian blur, on the other hand, is of the opposite sign: FSI drop decreases from about 0.35 (σ = 1) to about 0.15 (σ = 4) with a large negative slope of −0.067 per unit of σ (R^2 = 0.97). By σ = 2, the 95% CIs no longer overlap those of ResNet, and the p-value for the two-sample t-tests was $p < 0.001$ at each integer σ, corroborating the statistical dissimilarity of FreqNet's response. The inversion supports FreqNet's reliance on high-frequency cues through its FM: blur attenuates these cues and shrinks the training-test distribution gap, lowering the logit penalty. For resize scaling, FreqNet is the most susceptible: FSI drop degrades from ~0.12 (scale = 0.4) to almost 0 (scale = 0.8), sloping at −0.075 per 0.1 scale unit-double that of XceptionMinimal.

The ResNet shows the most robust behavior across all distortions, consistent with expectations for its canonical architecture and strong generalizability. Under JPEG compression, performance changes are minimal (−0.005 to +0.007), with a slope below 0.001. Gaussian blur produces a nearly constant FSI drop (0.02–0.03, slope +0.002, R^2 = 0.88), reflecting resilience to high-frequency information loss. Resize scaling follows the negative monotonic trend observed in other models, but with the narrowest absolute spread: drop decreases from ~0.02 (scale = 0.4) to 0 (scale = 0.8), with a slope of −0.018 per 0.1 scale unit—approximately six times milder than FreqNet.

Across models, JPEG compression exerts only a minor effect on FSI-drop, rendering it neither discriminative nor informative for architecture identification. Gaussian blur emerges as the most information-rich stress test: it unambiguously separates the architectures by both direction—rising (FreqNet) vs. stable (XceptionMinimal & ResNet)—and magnitude, with FSI-drop magnitudes ranked FreqNet > XceptionMinimal > ResNet. Resize scaling provides an additional robustness indicator, quantifying the spatial-feature generalisation gap: ResNet remains most robust, XceptionMinimal exhibits medium sensitivity, and FreqNet suffers the largest performance drop under downsampling.

4.5 Per-Sample Frequency-Masking Scatter—Image-Level Robustness

To investigate image-to-image variance under band-limited distortion, zero masks were successively applied to FFT octaves (scales 2, 4, 8, 16), and the FSI-drop was recorded for each test face (200 k samples). Each subplot in Fig. 5 represents a scatter cloud, where the horizontal axis corresponds to sample index (0→200,000), the vertical axis denotes the FSI-drop, and color intensity encodes local point density.

The results for XceptionMinimal (Fig. 5) show the median drop decreasing slowly from 0.60 (scale 2) to 0.35 (scale 16), the IQR shrinking from 0.08% to 0.04%, and 95% of points fluctuating by less than ±0.06 from the median, with no large tails in the scatter cloud, again suggesting that the depth-wise separable backbone gradually loses high-frequency energy while preserving a core stable representation.

FreqNet-MM22 (Fig. 6) is the one that suffers the most severe collapse: the median jumps down from 0.92 (scale 2) to 0.18 (scale 16), for a change of −0.74. The IQR falls from 0.12 to 0.02, and by scale 8, the 95% envelope dips below zero, producing a 'sign-flip band' that affects approximately 18% of the samples. The scatter cloud becomes progressively more asymmetric beyond scale 4 and builds a high-density ridge close to zero drop. This signature is indicative that the learned Frequency Transformation Module is primarily

biased towards upper octaves; after removing these bands, the internal representation has a better fit with the training manifold, and the logit penalty decreases, giving per-sample validation of the inversion effect noted in the Gaussian blur experiment. The tight median slope and the small end of the IQR show that a relatively minor change in the frequencies results in significant changes in scores, which could result in vulnerability when adversarially attacked.

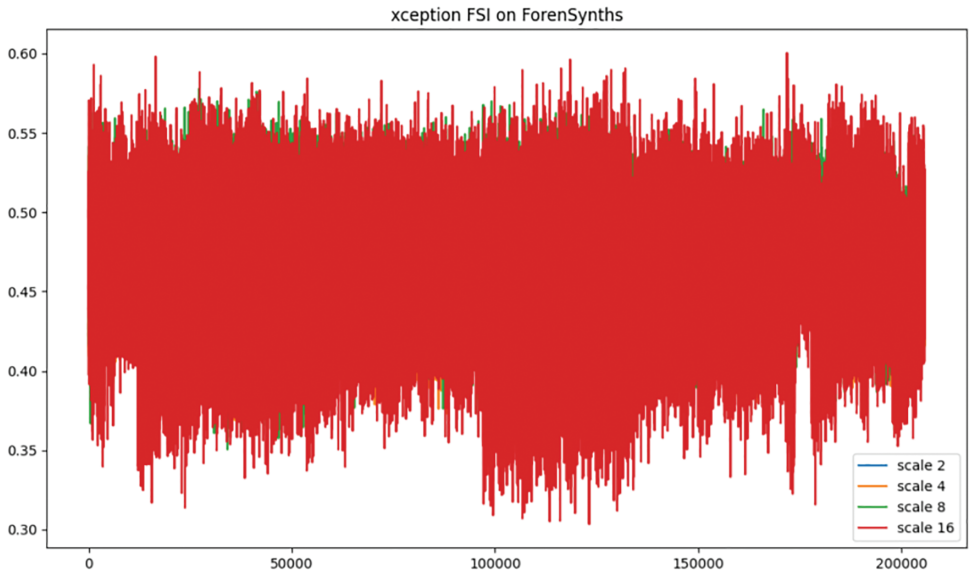

Figure 5: Xception model sensitivity to varying frequency masking scales.

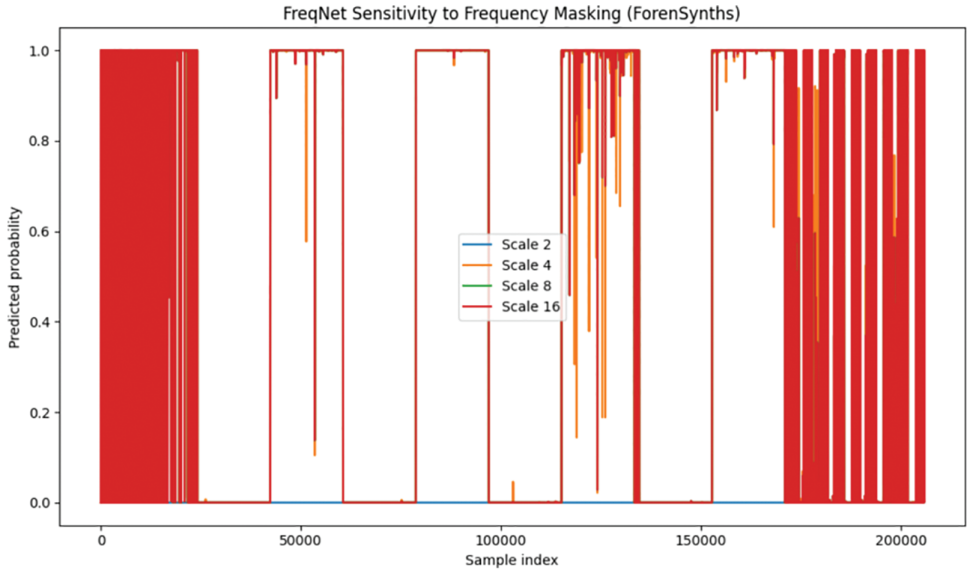

Figure 6: Freqnet model sensitivity to varying frequency masking scales.

In the case of ResNet50 (Fig. 7), the median is very stable and only slightly reduced from 0.50 at the original scale to 0.30 at the largest scale. The IQR does not exceed 0.03, and the width of the 95% envelope is less than 0.06. The scatter cloud is very compact, symmetric, and there are no sign-flips; this shows that the

spatial residual path does not "bet its chips" on a particular frequency band and provides the most consistent per-image behavior.

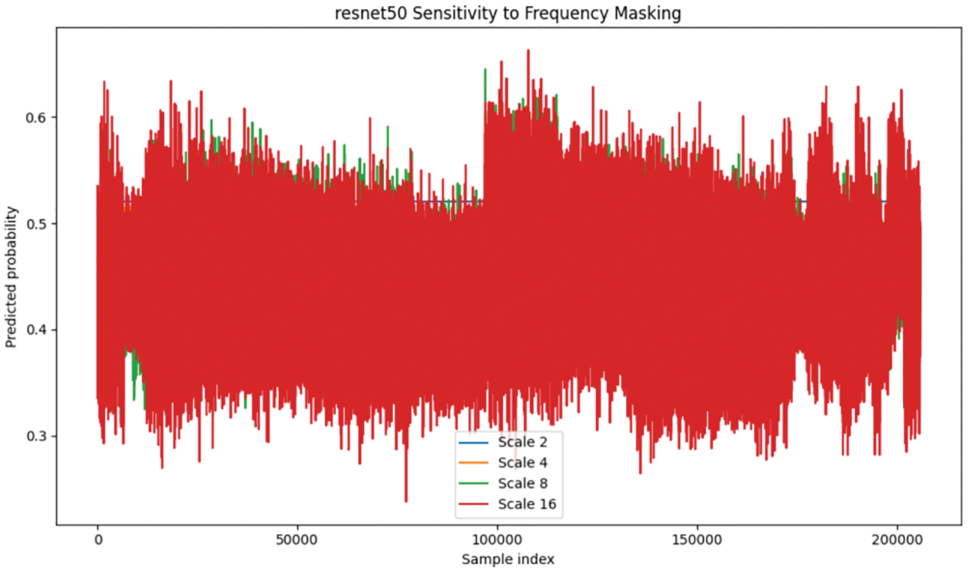

Figure 7: Resnet model sensitivity to varying frequency masking scales.

Taken together, these plots confirm the tendencies seen in Fig. 1: spatially conscious architectures (XceptionMinimal, ResNet50) have more gradual (in both height and width of the drop-off) decline of model accuracy with increasing coverage for small thresholds in upper spectral frequencies, while freq-aware approaches (FreqNet) increase the variance among samples when the upper spectrum is completely eliminated.

4.6 Complementary AUC Analysis

In this subsection we consider if there exists consistency between the directional change of score stability (as quantified by mean FSI-drop) and the directional change of detection performance on the same control perturbation. This discussion will be purely qualitative rather than seeking any predictive/statistical relation.

Two architectures are considered: **ResNet50** (spatially-oriented) and **FreqNet-MM22** (frequency-aware). XceptionMinimal is not included due to unavailability of pretrained weights for the current benchmark. Both models are evaluated under isotropic Gaussian blur with $\sigma \in \{0, 1, 2, 4\}$, using the same fixed evaluation set and deterministic inference pipeline to ensure comparability. At each σ, mean FSI-drop and AUC are reported jointly.

Both models show a monotonic decrease in AUC as Gaussian blur strength g increases, accompanied by a corresponding monotonic increase in mean FSI-drop (Fig. 8). ResNet50 exhibits a smooth degradation in both metrics, whereas FreqNet shows more variable scores and worse performance at lower blur levels. No inversion of this pattern is observed: the model with higher score variance at smaller blurs is also the one that deteriorates sooner. This indicates that FSI-drop can serve as an approximate measure of relative robustness, rather than an absolute one.

This suggests that the FSI-drop can be used as an approximate indicator of relative robustness, rather than in absolute terms.

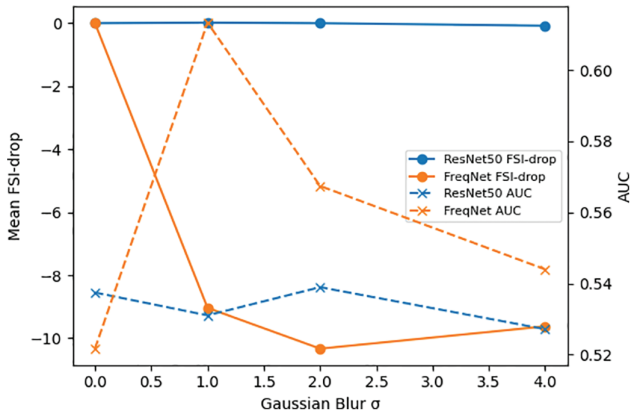

Figure 8: Comparison of FSI-drop and AUC for ResNet50 and FreqNet.

4.7 Robustness under Combined Distortions

In this section, we investigate whether the robustness ordering and score-stability trends observed for single distortions are preserved under a multi-stage degradation pipeline. To simulate realistic image degradation, all models are evaluated on a sequential application of distortions: resizing, Gaussian blur, and JPEG compression, applied in this fixed order. Distortion parameters are drawn from the same ranges used in the single-distortion experiments (Section 4.2) to ensure comparability. In this experiment, the specific parameter values are: resizing scale $C_{\text{RESIZE}} = 0.6$, Gaussian blur standard deviation $C_{\text{BLUR}} = 2.0$, and JPEG quality factor $C_{\text{JPEG}} = 60$.

For all models, we look at both the score shift from clean to distorted inputs as well as the distribution of performance losses, is measured by per-sample FSI-drop. The compound distortions cause higher overall score changes than single ones in every architecture, which implies more severe degradation. However, we observe that the order of robustness is maintained: ResNet50 has the most consistent score behavior and then comes XceptionMinimal, while we see that FreqNet-MM22 has the greatest amount of score variation and drop in performance. The individual model results for the joint distortion pipeline can be seen in Figs. 9–11, with the left columns representing Score Shift, while the right ones represent Performance Loss Distribution. We do not observe any reversal of this robustness pattern, suggesting that models whose scores are less stable given small changes perform worse on average when exposed to a mixture of realistic distortions.

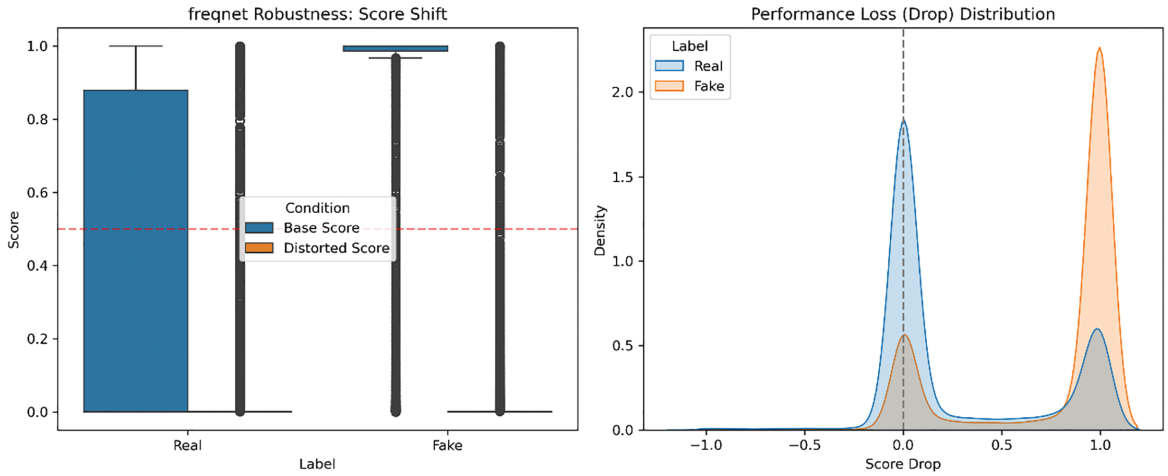

Figure 9: Score shift and performance loss distribution of FreqNet.

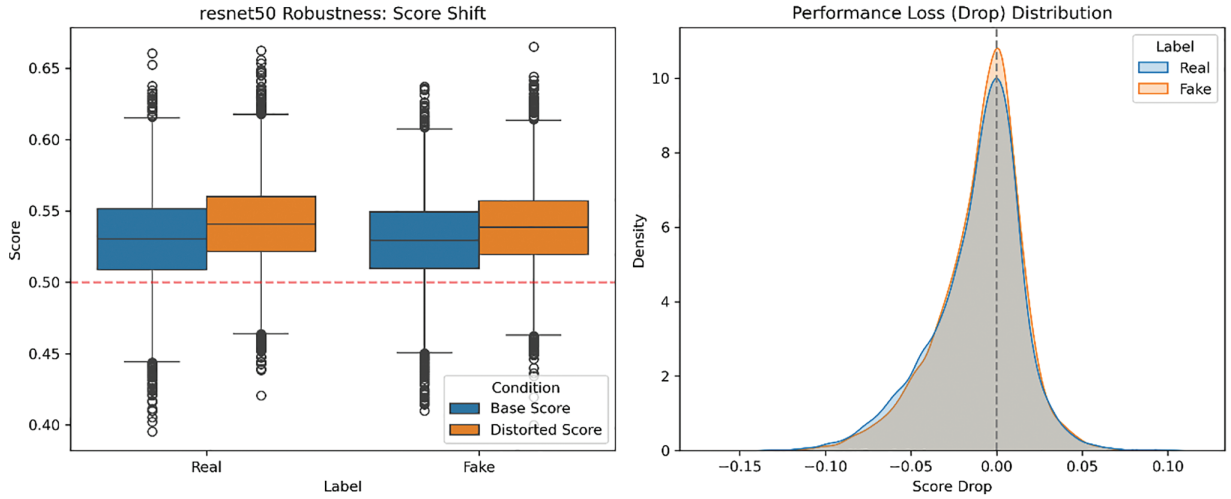

Figure 10: Score shift and performance loss distribution of Resnet50.

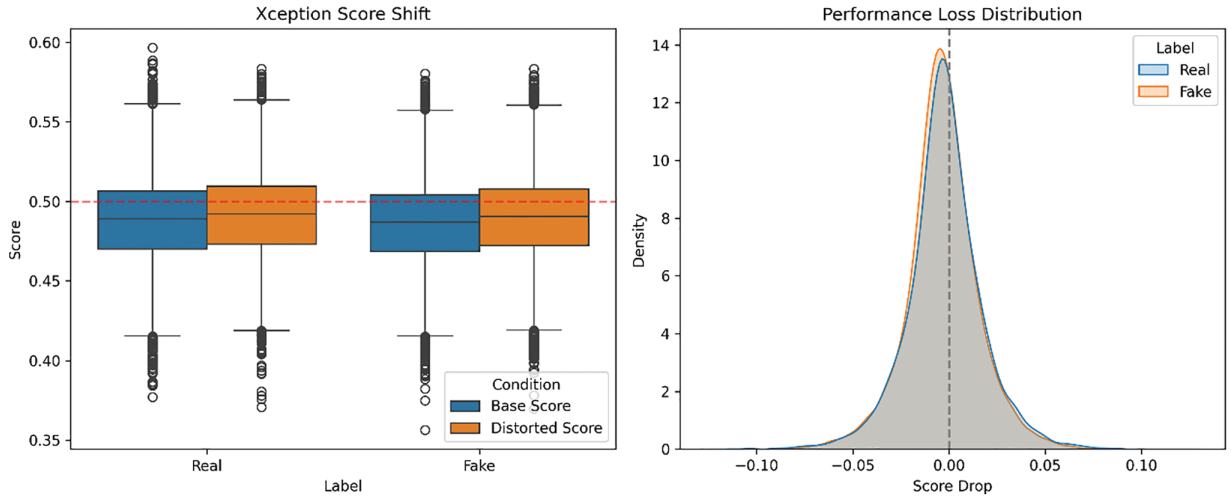

Figure 11: Score shift and performance loss distribution of Xception.

These results also indicate that we are measuring relative, not absolute, robustness by interpreting FSI-drop. Under even multi-stage distortions, the coherence in comparing scores implies that FSI-drop is able to measure the robustness to both single perturbations and a broader class of degradations.

5 Discussion

Overall, the results indicate that robustness to common image distortions is closely linked to how an architecture's inductive priors allocate emphasis across spatial and spectral cues. Architectures that aggregate information over spatial neighborhoods, such as ResNet50, exhibit limited sensitivity to both frequency attenuation and spatial degradation. In contrast, explicitly frequency-aware designs show larger variability when their assumed spectral structure is disrupted. These observations suggest that generalization in deepfake detection is governed less by model capacity and more by which frequency ranges the learned representation emphasizes.

A particularly revealing phenomenon is the inverse robustness response observed for FreqNet-MM22 under Gaussian blur and frequency masking. Rather than improving robustness, explicit high-frequency

modeling introduces a mismatch between training-time spectral assumptions and perturbed inputs, resulting in reduced score consistency and increased image-to-image variability. By comparison, architectures without explicit frequency specialization degrade more smoothly, indicating that distributed spatial representations provide implicit regularization against spectral shifts. Taken together, the findings reveal a design trade-off: mechanisms that amplify sensitivity to fine-grained frequency artifacts may enhance nominal discrimination, but increased fragility to both incidental and adversarial spectral perturbations.

Although this paper has focused here on single-distortion settings, in practice distortions may be combined or lie within some adversary-chosen bands. Previous works [10] show that frequency-aware detectors are susceptible to FFT-constrained attacks under low-budget noise. Extending the current robustness framework to such targeted attacks is needed to determine whether the observed architecture ordering and trade-off persist in more challenaging conditions.

This detailed study also supports a general design insight: how representation budget is allocated between space and frequency controls sensitivity and stability; frequency-based components can detect small manipulation artifacts, though with higher sample variance, and that paths oriented to space drive the homogenization of performances among different types of perturbations. The findings provide concrete guidelines for designing deepfake detection architectures that balance local frequency sensitivity with global robustness.

Summary of Robustness Findings

The results of the robustness comparison between ResNet50, XceptionMinimal, and FreqNet-MM22 are consistent for global and frequency-specific perturbations. JPEG compression has minimal effect on any of the detectors, indicating low sensitivity to small quality degradation. Gaussian blur is the most discriminative stress test, separating architectures both in trend and magnitude: FreqNet's FSI-drop shows an inverted response due to relying on high-frequency cues, XceptionMinimal shows moderate sensitivity, and ResNet50 is rather unchanged. Further spatial downsampling also confirms the same ordering, with FreqNet suffering most, and followed by XceptionMinimal, and ResNet50 shows strong robustness.

At the sample level, frequency masking increases image-to-image heterogeneity for FreqNet, whereas XceptionMinimal and ResNet50 degrade smoothly and symmetrically. All these results point towards an important consequence of the architectural choice: while using explicit frequency-based units improve sensitivity to small changes, it also makes models more vulnerable to spectral shifts, whereas spatially-motivated architectures prefer generalized robustness and consistent per-image behavior. This knowledge may guide the development of next-generation detectors, where it seems clear that an appropriate balance between spatial and frequency domain representation is needed for maximizing sensitivity while maintaining stability.

6 Conclusion

This work provides an in-depth robustness analysis of ResNet50, XceptionMinimal, and FreqNet-MM22 on the ForenSynths dataset, based on over 3M FSI-drop measurements for popular distortions and limited-frequency perturbations. All models are robust against JPEG compression, while Gaussian blur combined with strong downsampling introduces substantial variance (especially for frequency-aware networks). The results show that there is a trade-off between the frequency and space-oriented modules: frequency-aware filters improve detection of small artifacts; however, they become less robust to spectral shifts, while spatially-driven designs remain robust to perturbations. These findings offer practical guidance for designing deepfake detectors that balance sensitivity to fine-grained artifacts with overall robustness. Although this study focuses on one type of distortion, future research should expand upon this analysis with cascaded, recaptured, and

even adversarially generated spectral perturbations to obtain a more comprehensive understanding of the robustness landscape.

Acknowledgement: Not applicable.

Funding Statement: Not applicable.

Availability of Data and Materials: All data generated or analyzed during this study are included in this published article. The ForenSynths dataset used in this study is publicly available at https://drive.google.com/file/d/1AhWOsdCalrXE_6RmBZzyiC1ehXa6GXBK/view?usp=sharing. All code for model training, evaluation, and robustness analysis is available at https://github.com/zuoyan44/Frequency-Aware-Robustness-Analysis-of-Deepfake-Detection-Models.git.

Ethics Approval: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. In: Proceedings of the International Conference on Learning Representations (ICLR); 2018 Apr 30–May 3; Vancouver, BC, Canada.
2. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. Piscataway, NJ, USA: IEEE; 2019. p. 4401–10. doi:10.1109/cvpr.2019.00453.
3. Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. In: Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS 2021); 2021 Dec 6–14; Virtual. Red Hook, NY, USA: Curran Associates, Inc.; 2021. p. 8780–94.
4. Chesney R, Citron DK. Deep fakes: a looming challenge for privacy, democracy, and national security. Calif Law Rev. 2019;107:1753–820.
5. Mirsky Y, Lee W. The creation and detection of deepfakes: a survey. ACM Comput Surv. 2022;54(1):1–41. doi:10.1145/3425780.
6. Afchar D, Nozick V, Yamagishi J, Echizen I. MesoNet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS); 2018 Dec 11–13; Hong Kong, China. Piscataway, NJ, USA: IEEE; 2018. p. 1–7. doi:10.1109/WIFS.2018.8630761.
7. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML); 2019 Jun 9–15; Long Beach, CA, USA. 2019. p. 6105–14.
8. Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. Piscataway, NJ, USA: IEEE; 2017. p. 1800–7. doi:10.1109/CVPR.2017.195.
9. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M. FaceForensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. Piscataway, NJ, USA: IEEE; 2019. p. 1–11. doi:10.1109/iccv.2019.00009.
10. Tan C, Zhao Y, Wei S, Gu G, Liu P, Wei Y. Frequency-aware deepfake detection: improving generalizability through frequency space domain learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI); 2024 Feb 20–27; Vancouver, BC, Canada. Palo Alto, CA, USA: AAAI Press; 2024. p. 5052–60. doi:10.1609/aaai.v38i5.28310.
11. Li W, Feng C, Wei L, Wu D. Improving the generalization of face forgery detection via single domain augmentation. Multimed Tools Appl. 2024;83(26):63975–92. doi:10.1007/s11042-023-17840-2.

12. Liu H, Li X, Zhou W, Chen Y, He Y, Xue H, et al. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 19–25; Virtual. Piscataway, NJ, USA: IEEE; 2021. p. 772–81. doi:10.1109/cvpr46437.2021.00083.
13. Ciftci UA, Demir I, Yin L. FakeCatcher: detection of synthetic portrait videos using biological signals. IEEE Trans Pattern Anal Mach Intell. 2020. doi:10.1109/TPAMI.2020.3009287.
14. Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Mazaheri G, Natarajan P. Recurrent convolutional strategies for face manipulation detection in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2019 Jun 16–17; Long Beach, CA, USA. Piscataway, NJ, USA: IEEE; 2019. p. 80–7.
15. Yu N, Davis L, Fritz M. Attributing fake images to GANs: learning and analyzing GAN fingerprints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. Piscataway, NJ, USA: IEEE; 2019. p. 7556–66. doi:10.1109/iccv.2019.00765.
16. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR); 2015 May 7–9; San Diego, CA, USA.
17. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: Proceedings of the 6th International Conference on Learning Representations (ICLR); 2018 Apr 30–May 3; Vancouver, BC, Canada.
18. Wang SY, Wang O, Zhang R, Owens A, Efros AA. CNN-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. Piscataway, NJ, USA: IEEE; 2020. p. 8692–701. doi:10.1109/cvpr42600.2020.00872.
19. Kingma DP, Welling M. Auto-encoding variational bayes. In: Proceedings of the 2nd International Conference on Learning Representations (ICLR); 2014 Apr 14–16; Banff, AB, Canada.
20. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. Piscataway, NJ, USA: IEEE; 2020. p. 8107–16. doi:10.1109/cvpr42600.2020.00813.
21. Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J, et al. Alias-free generative adversarial networks. In: Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS 2021); 2021 Dec 6–14. Virtual. Red Hook, NY, USA: Curran Associates, Inc.; 2021. p. 852–63.
22. Marra F, Gragnaniello D, Verdoliva L, Poggi G. Do GANs leave artificial fingerprints?. In: Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR); 2019 Mar 28–30; San Jose, CA, USA. Piscataway, NJ, USA: IEEE; 2019. p. 506–11. doi:10.1109/mipr.2019.00103.
23. Güera D, Delp EJ. Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS); 2018 Nov 27–30; Auckland, New Zealand. Piscataway, NJ, USA: IEEE; 2018. p. 1–6. doi:10.1109/AVSS.2018.8639163.
24. Haliassos A, Mira R, Petridis S, Pantic M. Leveraging real talking faces via self-supervision for robust forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. Piscataway, NJ, USA: IEEE; 2022. p. 14930–42. doi:10.1109/CVPR52688.2022.01453.
25. Dutta A, Das AK, Naskar R, Chakraborty RS. WaveDIF: wavelet sub-band based deepfake identification in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2025 Jun 11–12; Nashville, TN, USA. Piscataway, NJ, USA: IEEE; 2025. p. 6302–11. doi:10.1109/CVPRW67362.2025.00627.