

USENIX

THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

On Improving Robustness of Deepfake Image Detectors

Abu Taib Mohammed Shahjahan, Mohammad Mannan, Abdessamad Ben Hamza, and Amr Youssef, *Concordia University*

https://www.usenix.org/conference/usenixsecurity26/presentation/shahjahan

This paper is included in the Proceedings of the 35th USENIX Security Symposium.

August 12–14, 2026 • Baltimore, MD, USA

ISBN 978-1-939133-58-8

Open access to the Proceedings of the 35th USENIX Security Symposium is sponsored by



جامعة الملك عبدالله للعلوم والتقنية

King Abdullah University of Science and Technology

ARTIFACT EVALUATED
USENIX
AVAILABLE

On Improving Robustness of Deepfake Image Detectors

Abu Taib Mohammed Shahjahan
Concordia University, Montreal, Canada

Mohammad Mannan
Concordia University, Montreal, Canada

Abdessamad Ben Hamza
Concordia University, Montreal, Canada

Amr Youssef
Concordia University, Montreal, Canada

Abstract

The rapid advancement of Generative AI has introduced remarkable opportunities while simultaneously raising critical concerns regarding content authenticity. While recent work has increasingly focused on improving the generalization of deepfake detectors across unseen generative models, their robustness against adversarial attacks remains limited. In particular, Abdullah et al. (IEEE SP 2024) evaluated eight detectors and demonstrated that most of them exhibit significant performance degradation under adversarial attacks. We also observed the same phenomenon by testing seven most recent state-of-the-art detectors. To address this problem, we propose a unified framework that integrates three complementary design principles without relying on adversarial training data: (i) higher-order statistical modeling in the frequency domain via Discrete Cosine Transform (DCT)-based moment pooling up to fourth order, (ii) content-agnostic feature representations derived from noise residuals, and (iii) cross-scene generalization enforced through patch-level semantic disruption. A key insight underpinning our approach is that adversarial attacks primarily operate on low-order statistics and visual semantics, leaving higher-order residual-frequency characteristics, particularly kurtosis, largely unconstrained. Extensive experiments demonstrate that our method consistently improves robustness across six architecturally diverse detectors. Notably, we achieve up to 88.9% reduction in recall degradation on current adversarial benchmarks, and improve the best-performing recent detector (Yang et al., IEEE CVPR 2025) from 81.9% to 97.15% accuracy under attack. Overall, our method provides a principled, architecture-agnostic approach for improving deepfake detection robustness against current attacks.

1 Introduction

The development of generative artificial intelligence (GenAI) has dramatically accelerated the synthesis of media content, such as images and videos. Within GenAI, a specific area of focus lies in the generation of fully synthetic images, commonly referred to as *deepfakes*. Unlike manual creation using graphics applications, well-trained deepfake generation models excel in both performance and speed of generation. Deepfake generation by Generative Adversarial Networks (GANs) and Diffusion Models (DMs) [17, 37, 45, 51, 55, 73] has attracted much public attention due to their visually striking outputs, which are nearly indistinguishable from genuine photographs, even to human observers [43].

Such progress, while remarkable, also introduces significant societal risks, as these synthetic images can be exploited for malicious purposes, including blackmailing, harassment, and large-scale disinformation campaigns. Diffusion model (DM)-based text-to-image frameworks, in particular, heighten these concerns by enabling adversaries to generate highly persuasive synthetic imagery aligned with specific narratives while requiring minimal technical expertise. As the realism and accessibility of these technologies continue to improve, recent studies have warned that deepfakes can erode public trust in digital media, amplify misinformation during politically sensitive events, and inflict significant psychological and societal harm [3]. The rapid proliferation of deepfake generation tools has also raised important legal and ethical concerns related to privacy violations, identity misuse, and the urgent need for effective regulatory frameworks to mitigate abuse [40]. These concerns are no longer theoretical, as harmful misuse of generative AI images has already caused severe real-world consequences, including suicide linked to AI-generated morphed nude images and videos [28], acute mental health impacts on young people [60], and political misinformation following the capture of Venezuelan President Maduro [59].

Due to increasing awareness of such threats and rapid advancements in generative techniques, several tools exist for fully synthetic image detection [7, 11, 12, 26, 29, 39, 49, 50, 52, 57, 61, 66, 67, 69, 70, 75]. Common approaches include detecting semantic inconsistencies in images, or capturing low-level artifacts introduced by the generative models for detection. For example, geometric inconsistencies in scenes as well as in shadows, perspective and lighting in an image

are used for detection [52]; low-level artifacts from generative models are captured in the spatial domain [65], or in the spectral domain [35]. Such detectors, as reported, often achieve very high accuracy in deepfake detection; e.g., 100% accuracy of FIRE [12] against well-known datasets.

Recently, deepfake detectors have encountered generalization challenges against adversarial attacks due to the lack of adversarial evaluation. Abdullah et al. [1] proposed a simple black-box attack to adversarial fake images by leveraging vision foundation models such as EfficientNet [58], ViT [15], CLIP-ResNet [47]. Their approach does not add any adversarial noise (in contrast to [16, 24]), which might degrade the quality of an image and a viewer might identify the image as fake just by looking at it. They only adversarially manipulate the content to match it to the given text prompt; e.g., a prompt that says "a smiling face" should craft an adversarial fake image with a smiling face. Guided by surrogate deepfake classifiers implemented using foundation models, they adversarially updated the weights of a fake image generator, which is then used to generate adversarial fake images. All eight detectors [2, 9, 23, 38, 44, 50, 54, 65] (from 2018–2024) they tested, suffered significant performance degradation against such adversarial attacks; e.g., the worst recall degradation of 88.35% observed for DCT [50] using the surrogate model CLIP-ResNet, and a degradation of 70.85% for CNN-F [65].

We evaluated adversarial attacks introduced by Abdullah et al. [1] against seven recent SOTA detectors [7, 11, 12, 29, 39, 57, 70] (all from 2024-2025), and found the similar performance degradation. For example, the accuracy of FIRE [12] on the ImageNet dataset [13] dropped from 100% to 50%, and the accuracy of Upsampling [57] on the DiffusionForensics [67] dataset dropped from 95.2% to 54.8%. To avoid such severe degradations, Abdullah et al. [1] suggested the use of adversarial training as an immediate strategy to improve resilience. However, such adversarial training datasets are often scarce—leaving the SOTA detectors brittle against these attacks. Even more critically, the attacker can adapt and use an adversarially trained surrogate to craft a new distribution of adversarial samples, and such an adaptive attack can still significantly degrade the performance of the adversarially trained defense, making adversarial training less effective in the long run.

To counter this challenge, we propose a new approach, by leveraging higher-order statistical modeling combined with content-agnostic feature representations and cross-scene generalization to improve robustness of SOTA detectors without requiring adversarial training. To date, as per our knowledge no deepfake detector has explicitly applied fourth-order statistics for GAN/diffusion image classification. We expect that fourth-order pooling can improve the performance of detectors as it is more sensitive to nonlinearity (compared to lower-order statistics) and strong asymmetry, which can potentially facilitate better generalization to new synthetic image generators. Abdullah et al. [1] showed content-agnostic features are promising to increase robustness against user-customized models. Zheng et al. [75] showed the need to break *semantic artifacts* for cross-scene generalization. We combine these suggestions in our proposed approach.

Instead of building a new detector, our goal is to show general effectiveness of the proposed methods against current adversarial attacks by integrating our approach into current SOTA detectors. Such integration is apparently feasible as evident from the six different detectors that we experimented with. We had to adjust the original network architectures of the selected detectors, in particular in D^3 [70] (due to its complex design) to incorporate our changes. Also, the integration of content-agnostic features and fourth-order statistics required high GPU resources and long training time. We conducted experiments with different order of statistics to find out which statistical order provides the desired robustness against adversarial attacks. In the end, by integrating our approach into six different types of detectors, we showed that our approach is adaptive, and could be extended to other types of deepfake detectors to improve their robustness against the adversarial attacks without the need for adversarial training.

Main contributions.

1. We evaluated adversarial attacks as introduced by Abdullah et al. [1] against seven most recent SOTA detectors, and found similar performance degradation with the maximum accuracy dropped from 98.9% (for the BOSSBase steganography dataset) to 48.7% for Directionality [39] (Usenix Sec'24), and the minimum accuracy dropped from 86.7% (for out-of-domain images) to 81.9% for D^3 [70] (CVPR'25).

2. To enhance robustness against adversarial attacks without adversarial training, we introduce a new approach, by combining three key components: (i) fourth-order statistics (i.e., kurtosis) in the frequency domain to construct a frequency-based representation for classifier design and evaluation in deepfake image detection; (ii) exploitation of directional statistics in deepfake images by applying random patch-level shuffling and rotation to improve cross-scene generalization; and (iii) content-agnostic feature representations.

3. Our proposed method improves accuracy of the best recent SOTA detector (from our tests) D^3 by Yang et al. [70], from 81.9% to 97.15%, against adversarial attacks. When applied to the five worst performing detectors against Abdullah et al. [1]'s attacks, our approach significantly mitigated the recall degradation against such attacks, with the maximum reduced from 88.35% to 9.8% and the minimum reduced from 34.05% to 23.3%.

2 Related Work

In this section, we discuss existing SOTA deepfake image detection approaches, and the recent attacks against them.

2.1 Existing Detection Approaches

Artifact-based detection methods. These methods rely on detecting intrinsic artifacts or frequency/spatial inconsistencies introduced by diffusion or generative models. Wang et al. [67] proposed the DIffusion REconstruction Error (DIRE), a spatial-domain representation that measures the reconstruction error between an input image and its diffusion-based reconstruction. Inspired by DIRE, Tran et al. [61] introduced DiffCoR, a detection method that converts both real and AI-generated images into the frequency domain using the Discrete Fourier Transform (DFT). Similarly, Chu et al. [12] proposed Frequency-guIded Reconstruction Error (FIRE) based on the observation that diffusion models struggle to reconstruct mid-band frequency components of real images. Karageorgiou et al. [29] investigated the spectral properties of real images and found that their spectral distributions exhibit invariant yet highly discriminative patterns. Complementarily, Ricker et al. [50] analyzed the frequency characteristics of diffusion model outputs, finding that such models produce fewer high-frequency artifacts. Tan et al. [57] proposed an artifact representation termed as neighboring pixel relationships (NPR), which models second-order statistics in the spatial domain to capture local *upsampling* artifacts among image pixels. Among these detectors, FIRE [12] demonstrated 100% accuracy against well-known datasets.

Learning-based detection methods. Beyond artifact-level cues, several approaches leverage deep architectures to automatically learn discriminative representations of synthetic content. Yang et al. [70] introduced the Deepfake Detector (D^3), a framework that incorporates an auxiliary branch trained on distorted images as discrepancy signals, thereby enhancing cross-generator generalization. Cheng et al. [11] proposed CO-SPY (COmbining Semantic and Pixel Features to Detect sYnthetic Images by AI), which integrates both semantic-level and pixel-level features for more robust synthetic image detection. Huang et al. [26] presented a vision-language-based framework named SIDA (Social media Image Detection, localization, and explanation Assistant) to detect tampered social media images. Among these detectors, D^3 [70] claimed the maximum accuracy of 97.9% (for in-domain images).

Pretrained-based detection methods. Several detectors have utilized the inherent structure of pretrained diffusion models to distinguish real from generated images. They aim to eliminate data maintenance and training altogether. Brokman et al. [7] developed a framework that integrates manifold analysis with diffusion model score functions. Zheng et al. [75] explored a patch-based approach designed to disrupt *semantic artifacts* in cross-scene generalization of AI-generated image detection. Cao et al. [8] investigated the internal dynamics of diffusion models and identified that visual artifacts often arise during the *mutation phase* of image generation, where certain spatial regions exhibit anomalous score dynamics over time; they utilized the intrinsic temporal characteristics of diffusion models themselves for more principled detection and correction of synthetic content. Among them, Brokman et al. [7] claimed the maximum accuracy of 83.9%.

Spatio-frequency statistical detection methods. Several studies explored the inherent statistical discrepancies between real and synthetic images across both spatial and frequency domains. Corvi et al. [49] employed second-order statistics—including the autocorrelation function and the power spectrum—to capture the characteristic deviations introduced by generative models. In a related direction, Wolter et al. [69] proposed a multi-scale spatio-frequency representation based on wavelet-packet coefficients, which jointly captures second-order spatial and frequency statistics. These studies collectively demonstrate that deviations in natural image statistics, when analyzed across the spatial–frequency spectrum, offer powerful cues for distinguishing diffusion/GAN-generated images from real ones. Wolter et al. [69] claimed to have an accuracy of 99.45% (no results available for Corvi et al. [49]).

Geometry and structural inconsistency-based detection methods. Sarkar et al. [52] proposed a scalable, data-driven framework leveraging three complementary projective geometry cues (perspective field, linear structures and vanishing lines, and object–shadow relationships) to detect geometric discrepancies in generated imagery. Specifically, they trained geometry-aware classifiers demonstrated robustness and generalization across a wide range of generative models, including the most recent diffusion-based architectures such as Stable Diffusion XL [45] and Kandinsky-v3 [64]. Complementing this perspective, Lorch et al. [39] explored the phenomenon of *directionality* in images, wherein horizontal and vertical pixel sequences display distinct statistical characteristics. These findings underscore the importance of incorporating geometric and structural reasoning for building robust, rotation-invariant detectors. Sarkar et al. [52] claimed AUC (Area Under the Curve) 0.97 for their detector.

Novelty of our techniques. We identified three approaches from the existing literature and statistical methods, which can significantly improve robustness of fake image detectors against known adversarial attacks. Our approaches are designed to be integrated with existing detectors to improve their performance against existing adversarial attacks. Our methods can also be adopted selectively—one or more at a time—by detectors as suitable for their architecture and available training resources. For example, in D^3 [70], we integrated all of our techniques, and for the other five detectors we used 1–2 techniques due to time and resource constraints (but still achieved significant performance improvement). To the best of our knowledge, no detector papers integrated the first four statistical moments, namely the mean, variance, skewness, and kurtosis in the frequency domain, and thus, did not exploit this multi-order statistical pooling to expose generative traces in fake adversarial images.

2.2 Existing Adversarial Attacks on Detectors

In addition to detection, recent research has investigated adversarial robustness and attack strategies targeting synthetic image detectors. Abdullah et al. [1] demonstrated that lightweight customization of large generative models allows attackers to generate numerous customized fake images, significantly increasing the threat landscape. They proposed improved generalization strategies using *content-agnostic* ensemble defenses that fuse foundation model features with frequency-based cues. Moreover, they showed that attackers can craft adversarial examples with *semantic properties* of an image for adversarial manipulation in a fully black-box setting without explicit noise injection. Eddoubi et al. [16] introduced the RAID dataset, containing transferable adversarial synthetic images created using highly transferable attack strategies. They demonstrated that ensemble adversarial attacks can achieve transferability comparable to white-box settings, thereby offering a standardized benchmark for evaluating detector robustness. Finally, Hou et al. [24] proposed StatAttack, a novel natural degradation-based attack that minimizes distribution-aware loss to bridge the feature distribution gap between real and fake images.

Defense	Acc	AP
Brokman et al. (ManifoldBias) [7]	0.500	0.706
Cheng et al. (CO-SPY) [11]	0.618	0.679
Chu et al. (FIRE) [12]	0.500	0.548
Karageorgiou et al. (SPAI) [29]	0.660	0.678
Lorch et al. (Directionality) [39]	0.487	0.485
Tan et al. (Upsampling) [57]	0.548	0.569
Yang et al. (D^3) [70]	**0.819**	**0.909**

Table 1: Performance of recent SOTA detectors tested against the surrogate StyleCLIP dataset [1] in terms of Accuracy (Acc) and Average Precision (AP)

3 Evaluation of Recent SOTA Detectors against New Attacks

Several new detectors in top academic venues came out after the publication of the attacks by Abdullah et al. [1]. However, these new detector papers do not provide any evaluation against these known attacks. To understand the effectiveness of such attacks against the newly designed detectors, we chose seven recent (published between 2024-2025) SOTA detectors [7, 11, 12, 29, 39, 57, 70], according to the following criteria. (1) Detection performance: we prioritized defenses that reported strong detection results. (2) Code availability: only models with publicly available checkpoints and training code were considered. (3) Content diversity: beyond standard face images, we included defenses evaluated on varied content types, such as artwork, illustrations, and images of different classes of objects. (4) Methodological variety: the selected defenses employ diverse approaches, allowing us to assess the robustness of different detection strategies comprehensively. We did not perform any fine-tuning on these seven models; rather, we utilized their officially published weights to evaluate them against the adversarial attack of Abdullah et al. [1]. The results were compared using accuracy (ACC) and average precision (AP), consistent with the metrics employed in the original publications. Note that we excluded the attacks from Eddoubi et al. [16] and Hou et al. [24] as their adversarial images are generated by adding noise, which may degrade the image quality significantly (see Sec. 5.1, under "Evaluating the quality of adversarial data").

Attack	Dataset	Acc	AP
Sec3	SDdataset	0.820	0.884
Sec3	StyleCLIP	0.808	0.896
Sec5.1 FM genimages	AbsoluteReality	0.873	0.923
Sec5.1 FM genimages	AnalogDiffusion	0.800	0.860
Sec5.1 FM genimages	DreamlikePhotoreal	0.871	0.933
Sec5.1 FM genimages	DreamShaper	0.893	0.948
Sec5.1 FM genimages	EpicDiffusion	0.842	0.903
Sec5.1 FM genimages	epiCRealism	0.831	0.891
Sec5.1 FM genimages	FotoAssistedDiffusion	0.839	0.901
Sec5.1 FM genimages	Haveall	0.867	0.923
Sec5.1 LoRA genimages	AddAestheticDetails	0.765	0.839
Sec5.1 LoRA genimages	AddDetails	0.765	0.833
Sec5.1 LoRA genimages	EnhanceImage	0.765	0.842
Sec5.1 LoRA genimages	ImproveContrast	0.826	0.897
Sec5.1 LoRA genimages	IncreaseSharpness	0.686	0.756
Sec5.1 LoRA genimages	ReduceImageNoise	0.773	0.845
Sec5.1 LoRA genimages	TweakBrightness	0.771	0.798
Sec5.1 LoRA genimages	TweakSkinTexture	0.758	0.826
Sec5.2 AdvImageswSurrogate	CLIPResNet	0.788	0.854
Sec5.2 AdvImageswSurrogate	EfficientNet	0.859	0.923
Sec5.2 AdvImageswSurrogate	ViT	0.813	0.902
Sec5.2 AdvtrainedCLIPResNet	advV2	0.752	0.832
Sec5.2 Dataset for Adv. training	ForDefenses	0.778	0.841
Sec5.2 Dataset for Adv. training	ForSurrogate	0.796	0.855
Sec5.2 Surrogate-StyleCLIP	StyleCLIP	0.819	0.909

Table 2: Performance of D^3 against all datasets of Abdullah et al. [1] in terms of Accuracy and Average Precision

Evaluating the SOTA detectors against the selected attacks required non-trivial efforts. A common issue with most of detectors' code is mismatched machine learning library versions causing the code to crash. We also had to make small changes to make the detectors run in our environments. In all cases, we first reproduced the published results of the tested detector before performing tests with the adversarial datasets.

Our results are summarized in Table 1. It is evident that D^3 [70] is the best detector against the selected adversarial attacks, while all other detectors performed very badly, with some detectors' results performing no better than random guessing (accuracy ≤ 0.5). To understand the performance of D^3 more thoroughly, we evaluated it against all types of test datasets provided by Abdullah et al. [1], including user-

customized generative models, and datasets generated by controlling content and image quality; see Table 2 for the results. Evidently, for all datasets, D^3 shows performance degradation, including the first two test datasets (labeled Sec3) which address the issue of controlling the content and image quality. It shows that performance of SOTA detectors might degrade against high quality fake test data, if the training fake dataset quality is not maintained, which is consistent with the findings of Abdullah et al. They also emphasized the need for generalization of the defenses against user-customized models by demonstrating performance degradation of their selected defenses tested against such models. In Table 2, under all the rows labeled Sec5.1, we also demonstrated performance degradation against two approaches – Full Model fine-tuning (FM [34]) and Low-Rank Adaptation (LoRA [25]) used by Abdullah et al. The rest of the results labeled Sec5.2 in Table 2 show performance degradation against different adversarial attacks.

Overall, we reaffirmed the significant performance degradation similar to Abdullah et al. [1] for the most recent SOTA detectors, and even the best performing detector (D^3), is affected by all of the test datasets. Such results highlight the need for robust new mechanisms to improve the detectors.

4 Methodology

In this section, we present a detector that attains robustness to content-manipulating adversaries without adversarial training.

4.1 Problem Formulation and Threat Model

Let $\mathcal{G} = \{G_i\}_{i=1}^{N}$ be a generator set composed of N different types of generators, and denote by $\mathcal{R}_i = \{r_1^{(i)}, r_2^{(i)}, \ldots, r_{N_i}^{(i)}\}$ and $\mathcal{F}_i = \{f_1^{(i)}, f_2^{(i)}, \ldots, f_{N_i}^{(i)}\}$ the real and fake samples of G_i, respectively, where N_i is the number of samples in each class. Our objective is to correctly distinguish between real and fake samples when presented with any image from the overall dataset $\mathcal{D} = \{\{\mathcal{R}_i\}_{i=1}^{N} \cup \{\mathcal{F}_i\}_{i=1}^{N}\}$.

For the threat model, we assume attackers are interested in using fully synthetic, high-quality deepfake images for malicious purposes. They have the resources to use open-source SOTA image generators by themselves, or can access commercial generators. They also exploit the best known adversarial techniques in their generated images to defeat the current deepfake detectors. We exclude partially modified images, which can also be used for malicious purposes.

4.2 Guiding Principles

We adhere to the following guiding principles in our design:

Disentangle semantic cues from generative traces. To develop a universal detector for AI-generated images, it is essential to consider not only generator artifacts but also the semantic artifacts inherent in visual content. Generator artifacts arise from the synthetic image generation process itself, whereas semantic artifacts emerge from the semantic composition of images, or from biases introduced during dataset curation and preprocessing. Notably, semantic artifacts exhibit properties analogous to generator artifacts and may inadvertently be exploited by detection models.

Prioritize content-agnostic evidence. To enhance generalization across diverse scenes, it is crucial to mitigate the overfitting of semantic artifacts during training [75]. Moreover, effective defenses should prioritize intrinsic imperfections in synthesized images rather than being influenced by shifts in content distribution. Consequently, incorporating content-agnostic features into detection frameworks is vital for improving robustness and cross-domain generalization [1].

4.3 Proposed Framework

4.3.1 Architecture Overview

The overall architecture of the proposed framework is illustrated in Figure 1 (integrated with $\mathbf{D}^3$). Conventional learning-based detection approaches often inadequately account for intrinsic characteristics of authentic images, which can lead to overfitting and limited generalization performance, particularly under distribution shifts [70]. To address this limitation, we adopt a pretrained-based detection paradigm, wherein feature representations are extracted from a model not explicitly optimized for generated image detection. Such representations benefit from less biased decision boundaries and improved robustness across diverse image domains.

To enhance cross-scene generalization and mitigate the influence of semantic *overfitting*, we disrupt global semantic structures while preserving localized artifact signatures. We partition the input image into non-overlapping patches and apply random transformations (shuffling and rotation), so that the model does not rely on high-level scene context (i.e., the semantic arrangement of objects). This forces the detector to extract universal, low-level features (e.g., DCT kurtosis, pixel gradient magnitudes, color covariance) that are invariant to semantic content. Here, high-level scene context refers to the global structural coherence and semantic relationships between objects, such as facial geometry, anatomical symmetry, and environmental spatial logic. Low-level features are defined as the stationary statistical properties of the image signal that persist regardless of spatial orientation—including frequency-domain distributions (e.g., DCT kurtosis), first-order spatial statistics (e.g., pixel gradient magnitudes), and inter-channel spectral dependencies (e.g., color covariance).

Prior studies have also demonstrated that the noise residual (or noise space) of an image, obtained by suppressing semantic content, can encode discriminative signatures valuable for deepfake detection [46]. Building upon this observation, we employ a state-of-the-art denoising method, MM-BSN [72],

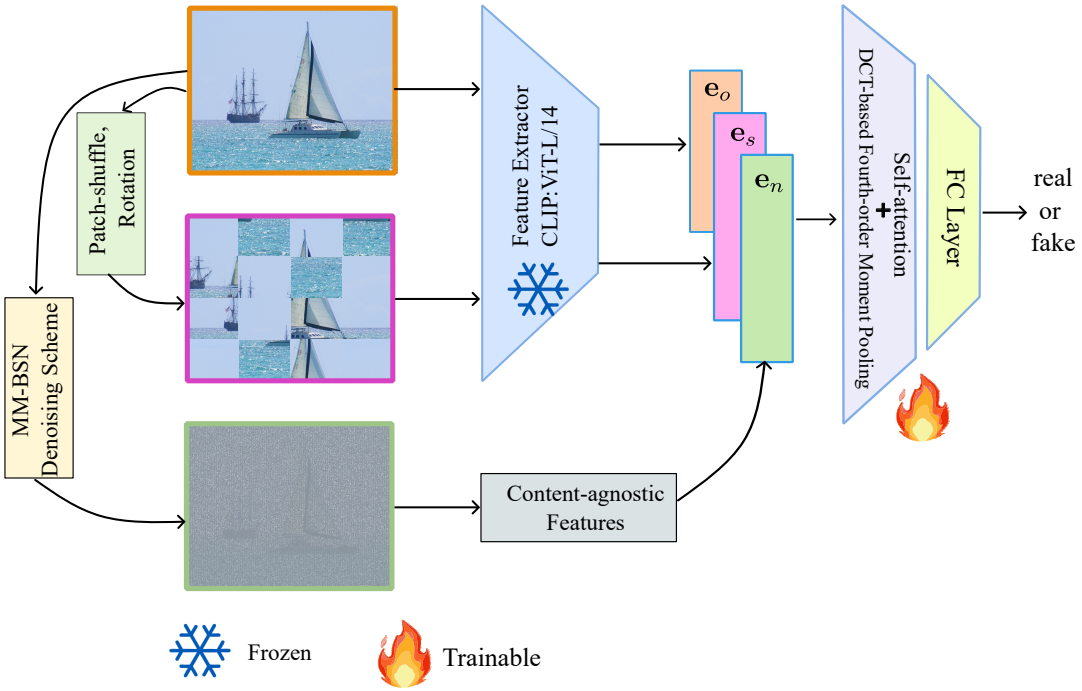


Figure 1: Overall framework of our enhanced deepfake detector as integrated with D^3. The original image and its patch-shuffled variant with our new addition of random rotation are fed into the pretrained vision model for feature extraction. We add a new MM-BSN denoising scheme as the third branch to extract content-agnostic features from the same input image. We then utilize the self-attention module (as in D^3) to encourage the learning of shared artifacts between the features, which are then concatenated with our newly added frequency-aware fourth-order statistics captured with DCT-based fourth-order moment pooling. Finally, we use a single, fully-connected (FC) layer linear classifier to get the final prediction (as in D^3).

to extract noise residuals for all training images. Features extracted from these residuals serve as content-agnostic representations, further strengthening generalization by reducing sensitivity to image semantics. Within our architecture, the original input image, along with its patch-shuffled, and rotated variant, is fed into a pretrained CLIP ViT-L/14 model [47] to obtain visual embeddings. Features are extracted from the penultimate layer of the CLIP network, as this layer preserves richer and more detailed visual information compared to the final embedding layer. In parallel, content-agnostic features are extracted from the corresponding noise residuals of the same input image. Subsequently, a self-attention mechanism, together with frequency-aware fourth-order statistics, is employed to enable effective interaction and discrepancy learning among the three feature modalities. This design allows the model to emphasize inconsistencies indicative of synthesized content while suppressing semantic bias. Finally, a linear classifier aggregates the fused representations to predict binary real or fake labels.

4.3.2 Framework Components

We now provide the details of the components of our architecture (see Figure 1).

Triple-path discrepancy learning. To encourage the model to capture deeper and more robust artifacts, we introduce two additional branches in conjunction with the primary input pathway. Specifically, the first branch processes the original input image, the second branch operates on a patch-shuffled and rotated variant of the same image, and the third branch processes content-agnostic features derived from the corresponding noise residual. This multi-branch design is intended to reduce reliance on semantic cues while promoting the learning of invariant and generalizable discriminative patterns.

Formally, let an original image $\mathbf{X}_o$ be sampled from the source dataset $\mathcal{D}_s = \left\{ \{R_i\}_{i=1}^k \cup \{F_i\}_{i=1}^k \right\}$, where R_i and F_i denote real and synthesized images, respectively, and $\mathbf{X}_o \in \mathcal{G}_s \subset \mathcal{D}_s$. Two visual embeddings are extracted using a pretrained CLIP backbone. The first embedding is obtained from the augmented original image as

$$\mathbf{e}_o = \text{CLIP}^*\big(\text{AUG}(\mathbf{X}_o)\big). \tag{1}$$

Here, the operator $\text{AUG}(\cdot)$ represents standard data augmentation strategies commonly adopted for generated image detection, such as random cropping, resizing, and color jittering [65]. The second embedding is generated from a patch-shuffled version of the image, which suppresses superficial

and semantic artifacts:

$$\mathbf{e}_s = \text{CLIP}^*\big(\text{PS}(\text{AUG}(\mathbf{X}_o))\big). \tag{2}$$

In the patch-shuffling operation $\text{PS}(\cdot)$, the image is divided into fixed-size patches that are randomly permuted and independently subjected to random rotations (e.g., 90°, 180°, or 270°). This operation effectively disrupts global spatial structure while preserving local statistics. The embeddings $\mathbf{e}_o$ and $\mathbf{e}_s$ are D-dimensional visual feature vectors extracted from the penultimate layer of the pretrained CLIP model, denoted by CLIP*, which retains richer visual representations than the final projection layer.

Content-agnostic residual extraction. We employ the multi-mask blind-spot network (MM-BSN) [72] to estimate a noise residual map from a noisy observation $\mathbf{X}_o$. MM-BSN was selected over alternatives (e.g., DnCNN [74] and Blind2Unblind [68]) as it operates in a self-supervised manner, eliminating the need for clean reference images during training, and it has demonstrated state-of-the-art performance on real-world noise estimation benchmarks. To effectively disrupt spatial dependencies in the noise while preserving fine-grained texture information, MM-BSN adopts a multi-mask strategy that fuses features extracted from multiple masked convolutional filters. Following the multi-branch feature aggregation paradigm introduced by Sun et al. [56], features from parallel masked pathways are jointly integrated to predict a spatial noise residual map $\hat{\mathbf{V}} \in \mathbb{R}^{H\times W\times 3}$ as follows:

$$\hat{\mathbf{V}} = \mathcal{F}_{\text{fuse}}\big(\text{Concat}(f_{m_1}(\mathbf{X}_o), f_{m_2}(\mathbf{X}_o), \ldots, f_{m_k}(\mathbf{X}_o))\big), \tag{3}$$

where $f_{m_i}(\cdot)$ denotes the i-th masked processing branch. The hierarchical fusion network $\mathcal{F}_{\text{fuse}}$ comprises concatenation-based dilated convolutional layers (CDCLs) that aggregate multi-scale spatial context. By incorporating a limited number of dilated convolutional layers (DCLs) within each CDCL block, the effective receptive field is expanded to capture surrounding contextual information while avoiding the parameter growth associated with naive stacking of standard convolutions. This design facilitates efficient interaction among different mask pathways, thereby reinforcing texture recovery while maintaining a compact model footprint.

Unlike feature representations extracted directly from the original image, the noise artifacts require a compact yet expressive encoding. Inspired by the projection head design used by Chen et al. [10], we introduce a noise projection module $\mathcal{P}_{\text{noise}}$ to map the spatial residual features into a low-dimensional latent space. Specifically, the noise embedding $\mathbf{e}_n \in \mathbb{R}^D$ is defined as

$$\mathbf{e}_n = \mathcal{P}_{\text{noise}}(\hat{\mathbf{V}}), \tag{4}$$

where $\mathcal{P}_{\text{noise}}(\cdot) = \text{MLP}(\text{GAP}(\cdot))$. Here, GAP denotes Global Average Pooling, which collapses the spatial dimensions $H \times W$ into a channel-wise mean vector, and MLP represents a multi-layer perceptron that projects the resulting vector into a D-dimensional embedding space. Consequently, $\mathbf{e}_n$ encodes global noise characteristics as a compact feature vector, matching the dimensionality of the CLIP-based embeddings $\mathbf{e}_o$ and $\mathbf{e}_s$.

This tri-path representation is designed to enhance robustness and generalization to unseen generators and post-processing manipulations. Building upon the complementary embeddings $\mathbf{e}_o$, $\mathbf{e}_s$, and $\mathbf{e}_n$, the subsequent objective is to integrate their heterogeneous feature representations into a unified invariant space. To this end, we introduce a dual-branch fusion module that jointly exploits self-attention mechanisms and fourth-order moment pooling, enabling the model to capture both relational dependencies and higher-order distributional consistencies across the three feature pathways.

Self-attention and 4th-order invariant feature extraction. Following the triple-path design described above, each input image $\mathbf{X}_o$ gives rise to three complementary feature representations: a semantic-preserving embedding from the original image branch $\mathbf{e}_o$, an artifact-sensitive embedding from the patch-shuffled branch $\mathbf{e}_s$, and a content-agnostic noise-based embedding $\mathbf{e}_n$ obtained via MM-BSN-based denoising and residual extraction. All three representations are extracted from the penultimate layer of the pretrained CLIP backbone and are projected into a common feature space of dimensionality D. Consequently, $\mathbf{e}_o$, $\mathbf{e}_s$, and $\mathbf{e}_n$ are vectors in $\mathbb{R}^D$ (with $D = 1024$), enabling direct comparison and joint modeling. These embeddings are then stacked and forwarded to the attention and pooling modules for feature fusion. Specifically, the three branch-specific representations are concatenated along the branch dimension to form a unified feature tensor:

$$\mathbf{E} = \text{Concat}(\mathbf{e}_o, \mathbf{e}_s, \mathbf{e}_n) \in \mathbb{R}^{D\times 3}. \tag{5}$$

To capture higher-order invariant statistics across branches, a linear dimensionality reduction operator $\mathcal{R} : \mathbb{R}^D \to \mathbb{R}^{D/4}$ is first applied independently to each branch embedding. Note that the reduction from $D = 1024$ to $D/4 = 256$ before DCT-based pooling serves two purposes: it reduces the memory footprint of the fourth-order moment computation (which scales as $O(D^2)$ in the covariance terms) to a tractable size, and it acts as a mild bottleneck that encourages the representation to retain only the most statistically stable features before moment computation. The reduced representation is then processed by a *DCT-based fourth-order moment pooling* module, which performs a DCT-II transform along the branch dimension and computes the first four statistical moments, namely the mean, variance, skewness, and kurtosis, of the transformed coefficients:

$$\tilde{\mathbf{E}} = \mathcal{R}(\mathbf{E}), \quad \mathbf{f}_4 = \mathcal{P}_4^{\text{DCT}}(\tilde{\mathbf{E}}) \in \mathbb{R}^D. \tag{6}$$

This operation encodes frequency-aware fourth-order statistics while remaining invariant to branch ordering, thereby enhancing robustness to viewpoint and representation perturbations. In parallel, the original stacked feature tensor $\mathbf{E}$ is fed

into a Transformer-based self-attention module that operates along the branch dimension:

$$\mathbf{f}_{\text{attn}} = \mathcal{A}(\mathbf{E}) \in \mathbb{R}^D, \tag{7}$$

where attention weights are learned to model cross-branch dependencies and enforce consistency among the original, patch-shuffled, and noise-derived views. The resulting descriptors are ℓ_2-normalized, yielding $\hat{\mathbf{f}}_{\text{attn}}$ and $\hat{\mathbf{f}}_4$, and then concatenated to construct the final multi-view invariant representation:

$$\mathbf{f}_{\text{fused}} = \text{Concat}(\hat{\mathbf{f}}_{\text{attn}}, \hat{\mathbf{f}}_4). \tag{8}$$

Intuitively, the original branch embedding $\mathbf{e}_o$ preserves semantic fidelity, the patch-shuffled embedding $\mathbf{e}_s$ enforces robustness by disrupting spatially correlated shortcuts, and the noise-based embedding $\mathbf{e}_n$ introduces content-independent statistical cues that emphasize generator-specific artifacts. By jointly modeling transformer-based self-attention and DCT-based 4th-order statistics, the proposed fusion strategy achieves invariance to spatial perturbations while remaining sensitive to subtle statistical traces characteristic of synthesized imagery.

Classification head. The fused representation is passed through a linear classification layer to produce the final prediction:

$$\hat{\mathbf{y}} = \sigma(\mathbf{W}\mathbf{f}_{\text{fused}} + \mathbf{b}), \tag{9}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, $\mathbf{W}$ is the learnable weight matrix of the fully-connected (FC) layer, and $\mathbf{b}$ is the bias term. The model parameters of the classifier are optimized using the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{M}\sum_{i=1}^{M}\left[y_i \log \hat{y}_i + (1-y_i)\log\left(1-\hat{y}_i\right)\right], \tag{10}$$

where M is the batch size, $y_i \in \{0,1\}$ is the ground-truth label of the i-th sample, and $\hat{y}_i$ is the prediction. During training, the CLIP backbone remains frozen, while the self-attention module and classification head are optimized, ensuring that learning focuses on invariant relational and higher-order statistical representations. This design plugs into discrepancy-based frameworks (e.g., $\mathbf{D}^3$) with minimal changes and avoids adversarial training, yet targets the very weaknesses exposed by black-box surrogate attacks.

Why the combination improves robustness? Patch shuffle and rotations suppress scene semantics and orientation cues responsible for cross-scene failures, compelling reliance on universal, low-level regularities; MM-BSN contributes content-agnostic evidence; and fourth-order frequency pooling emphasizes heavy-tailed spectral irregularities harder to steer via surrogate-guided semantic editing. Together, these modules directly counteract the vulnerabilities highlighted in recent robustness studies while remaining training- and architecture-agnostic [72, 75].

5 Experiments and Results

5.1 Datasets, Quality, and Evaluation Metrics

Datasets for training. To assess the performance of our proposed framework, initially we considered experiments on both the UniversalFakeDetect (UFD) [44] and GenImage [76] datasets. The UFD dataset comprises approximately 720K training images, evenly split between 360K authentic samples sourced from LSUN [71], LAION-400M [53], and ImageNet [13]; and 360K synthetic images produced by a diverse set of generative models (e.g., ProGAN [30], StyleGAN [31], GLIDE [42], and DALL·E [48]). However, later we considered GenImage [76] dataset only for our training and evaluation as it is more recent, and robust in terms of quality, number of images, types of generators, and content variations. The GenImage dataset consists of approximately 2.68 million samples, divided into 1.33 million real and 1.35 million synthetic images. For each generator type, 50,000 images are reserved for testing. The synthetic subset is produced by eight different generative models, including one GAN (BigGAN [6]) and seven diffusion-based models: Stable Diffusion V1.4 [51], Stable Diffusion V1.5, GLIDE [42], VQDM [21], Wukong [20], ADM [14], and Midjourney [41].

We extended the GenImage dataset by adding StyleCLIP datasets (labeled as Sec3, non-adversarial data) from Abdullah et al. [1] as a separate class, which has a balanced set of 16,000, 2,000 and 2,000 images across both fake and real classes for training, validation, and testing, respectively. For training the modified D^3 model with our methodology, we used only 36,000 images from the extended GenImage dataset (due to time and resource constraints), randomly selected in equal number from each of its classes, and for validation we used all the given data. We maintained content variations even though selecting only a limited number of images for training (36,000 out of 2.68 million). Importantly, we did not use any adversarial data for training.

Datasets for testing. To test our trained models against adversarial attacks, initially we considered using two datasets RAID [16] and Abdullah et al. [1]. RAID provides approximately 72,000 adversarially-perturbed images derived from multiple text-to-image models, designed to systematically test how well image-generation detectors withstand deliberate evasion attempts. The dataset is constructed by attacking an ensemble of seven SOTA detectors, and includes examples that transfer across unseen detectors, thereby revealing significant vulnerabilities in current detection systems [16].

Abdullah et al. [1] introduced a dataset of *adversarial fake images* generated by leveraging large vision foundation models to perform *content-aware* manipulations rather than classic pixel-level perturbations. Given an already deceptive synthetic image, the attacker specifies a semantic attribute via a text prompt (e.g., "a smiling face"), and uses the foundation model to alter that attribute while preserving the image's

realism. Crucially, this procedure does not rely on additive adversarial noise; instead, it modifies high-level semantics to evade detectors that were trained on non-semantic corruptions (cf. RAID's noise-based adversarial perturbations). Empirical evaluation [1] shows that these content-aware, semantically guided edits significantly degrade detector performance.

Evaluating the quality of adversarial data. Abdullah et al. [1] used the Kernel Inception Distance (KID) [5] metric to measure synthetic image quality. KID measures the distribution distance between the real and fake image sets. The values are unbounded, and smaller values (closer to zero) indicate better synthetic image quality, i.e., fake images match the distribution of the real images. Abdullah et al. obtained a KID score of 0.008 from their SD dataset. For the RAID [16] dataset, we computed the KID score of 0.026, which is far larger (over 3.25x compared to [1]), indicating degraded image quality (which we also manually verified). So, we finally excluded the RAID dataset for the evaluation of our models and the baselines. We also excluded the adversarial noise (perturbations) based StatAttack [24] due to the same reason (also as their dataset is not public, we could not compute the corresponding KID score).

Evaluation metrics. We employ classification accuracy and average precision (AP) to evaluate the performance of the selected detectors, using a fixed threshold of 0.5 when calculating the accuracy score for each detector, following D^3 [70]. AP measures how well a model distinguishes between real and synthetic images, independent of any particular decision threshold. Some prior studies [36, 44, 65, 67] report the mean Average Precision (mAP) by computing the AP for each image generator separately and then taking their average. However, this simple averaging approach overlooks the considerable variability in optimal classification thresholds across different generators [18,65]. In realistic deployment scenarios, users cannot feasibly fine-tune detection thresholds for every unseen generator. We thus adopt a unified evaluation metric by computing a global AP that enforces consistent decision standards across all tested generators, following D^3 [70].

5.2 Implementation and Results

In this section we provide implementation details to incorporate our changes to the six target detectors, and results of our proposed improvement. We first discuss D^3 [70], which has been selected among the seven recent SOTA detectors for its comparatively best performance (see Sec. 3) against adversarial attacks. Then we discuss the five worst performing detectors selected from the eight detectors of Abdullah et al. [1] for improvement by implementing our techniques.

Improvement of D^3. Following its original design [70], in our experiments, we utilize the CLIP ViT-L/14 model, pretrained on the extensive WebImageText dataset, as the backbone for extracting patch-level representations. To preserve rich information, we use the penultimate layer features prior to any dimensionality reduction. With the backbone frozen, the whole network is trained on a single NVIDIA A100 80GB GPU with a batch size of 64, except for the experiment with fourth-order statistics, for which we used a H100 96GB GPU due to more GPU memory demand. During training, we employ the Adam optimizer [32] with an initial learning rate of 1×10^{-4} and no weight decay. All input images are first resized to 256×256 and subsequently randomly cropped to 224×224(native resolution of the backbone model CLIP: ViT-L/14). The data augmentation strategy follows prior work [44, 65], incorporating Gaussian blur and JPEG compression to enhance robustness. Specifically, the JPEG compression quality is uniformly sampled within the range of 30 to 100, while the Gaussian blur parameter σ is randomly selected from the interval $[0, 3]$, each applied with a probability of 0.5. During validation and testing, all images are resized to 224×224 without applying any data augmentation to ensure the evaluation is deterministic. For training the modified D^3 up to third-order pooling, it took around 24 hours per epoch in an A100 80GB GPU; for fourth-order pooling it took around 14 hours per epoch in a H100 96GB GPU.

As summarized in Table 3, our proposed modifications in D^3 consistently outperform the original model across all evaluated adversarial datasets. Specifically, we achieved a substantial improvement in classification accuracy, with an average gain of $+16.72\%$ over the baseline. Surprisingly, the most significant enhancement is observed for the *advV2* dataset, which is created by Abdullah et al. using the most effective surrogate model shown in their work (CLIP-ResNet), where our model yields a $+22.74\%$ increase in accuracy, demonstrating superior robustness against CLIP-based adversarial perturbations. Furthermore, the consistent gains in Average Precision (AP), averaging $+9.94\%$, indicate that our model maintains a more reliable ranking of fake samples even under sophisticated surrogate-model attacks. These results suggest that by integrating fourth-order frequency statistics with self-attentive cross-branch modeling, our architecture successfully mitigates adversarial attack without adversarial training that remains elusive to recent SOTA defense mechanisms.

Improvement of weak detectors. We selected the five worst-performing detectors from Abdullah et al. [1] to test if those could be improved by implementing our proposal. The five detectors are: DCT [50], DE-FAKE [54], CNN-F [65], PatchForensics [9] and Resynthesis [23]. First, we tried to reproduce the results of Abdullah et al. We could reproduce all the results except DE-FAKE [54] and Patch-Forensics [9]. For DE-FAKE [54], the maximum difference between Abdullah et al. [1] and our reproduced result is 64% for the EfficientNet dataset. For Patch-Forensics [9], the maximum difference between [1] and our reproduced result is 60.44% for the ViT dataset. We attempted to clarify this discrepancy via personal communication with the authors. At the time of this submission, the discrepancy remains unresolved. For the other three models DCT [50], CNN-F [65] and Resynthesis [23],

Attack	Dataset	Acc		AP	
		D^3	Ours	D^3	Ours
Sec5.2 AdvImages w/ Surrogate Models	CLIPResNet	0.788	0.9120	0.854	0.9466
Sec5.2 AdvImages w/ Surrogate Models	EfficientNet	0.859	0.9480	0.923	0.9673
Sec5.2 AdvImages w/ Surrogate Models	ViT	0.813	0.9285	0.902	0.9635
Sec5.2 Advtrained CLIPResNet	advV2	0.752	0.9230	0.832	0.9549
Sec5.2 Dataset for Adv. training	ForDefenses	0.778	0.9245	0.841	0.9540
Sec5.2 Dataset for Adv. training	ForSurrogate	0.796	0.9275	0.855	0.9506
Sec5.2 surrogate-StyleCLIP-dataset	StyleCLIP	0.819	0.9715	0.909	0.9795

Table 3: Performance of our model vs. the original D^3 defense, tested against adversarial datasets (Abdullah et al.)

Surrogate	DCT [50]			DE-FAKE [54]				CNN-F [65]			Patch-Forensics [9]				Resynthesis [23]		
Model	SP24	Ours	Imp.	SP24	Org.	Ours	Imp.	SP24	Ours	Imp.	SP24	Org.	Ours	Imp.	SP24	Ours	Imp.
EfficientNet	57.43	27.81	**51.6%**	75.76	26.90	3.63	**86.5%**	40.06	15.44	**61.5%**	28.37	43.43	30.89	**28.9%**	44.58	23.21	**47.9%**
ViT	53.41	16.76	**68.6%**	78.43	28.57	8.48	**70.3%**	32.23	10.50	**67.4%**	13.47	34.05	23.30	**31.6%**	37.39	16.22	**56.6%**
CLIP-ResNet	88.35	9.80	**88.9%**	80.04	40.49	11.63	**71.3%**	70.85	15.03	**78.8%**	40.96	72.76	58.60	**19.5%**	73.96	24.93	**66.3%**

Table 4: Evaluation of the defenses on adversarial fake images created using Abdullah et al. [1] (labeled as SP24). We use "Org." for our reproduction of SP24 attacks, where our results significantly differ with SP24. We report ΔR for the fake class in %, where lower is better. The "Imp." columns show the percentage reduction in recall degradation achieved by our modified model relative to the baseline.

we could reproduce the result with a deviation of under 3% from the result of Abdullah et al., which could be attributed to the version differences of machine learning libraries. Table 4 summarizes our results. We then implemented our proposal into the five selected models. For a fair comparison, we used StyleCLIP datasets (labeled as Sec3, non-adversarial) from Abdullah et al. [1] for retraining each model following the procedure in their paper. Our results are summarized in Table 4, showing clear performance improvement for each defense after implementing our proposal. We also depict our modifications compared to the original models, except for DCT and CNN-F, as their papers lack architectural figures.

In DCT [50], we replaced the raw log-DCT coefficients with multi-band DCT features, computing patch-wise fourth-order spectral statistics (mean, variance, skewness, kurtosis) over low, mid, and high frequencies, yielding a compact, distribution-aware representation for diffusion deepfake detection. We retrained the model using a single A100 80 GB GPU within 24 hours for the best result. For the modified DCT, the maximum improvement of 88.9% reduction in recall degradation is observed in the case of CLIP-ResNet, and the minimum improvement of 51.6% reduction in recall degradation is observed in the case of EfficientNet.

For DE-FAKE [54], we augmented it by integrating patch-based DCT fourth-order spectral statistics (see Figure 5, in the appendix). Images are first decomposed into patches; after that low, mid and high frequency bands are summarized using mean, variance, skewness, and kurtosis, which are concatenated with CLIP image–text embeddings for enhanced diffusion deepfake detection. We retrained the model using a single A100 80 GB GPU within 24 hours for the best result. For the modified DE-FAKE, the maximum improvement of 86.5% reduction in recall degradation is observed for EfficientNet, and the minimum improvement of 70.3% reduction in recall degradation is observed for ViT.

For CNN-F [65], we extended the original detector by incorporating patch-based DCT fourth-order spectral statistics. Images are first decomposed into fixed-size blocks, and their DCT coefficients are computed using mean, variance, skewness, and kurtosis, which are fused with CNN features to improve robustness against generative artifacts. The model was then retrained using a single A100 80 GB GPU, which took a total runtime of 1 day and 15 hours, and we got the best result at the 23rd epoch. For the modified CNN-F, the maximum improvement of 78.8% reduction in recall degradation is observed for CLIP-ResNet, and the minimum improvement of 61.5% reduction in recall degradation for EfficientNet.

For Resynthesis [23], we extended the detector by augmenting spatial residual features with global frequency-domain statistics derived from the DCT, capturing fourth-order distributional shifts introduced by diffusion and GAN-based generators (see Figure 6, in the appendix). Unlike the original model, our DCT statistics are non-learned, global descriptors extracted from reconstruction residuals and fused with Convolutional Neural Network (CNN) features. The model was then retrained for 100 epochs (following the original paper)

using a single A100 80 GB GPU, which took a total runtime of 4 days and 3 hours. For the modified Resynthesis, the maximum improvement of 66.3% reduction in recall degradation is observed for CLIP-ResNet, and the minimum improvement of 47.9% reduction in recall degradation for EfficientNet.

For Patch-Forensics [9], we extended the original architecture by integrating a DCT-based statistical branch that captures fourth-order frequency statistics (mean, variance, skewness, and kurtosis), computed over the 2D DCT coefficients of the input image (see Figure 4, in the appendix). These global frequency descriptors are then fused with Convolutional Neural Network (CNN) features, while patch shuffling is introduced for data augmentation to emphasize local artifact inconsistencies and improve generalization. The model was then retrained for 1000 epochs (following the original paper) using a single A100 80 GB GPU, which took a total runtime of 4 days and 6 hours. For the modified Patch-Forensics, the maximum improvement of 31.6% reduction in recall degradation is observed for ViT, and the minimum improvement of 19.5% reduction in recall degradation for CLIP-ResNet.

5.3 Ablation Study

We performed ablation studies only for D^3 [70], where we integrated all of our methods.

Effects of different disruptions. Unlike D^3 [70], we applied various disruptions, e.g., patch-shuffling, random rotations (e.g., 90, 180, 270), flipping of the image at the patch level. Table 5 summarizes the results for integrating our techniques in D^3 only. We studied their effects on the model performance without introducing the third branch of content-agnostic features in the original model architecture of D^3. It is evident that random flip does not improve the result, while patch-shuffling alone provides the best outcome. Although no performance gain came from adding random rotation, we still kept it for robustness against the use of directionality in images to create deepfakes, following Lorch et al. [39]. We did not conduct any ablation studies after adding the third MM-BSN branch as integrating it for content-agnostic features is highly resource-intensive and increases training time very significantly.

The D^3 integration took approximately 14 hours/epoch on an Nvidia H100-96GB, and achieving convergence required over two weeks. A full ablation isolating the MM-BSN branch would require training four model variants (with/without MM-BSN crossed with with/without fourth-order-statistics), each taking comparable time, amounting to roughly 6–8 weeks of additional H100-time. Similarly, the five weaker detectors may take approximately 2–3 months of additional A100/H100-time. **Note:** Table 4 shows that DCT fourth-order statistics are the critical driver, yielding up to 88.9% recall degradation reduction (for the DCT detector [50], without MM-BSN/triple-path learning).

Pooling order. Table 6 summarizes the effects of different pooling orders on the modified D^3 model [70] performance against adversarial attacks. The results show that fourth-order pooling consistently achieves the best performance.

Defense	Acc	AP
Patch-Shuffling	**0.819**	**0.909**
Patch-Shuffling + Random Rotations	0.812	0.902
Patch-Shuffling + Random Rotations + Flip	0.795	0.881

Table 5: Effects of different disruption strategies evaluated on the Surrogate_StyleCLIP dataset [1]

Pooling Order	Acc	AP
First-order statistics	0.8530	0.9257
Second-order statistics	0.7900	0.8800
Third-order statistics	0.7960	0.7725
Fourth-order statistics	**0.9715**	**0.9795**

Table 6: Performance of our defense evaluated on the Surrogate_StyleCLIP dataset [1] w.r.t. different pooling orders

6 Discussion

6.1 Variations in Performance Gain

Our proposal clearly improved all the SOTA detectors we experimented with (see Table 4). However, the magnitude of improvement varies according to the baseline's architectural priors. Our method provides the most significant gains (up to 88.9% reduction in ΔR) for DCT [50], as it uses mean (first-order statistics), variance and the power spectrum (second-order statistics). The attacks of Abdullah et al. [1] force the adversarial generator to prioritize specific semantic features (e.g., a smile) while satisfying the "must look real to the surrogate" constraint. The generator often produces textures that are "cleaner" or more "standardized" than a raw fake image, against which DCT [50] is easily fooled as these semantic changes shift the first- and second-order statistics into the "real" range (i.e., DCT looks for specific spectral peaks or mean/variance shifts). Similarly, CNN-F [65], which typically relies on first- or second-order spectral statistics, shows significant improvement with our updates in the network architecture. By integrating fourth-order moments, our model identifies high-order statistical anomalies in the residual domain that are not easily masked by adversarial optimization.

Conversely, the performance gain over Patch-Forensics [9] is relatively more moderate, ranging from 19.5% to 31.6%. We attribute this behavior to a partial overlap in inductive biases; Patch-Forensics is specifically designed to detect local textural inconsistencies by restricting the classifier's receptive field to small image patches. Our architecture adopts a similar strategy by extracting $K = 16$ random patches of 224×224

resolution, processed via a ResNet50 backbone [22]. This patch-wise approach inherently avoids the global semantic shortcuts that allow adversarial attacks to fool whole-image detectors. However, our model achieves superior robustness by augmenting these spatial features with a dedicated DCT-based branch. While Patch-Forensics relies on spatial RGB distributions, our model incorporates a 20,480-dimensional frequency feature vector derived from 8×8 and 16×16 block-wise DCT transforms. By calculating the mean, variance, skewness, and kurtosis across these blocks, our classifier captures high-order statistical anomalies, such as the spectral "spikes" and asymmetric residual distribution characteristics of GAN-based upsampling [19,65], which are often smoothed in the spatial domain to evade local texture classifiers. The fusion of patch-based spatial features with these multi-order frequency invariants allows our model to maintain high detection accuracy even when adversarial manipulations successfully preserve local pixel coherence.

6.2 Possible Reasons for Our Improvements

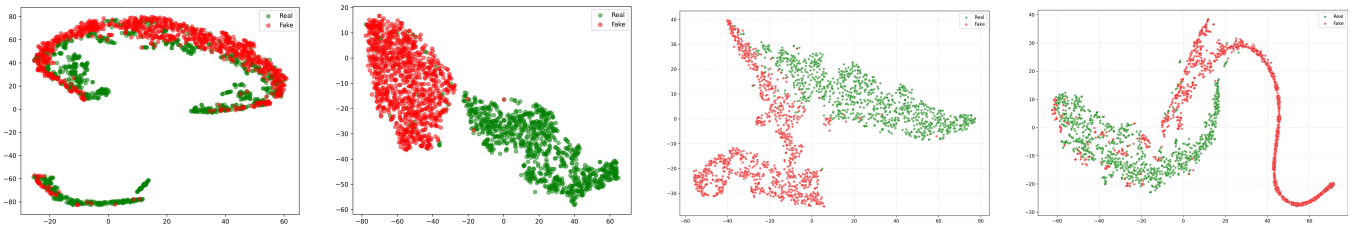
Even though Abdullah et al. [1] use StyleCLIP (a foundation model-based interface), the actual image's pixels are still being produced by the underlying GANs, i.e., StyleGAN2, which is their primary tool for crafting adversarial images. StyleGAN2, like most GANs, utilizes upsampling layers (e.g., transposed convolutions) that inherently introduce periodic correlations between pixels [65]. As demonstrated by Frank et al. [19], these structural footprints manifest as distinct spectral spikes in the frequency domain. While the adversarial attacks of Abdullah et al. [1] update the generator's weights to modify semantic content and evade CLIP-based surrogates, the structural nature of the underlying upsampling architecture remains unchanged, which inherently introduces *periodic artifacts*, in the form of periodic pixel correlations that manifest as high-frequency spikes in the residual frequency domain [19, 65], which we used to identify adversarial fake.

The attack by Abdullah et al. [1] is guided by surrogate classifiers (e.g., EfficientNet, ViT, or CLIP), which primarily focus on semantic features and low-to-mid frequency energy (first- and second-order statistics), leaving higher-order properties of residual-frequency distributions largely unconstrained. Our framework exploits this blind spot by (i) integrating semantic disruption (patch shuffling and rotation), which reduces reliance on global contextual cues and prevents the model from exploiting high-level semantics; (ii) content-agnostic residual features, which suppress semantic information and emphasize generator-induced noise patterns; and (iii) fourth-order frequency statistics, which capture leptokurtic residual distributions that persist under semantic manipulation. Importantly, the observed improvements are not attributable to increased feature dimensionality or model capacity. As shown in Table 6, performance varies non-monotonically with statistical order: second- and third-order statistics do not outperform first-order features, while fourth-order statistics yield a substantial gain. This demonstrates that the improvement arises from the specific information captured by kurtosis, rather than from simply adding more features. As established in prior frequency analyses [19, 65], periodic artifacts from GAN upsampling layers manifest as high-magnitude outliers/spikes in the residual frequency domain, creating a distribution, which is both *leptokurtic* and asymmetric [27]. Such distributions exhibit a sharper central peak and "heavy tails," indicating a higher probability of extreme values or outliers compared to a normal distribution. In our framework, these outliers correspond to the spectral spikes induced by generative upsampling layers. From a theoretical perspective (see Appendix A), the first four moments provide a minimal characterization of distributional shape, with kurtosis capturing tail heaviness and outlier prevalence. In our setting, adversarial images exhibit leptokurtic residual-frequency distributions due to structural artifacts introduced by generative models. Our model is effective because while the mean and variance may be manipulated to appear real, the skewness and kurtosis remain characteristically fake, allowing our multi-order statistical pooling to expose the generative traces that semantic-based attacks overlook (see Appendix B for differences in average kurtosis values between real and synthetic images).

6.3 Feature Space Visualization via t-SNE

With two of the selected detectors Resynthesis [23] and CNN-F [65], we conducted a more in-depth analysis of the learned feature space to evaluate the impact of our DCT-based fourth-order statistical features on latent space representation. For this analysis, we utilized t-SNE [62] to visualize the extracted features; see Figure 2. For Resynthesis (Figures 2a-2b), real and synthetic images form more separable clusters in the learned feature space in the case of DCT-based fourth-order statistics features. The t-SNE visualizations reveal that the DCT-based fourth-order statistics-enhanced model learns a more discriminative feature space, where real and synthetic images form compact and well-separated clusters. In contrast, the original baseline model (Figure 2a) exhibits significant overlaps between classes, consistent with its lower accuracy and recall. The DCT-based fourth-order statistics representation (Figure 2b) depicts a clearer separation between real and fake samples compared to the original spatial-domain features, consistent with the improved recall and accuracy.

We also compare the t-SNE embeddings of the baseline CNN-F against our proposed model. Figure 2c shows that the original model exhibits significant inter-class entanglement when processing the adversarial dataset; a substantial number of fake samples (red) is scattered within the real (green) cluster, indicating a failure to extract discriminative features under perturbation. This mixing directly translates to recall degradation, as the decision boundary cannot effectively isolate the manipulated samples. In contrast, Figure 2d reveals



(a) Original (Resynthesis) (b) 4th-order DCT (Resynthesis) (c) Original (CNN-F) (d) 4th-order DCT (CNN-F)

Figure 2: t-SNE visualization of feature embeddings across two detectors. From left to right: original and DCT-based fourth-order representations for Resynthesis and CNN-F.

that by integrating frequency-domain statistical priors, our model successfully pushes a large majority of the fake samples into a distinct, high-density cluster with significantly less real class leakage. There is relatively clear separation between the green and red distributions, suggesting that our approach successfully recaptures discriminative signals that the baseline model misses, effectively mitigating the degradation of recall in adversarial scenarios.

6.4 Other Considerations

Evaluation against other attacks. We further evaluated our enhanced-D^3 (realistic GenImage datasets) on the RAID benchmark [16], which includes a substantial proportion of abstract art images. This evaluation yielded a modest improvement, with AP increasing from 0.857 to 0.8725. It is important to note that RAID does not provide a corresponding non-adversarial training set. Evaluation on StatAttack [24] was not possible due to the unavailability of the dataset. In contrast, the attacks introduced by Abdullah et al. [1] are practically relevant: semantic/content-aware black-box evasion without additive pixel noise, specifically designed to preserve perceptual realism while defeating detectors. By comparison, noise-based perturbations (RAID/StatAttack) tend to degrade visual quality (over 3×worse), thereby posing a lesser threat.

Robustness against common manipulations. Our training pipeline incorporates JPEG compression and Gaussian blur, each applied with probability 0.5 (Sec. 5.2). The fourth-order DCT statistics are computed over frequency coefficients rather than raw pixels, adding inherent robustness to mild spatial perturbations (cropping/screenshotting). We tested our enhanced D^3 against the Surrogate-StyleCLIP-dataset with a few manipulations; the average-precision values are as follows: 0.9794 (75% JPEG-compression), 0.9470 (80% center-crop), 0.9830 (screen-shot-simulation); the baseline AP is 0.9795.

Overfitting mitigation. Our design mitigates overfitting to shuffling-induced artifacts through several mechanisms. First, patch-shuffling is applied randomly at training time with varying patch boundaries, thereby preventing the model from learning fixed discontinuity patterns. Second, because patch-shuffling is applied equally to both real and fake images, any artifact introduced by shuffling is class-independent, i.e., non-discriminative. Third, as shown in Table 5, patch shuffling alone yields an accuracy of 0.819 on the StyleCLIP adversarial dataset, identical to that of the original D^3, indicating that performance gains are not driven by spurious discriminative cues introduced by shuffling. Finally, during inference, the full unshuffled image is also processed, so the model never relies exclusively on the shuffled representation.

Adversarial attacks against the proposed features. We performed some preliminary experiments. By explicitly minimizing the error between the adversarial and authentic kurtosis across radial-frequency bands, we forced the optimizer to sculpt more organic deepfakes, resulting in a ∼30% recall degradation (against our improved DCT detector). This needs more experiments for validation.

7 Conclusion

By evaluating seven recent state-of-the-art detectors, we reconfirm the serious performance degradation reported by Abdullah et al. [1] under adversarial attacks. Our proposed framework is designed to enhance the robustness of existing state-of-the-art detectors against such known adversarial attacks. Our results demonstrate that, when trained on datasets with sufficient semantic diversity and high visual fidelity, the proposed modifications can substantially improve detector robustness without requiring adversarial training data. In particular, incorporating diverse semantic content within training datasets helps mitigate reliance on spurious semantic artifacts, while ensuring that each class includes samples generated by a wide range of high-quality generative models. Furthermore, our findings highlight the importance of emphasizing content-agnostic representations and higher-order statistical features in detector design. While our work provides a principled approach toward this direction, it also reveals several open challenges that warrant further investigation.

Acknowledgments

We sincerely thank the anonymous Usenix Security 2026 reviewers for their insightful suggestions and comments. For computing and GPU support, we are grateful to David Lie (University of Toronto), and the high-performance computing facility (*Speed*) at Concordia University. This research is funded by an NSERC-CSE Research Communities Grant. Researchers funded through the NSERC-CSE Research Communities Grants do not represent the Communications Security Establishment Canada or the Government of Canada. Any research, opinions or positions they produce as part of this initiative do not represent the official views of the Government of Canada.

Ethical Considerations

We begin our ethical analysis by identifying the primary stakeholders impacted by our work. They include: (i) victims of deepfake abuse (e.g., individuals targeted by harassment, blackmail, non-consensual intimate image abuse, or political disinformation); (ii) society at large (general public affected by misinformation and erosion of trust in digital media); (iii) developers and users of deepfake detection systems (e.g., media sharing platforms, journalists, and forensic analysts); (iv) generative AI companies and content moderation teams; (v) malicious actors who create adversarial deepfakes; and (vi) the research community working on the problem of deepfake detection.

Our research was guided by core principles for ethical computing research [33]. The ethical principles we considered are: (i) Beneficence: Our work seeks to improve the robustness of SOTA deepfake detectors against content-manipulating adversarial attacks without requiring adversarial training, thereby reducing the real-world harm caused by high-quality synthetic images used for disinformation, harassment, and blackmail; (ii) Respect for Persons: No human subjects, private data, or deceptive experiments were involved. We used only publicly available datasets; (iii) Justice: We aim for fair detection performance across diverse content types. However, we acknowledge that detection systems may still inherit biases from training data or model architectures, potentially leading to uneven performance across demographics or image styles. We did not measure or mitigate demographic bias as the GenImage and StyleCLIP datasets we used are not annotated with demographic attributes, which will require substantial additional data collection and labeling effort beyond the scope of this work. GenImage uses ImageNet object/animal classes, and StyleCLIP is a face-only dataset with no demographic labels such as skin tone, ethnicity, or identity. Future work should audit deepfake detectors using demographically stratified benchmarks, particularly given evidence in face-recognition literature that GAN artifacts can correlate with skin tone due to training data biases. We view this as an important open problem; and (iv) Respect for Law and Public Interest: This research supports societal trust in digital media and does not involve unauthorized system access or terms-of-service violations.

The harms we identified fall into two categories. Tangible harms include: (i) Dual-use risk: Our defense techniques could be analyzed by attackers to develop more sophisticated evasion methods; (ii) Over-reliance: Users may place excessive trust in automated detectors, leading to insufficient manual verification of critical content; and (iii) Resource disparity: Our approach requires significant GPU resources for training, which may limit adoption by under-resourced researchers or organizations.

Regarding rights-based harms, no direct violations of privacy, consent, or autonomy occurred, as all experiments used public datasets and no human participants.

We implemented several measures to address these risks. Our key mitigations are: (i) We exclusively used publicly available, non-sensitive datasets (GenImage and StyleCLIP), and pretrained models; (ii) We open-source our code and improved models to promote transparency, reproducibility, and equitable access; (iii) We explicitly avoided low-quality adversarial datasets (e.g., RAID) to prevent misleading evaluations; (iv) We did not perform adversarial training, reducing the risk of exposing harmful content during training; and (v) All attacks we tested on new detectors are already public [1]. Our attack experiments were performed locally, in a controlled lab setting within our own lab machines, or in our school's computing cluster. We did not experiment with live systems or services. We also do not propose any new attack methodologies for deepfake image generation, or bypassing the current SOTA detectors. Our techniques are designed to help defenders to improve robustness of their deepfake image detectors.

Despite these efforts, some challenges remain. The unmitigated/ residual risks include: (i) An ongoing adversarial arms race may continue, as attackers can adapt to our improvements; and (ii) Generalization may be limited in certain cultural or demographic contexts.

After weighing these factors, we decided to conduct and publish this research because the potential benefits (significantly stronger defenses against current adversarial deepfake attacks) substantially outweigh the associated risks. Although publishing robustness evaluations and defense strategies may indirectly help adversaries understand current detector limitations, the attacks evaluated in this work are already publicly known [1], and our work does not introduce new attack methodologies or provide operational guidance for misuse. Instead, the primary contribution is improving defensive robustness against existing threats. Prior work has documented broad societal risks of deepfake technologies, including blackmail, harassment, erosion of trust in digital media, and large-scale disinformation [4, 63]. Legal and ethical analyses further highlight concerns related to privacy violations and iden-

tity misuse, underscoring the need for effective regulatory responses [40]. These findings are consistent with reported real-world cases of deepfake misuse, involving blackmail with AI-generated intimate imagery [4], psychological harm [3], and political misinformation campaigns [63]. In light of these findings, we believe that openly advancing and evaluating defensive techniques serves the public interest and supports more trustworthy deployment of deepfake detection systems. Finally, we confirm our legal compliance. No IRB approval was required, as this work does not involve human subjects or any collection/use of personal information. All datasets are publicly available for research use. We did not test on live systems, violate terms of service, or engage in unauthorized access.

Open Science

Code and datasets are available at: https://doi.org/10.5281/zenodo.20327568. We release the six detectors integrated in our proposed framework. Detailed instructions for building and running the code are included in the `README.md` file.

The size of our training subset (36,000 images) is 13.2 GB, which exceeds the storage limitations of the repository platform. However, we provide sample images sufficient for running the code, though not for full reproducibility of our results. The complete dataset can be made available upon request (e.g., via Google Drive or other file-sharing services).

References

[1] Sifat Muhammad Abdullah, Aravind Cheruvu, Shravya Kanchi, Taejoong Chung, Peng Gao, Murtuza Jadliwala, and Bimal Viswanath. An analysis of recent advances in deepfake image detection in an evolving threat landscape. In *IEEE Symposium on Security and Privacy (SP)*, pages 91–109, 2024.

[2] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security*, pages 1–7, 2018.

[3] Sami Alanazi, Seemal Asif, Antoinette Caird-Daley, and Irene Moulitsas. Unmasking deepfakes: a multidisciplinary examination of social impacts and regulatory responses. *Human-Intelligent Systems Integration*, pages 131–153, 2025.

[4] Sami Alanazi, Seemal Asif, and Irene Moulitsas. Examining the societal impact and legislative requirements of deepfake technology: A comprehensive study. *International Journal of Social Science and Humanity*, 2024.

[5] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[7] Jonathan Brokman, Amit Giloni, Omer Hofman, Roman Vainshtein, Hisashi Kojima, and Guy Gilboa. Manifold induced biases for zero-shot and few-shot detection of generated images. In *International Conference on Learning Representations*, pages 1–26, 2025.

[8] Yu Cao, Zengqun Zhao, Ioannis Patras, and Shaogang Gong. Temporal score analysis for understanding and correcting diffusion artifacts. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2025.

[9] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? Understanding properties that generalize. In *Proc. European Conference on Computer Vision*, 2020.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. International Conference on Machine Learning*, 2020.

[11] Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. CO-SPY: Combining semantic and pixel features to detect synthetic images by AI. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 13455–13465, 2025.

[12] Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. FIRE: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2025.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[14] Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In

International Conference on Learning Representations, 2021.

[16] Hicham Eddoubi, Jonas Ricker, Federico Cocchi, Lorenzo Baraldi, Angelo Sotgiu, Maura Pintor, Marcella Cornia, Lorenzo Baraldi, Asja Fischer, Rita Cucchiara, and Battista Biggio. RAID: A dataset for testing the adversarial robustness of AI-generated image detectors. arXiv preprint (June 9, 2025), https://arxiv.org/abs/2506.03988.

[17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. International Conference on Machine Learning*, 2024.

[18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2), 2010.

[19] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proc. International Conference on Machine Learning*, pages 3247–3258, 2020.

[20] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, Chunjing Xu, and Hang Xu. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In *Advances in Neural Information Processing Systems (NeurIPS'22)*, 2022.

[21] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Vector quantized diffusion model for text-to-image synthesis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[23] Yang He, Ning Yu, Margret Keuper, and Mario Fritz. Beyond the spectrum: Detecting deepfakes via re-synthesis. In *Proc. International Joint Conference on Artificial Intelligence*, 2021.

[24] Yang Hou, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, and Jianjun Zhao. Evading deepfake detectors via adversarial statistical consistency. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 12271–12280, 2023.

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[26] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 28831–28841, 2025.

[27] Nils Hulzebosch, Sarah Ibrahimi, and Marcel Worring. Detecting CNN-generated facial images in real-world scenarios. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[28] IndiaTimes.com. Faridabad AI shocker: College student dies by suicide; friend made obscene deepfakes, blackmailed him. News article (Oct 27, 2025), https://timesofindia.indiatimes.com/city/gurgaon/blackmailed-by-ai-generated-pictures-student-dies-by-suicide-in-faridabad/articleshow/124833566.cms. Accessed: Feb 4, 2026.

[29] Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution AI-generated image detection by spectral learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 18706–18717, 2025.

[30] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[31] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[32] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations*, 2015.

[33] Tadayoshi Kohno, Yasemin Acar, and Wulf Loh. Ethical frameworks and computer security trolley problems: Foundations for conversations. In *32nd USENIX Security Symposium*, pages 5145–5162, 2023.

[34] Hao Li, Pratik Chaudhari, Hao Yang, Michael Lam, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Rethinking the hyperparameters for fine-tuning. In *International Conference on Learning Representations*, 2020.

[35] Yanhao Li, Quentin Bammey, Marina Gardella, Tina Nikoukhah, Jean-Michel Morel, Miguel Colom, and Rafael Grompone Von Gioi. Masksim: Detection of synthetic images by masked spectrum similarity analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

[36] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *Proc. European Conference on Computer Vision*, 2022.

[37] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

[38] Zhengzhe Liu, Xiaojuan Qi, and Philip Torr. Global texture enhancement for fake face detection in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 8060–8069, 2020.

[39] B. Lorch and R. Böhme. Landscape more secure than portrait? Zooming into the directionality of digital images with security implications. In *Proc. 33rd USENIX Security Symposium*, pages 6903–6920, 2024.

[40] Edvinas Meskys, Julija Kalpokiene, Paul Jurcys, and Aidas Liaudanskas. Regulating deep fakes: legal and ethical considerations. *Journal of Intellectual Property Law & Practice*, 15(1):24–31, 2020.

[41] Midjourney, Inc. Midjourney: Proprietary text-to-image generation system. https://www.midjourney.com/, 2022. Online; accessed Oct. 23, 2025.

[42] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proc. International Conference on Machine Learning*, 2022.

[43] Sophie J. Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 119(8), 2022.

[44] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[45] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

[46] Jiameng Pu, Neal Mangaokar, Bolun Wang, Chandan K. Reddy, and Bimal Viswanath. NoiseScope: Detecting deepfake images in a blind setting. In *Proc. Annual Computer Security Applications Conference*, 2020.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning*, 2021.

[48] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proc. International Conference on Machine Learning*, pages 8821–8831, 2021.

[49] R.Corvi, D.Cozzolino, G.Poggi, K.Nagano, and L.Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[50] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. In *Proc. Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2024.

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[52] A. Sarkar, H. Mai, A. Mahapatra, S. Lazebnik, D. A. Forsyth, and A. Bhattad. Shadows don't lie and lines can't bend! Generative models don't know projective geometry...for now. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[53] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. In *NeurIPS Workshop on Machine Learning for Creativity and Design*, 2021.

[54] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models. In *Proc. ACM SIGSAC Conference on Computer and Communications Security*, page 3418–3432, 2023.

[55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In

Proc. International Conference on Machine Learning, pages 2256–2265, 2015.

[56] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 569–578, 2019.

[57] C. Tan, H. Liu, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[58] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proc. International Conference on Machine Learning*, volume 97, pages 6105–6114. PMLR, 2019.

[59] TheGuardian.com. AI images of Maduro capture reap millions of views on social media. News article (Jan 6, 2026), https://www.theguardian.com/technology/2026/jan/05/maduro-venezuela-ai-images. Accessed: Feb 4, 2026.

[60] TheGuardian.com. The rise of deepfake pornography in schools. News article (Dec 2, 2025), https://www.theguardian.com/society/ng-interactive/2025/dec/02/the-rise-of-deepfake-pornography-in-schools. Accessed: Feb 4, 2026.

[61] V. N. Tran, P. Choi, H. S. Le, S. H. Lee, and K. R. Kwon. DiffCoR: Exposing AI-Generated image by using stable diffusion model based on consistent representation learning. *IEEE Open Journal of the Computer Society*, 2025.

[62] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[63] Namosha Veerasamy and Heloise Pieterse. Rising above misinformation and deepfakes. In *International Conference on Cyber Warfare and Security*, volume 17, pages 340–348, 2022.

[64] Arkhipkin Vladimir, Viacheslav Vasilev, Andrei Filatov, Igor Pavlov, Julia Agafonova, Nikolai Gerasimenko, Anna Averchenkova, Evelina Mironova, Bukashkin Anton, Konstantin Kulikov, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 3: Text-to-image synthesis for multifunctional generative framework. In *Proc. Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 475–485, 2024.

[65] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot...for now. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[66] Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7278–7287, 2023.

[67] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li. DIRE for diffusion-generated image detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 22445–22455, 2023.

[68] Zejin Wang, Jiazheng Liu, Guoqing Li, and Hua Han. Blind2Unblind: Self-supervised image denoising with visible blind spots. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2027–2036, 2022.

[69] M. Wolter, F. Blanke, R. Heese, and J.Garcke. Wavelet-packets for deepfake image analysis and detection. *Machine Learning*, pages 4295–4327, 2022.

[70] Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. D^3: Scaling up deepfake detection by learning from discrepancy. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2025.

[71] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint (Jun 4, 2016), https://arxiv.org/abs/1506.03365.

[72] Dan Zhang, Fangfang Zhou, Yuwen Jiang, and Zhengming Fu. MM-BSN: Self-supervised image denoising for real-world with multi-mask. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 4189–4198, 2023.

[73] Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4K: Ultra-high-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2025.

[74] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 2017.

[75] Chende Zheng, Chenhao Lin, Zhengyu Zhao, Hang Wang, Xu Guo, Shuai Liu, and Chao Shen. Breaking semantic artifacts for generalized AI-generated image

detection. In *Advances in Neural Information Processing Systems*, 2024.

[76] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. In *Advances in Neural Information Processing Systems*, 2024.

Appendix

A Theoretical Justification For First–Fourth Moments

The following theorem motivates first–fourth moments as the minimal set capturing location/scale plus the first two non-Gaussian shape descriptors.

Theorem 1 (Why first–fourth moments suffice, and why kurtosis matters)**.** *Let X denote a high-frequency DCT coefficient drawn from the residual-frequency representation used by our detector, and let P_r and P_f be the distributions of X for real and adversarially generated images, respectively. Define standardized moments*

$$\mu_k \;=\; \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^k\right], \quad \mu=\mathbb{E}[X],\ \sigma^2=\mathrm{Var}(X)>0.$$

Assume: (i) P_r is approximately Gaussian after standardization (sensor-noise-dominated residuals), and (ii) P_f differs from P_r primarily by sparse, high-magnitude spectral outliers (a "spike" deviation), i.e., a stylized model $Z\sim(1-\varepsilon)\mathcal{N}(0,1)+\varepsilon\mathcal{N}(0,\tau^2)$ with $\varepsilon\ll 1$ and $\tau^2\gg 1$ captures the dominant departure in the tail of standardized coefficients. Then:

(a) First–fourth moments are the right set. *Under (i), P_r is fully characterized (up to estimation error) by $\mu_1=0$ and $\mu_2=1$ and has $\mu_3=0$, $\mu_4=3$, so the first four moments capture all low-order non-Gaussian departures from P_r via μ_3 (asymmetry) and μ_4 (tail-heaviness).*

(b) Kurtosis is especially important. *Under (ii), the excess kurtosis satisfies*

$$\kappa := \mu_4-3 \;=\; \Theta(\varepsilon\tau^4), \quad \textit{whereas} \quad \mathrm{Var}(Z) \;=\; 1+\Theta(\varepsilon\tau^2).$$

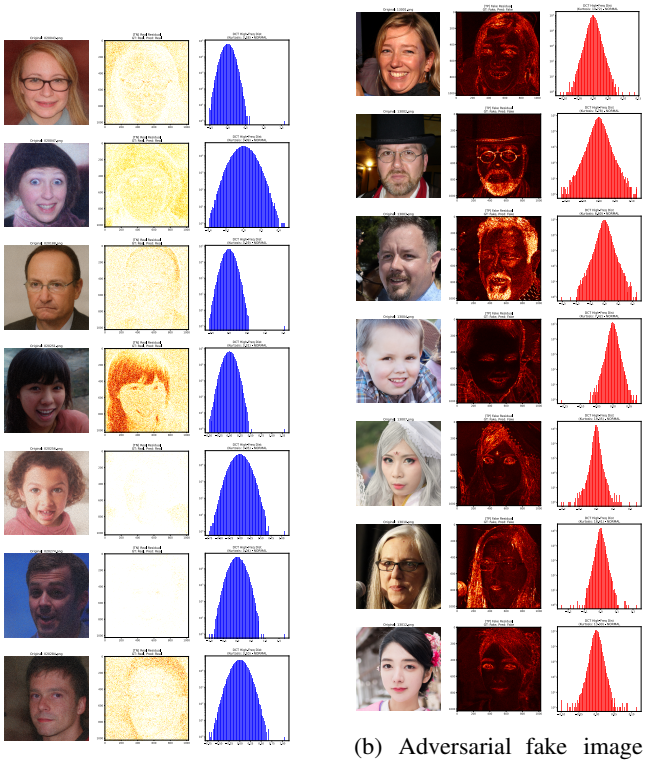
Hence, for rare but large outliers, μ_4 grows parametrically faster than any first- or second-order statistic, making kurtosis the lowest-order *moment that sharply exposes spike-induced tail deviations.*

(c) Higher moments are unnecessary (and often undesirable) at patch scale. *For n coefficient samples (e.g., from a fixed DCT block/band), the empirical moment estimator $\widehat{\mu}_k=\frac{1}{n}\sum_{i=1}^n Z_i^k$ satisfies*

$$\mathrm{Var}(\widehat{\mu}_k) \;=\; \frac{1}{n}\Big(\mathbb{E}[Z^{2k}]-(\mathbb{E}[Z^k])^2\Big),$$

so reliable estimation of μ_k requires control of the $2k$-th moment. In heavy-tailed settings, $\mathbb{E}[Z^{2k}]$ grows rapidly with k, implying that moments beyond fourth order ($k\geq 5$) become statistically unstable at practical n, while μ_4 remains the highest-order moment that is both tail-sensitive and reliably estimable.

Proof sketch. Part (a) follows from Gaussian standardized-moment identities and the fact that μ_3,μ_4 are the first moments that quantify non-Gaussian asymmetry and tail behavior. Part (b) follows by direct computation of $\mathbb{E}[Z^2]$ and $\mathbb{E}[Z^4]$ under the spike mixture, showing κ scales as $\varepsilon\tau^4$ while variance scales as $\varepsilon\tau^2$. Part (c) follows from $\mathrm{Var}(\widehat{\mu}_k)=\mathrm{Var}(Z^k)/n$ and the rapid growth of $\mathbb{E}[Z^{2k}]$ for heavy tails, which makes high-order moment estimation noisy under fixed sample budgets.



(a) Real image residuals

(b) Adversarial fake image residuals

Figure 3: Comparison of spatial and frequency characteristics of residuals across multiple samples. (a) Residuals from real images exhibit near-Gaussian behavior, consistent with low-amplitude, stochastic noise. (b) Residuals from adversarially generated images display consistently heavy-tailed (leptokurtic) distributions, indicative of structured high-frequency artifacts.

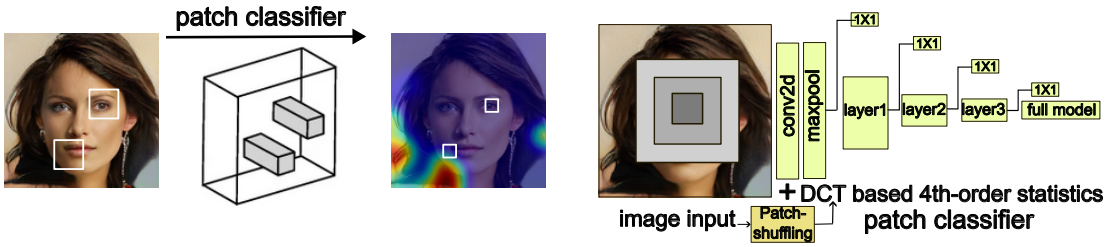


Figure 4: Modified Patch-Forensics architecture

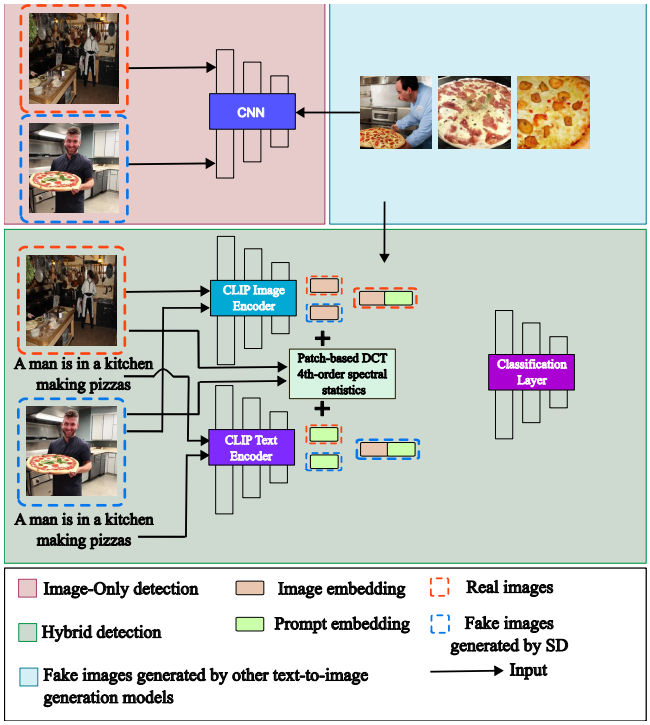


Figure 5: Modified DE-FAKE architecture

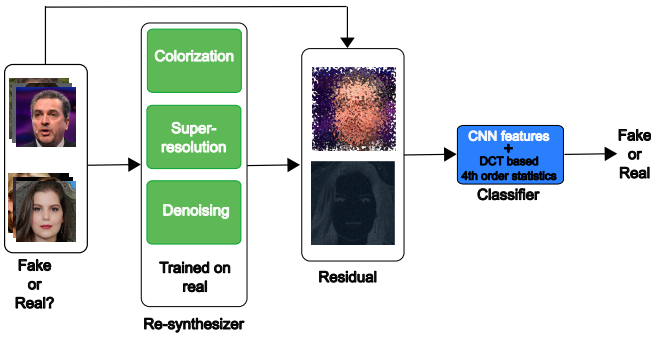


Figure 6: Modified Resynthesis architecture

B Analysis of Higher-Order Statistical Moments and Other Architectural Figures

Figure 3a presents the spatial residuals obtained from seven true negative (real) samples. Across multiple instances, these residuals exhibit low-amplitude, spatially unstructured behavior that is characteristic of stochastic noise. Consistent with this qualitative observation, the corresponding high-frequency DCT coefficient distributions yield kurtosis values of 3.18, 3.09, 3.19, 3.31, 3.05, 3.04, and 3.10, all of which closely approximate the kurtosis of a normal distribution ($\kappa \approx 3$). The average kurtosis across all 741 true negative samples (detected by our modified Resynthesis) is 3.17, indicating that the reconstruction errors in real images are dominated by random sensor noise and lack systematic high-frequency structure.

In contrast, the seven adversarial fake samples shown in Figure 3b exhibit markedly different higher-order statistical behavior. Their high-frequency DCT residuals display substantially elevated kurtosis values of 14.72, 5.78, 6.60, 7.42, 19.35, 18.41, and 13.30, reflecting strongly leptokurtic, heavy-tailed distributions. This pronounced deviation from Gaussianity is indicative of structured and sporadic high-frequency artifacts. The average kurtosis across all 925 true positive/fake samples (detected by our modified Resynthesis) reaches 22.4, underscoring a clear statistical separation from real images. Such heavy-tailed residual distributions are consistent with known artifacts introduced by generative synthesis pipelines. While these adversarial images may appear visually indistinguishable from real images to human observers, the elevated higher-order moments in the frequency domain reveal persistent generative traces. Crucially, these signatures are reliably captured by our proposed model, enabling robust discrimination between authentic and adversarial content beyond what is possible through first- and second-order statistics alone.