Comparative Analysis of GAN- and Diffusion-Generated Images for Generalizable Deepfake Detection

Sam Laborde-Balen

Master of Data Science
School of Informatics
University of Edinburgh
2025

Abstract

The rise of photorealistic images from GANs and diffusion models challenges deepfake detection, as current methods often fail on images produced by unseen generative architectures. This dissertation examines shared and distinct visual patterns of GAN- and diffusion-generated images to address these limitations and inform the design of more generalizable detectors. We compare images from six leading generative models with real photographs across ImageNet, Faces, and Bedrooms domains using the ArtiFact dataset [23], focusing on four feature families: color moments, high-frequency noise, frequency descriptors, and texture embeddings.

Across all domains and descriptors, diffusion models more closely match natural image statistics than GANs. GANs tend to boost brightness, compress color variance, concentrate high-frequency noise around edges, produce steeper-than-natural spectral slopes with depleted high-frequency energy, and yield flatter texture embeddings. Diffusion images partially restore color diversity, noise dispersion, spectral regularity, and texture complexity. This Real $>$ Diffusion $>$ GAN hierarchy appears consistently across datasets, revealing architecture-specific fingerprints that can guide robust detection strategies.

In classification experiments with Random Forests, models trained on real and diffusion images generalize more effectively to unseen GAN samples than the reverse (+8%), though the difference is not statistically significant. Frequency features, while also not statistically superior, show the most consistent performance across domains (+7–13% over other modalities). These results suggest that pairing frequency-based cues with real+diffusion training data offers a lightweight, architecture-agnostic path toward more robust deepfake detection.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics Policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified. Minor language refinement was supported by AI-assisted grammar tools, with all research design, analysis, and interpretation carried out solely by the author.

(*Sam Laborde-Balen*)

Acknowledgements

I would like to express my sincere gratitude to my supervisor Rik Sarkar, for his invaluable guidance, insightful feedback, and unwavering support throughout this research endeavor. I am also deeply grateful to my parents for their constance encouragement and support throughout this journey.

Table of Contents

Chapter 1

Introduction

The rapid progression of generative models, especially Generative Adversarial Networks (GANs) and diffusion models, has resulted in the extensive production of highly realistic synthetic images. This increase poses a significant challenge: accurately identifying and differentiating synthetic images across various generative frameworks. Recent detection models frequently exhibit constrained cross-architecture generalization, diminishing their practical efficacy and lowering confidence in digital media [5]. To establish resilient detection frameworks, it is crucial to recognize both common and architecture-specific artifacts among images produced by these leading models.

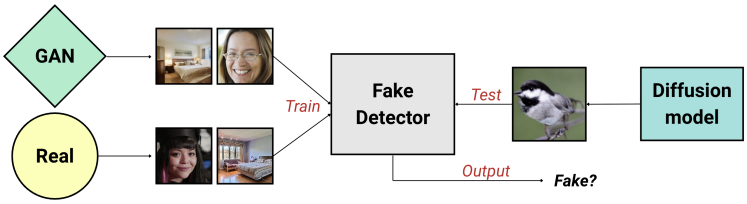


Figure 1.1: Using real and just one generative model images, can we detect images from a different type of generative model as fake?

GANs and diffusion models are fundamentally distinct in their generative methodologies (see Chapter 2). GANs produce images via adversarial training between generator and discriminator networks, commonly resulting in artifacts like checkerboard patterns or frequency anomalies [38]. In contrast, diffusion models iteratively denoise random noise, yielding smoother textures with fewer adversarial patterns [12]. Differentiating images from these paradigms is crucial for understanding generative behaviors and improving forensic methods.

1.1 Motivation

The emergence of lifelike synthetic visuals raises substantial concerns regarding media credibility and digital veracity [35, 26]. Despite progress in deepfake detection, most techniques struggle to generalize across different generative architectures [5, 3]. Previous research [21] has recognized certain similarities between GANs and diffusion models, yet the precise nature of these overlaps remains unclear. A comprehensive investigation is needed to identify universal and architecture-specific fingerprints for improved generalizable detection.

Several factors contribute to the difficulty of this problem. First, the diversity of handcrafted descriptors (e.g., color, noise, frequency, texture) makes it unclear which features best capture synthetic artifacts across models. Second, the visual fidelity of modern generative models has reached a level where artifacts are often imperceptible to the human eye. Furthermore, existing detection methods often overfit to a specific architecture or dataset, lacking the ability to generalize across domains or model families. These challenges, combined with a lack of standardized benchmarks and limited interpretability of detection models, have hindered progress in understanding and bridging the gap between GANs and diffusion models. Nonetheless, prior studies indicate that most state-of-the-art detectors rely on frequency features and are typically trained either on real and GAN-generated images or on a mixture of real, GAN, and diffusion images. Yet, the rationale for choosing one training setup over the other remains unclear [20, 33].

This dissertation examines the unique and shared visual and statistical characteristics of images generated by GANs and diffusion models, offering insights to enhance the development of more effective detection models.

1.2 Research Objective

This research aims to investigate the statistical and visual differences between images generated by GANs and diffusion models, with the goal of improving generalizable deepfake detection. Using the ArtiFact dataset, which includes a diverse set of generative architectures and visual domains, we evaluate how well various feature representations capture model-specific and universal artifacts.

The core research questions guiding this work are:

- Which statistical signatures among color distributions, noise patterns, frequency spectra, and texture features distinguish GAN- from diffusion-generated images, and which of these are universal across generative models?
- Which training configurations between training on real and GAN images to test on diffusion, or on real and diffusion images to test on GAN yield the highest generalization performance across unseen generative models?
- Which descriptor among color distributions, noise patterns, frequency features, and texture & structure metrics most effectively supports robust generalization across unseen generative architectures?

1.3 Contributions and Outcomes

The dissertation offers a systematic comparative analysis of statistical signatures distinguishing GAN- and diffusion-generated images. This analysis was conducted using a diverse set of hand-crafted features, including color histograms, local and wavelet-based noise descriptors, spectral transforms (FFT, DCT), and texture-based embeddings (LBP, GLCM, and VGG activations) to evaluate their effectiveness in supporting generalizable detection. Core outcomes and contributions include:

- **Statistical Signature Identification:** Uncovering consistent, domain-invariant differences between GAN- and diffusion-generated images across multiple statistical descriptors. GANs tend to increase brightness and suppress color diversity, yielding compressed histograms and unnatural color saturation. They also exhibit concentrated high-frequency noise near edges, spectral profiles with steeper-than-natural slopes and depleted high-frequency energy, and flattened texture embeddings, indicating reduced local structural complexity. In contrast, diffusion-generated images partially recover natural characteristics, displaying intermediate levels of color variance, dispersed noise patterns, smoother spectral decay, and more complex textures, although still distinguishable from real photographs. These consistent Real $>$ Diffusion $>$ GAN patterns suggest that each generative architecture introduces characteristic statistical distortions, which can serve as reliable fingerprints for model-specific or universal deepfake detection.

- **Generalization Strategy:** Showing that training classifiers on genuine and diffusion-generated images enhance generalization to unseen GAN samples by 8% relative to the reverse setup, though the difference is not statistically significant according to a paired t-test. This asymmetry highlights diffusion-based training as a promising path for robust, model-agnostic detection.
- **Feature Evaluation:** Demonstrating that frequency-domain descriptors, while not statistically superior according to a paired t-test, yield the most stable performance across domains and generative models, achieving a relative gain of 7 to 13% in average G-score over other feature types. This consistency makes them strong candidates for lightweight and architecture-agnostic detection systems.

Together, these findings uncover both universal and model-specific fingerprints in synthetic imagery and lay the groundwork for more resilient and adaptable detection methods capable of keeping pace with evolving generative architectures.

1.4 Outline

The structure of this project is described below:

- **Chapter 1 - Introduction** provides an overview of the project.
- **Chapter 2 – Background & Dataset** introduces GANs and diffusion models, and outlines the dataset used in this study.
- **Chapter 3 – Color Distribution Analysis** compares color statistics of real, GAN, and diffusion images, highlighting distinctive patterns.
- **Chapter 4 – Noise Pattern Analysis** analyzes frequency-domain noise and spectral features of GAN and diffusion-generated images.
- **Chapter 5 – Frequency Domain Analysis** explores spectral slopes α and high-frequency energy ratios in real and synthetic images.
- **Chapter 6 – Texture & Structure Analysis** examines texture and structural differences between real, GAN and diffusion images.
- **Chapter 7 – Generalization Evaluation** evaluates classifiers' ability to generalize across real and synthetic images from various generative architectures.
- **Chapter 8 – Conclusion** discusses findings, limitations, and future work.

Chapter 2

Background & Dataset

2.1 Introduction to Image Generative Models

The development of high-quality synthetic imagery has revolutionized computer vision, enabling deepfake generation, content production, and data augmentation. Generative models are at the core of this revolution, by learning to approximate the true image distribution, $p_{\text{data}}(x)$, they are able to create images that are frequently indistinguishable from real ones.

There have been several notable innovations along the way to the models of today. VAEs offered early latent-variable frameworks but produced blurry outputs. Autoregressive models improved likelihoods but were too slow for practical use. Then came GANs, with adversarial training yielding visually sharp images but at the cost of instability and mode collapse [9]. The new standard for image fidelity is diffusion models, which use iterative denoising to achieve incredibly realistic results but at a higher computational cost [12]. Our ability to replicate natural images improved with each model generation, but new trade-offs and artifacts were also introduced.

Understanding these design differences is essential for detection. In the sections that follow, we review the fundamentals of GANs (Section 2.2) and diffusion models (Section 2.3), and introduce the dataset used in this study (Section 2.4).

2.2 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs), developed by Goodfellow et al. [9], are a prominent category of generative models based on adversarial optimization. A GAN includes two adversarial neural networks: a Generator that generates images from noise, and a Discriminator that differentiates between authentic and artificial samples. Through iterative adversarial training, the Generator learns to produce increasingly realistic outputs, while the Discriminator sharpens its ability to detect synthetic artifacts.

This adversarial dynamic leads to sharp and high-resolution images, but GANs are also prone to instability, mode collapse, and architecture-specific artifacts such as checkerboard patterns, unnatural texture repetitions, or spectral gaps, especially due to upsampling operations and convolutional biases. These traits vary across GAN architectures and provide useful cues for detection methods based on color, frequency, or noise irregularities.

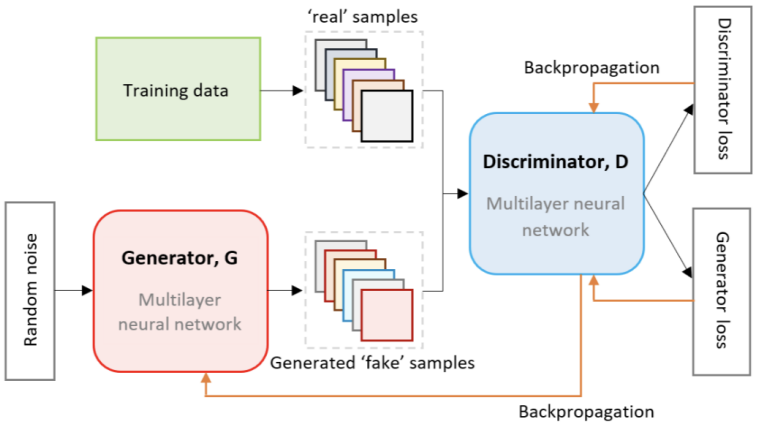


Figure 2.1: In a GAN, a Generator transforms a noise vector into synthetic data, while a Discriminator distinguishes those samples from real images; the Discriminator's classification loss is then used adversarially to iteratively improve both networks, sharpening the Discriminator's detection and boosting the Generator's realism [17].

2.3 Diffusion models

Diffusion models have recently emerged as a state-of-the-art approach to generative image modeling, with superior sample fidelity compared to GANs in many domains. Initially proposed by Sohl-Dickstein et al. [28] and later improved by Ho, Jain, and

Abbeel [12], these models generate images by reversing a gradual noising process through learned denoising steps.

In contrast to GANs, which transform noise into data in one step, diffusion models perform a series of iterative steps, progressively enhancing the image's realism. This results in improved training stability and increased sample diversity. However, the iterative nature of sampling introduces subtle temporal and structural artifacts, such as overly smooth regions or misaligned high-frequency details.

These differences in generative process, noise behavior, and spectral signatures are critical for deepfake detection. As later analyses (Chapters 3–7) will show, diffusion images often lie between GAN and real samples in feature space, making them a valuable training source for generalization.

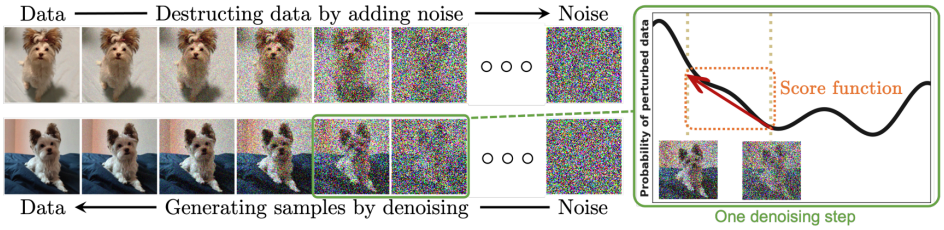


Figure 2.2: Illustration of a diffusion model [34]. The forward process gradually adds noise to real data. The reverse process learns to denoise and reconstruct samples step-by-step, guided by a score function that estimates the gradient of the data distribution.

2.4 Dataset

This study uses a portion of the ArtiFact dataset [23], which is ideal for assessing image forensic methodologies across various generative architectures. The dataset comprises synthetic images generated by six models: three based on Generative Adversarial Networks (StyleGAN2 [16], BigGAN [4], Gansformer [13]) and three based on diffusion processes (DDPM [12], Palette [25], and VQ-Diffusion [10]). Each synthetic subset is matched with authentic photos from the same domain, facilitating consistent and controlled comparisons. The dataset is organized into three visual domains:

- **ImageNet domain:** consisting of 10,000 photos generated by BigGAN, 10,000 by VQ-Diffusion, and 10,000 authentic samples, all sourced from the 1,000-

class ImageNet taxonomy. Its comprehensive semantic scope and class-balanced architecture render it a distinctly varied benchmark for assessing detection models over a broad spectrum of content.

- **Face domain:** This subset comprises 6,000 real facial images, supplemented by 6,000 synthetic images generated using StyleGAN2 (GAN) and 6,000 using Palette (diffusion). Both generative models are acknowledged for their ability to generate highly realistic faces, making this field particularly pertinent for forensic applications. The prevalence of facial alteration in identity fraud and impersonation crimes renders this category crucial for evaluating the effectiveness of deepfake detection in sensitive contexts.

- **Bedroom domain:** including 5,000 real bedroom scenes, alongside 5,000 GAN-generated samples from Gansformer and 768 diffusion-generated images from DDPM. It exhibits a marked imbalance between GAN and diffusion images, which could affect classifier training in Chapter 7. However, this limitation is less critical for feature-based analyses, where statistical descriptors are aggregated across samples and do not depend on classifier balance.

Table 2.1 summarizes the number of images per model and domain. Despite the imbalance in the bedroom subset, the inclusion of multiple GAN and diffusion models per domain, especially in ImageNet, makes the data set well suited for our comparative analysis. It enables systematic evaluation of both model-specific artifacts and generalization across generative paradigms, which is central to this study's goals.

Table 2.1: Dataset summary grouped by domain.

Domain	**Model (Type)**	**Synthetic**	**Real**
Bedroom	Gansformer (GAN)	5,000	5,000
	DDPM (Diffusion)	768	
Faces	StyleGAN2 (GAN)	6,000	6,000
	Palette (Diffusion)	6,000	
ImageNet	BigGAN (GAN)	10,000	10,000
	VQ-Diffusion (Diffusion)	10,000	

Chapter 3

Color Distribution Analysis

Color composition is a key property that distinguishes real photographs from synthetic images. In natural scenes, pixel color distributions, summarized by the first four moments (mean, variance, skewness, kurtosis) of each channel, reflect both physical scene factors (illumination, materials, sensor response) and natural lighting statistics [37]. Vision and forensics research shows real images have characteristic color-moment signatures: means and variances within predictable ranges, higher moments capturing saturation extremes, and broad, multi-modal chrominance distributions [36]. In contrast, GANs and diffusion models learn statistical color patterns from data but do not model the underlying color formation process, often yielding histograms with sharper peaks, reduced variance, and less hue diversity, especially out-of-domain [32, 1, 19].

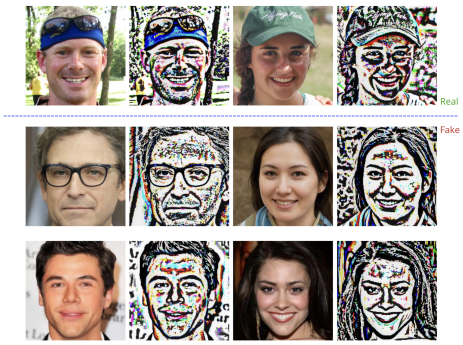


Figure 3.1: Comparison between the original image and the color difference map. From top to bottom: Real image, GAN (StyleGAN2), and Diffusion (Palette) generated results.

In this chapter, we examine how GAN and diffusion generated images deviate from real photographs in color-moment statistics across three color spaces: RGB, HSV, and YCbCr. We focus on the first four statistical moments (mean, variance, skew-

ness, kurtosis), quantifying systematic shifts in brightness, saturation, and channel variance that may serve as model-agnostic cues for image authenticity. We hypothesize that real images show broader, more diverse color-moment distributions, with GAN outputs most compressed and diffusion images partially restoring natural diversity.

To test this, we extract per-image moment features in RGB, HSV, and YCbCr, and conduct both quantitative and qualitative analyses. This includes statistical comparisons, histogram inspection, and PCA visualization to identify distinctive patterns. As will be shown, both GAN and diffusion models leave different yet overlapping fingerprints. For example, GANs tend to boost red channel means while reducing variance, whereas diffusion models apply subtler, domain-specific shifts and often increase saturation. These consistent deviations validate the discriminative power of color-moment features as robust, architecture-agnostic indicators of synthetic imagery.

3.1 Quantitative Results

3.1.1 HSV Color Metrics

We analyze three key HSV statistics: brightness (V), saturation (S), and hue diversity (H), which capture complementary aspects of color fidelity in generated imagery.

Table 3.1: HSV channel moments for hue, saturation and brightness across Real, GAN and Diffusion images. Highlighted entries indicate key deviations from real images.

Dataset	Model	μ_H	σ_H^2	skew_H	kurt_H	μ_S	σ_S^2	μ_V	σ_V^2
Bedroom	Real	42.3	2392	1.46	0.80	77.2	4593	150.6	4193
	Gansformer	33.0	1352	2.22	4.18	67.8	3301	**155.3**	3100
	Diffusion	**58.5**	2443	0.64	-0.83	**71.4**	4075	**152.4**	4177
Face	Real	**41.0**	2800	1.43	0.62	93.2	3513	138.9	5016
	StyleGAN2	27.6	1854	2.38	4.47	94.0	2570	**144.1**	4392
	Palette	29.8	2155	2.16	3.30	**99.6**	3074	**143.0**	5436
ImageNet	Real	39.7	2212	1.22	0.44	70.4	6232	97.2	7066
	BigGAN	49.5	2228	1.06	-0.03	77.7	4189	**130.6**	4040
	VQ-Diffusion	**57.8**	2896	0.78	-0.77	**90.1**	5372	**135.9**	4803

Table 3.1 summarizes mean HSV metrics across real, GAN, and diffusion-generated images. In terms of brightness (V), GAN and diffusion images show similarly elevated means compared to real images; for instance, GAN and diffusion brightness exceed real imagery by about 2–40 units, from 97.2 (real) to 130.6 (GAN) and 135.9 (diffusion) in

the ImageNet domain. Regarding saturation (S), diffusion images consistently show higher means, such as an increase from 70.4 (real) to 90.1 (diffusion) in ImageNet, whereas real and GAN images remain generally comparable across datasets. Lastly, hue (H) shows dataset-dependent trends: diffusion images have higher means in Bedroom (58.5 vs. 42.3 real, 33.0 GAN) and ImageNet (57.8 vs. 39.7 real, 49.5 GAN), whereas in the Face domain, real images hold the highest hue values (41.0 vs. 27.6 GAN and 29.8 diffusion), indicating no uniform pattern for hue diversity across domains.

3.1.2 Tonal Range and Contrast (RGB)

We assess tonal range by computing the mean and variance of R, G, and B channels in each image (Table 3.2). GAN outputs show consistently lower channel variances, indicating reduced tonal range and flatter contrast versus real photographs. Diffusion models recover much of the real-image variance, typically within 5–15% of real.

Table 3.2: RGB channel means and variances for Real, GAN and Diffusion images.

Dataset	Model	μ_R	σ_R^2	μ_G	σ_G^2	μ_B	σ_B^2
Bedroom	Real	146.5	4268	130.7	4580	116.5	5039
	GAN	**153.3**	**3074**	**139.4**	**3513**	**122.6**	**3968**
	Diffusion	139.9	4498	135.5	4450	128.5	4915
Face	Real	133.5	5134	109.2	4248	97.8	4296
	GAN	**141.6**	**4474**	**111.6**	**3411**	97.4	**3327**
	Diffusion	141.0	5511	108.6	4120	**93.8**	3886
ImageNet	Real	89.5	6635	85.6	6209	75.8	5937
	GAN	**120.9**	**4065**	**115.2**	**3821**	**101.5**	**4434**
	Diffusion	**124.4**	4970	115.1	4653	101.9	5168

Across all domains, synthetic images exhibit upward shifts in RGB means compared to real photographs. In ImageNet, the R-channel mean rises from 89.5 (real) to 120.9 (GAN) and 124.4 (diffusion), with G and B up by roughly +30 and +26 units. In bedroom scenes, both GAN (G: +8.7, B: +6.1) and diffusion ($G: +4.8$, $B: +12.0$) demonstrate increases in G and B. Yet, diffusion reduces R from 146.5 to 139.9. Facial portraits exhibit minor alterations: both GAN and diffusion increase R by around 8 units, although G and B remain near authentic values ($G: 109.2 \Rightarrow 111.6/108.6$; $B: 97.8 \Rightarrow 97.4/93.8$). Generally, the red channel exhibits the most significant shifts, reaching +34.9 in ImageNet diffusion and rendering it a potent synthetic indicator. Furthermore, GANs generate a consistent enhancement across channels, while diffusion introduces subtle, domain-specific modifications, occasionally attenuating red or

amplifying blue, exposing complimentary color biases.

In summary, synthetic photos systematically deviate from genuine photographs in tone range and color balance. GANs produce diminished contrast with lower channel variances, while diffusion models largely reinstate natural variance levels. Both approaches elevate RGB means, particularly in the red channel; however, GANs apply a more uniform enhancement, while diffusion models make more domain-specific adjustments. The combined mean–variance and channel-bias patterns provide reliable, interpretable indicators for distinguishing authentic from manufactured images.

3.1.3 Chrominance Signatures (YCbCr)

We next analyze chrominance artifacts in the YCbCr color space by comparing mean and variance of the Y, Cr, and Cb channels for each dataset and generative model.

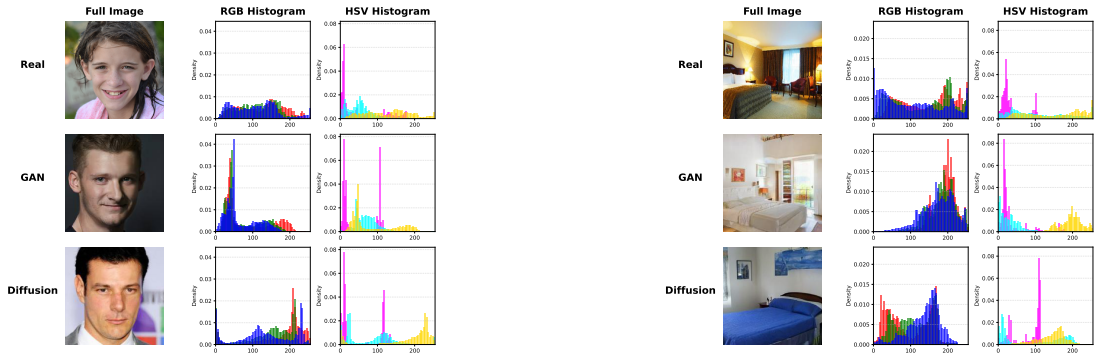
Table 3.3: YCbCr channel means and variances for Real, GAN and Diffusion images.

Dataset	Model	μ_Y	σ^2_Y	μ_{Cr}	σ^2_{Cr}	μ_{Cb}	σ^2_{Cb}
Bedroom	Real	133.8	4315	137.1	203	118.2	206
	GAN	141.6	3318	136.3	102	117.2	116
	Diffusion	136.0	4235	130.8	260	123.8	260
Face	Real	115.2	4221	141.1	291	118.2	207
	GAN	119.0	3527	144.2	193	115.8	125
	Diffusion	116.6	4296	145.4	218	115.1	142
ImageNet	Real	85.7	6068	130.7	200	122.4	226
	GAN	115.3	3673	131.9	248	120.2	272
	Diffusion	116.4	4367	133.7	382	119.8	384

There is no clear pattern across our subsets. When comparing GAN-generated images to real photos, the Bedroom and Face datasets have less variation in chrominance (Cr, Cb), while the ImageNet datasets have more variation. Diffusion models also exhibit results that are not consistent. They often show more variation than real images (in Bedroom and ImageNet), and sometimes they show less (as observed in Face). Similarly, mean Cr and Cb change inconsistently: GANs can both slightly increase or decrease chroma peaks depending on the domain, and diffusion outputs follow no uniform trend either. Consequently, neither chrominance means nor variances in the YCbCr space serve as reliable, dataset-agnostic markers to distinguish GAN, diffusion, and real images.

3.2 Qualitative Results

To complement the quantitative statistics, we visualize two aspects of color fidelity: representative histograms for bedroom scenes, and a joint PCA embedding across datasets.



(a) Face portraits (b) Bedroom scenes

Figure 3.2: Bedroom scenes and face portraits with RGB and HSV histograms for Real, GAN, and Diffusion images. GANs show brighter, less diverse color distributions with pronounced hue peaks; Diffusion shows better match the broader profiles of real images.

Figure 3.2 compares RGB and HSV histograms for typical Real, GAN, and Diffusion images. GAN outputs exhibit more pronounced, smaller RGB peaks shifted towards higher intensities, indicating enhanced brightness and diminished tone diversity. Their HSV histograms show distinct, narrow hue peaks and constricted mid-saturation areas, indicating restricted color diversity. Conversely, despite persistent peaks diffusion images display wider, more organic histogram profiles that more closely match those of authentic photographs in both RGB and HSV color spaces.

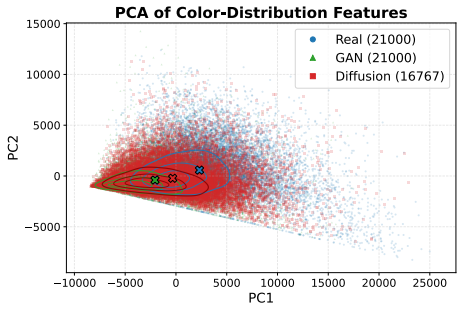


Figure 3.3: PCA embedding of 36-dimensional color features for Real, GAN, and Diffusion images. Density contours (50%, 75%, 90%) and centroids (“×”) show real images span the largest area, diffusion is intermediate, and GANs are tightly clustered.

We processed each image's 36-dimensional color-moment signature (RGB, HSV, YCbCr) into a two-dimensional PCA plot to highlight color characteristic differences across our three image types (Figure 3.3). Real images disperse throughout the plot, defining a wide spectrum of hues and contrasts. In contrast, GAN-generated photos converge into a dense cluster, revealing their limited, peak-dominant color profiles. Diffusion samples occupy an intermediate position between both extremes. They are neither as distributed as real images nor as restricted as GANs, demonstrating their partial restoration of genuine color diversity.

These figures support the quantitative results: GANs produce sharp, bright RGB peaks and narrow HSV spikes, while diffusion images more closely mirror the broader, multi-modal shapes of real photographs. In PCA space (Figure 3.3), real images spread widely, GANs cluster tightly, and diffusion samples lie in between, visually confirming that diffusion partially restores natural color diversity lost by GANs.

3.3 Discussion & Summary

Both GAN and diffusion models leave different yet overlapping color-moment fingerprints. In HSV, both increase brightness by up to 40 units, but only diffusion consistently raises saturation. In RGB, GANs uniformly boost channel means, especially red, while reducing variance, whereas diffusion restores variance and applies subtler, domain-specific shifts. YCbCr means and variances are too inconsistent to rely on. Qualitative histograms and PCA plots mirror and confirm these findings. These consistent mean–variance–shape deviations validate our hypothesis and establish color-moment features as robust, model-agnostic cues for detecting synthetic images.

This approach has some limitations; it relies on global color statistics, potentially missing localized or context-dependent artifacts, and does not capture spatial color correlations or textures. Our results also depend on the calibration of input images, and advances in generative modeling may eventually narrow these color gaps.

While color cues are powerful forensic signals, they address only one dimension of authenticity. The next section turns to high-frequency noise analysis, using local and wavelet-based descriptors, to evaluate whether the lack of sensor-like noise in synthetic images offers further, model-agnostic evidence for deepfake detection.

Chapter 4

Noise Pattern Analysis

The fine-grained fluctuations in real photographs, known as *high-frequency noise*, arise from the stochastic nature and physical constraints of image acquisition. In digital sensors, this noise stems from photon shot noise, thermal and electronic noise, and downstream amplification, producing unpredictable grain patterns throughout an image [15]. Such noise typically exhibits known statistical properties: it is largely uncorrelated, spatially stationary at local scales, and spans multiple spatial scales [18]. Generative models such as GANs and diffusion models synthesize images without explicitly modeling these sensor processes, often failing to reproduce authentic high-frequency noise patterns. Prior multimedia forensics research has shown that sensor noise is a reliable indicator for source identification and deepfake detection [31].

In this chapter, we examine how synthetic images diverge from real ones in their noise characteristics across Bedroom, Face, and ImageNet datasets, identifying both unique and shared artifacts in GAN and diffusion outputs that could serve as model-agnostic cues for authenticity. We hypothesize that authentic images have measurable, dataset-independent noise characteristics distinct from synthetic images, and that diffusion outputs will exhibit noise levels intermediate between GAN and real images.

To test this hypothesis, we compute two complementary noise metrics: local noise via sliding-window standard deviation and wavelet noise energy from single-level Haar decomposition. We then apply statistical tests, bar plots, noise-map visualizations, and PCA embeddings to assess differences and separability across categories. As will be shown, real photographs consistently exhibit the highest noise levels, GANs the lowest, and diffusion outputs fall in between, a hierarchy (Real $>$ Diffusion $>$ GAN)

that persists across datasets and metrics, with narrower gaps in smoother domains such as faces. These consistent, statistically significant patterns confirm local and wavelet-based noise features as lightweight, interpretable, and model-agnostic indicators for distinguishing real from synthetic imagery.

4.1 Noise-Pattern Estimation

4.1.1 Local Noise Estimation

To capture pixel-level fluctuations, we compute the local noise at each position (i, j) as the standard deviation over a $K \times K$ window centered at that pixel [39]. With $K = 7$ ($r = 3$) and reflective padding to preserve full-image coverage, we define the local mean and standard deviation as:

$$\mu_{i,j} = \frac{1}{K^2} \sum_{u=-r}^{r} \sum_{v=-r}^{r} I(i+u, j+v), \tag{4.1a}$$

$$\sigma_{i,j} = \sqrt{\frac{1}{K^2} \sum_{u=-r}^{r} \sum_{v=-r}^{r} \big(I(i+u, j+v) - \mu_{i,j}\big)^2}. \tag{4.1b}$$

4.1.2 Wavelet-Based Local Noise Energy

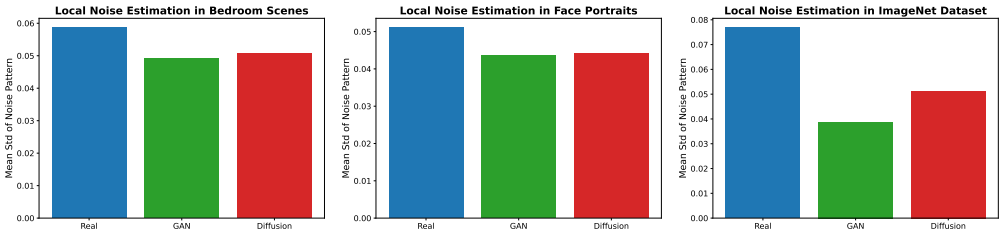
To characterize localized high-frequency signal variations, we apply a single-level 2D discrete wavelet transform using the Haar (Daubechies-1) basis, resulting in detail coefficient maps $\{c_H, c_V, c_D\}$ [29]. These capture abrupt spatial changes over small neighborhoods, rather than the global frequency distribution. The local noise energy at each position (i, j) is defined as:

$$E_{i,j} = \sqrt{c_H(i,j)^2 + c_V(i,j)^2 + c_D(i,j)^2}. \tag{4.2}$$

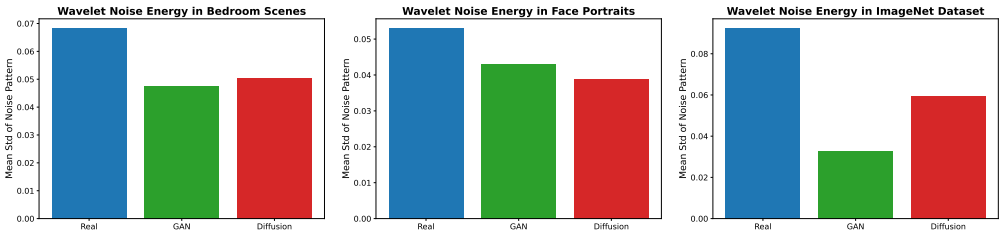
4.2 Quantitative Results

Figure 4.1 shows that, regardless of whether we measure local standard deviation or wavelet-based energy, real images consistently exhibit the highest levels of high-frequency noise, GANs the lowest, and diffusion models fall in between. In the ImageNet domain, the local-noise mean rises from about 0.077 in real images to just 0.039 in GANs and 0.051 in diffusion outputs. The wavelet-noise energies follow a similar pattern, measuring 0.093 for real, 0.033 for GANs, and 0.060 for diffusion.

Bedroom scenes and face portraits also exhibit the Real > Diffusion > GAN trend, although the size of the gaps differs. In ImageNet, for instance, the abundance of fine textures may amplify synthetic images' grain deficiencies, producing the widest separation. In contrast, smooth skin regions in face portraits may partially conceal the lack of authentic noise, resulting in a much smaller gap between real and generated samples.



(a) Bedroom (Local) (b) Face (Local) (c) ImageNet (Local)



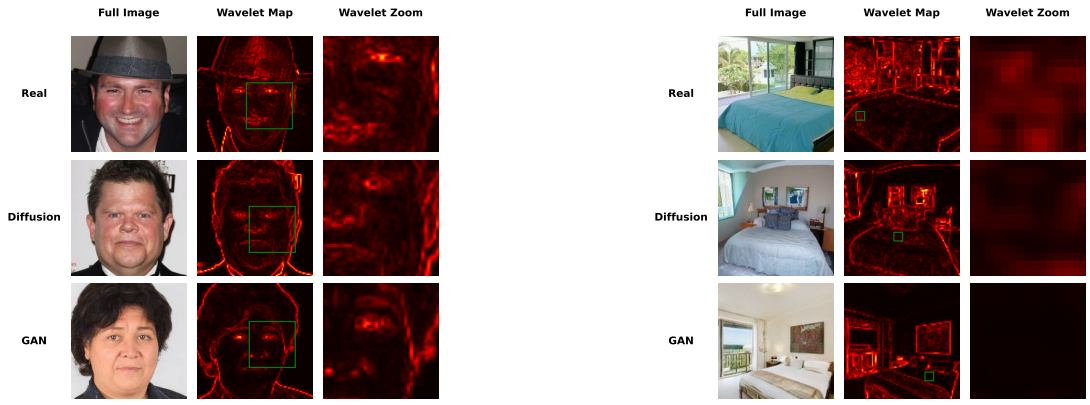
(d) Bedroom (Wavelet) (e) Face (Wavelet) (f) ImageNet (Wavelet)

Figure 4.1: Mean ± SE of local noise (7×7 std) and wavelet noise energy for Real, Diffusion, and GAN images across the Bedroom, Face, and ImageNet datasets. Real images exhibit the highest noise levels, Diffusion are intermediate, and GAN the lowest.

These findings reveal two complementary model-agnostic cues. First, the uniquely low noise imprint of GANs serves as a clear "synthetic fingerprint", reflecting their failure to capture the stochastic physics of real sensors. Second, diffusion models, by design, recover some of this randomness, yielding intermediate noise levels that nevertheless fall short of authentic photographs. These differences are statistically confirmed by Welch's t-tests across modality pairs ($p < 10^{-3}$) and datasets. Crucially, both local- and wavelet-based metrics produce the same hierarchy of noise magnitudes across all three domains, demonstrating their reliability as simple yet powerful features for distinguishing real from synthetic imagery.

4.3 Qualitative Results

Figure 4.2 presents three bedroom scenes and three face portraits, each a real photograph, a diffusion model output, and a GAN sample, paired with their wavelet-noise energy maps. In the GAN images, noise is tightly confined to prominent edges and structural boundaries, leaving smooth regions mostly artifact-free. In contrast, diffusion samples exhibit both edge-highlighting and a fine 'speckle' of noise across flatter surfaces. Real photographs display the richest and most uniform noise distribution, with sensor-like grain appearing consistently in both textured and smooth areas. These visual patterns mirror our quantitative ranking of noise magnitude (Real $>$ Diffusion $>$ GAN), confirming that diffusion models recover some, but not all, of the natural high-frequency content lost by adversarial generators.



(a) Face portraits

(b) Bedroom scenes

Figure 4.2: Wavelet-noise energy maps for three bedroom scenes and three face portraits. Real images exhibit the most pervasive noise, diffusion images are intermediate, and GAN outputs concentrate noise along edges, following Real $>$ Diffusion $>$ GAN.

We assess global separability by embedding noise maps via PCA (Figure 4.3). Real images spread into a broad cloud, reflecting rich, varied noise. In contrast, GAN samples form a tight cluster near the origin, indicating uniformly low noise. Diffusion outputs lie between these extremes, overlapping both distributions.

These qualitative observations mirror our quantitative trends: real photographs show pervasive, multi-scale noise, GAN outputs concentrate minimal noise along edges, and diffusion samples lie squarely in between with moderate speckled patterns. Together, the noise-energy maps and PCA embedding visually confirm the consistent Real $>$ Diffusion $>$ GAN hierarchy found in our statistical analyses.

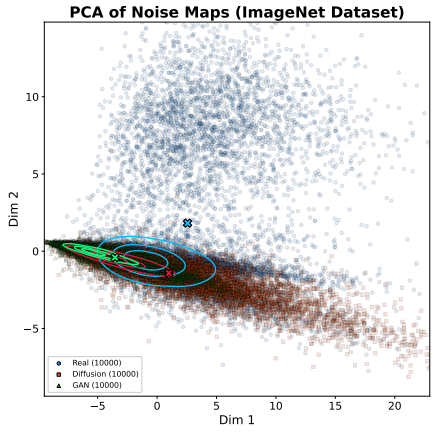


Figure 4.3: PCA embedding of flattened wavelet-noise maps for Real, GAN, and Diffusion images. Density contours (50%, 75%, 90%) and centroids ("×") show real images span the largest area, diffusion is intermediate, and GANs are tightly clustered, reflecting the separability of noise patterns across categories.

4.4 Discussion & Summary

Our results show that authentic photos exhibit richer and more spatially pervasive high-frequency noise than synthetic images, making it a reliable authenticity cue. GAN outputs lack the distinctive, multi-scale grain of real photographs, concentrating residual noise along sharp edges, a signature of adversarial upsampling. Diffusion models partly recover natural randomness with mid-level noise that bridges the gap between GANs and real images, yet still fall short of genuine sensor complexity. This persistent hierarchy (Real $>$ Diffusion $>$ GAN), observed across scenes and metrics, highlights noise-pattern features as an effective, lightweight, and model-agnostic tool for distinguishing real from synthetic imagery.

A limitation of this study is its focus on grayscale, single-scale noise; future work should extend to multi-scale, color, and sensor-specific analyses, and track models as they evolve to better imitate physical sensor noise.

Building on these findings, the next chapter moves from *localized noise patterns*, captured here via sliding-window and wavelet-based measures, to *global frequency characteristics* in the Fourier domain. This shift enables the analysis of complementary cues, such as spectral slopes and high-frequency energy ratios, that reveal broader structural regularities and generative artifacts beyond local noise.

Chapter 5

Frequency Domain Analysis

Natural images display distinct spatial frequency statistics, characterized by a heavy-tailed power spectrum where Fourier magnitudes diminish approximately as $1/f^{\alpha}$, typically $\alpha \approx 2$ [7, 24, 30]. This reflects the multi-scale, fractal nature of natural scenes and the mechanics of image formation. Synthetic images generated by GANs and diffusion algorithms acquire frequency statistics from training data but often diverge from these natural patterns. GANs can produce frequency anomalies such as excessive high-frequency energy, periodic patterns, or spectrum gaps due to convolutional design and upsampling [6, 8], while diffusion models typically yield more natural spectra yet may still struggle to reproduce the precise $1/f^{\alpha}$ decay.

This chapter compares the global frequency profiles of real, GAN, and diffusion images across Bedroom, Face, and ImageNet datasets to identify unique artifacts and shared signatures as model-agnostic cues for authenticity. We hypothesize that real images follow a smooth power-law decay, GANs display altered spectral slopes and reduced high-frequency content, and diffusion models occupy an intermediate position, aligning more closely with real data in some domains but less consistently in others.

We test this using two FFT-based metrics described in Section 5.1. These are applied through statistical analyses and visual inspections to evaluate differences across categories. Results show that real photographs generally have the lowest slopes and highest HF ratios, GANs the steepest slopes and lowest HF ratios, and diffusion models in between. This Real $>$ Diffusion $>$ GAN pattern holds in Bedroom and ImageNet, while faces deviate, with diffusion sometimes surpassing GANs in both metrics. These domain-dependent trends confirm the metrics as compact, interpretable, and statistically

significant signatures for distinguishing real from synthetic images.

5.1 Frequency-Domain Estimation

We compute the centered 2D discrete Fourier transform (DFT) of each grayscale image $I\colon [1,H]\times[1,W]\to[0,1]$ via the fast Fourier transform:

$$F = \text{fftshift}(\text{fft2}(I)), \tag{5.1a}$$

$$M = \log\bigl(1+|F|\bigr), \tag{5.1b}$$

where M is the log-magnitude spectrum. The radial frequency profile $\rho(f)$ is derived by averaging M over concentric rings around the frequency origin. We then fit a power-law decay model in log-log scale:

$$\log\rho(f) \sim -\alpha\log f, \tag{5.2}$$

where α is the spectral slope describing the characteristic decay of natural images [30]. Additionally, the high-frequency (HF) energy ratio is defined as the proportion of spectral energy in the upper third of frequencies. Each image is summarized by a spectral slope (α) and a HF energy ratio for statistical analysis.

5.2 Quantitative Results

We analyze frequency-domain behavior of real and synthetic images using two complementary metrics: the log-log spectral slope (α), quantifying decay rate in the azimuthally averaged power spectrum, and the high-frequency (HF) energy ratio, measuring the proportion of spectral energy beyond the upper third of radial frequencies.

Figure 5.1 demonstrates that synthetic images uniformly display elevated spectral slopes compared to authentic images across all domains, indicating a more gradual loss of frequency energy. This suggests both GANs and diffusion models overrepresent mid-to-high frequencies in the log-log spectral domain. The difference is most pronounced for GANs on ImageNet, where the mean GAN slope is ($\bar{\alpha} = 0.895$) versus ($\bar{\alpha} = 0.501$) for real samples. Diffusion models generally fall between real and GAN samples in terms of slope, most clearly in the Bedroom and ImageNet domains. Nonetheless, this pattern is not applicable to facial pictures, as diffusion samples demonstrate the most pronounced slope overall ($\bar{\alpha} = 0.786$), exceeding both actual and GAN distributions.

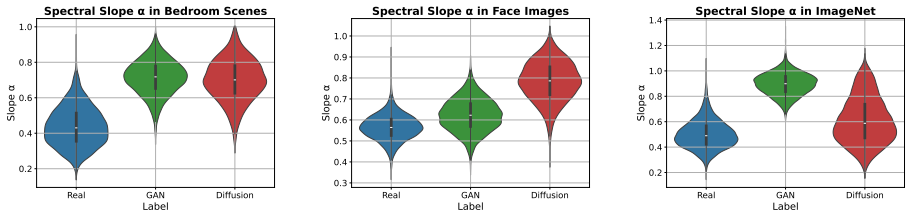


Figure 5.1: Violin plots of log-log spectral slope α for Real, GAN, and Diffusion images across Bedroom, Face, and ImageNet datasets. GAN images show higher α than real ones, indicating slower frequency decay and excess mid-to-high frequencies. Diffusion images have intermediate slopes, often closer to real data.

A similar pattern emerges in the high-frequency energy ratio (Fig. 5.2). Real images preserve the most high-frequency content, while GANs retain the least and diffusion images lie between them. This holds in Bedroom and ImageNet, although the Face dataset deviates: diffusion images show the highest HF energy, followed by GANs, with real faces lowest.

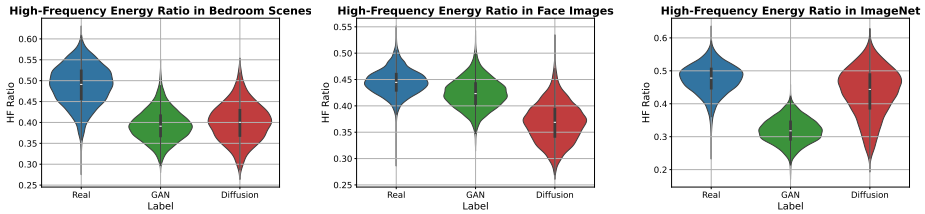


Figure 5.2: Violin plots of high-frequency energy ratios for Real, GAN, and Diffusion images across the three datasets. Real photos show higher HF ratios, reflecting richer fine detail. Overall, GANs have the lowest ratios; diffusion models fall in between, retaining more high-frequency content than GANs but less than real images.

The results have been confirmed by the log-log radial profiles presented in Figure 5.3. Genuine photographs regularly display natural ($1/f^{\alpha}$) degradation, while artificial image spectra often experience abrupt flattening or collapse in the mid-frequency band. The spectral properties of GAN and diffusion images vary across domains, revealing a discrepancy in the replication of natural frequency statistics by both models.

Collectively, figures 5.1, 5.2, 5.3 indicate that artificial images diverge from the natural $1/f^{\alpha}$ frequency decay observed in authentic photographs. GANs generally exhibit higher spectral slopes and less high-frequency energy, signifying an overemphasis on mid-frequency components and a loss of fine detail. On the other hand, diffusion models

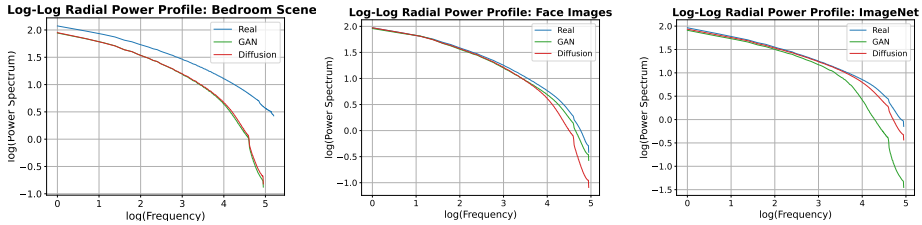


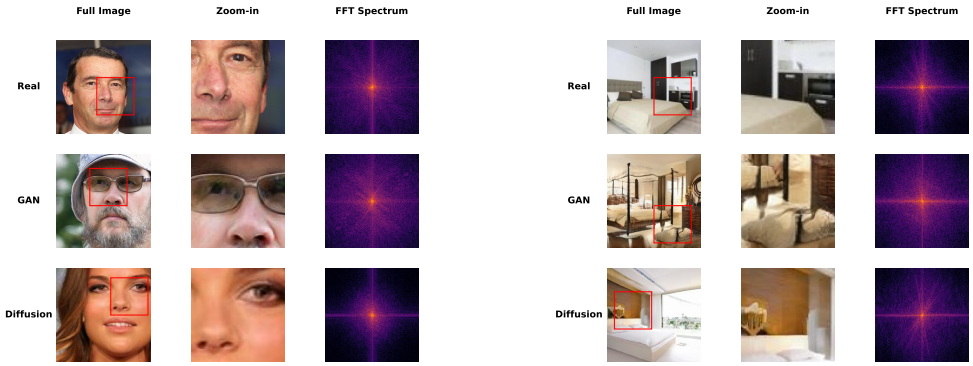
Figure 5.3: Average log-log radial power profiles per dataset. Real images show slower spectral decay (more high-frequency power) than GAN and diffusion images, which have a sharper drop-off, reflecting loss of fine detail and typical spectral patterns in synthetics.

often occupy an intermediate position, closer to real images in Bedroom and ImageNet domains, but exhibit less consistent behavior in facial imagery, where they sometimes exaggerate spectral decay and fine detail. These domain-specific deviations highlight the limitations of each generative approach. Statistical comparisons via Welch's t-tests confirm these differences as highly significant across modality pairs ($p < 10^{-3}$). Overall, spectral slope and high-frequency energy ratio provide a compact, interpretable frequency-domain signature for distinguishing real from synthetic images.

5.3 Qualitative Results

To complement quantitative findings, we present a qualitative comparison of real, GAN, and diffusion bedroom and face images in Fig. 5.4. Each row shows a full-resolution RGB image, a zoomed region, and the log-magnitude spectrum of the grayscale FFT.

The genuine images show coherent geometry and fine textures, with a smooth, isotropic FFT decay characteristic of natural $1/f^{\alpha}$ distributions. In contrast, GAN-generated images often exhibit structural glitches or repeating motifs in zoomed regions (e.g., glitches around eyes in faces, faint grid-like textures in bedroom objects) and corresponding spectral irregularities, such as weak cross-shaped streaks or uneven banding from upsampling and generator artifacts. Diffusion images appear smoother and closer to real samples in their FFT spectra, yet still display mild anisotropies or directional peaks (most noticeable in the bedroom patch), indicating only a partial restoration of natural frequency behavior. These visual results confirm that while both GANs and diffusion models can achieve high visual fidelity, diffusion models better reproduce the spatial and spectral characteristics of genuine photographs.



(a) Face portraits (b) Bedroom scenes

Figure 5.4: Visual comparison of local structure and frequency content for bedroom and face images. Real images show fine detail and a natural frequency spectrum, while GAN and diffusion images exhibit structural artifacts and spectral deviations.

5.4 Discussion & Summary

Our results reveal persistent and statistically significant disparities in frequency statistics between authentic and manufactured pictures. Real photographs follow natural $1/f^{\alpha}$ decay, characterized by smooth spectral slopes and high high-frequency energy ratios. GANs consistently deviate from this pattern, showing steeper spectral slopes and reduced high-frequency content, which are likely arising from upsampling operations and training instabilities. Diffusion models generate spectra closer to genuine images overall, especially in the Bedroom and ImageNet domains, but their behavior varies: in the face domain, they exaggerate spectral decay and retain excessive high-frequency energy, surpassing even GANs. These inconsistencies suggest domain-dependent limitations in current generative models. Radial profiles and FFT visualizations confirm these findings, highlighting structural artifacts absent in real images. Taken together, spectral slope and HF energy ratio form a compact and interpretable frequency signature that reliably separates real and synthetic images.

This study focuses on global grayscale frequency statistics, which miss localized or multiscale artifacts. Future research may investigate region-aware or patch-based spectral attributes to more effectively solve spatial inconsistencies as generative models advance.

Building on these findings, the next chapter explores texture and structure using spatial-domain metrics to detect artifacts not captured by frequency analysis.

Chapter 6

Texture & Structure Analysis

Genuine photographs contain coherent and organized textures shaped by the physics of surfaces, lighting, and materials [22, 14]. These textures exhibit statistical dependencies within small spatial regions that natural image formation preserves. GANs and diffusion models, however, often struggle to replicate fine-grained texture statistics and spatially coherent edges. GANs are prone to inconsistencies from limited receptive fields and latent space entanglement [38, 2], while diffusion methods alleviate some of these issues yet can produce overly smooth or over-regularized details in complex areas.

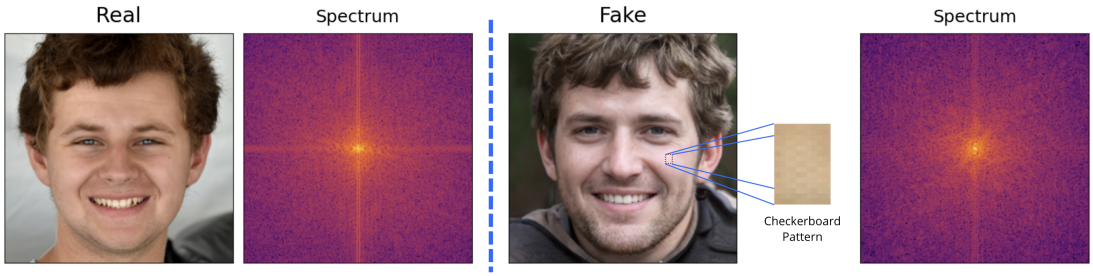


Figure 6.1: The spectra of a Real face image and a Fake face image generated by a GAN (StyleGAN2). Note the checkerboard artifact in the zoomed out details.

This chapter examines whether and how real, GAN, and diffusion images differ in their fine-grained textures and structural layouts across the Bedroom, Face, and ImageNet datasets. We aim to identify both shared artifacts and model-specific fingerprints that can act as model-agnostic authenticity cues. We hypothesize that real images display richer local texture variation and more coherent structural patterns, GANs show reduced texture complexity, and diffusion outputs fall in between, matching real data more closely in some domains than others.

To test this, we extract two complementary types of descriptors: (i) handcrafted gray-level co-occurrence matrix (GLCM) metrics (contrast, homogeneity, energy, correlation) and (ii) deep convolutional activations from VGG16's `conv1_2` layer. These features are analyzed via dimensionality reduction, statistical summaries, and linear SVM classification, supported by qualitative comparisons of texture-rich regions and gradient maps. As will be shown, real photographs form a tight, separable feature cluster, GANs occupy a distinct region with oversimplified textures, and diffusion images lie between the two, partially recovering structure but still lacking the full variability of natural textures.

6.1 Texture & Structure Estimation

To quantify structural and textural differences between real and synthetic images, we extract a combination of handcrafted and learned descriptors. These include gray-level co-occurrence matrix (GLCM) features and deep convolutional activations from a pretrained VGG16 network. Each representation captures complementary cues: handcrafted features are designed to characterize low-level spatial dependencies, while deep features implicitly encode higher-order structure and perceptual texture information.

6.1.1 Gray-Level Co-Occurrence Matrix Features

We compute GLCM descriptors as in [11], which quantify the joint probability of intensity transitions between neighboring pixels. From the co-occurrence matrix $P(i, j)$, we extract four standard metrics: contrast, homogeneity, energy, and correlation. For example, contrast and homogeneity are computed as:

$$\text{Contrast} = \sum_{i,j} (i-j)^2 \cdot P(i,j), \qquad \text{Homogeneity} = \sum_{i,j} \frac{P(i,j)}{1+|i-j|}$$

These features capture second-order texture properties such as repetition, smoothness, and directional consistency.

6.1.2 Deep Texture Features

To capture perceptual texture representations beyond handcrafted features, we extract deep activations from the `conv1_2` layer of a VGG16 network pretrained on ImageNet [27]. Each image is resized to 224×224 and normalized using standard ImageNet

statistics. The resulting activation tensor $F \in \mathbb{R}^{C\times H\times W}$ is average-pooled over spatial dimensions to yield a 64-dimensional descriptor:

$$\bar{f}_k = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_k(i,j), \quad k = 1,\ldots,C$$

These feature vectors reflect low-level visual structure, such as edge orientation and texture complexity, while remaining robust to color and illumination changes.

6.1.3 Linear SVM Classification

To evaluate the discriminability of deep texture features, we train a linear support vector machine (SVM) to classify images into three categories: real, GAN-generated, and diffusion-generated. Feature vectors are standardized and split into 70% training and 30% test sets. The SVM is implemented using scikit-learn's `SVC` with a linear kernel and default regularization. Classification accuracy and confusion matrices provide empirical evidence of feature separability in the texture space.

6.2 Quantitative Results

6.2.1 Handcrafted Texture Descriptors (GLCM)

We commence by analyzing the four GLCM metrics that are *contrast*, *homogeneity*, *energy*, and *correlation* among actual, GAN, and diffusion images to reveal systemic textural disparities based on local intensity connections.

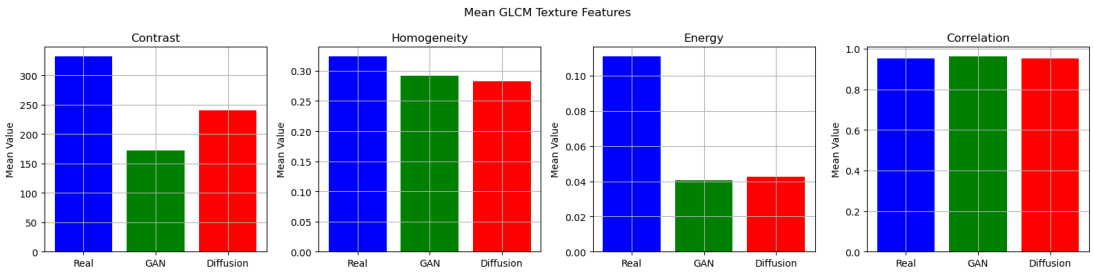


Figure 6.2: Mean GLCM features across modalities. Real images exhibit higher contrast, energy, and homogeneity, whereas GAN and Diffusion samples show reduced texture diversity. Yet, correlation differences are minimal.

As shown in Figure 6.2, real images exhibit the highest average *contrast*, reflecting richer local intensity variation. Both GAN and diffusion outputs exhibit significantly

lower *energy*, marginally decreased *homogeneity*, and comparable *correlation*, signifying smoother and more uniform textures with diminished co-occurrence diversity. Although diffusion samples preserve co-occurrence patterns more effectively than GANs, neither replicates the complete variety of natural textures. This demonstrates that, while synthetic models encapsulate certain spatial structures, they inadequately represent the complexity of real images.

Overall, GLCM features provide an interpretable, low-cost way to reveal texture discrepancies between real and synthetic images, supporting the hypothesis that handcrafted descriptors expose differences in local structural complexity even when global appearance seems convincing.

6.2.2 Deep Structure Features

We also use the `conv1_2` activations from VGG16 that is compressed into 64-dimensional embeddings to capture learned structural cues (edges, micro-textures) that complement handcrafted descriptors in distinguishing real from synthetic images.

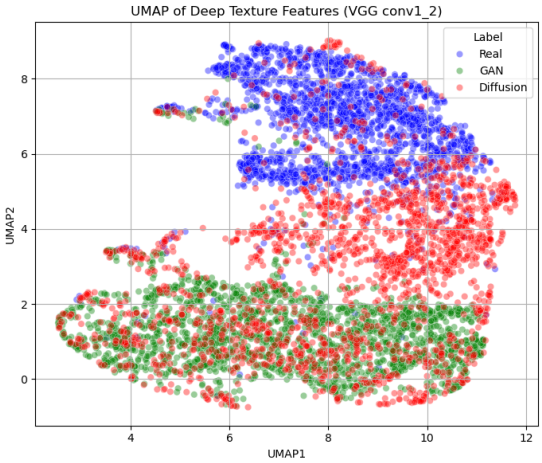


Figure 6.3: UMAP projection of VGG `conv1_2` features. Real and GAN images form distinct clusters, while diffusion images overlap both, indicating structural similarity.

Fig. 6.3 shows clear separation: real images cluster tightly in one region, GAN samples group distinctly elsewhere, and diffusion samples scatter between them, partially overlapping both. This supports that diffusion models produce structural features closer to real data than GANs, though residual artifacts remain. Early convolutional activations thus effectively separate synthetic and real images.

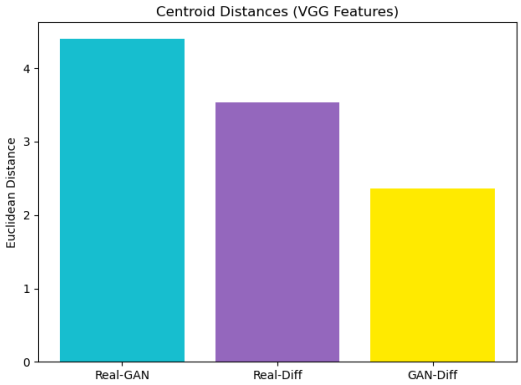


Figure 6.4: Euclidean distances between feature centroids, largest between Real and GAN, intermediate for Real and Diffusion, smallest for GAN and Diffusion, showing GAN and Diffusion are closer structurally.

Quantifying this, Euclidean distances between class centroids (Fig. 6.4) show the largest gap between real and GAN images, a smaller gap for real and diffusion, and the smallest between GAN and diffusion, highlighting their structural similarity.

We further assess discriminability using a linear SVM trained on all datasets. Figure 6.5 illustrates that real images attain an accuracy of 89.3%, with misclassifications roughly evenly distributed across GAN (91) and diffusion (72). GANs are classified with 82.4% accuracy, but 220 are mistaken for diffusion, reflecting their structural resemblance. Diffusion images achieve an accuracy of 75.1%, with misclassifications predominantly associated with GAN (257) and some with real images (100), situating diffusion inside a hybrid feature space that bridges GAN and real textures.

In summary, deep texture features reveal consistent structural differences between real and synthetic images. GANs and diffusion models occupy separable yet overlapping regions. Real images form a tight, discriminable cluster, while diffusion samples lie in between, causing classification ambiguity. This variability in synthetic structural features across architectures limits deepfake detectors' generalization.

6.3 Qualitative Results

To visually illustrate texture and structure differences across modalities, we present qualitative comparisons in Figures 6.6a and 6.6b. Each figure shows three rows (Real, GAN, Diffusion) and three columns: the full image with a red zoom box, a zoomed crop

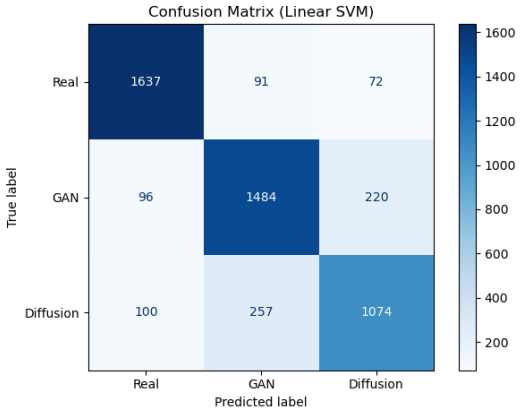
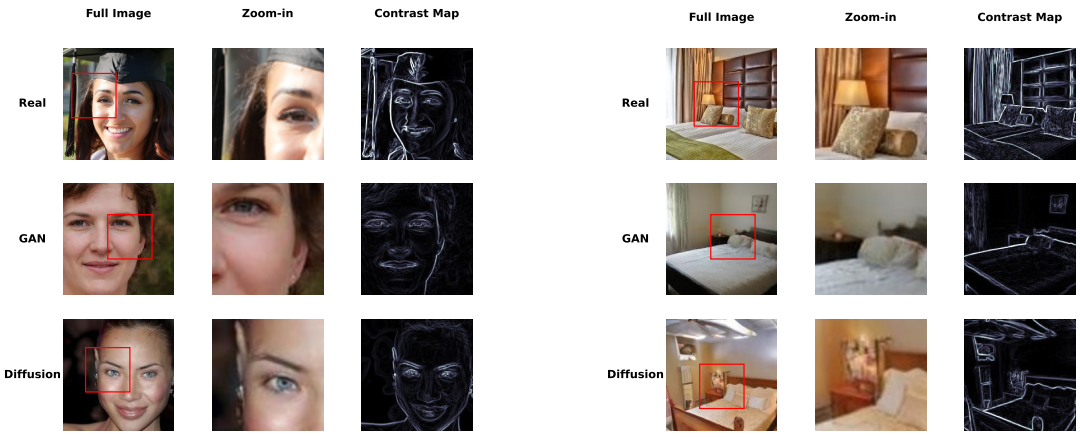


Figure 6.5: Confusion matrix for SVM classifier using VGG `conv1_2` features, showing strong Real vs. synthetic discrimination, with most errors between GAN and Diffusion.

of a texture-rich region, and a contrast map from the Sobel gradient. These underscore disparities in local textural complexity and structural diversity.



(a) Face portraits

(b) Bedroom scenes

Figure 6.6: Qualitative texture comparisons of Real, GAN, and Diffusion images using zoomed regions and contrast maps. Real images reveal fine, pervasive texture; GANs suppress detail except along edges; Diffusion images partially recover texture with more spatial regularity.

In both domains, real images show rich microtexture and detailed spatial structure, clearly visible in zoomed regions and contrast maps. Gradients are stronger and more spatially diverse, indicating fine edge transitions and natural surface irregularities. In contrast, GAN images exhibit texture oversimplification, with blocky patterns or repeated motifs and limited gradient variation, especially in faces where texture flatness

is more pronounced. Diffusion samples are intermediate: structural richness improves over GANs in bedrooms but lacks real high-frequency variability; in faces, contrast patterns resemble GANs, indicating diffusion still underperforms in fine-scale facial textures.

These qualitative findings validate and contextualize quantitative patterns. Visual anomalies like edge blurriness, repeated patterns, and minor gradient variation consistently signal a synthetic origin and highlight the necessity of examining both global statistics and localized structures in the development of detection classifiers.

6.4 Discussion & Summary

Our study reveals distinct discrepancies in texture and structure between authentic and artificial images. GLCM metrics indicate that although both GAN and diffusion outputs do not accurately replicate authentic textures, diffusion images demonstrate marginally greater contrast than GANs; yet, their overall co-occurrence statistics remain relatively suppressed compared to real photographs. In parallel, deep features from VGG16's conv1_2 layer place real images in a tight, separable cluster in UMAP space and yield the highest classification accuracy with a linear SVM. GAN and diffusion samples overlap more closely, with diffusion sitting nearer to the real cluster but still distinguishable. These patterns, visually reinforced in Figures 6.6a and 6.6b, suggest that real photographs occupy a cohesive manifold of rich, varied textures, whereas each generative model imprints its own, model-specific signature.

This texture–structure analysis is constrained by its reliance on local, early-layer features and grayscale statistics, thereby overlooking the richer color textures and multi-scale patterns addressed in prior chapters. Global semantic and layout discrepancies are overlooked, suggesting an opportunity for integrating higher-level and multi-scale descriptors in future works.

The next section builds on these insights by analyzing counterfeit image classifiers trained on independently color, noise, frequency, and structure feature types to evaluate their robustness and cross-model generalizability.

Chapter 7

Generalization Evaluation

As explained in Chapter 1, deepfake detectors often excel on the architectures they are trained on but fail dramatically when faced with novel generative models [5, 3]. This lack of cross-architecture generalization undermines media reliability and digital authenticity, as highlighted in recent studies calling for universal fake-image detectors [21].

To address this, we evaluate how different handcrafted feature sets and training regimes affect robustness to unseen generators. Building on our methodology (Chapter 7.1), we extract four complementary descriptors: *color distributions*, *noise patterns*, *frequency statistics*, and *texture & structure embeddings*, and train Random Forest classifiers under two cross-architecture setups (Real+GAN $\Rightarrow$ Diffusion and Real+Diffusion $\Rightarrow$ GAN). We measure both in-domain and out-of-domain performance using our Generalization (*G-*) score, apply statistical tests (Friedman and Wilcoxon signed-rank), and analyze intra-modality feature importance. Based on previous research [20, 33], we hypothesize that Real+GAN $\Rightarrow$ Diffusion configuration will outperform the reverse in detecting unseen diffusion images. We further expect frequency-domain features to achieve the highest average *G-scores* by exploiting global spectral patterns that persist across architectures.

However, results show a consistent, though not statistically significant, advantage for Real+Diffusion $\Rightarrow$ GAN training, reflecting the intermediate feature-space position of diffusion outputs. Frequency features indeed deliver the most stable generalization (mean G-score 0.971, lowest variance), while color features top performance in specific datasets but remain less consistent. Across all modalities, broader-context features such as HSV histograms, wavelet noise measures, global spectral slope, and deep

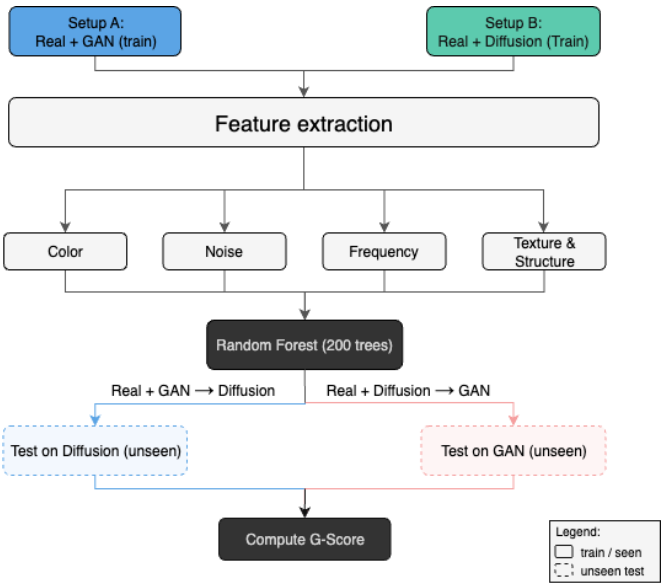


Figure 7.1: Generalization evaluation pipeline comparing two setups: (**A**) Real + GAN $\rightarrow$ Diffusion and (**B**) Real + Diffusion $\rightarrow$ GAN. Both use the same handcrafted features (color, noise, frequency, texture & structure) and a Random Forest (200 trees). Dashed boxes indicate unseen test domains, with the G-score measuring generalization.

VGG activations are the strongest contributors. These findings highlight the combined importance of diffusion-based training and globally oriented descriptors for building detectors that generalize to future generative models.

7.1 Methodology

7.1.1 Classifier and Training Setup

This study employs **Random Forest classifiers** consisting of **200 trees**. They are implemented via the `scikit-learn` library, utilizing class-balanced weights and fixed random seeds to guarantee reproducibility. Random Forest architecture selected due to its robustness, interpretability, and capacity for handling tabular input features. The exact same classifier model and hyperparameters are utilized across all tests to guarantee comparability.

To evaluate generalization, we consider two training configurations per dataset:

- **Real + GAN $\Rightarrow$ Diffusion:** Trained on real and GAN-generated images, tested on diffusion-generated images.

- **Real + Diffusion ⇒ GAN:** Trained on real and diffusion-generated images, tested on GAN-generated images.

Each setup simulates a generalization scenario, across generative architectures. All training sets are balanced between real and fake images, except in the Bedroom dataset which includes only 768 diffusion samples. We apply SMOTE to address it before training.

Validation sets are drawn from the training domain to measure **seen accuracy**, while **unseen accuracy** is computed on the held-out generative model. To quantify generalization, we report a **bounded G-score**:

$$G = \max\left(0, \min\left(1, 1 - \frac{A_{\text{seen}} - A_{\text{unseen}}}{A_{\text{seen}} + \varepsilon}\right)\right),$$

where A_seen and A_unseen are seen and unseen accuracies respectively, and $\varepsilon = 10^{-8}$ ensures numerical stability. This formulation constrains G to the range [0,1], with higher values indicating better generalization and penalizing models that overfit to the training distribution.

7.1.2 Evaluated Feature Modalities

We divide the retrieved descriptors into four categories, each addressing a distinct statistical characteristic of both genuine and synthetic content. Comprehensive extraction details are provided in previous methodology sections.

- **Color Distributions**: Color features are computed as concatenated histograms across RGB, HSV, and YCrCb color spaces, capturing brightness, saturation, and chrominance characteristics.
- **Noise Patterns** (Section 4.1): We extract local noise residuals using a window-based standard deviation filter and wavelet-based energy from Haar subbands, both designed to capture fine-grained noise inconsistencies.
- **Frequency Features** (Section 5.1): From the 2D Fourier spectrum, we compute the spectral slope α and the high-frequency (HF) energy ratio, summarizing the global distribution of spatial frequencies.
- **Texture & Structure** (Section 6.1): This includes handcrafted texture features such as LBP and GLCM metrics, as well as deep descriptors from the early layers of a pre-trained VGG16 network.

7.2 Results and Discussion

This section delivers a comparative assessment of generalization performance across several feature modalities and training configurations, focusing on the *G-score* metric as described in Section 7.1. Our objective is to determine which feature types and training methods facilitate effective deepfake detection across generative models.

7.2.1 Training Setup Comparison

This section evaluates how training configuration affects generalization across generative architectures. We compare the two setups: Real+GAN $\Rightarrow$ Diffusion and Real+Diffusion $\Rightarrow$ GAN.

As shown in Table 7.1, training on diffusion-generated images generally yields stronger generalization to unseen GAN content than the reverse configuration, with an overall *G-score* of 0.9171 compared to 0.8462. This benefit is most apparent in the ImageNet dataset, our most balanced and varied benchmark, where diffusion-based training excels across all four feature types, achieving a 36% relative improvement (see Table 7.2). A comparable pattern appears in the Bedroom dataset, notwithstanding its significant class imbalance (768 diffusion images versus over 5,000 GAN images), where diffusion-based training yields a 16% relative enhancement, indicating that it extracts more generalizable features even in conditions of data scarcity. The lone exception is the Face dataset, where Real+GAN $\Rightarrow$ Diffusion achieves the highest G-score (0.9894), although it is outperformed by the reverse configuration with a 24% relative gain which likely reflects domain-specific effects such as StyleGAN2's alignment priors or Palette's lighting regularities. These empirical trends mirror the global pattern described in earlier chapters (3–6), suggesting that diffusion-based training helps classifiers learn broader, less overfit decision boundaries.

Table 7.1: Average *G-scores* across datasets (Bedroom, Face, ImageNet) for each training setup. The final columns report the overall mean, standard deviation, and one-sided t-test p-value comparing the first two training setups.

Training Setup	Tested on	Bedroom	Face	ImageNet	Mean	Std Dev	One-sided p-value (rg vs. rd)
Real+GAN	Diffusion	0.8007	**1.0000**	0.7378	0.8462	0.1118	0.1688
Real+Diffusion	GAN	**0.9548**	0.7966	**1.0000**	**0.9171**	**0.0872**	0.1688

We conducted a paired t-test on the 12 matched *G-scores* (4 descriptors $\times$ 3 datasets) to assess the statistical significance of the observed difference. The resulting one-sided p-value of 0.1688 indicates no statistical significance, yet the empirical advantage of diffusion-based training, especially in ImageNet and Bedroom, remains noteworthy.

Table 7.2: Per-dataset G-scores by feature type and training setup, with each row's mean and standard deviation. Bold marks the best score or lowest variance.

Dataset	Setup	Color	Noise	Freq.	Texture	Mean	Std Dev
ImageNet	R+G⇒D	0.9108	0.6784	0.6481	0.7137	0.7377	0.1116
	R+D⇒G	**0.9883**	**1.0000**	**1.0000**	**1.0000**	**0.9971**	**0.0051**
Face	R+G⇒D	**0.9578**	**1.0000**	**1.0000**	**1.0000**	**0.9894**	**0.0183**
	R+D⇒G	0.9509	0.7575	0.6720	0.8058	0.7966	0.1012
Bedroom	R+G⇒D	0.6120	0.7900	0.9950	0.8057	0.8007	0.1356
	R+D⇒G	**0.7220**	**1.0000**	**1.0000**	**1.0000**	**0.9305**	**0.1204**

In summary, while the Real+Diffusion $\Rightarrow$ GAN configuration does not attain statistical significance compared to Real+GAN $\Rightarrow$ Diffusion, its consistent performance, robustness to imbalance, and conformity with feature-space theory render it a formidable contender for robust cross-architecture deepfake detection.

7.2.2 Feature Modality Comparison

To evaluate the relative utility of different feature types for cross-architecture deepfake detection, we computed the G-score for each modality: color, noise, frequency, and texture & structure, averaged across datasets and training setups. Table 7.3 shows that frequency features attained the highest overall G-score (0.971), representing an improvement of roughly 8% over texture & structure and noise descriptors, and 13% over color features, with an exceptionally low standard deviation (0.025) indicating highly consistent generalization performance across datasets.
While color features led in the Face and ImageNet datasets, they performed worst in Bedroom (0.667). Frequency features, despite being the top performer in only one dataset (Bedroom), maintained strong results in all three, explaining their superior mean G-score and stability.

A *Friedman test* across modalities yielded $\chi^2 = 1.4$ and $p = 0.706$, indicating no significant rank differences. All pairwise *Wilcoxon signed-rank tests* produced p-values

Table 7.3: Summary of *G-scores* across feature modalities. Columns report mean and standard deviation across datasets, along with average pairwise Wilcoxon p-values.

Feature Type	Bedroom	Face	ImageNet	Mean	Std Dev	Avg p-value
Color	0.667	**0.967**	**0.950**	0.861	0.169	0.656
Noise	0.933	0.897	0.864	0.898	0.035	0.760
Frequency	**0.998**	0.965	0.949	**0.971**	**0.025**	0.438
Texture & Structure	0.913	0.933	0.859	0.902	0.038	0.708

above 0.43 (Table 7.3).

These findings suggest that, although no modality is statistically dominant, frequency-based descriptors offer the most stable cross-domain performance, making them attractive for architecture-agnostic deepfake detection pipelines.

7.2.3 Intra-Modality Feature Importance

We examined the relative contributions of sub-feature groups within each modality by taking advantage of the interpretability benefit of Random Forest classifiers. Table 7.4 presents the mean feature importance scores for the Bedroom, Face, and ImageNet datasets.

Table 7.4: Average feature importance within each modality across datasets, based on Random Forest classifiers (200 trees).

Modality	Top Feature	Importance (avg %)	Other Features (avg %)
Color	HSV Histogram	**78.0**	RGB (19.2), YCrCb (3.0)
Noise	Wavelet Features	**71.2**	Local Noise (28.8)
Frequency	Spectral Slope	**51.8**	HF Ratio (48.2)
Texture & Structure	VGG Activations	**66.7**	LBP (24.3), GLCM (9.0)

A clear trend emerges, features capturing broader spatial or statistical context dominate their respective modalities. In the color space, HSV histograms frequently surpass RGB and YCrCb, highlighting the significance of hue and saturation. Wavelet-based descriptors dominate local residuals for noise features, presumably owing to their enhanced sensitivity to fine-scale textures. In the frequency domain, both spectral slope and high-frequency ratio are significant, with a slight advantage for slope, suggesting that global decay and local energy patterns provide complimentary insights. Within the texture category, deep VGG activations clearly outperform handcrafted descriptors like

LBP and GLCM, highlighting the value of mid-level learned representations.

These intra-modality discoveries underline the specific features that most significantly improve classifier performance within each descriptor category. This comprehension may direct future feature design and dimensionality reduction techniques in the development of lightweight or interpretable deepfake detectors.

7.3 Discussion & Summary

Through this Generalization Evaluation chapter, we discover that training on Real+Diffusion to detect GANs achieves the highest average generalization. This outcome is intuitive, as diffusion outputs occupy an intermediate position between real and GAN images in feature space. However, this result contradicts our initial hypothesis, supported by previous research [20, 33], that GAN samples would provide stronger training signals for detecting unseen diffusion images. While color descriptors achieved the highest average G-scores in two of three datasets, their high variability limited their overall reliability. In contrast, frequency-based features, capturing global spectral decay and high-frequency energy, were remarkably consistent, averaging a G-score of 0.971 compared to 0.861 for color features, supporting our original hypothesis. Within each feature family, those capturing broader context (HSV over RGB/YCrCb, wavelet noise over local residuals, global spectral slope over HF ratio, and deep VGG activations over handcrafted texture) contributed most to classifier performance, suggesting a focus on global rather than purely local statistics for robust deepfake detection.

Nevertheless, our study has few limitations. We only evaluate handcrafted features with Random Forests; deep or end-to-end models may yield different results. We test just six GAN and diffusion variants, therefore new architectures could shift these patterns. The G-score metric ignores real-world priors and adversarial attacks, and, despite SMOTE, data imbalances may still bias outcomes. Finally, our results are not statistically significant.

Overall, our results highlight diffusion-based training and frequency-domain descriptors as key ingredients in crafting deepfake detectors that generalize effectively across unseen generative architectures.

Chapter 8

Conclusions

In this study, we characterized and compared signatures distinguishing images synthesized by GANs and diffusion models, and evaluated which handcrafted features and training strategies support generalization to unseen architectures. Across four complementary analyses—*color distributions*, *high-frequency noise patterns*, *global frequency spectra*, and *texture–structure embeddings*—we observed a Real $>$ Diffusion $>$ GAN hierarchy: real images exhibit the richest color diversity, broadly distributed sensor-like noise, smooth $1/f^{\alpha}$ spectral decay, and complex textures, while GAN images compress these characteristics and diffusion images lie intermediate. In the color domain, GANs boost brightness and compress channel variance, while diffusion models partially recover natural variances; in the noise domain, GANs concentrate grain around sharp edges, diffusion models restore mid-level randomness; spectrally, GANs exhibit steep slopes and reduced high-frequency energy, while diffusion spectra resemble real power-laws but vary by domain; spatially, GLCM and VGG activations indicate diffusion textures better approximate real variability than GANs, though both synthetic types lack real-image complexity. These findings provide a concrete characterization of overlaps and distinctions between GAN and diffusion fingerprints, expanding on prior observations of cross-architecture similarities in previous research [21].

Building on these findings, we trained Random Forest classifiers using four handcrafted feature sets under two cross-architecture scenarios. Incorporating diffusion samples into training improved generalization to unseen GAN images, yielding an 8% relative G-score gain (0.9171 vs. 0.8462). This was clear in the ImageNet (36%) and Bedroom (16%) datasets, whereas the Face dataset showed a 24% advantage for the reverse setup. These trends reflect diffusion's intermediate position in feature space and its

ability to shift decision boundaries away from GAN-specific artifacts. Among features, frequency-based characteristics were most stable (mean G-score 0.971, std. 0.025), outperforming other descriptors by 8–13%, though color histograms achieved the highest scores on most benchmarks.

Together, these results reveal which statistical cues best differentiate generative models and show that diffusion-generated training data with frequency-based features offers a lightweight, architecture-agnostic approach to robust deepfake detection.

8.1 Limitations

Several limitations constrain our findings. Most descriptors operate globally and in grayscale, potentially missing localized, color-sensitive, or multi-scale artifacts. Texture and structure metrics rely on local descriptors and overlook semantic inconsistencies or global spatial patterns. While interpretable, handcrafted features may miss subtler cues detectable by deep models, especially as generative techniques evolve. The evaluation is limited in scope: we used only Random Forests and six generative models, so other classifiers might behave differently. G-score does not account for adversarial manipulation or real-world priors, and class imbalances (notably in Bedroom) may skew results despite SMOTE. Lastly, although diffusion-based training and frequency features performed best on average, these differences lacked statistical significance, underscoring the need for broader validation.

8.2 Future Work

Future research could extend this study in several directions. First, scaling to a larger, more diverse corpus with newer GANs and diffusion variants, higher resolutions, and broader semantic domains would test the robustness of current findings. Second, moving beyond handcrafted features to joint multi-modal descriptors or learned embeddings via contrastive or self-supervised training may capture more nuanced differences. Third, evaluating end-to-end or hybrid architectures using domain adaptation, adversarial training, or meta-learning could further enhance cross-architecture generalization. Finally, incorporating real-world noise sources such as compression artifacts, sensor metadata, and adversarial perturbations will be essential for building practical, resilient detection systems.

Bibliography

[1] Mahmoud Afifi, Marcus A. Brubaker, and Michael S. Brown. *HistoGAN: Controlling Colors of GAN-Generated and Real Images via Color Histograms*. 2021. arXiv: 2011.11731 [cs.CV]. URL: https://arxiv.org/abs/2011.11731.

[2] David Bau et al. "GAN Dissection: Visualizing and Understanding Generative Adversarial Networks". In: *CoRR* abs/1811.10597 (2018). arXiv: 1811.10597. URL: http://arxiv.org/abs/1811.10597.

[3] Jordan J. Bird and Ahmad Lotfi. "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images". In: *IEEE Access* 12 (2024), pp. 15642–15650. DOI: 10.1109/ACCESS.2024.3356122.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. 2019. arXiv: 1809.11096 [cs.LG]. URL: https://arxiv.org/abs/1809.11096.

[5] Davide Cozzolino et al. *ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection*. 2019. arXiv: 1812.02510 [cs.CV]. URL: https://arxiv.org/abs/1812.02510.

[6] Ricard Durall, Margret Keuper, and Janis Keuper. *Watch your Up-Convolution: CNN Based Generative Deep Neural Networks are Failing to Reproduce Spectral Distributions*. 2020. arXiv: 2003.01826 [cs.CV]. URL: https://arxiv.org/abs/2003.01826.

[7] David J. Field. "Relations between the statistics of natural images and the response properties of cortical cells". In: *J. Opt. Soc. Am. A* 4.12 (Dec. 1987), pp. 2379–2394. DOI: 10.1364/JOSAA.4.002379. URL: https://opg.optica.org/josaa/abstract.cfm?URI=josaa-4-12-2379.

[8] Joel Frank et al. "Leveraging frequency analysis for deep fake image recognition". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org, 2020.

[9] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML]. URL: https://arxiv.org/abs/1406.2661.

[10] Shuyang Gu et al. *Vector Quantized Diffusion Model for Text-to-Image Synthesis*. 2022. arXiv: 2111.14822 [cs.CV]. URL: https://arxiv.org/abs/2111.14822.

[11] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. "Textural Features for Image Classification". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3.6 (1973), pp. 610–621. DOI: 10.1109/TSMC.1973.4309314.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG]. URL: https://arxiv.org/abs/2006.11239.

[13] Drew A. Hudson and C. Lawrence Zitnick. *Compositional Transformers for Scene Generation*. 2021. arXiv: 2111.08960 [cs.CV]. URL: https://arxiv.org/abs/2111.08960.

[14] B. Julesz. "Visual Pattern Discrimination". In: *IRE Transactions on Information Theory* 8.2 (1962), pp. 84–92. DOI: 10.1109/TIT.1962.1057698.

[15] Thibaut Julliand, Vincent Nozick, and Hugues Talbot. "Image Noise and Digital Image Forensics". In: vol. 9569. Mar. 2016, pp. 3–17. ISBN: 978-3-319-31959-9. DOI: 10.1007/978-3-319-31960-5_1.

[16] Tero Karras et al. *Analyzing and Improving the Image Quality of StyleGAN*. 2020. arXiv: 1912.04958 [cs.CV]. URL: https://arxiv.org/abs/1912.04958.

[17] Claire Little et al. *Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study*. Dec. 2021. DOI: 10.48550/arXiv.2112.01925.

[18] J. Lukas, J. Fridrich, and M. Goljan. "Digital camera identification from sensor pattern noise". In: *IEEE Transactions on Information Forensics and Security* 1.2 (2006), pp. 205–214. DOI: 10.1109/TIFS.2006.873602.

[19] Scott McCloskey and Michael Albright. *Detecting GAN-generated Imagery using Color Cues*. 2018. arXiv: 1812.08247 [cs.CV]. URL: https://arxiv.org/abs/1812.08247.

[20] Zheling Meng et al. *Artifact Feature Purification for Cross-domain Detection of AI-generated Images*. 2024. arXiv: 2403.11172 [cs.CV]. URL: https://arxiv.org/abs/2403.11172.

[21] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. *Towards Universal Fake Image Detectors that Generalize Across Generative Models*. 2024. arXiv: 2302.10174 [cs.CV]. URL: https://arxiv.org/abs/2302.10174.

[22] Javier Portilla and Eero P. Simoncelli. "A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients". In: *International Journal of Computer Vision* 40.1 (2000), pp. 49–70. DOI: 10.1023/A:1026553619983. URL: https://doi.org/10.1023/A:1026553619983.

[23] Md Awsafur Rahman et al. "Artifact: A Large-Scale Dataset With Artificial And Factual Images For Generalizable And Robust Synthetic Image Detection". In: *2023 IEEE International Conference on Image Processing (ICIP)*. 2023, pp. 2200–2204. DOI: 10.1109/ICIP49359.2023.10222083.

[24] Daniel L. Ruderman and William Bialek. "Statistics of natural images: Scaling in the woods". In: *Phys. Rev. Lett.* 73 (6 Aug. 1994), pp. 814–817. DOI: 10.1103/PhysRevLett.73.814. URL: https://link.aps.org/doi/10.1103/PhysRevLett.73.814.

[25] Chitwan Saharia et al. *Palette: Image-to-Image Diffusion Models*. 2022. arXiv: 2111.05826 [cs.CV]. URL: https://arxiv.org/abs/2111.05826.

[26] Alberto Sanchez-Acedo et al. "The challenges of media and information literacy in the artificial intelligence ecology: deepfakes and misinformation". In: *Communication Society* (2024). DOI: 10.15581/003.37.4.223-239.

[27] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV]. URL: https://arxiv.org/abs/1409.1556.

[28] Jascha Sohl-Dickstein et al. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. 2015. arXiv: 1503.03585 [cs.LG]. URL: https://arxiv.org/abs/1503.03585.

[29] Madhur Srivastava, C. Lindsay Anderson, and Jack H. Freed. "A New Wavelet Denoising Method for Selecting Decomposition Levels and Noise Thresholds". In: *IEEE Access* 4 (2016), pp. 3862–3877. DOI: 10.1109/ACCESS.2016.2587581.

[30] Antonio Torralba and Aude Oliva and. "Statistics of natural image categories". In: *Network: Computation in Neural Systems* 14.3 (2003). PMID: 12938764, pp. 391–412. DOI: 10.1088/0954-898X_14_3_302. eprint: https://doi.org/10.1088/0954-898X_14_3_302. URL: https://doi.org/10.1088/0954-898X_14_3_302.

[31] Luisa Verdoliva. "Media Forensics and DeepFakes: An Overview". In: *IEEE Journal of Selected Topics in Signal Processing* 14.5 (2020), pp. 910–932. DOI: 10.1109/JSTSP.2020.3002101.

[32] Yujin Wang, Mike A. Webster, and Daniel S. Joyce. "Color statistics of images created by generative AI". In: *J. Opt. Soc. Am. A* 42.5 (May 2025), B76–B80. DOI: 10.1364/JOSAA.545030. URL: https://opg.optica.org/josaa/abstract.cfm?URI=josaa-42-5-B76.

[33] Shilin Yan et al. *A Sanity Check for AI-generated Image Detection*. 2025. arXiv: 2406.19435 [cs.CV]. URL: https://arxiv.org/abs/2406.19435.

[34] Ling Yang et al. "Diffusion Models: A Comprehensive Survey of Methods and Applications". In: *ACM Computing Surveys* 56 (2022), pp. 1–39. DOI: 10.1145/3626235.

[35] Matthew Yates et al. "Evaluation of synthetic aerial imagery using unconditional generative adversarial networks". In: *ISPRS Journal of Photogrammetry and Remote Sensing* (2022). DOI: 10.1016/j.isprsjprs.2022.06.010.

[36] Sergej Yendrikhovskij. "Computing color categories from statistics of natural images". In: *Journal of Imaging Science and Technology* 45 (Sept. 2001).

[37] Hui Yu et al. "Color texture moments for content-based image retrieval". In: *Proceedings. International Conference on Image Processing*. Vol. 3. 2002, 929–932 vol.3. DOI: 10.1109/ICIP.2002.1039125.

[38] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. "Detecting and Simulating Artifacts in GAN Fake Images". In: *2019 IEEE International Workshop on Information Forensics and Security (WIFS)* (2019), pp. 1–6. DOI: 10.1109/WIFS47025.2019.9035107.

[39] Ziqi Zhu et al. "Dynamic texture modeling and synthesis using multi-kernel Gaussian process dynamic model". In: *Signal Processing* 124 (2016). Big Data Meets Multimedia Analytics, pp. 63–71. ISSN: 0165-1684. DOI: https://doi.org/

10.1016/j.sigpro.2015.10.025. URL: https://www.sciencedirect.com/science/article/pii/S0165168415003680.

Appendix A

Appendix

A.1 Software and Implementation

The implementation of this project relied on a Python-based environment combining scientific computing, image processing, and machine learning libraries. NumPy and pandas were used extensively for numerical operations and data management, while Matplotlib and Seaborn supported the creation of plots and visualizations. Image preprocessing and feature extraction leveraged scikit-image and OpenCV, complemented by PyWavelets for wavelet decomposition in noise analysis. Machine learning experiments were conducted using scikit-learn, which provided tools for feature scaling, classifier training (e.g., Random Forest), and model evaluation. SciPy facilitated statistical testing, including the Friedman test, Wilcoxon signed-rank test, and Welch's t-test. Jupyter Notebooks served as the main environment for exploratory analysis and iterative development, ensuring a clear and reproducible research workflow.

A.2 Glossary

This glossary defines key technical terms used throughout the dissertation to aid reader understanding, particularly for concepts in recommendation systems, graph learning, and deep learning.

Checkerboard Pattern An artifact in synthetic images, often caused by certain upsampling methods such as transposed convolution, that appears as a grid-like pattern of alternating light and dark blocks. *Example:* In some GAN outputs, checkerboard patterns can be seen in flat or uniform regions of an image due to

uneven kernel overlap during upsampling.

Chrominance The component of a color signal that represents color information independently of brightness, typically encoded as two values (e.g., Cb and Cr in the YCbCr color space). *Example:* Differences in chrominance variance between real and synthetic images can indicate the presence of generative artifacts.

Cross-Architecture Generalization The ability of a detection model trained on data from one generative architecture to accurately identify synthetic content from a different, unseen architecture. *Example:* A classifier trained on real and GAN images that can also detect diffusion-generated images without retraining.

Deepfake Synthetic media, often created using generative models, that convincingly imitates real images, videos, or audio, typically for deceptive purposes. *Example:* A face-swapped video generated by a GAN to impersonate a public figure.

Discriminator The component of a generative model, particularly in GANs, that evaluates whether an input is real (from the dataset) or fake (produced by the generator), guiding the generator's improvement through feedback. *Example*: In a GAN, the discriminator learns to distinguish genuine images from those created by the generator.

Friedman Test A non-parametric statistical test used to detect differences in performance across multiple related samples or experimental conditions. *Example:* Applied to compare the G-scores of different feature modalities across datasets in deepfake detection experiments.

Generator The component of a generative model that creates synthetic data, typically from random noise or a latent representation, aiming to mimic the distribution of real data. *Example:* In a GAN, the generator produces images intended to fool the discriminator.

Handcrafted Feature A manually designed descriptor that captures specific characteristics of data, based on domain knowledge rather than learned automatically by a model. *Example:* Color histograms, wavelet-based noise measures, and Gray-Level Co-occurrence Matrix (GLCM) statistics used to detect synthetic image artifacts.

Hyperparameters Configurable settings of a machine learning model or algorithm that are set before training and influence the learning process and performance. *Example:* The number of trees in a Random Forest classifier or the window size in local noise estimation.

Model-Agnostic Detection A detection approach designed to identify synthetic content regardless of the specific generative model or architecture that produced it. *Example:* Using frequency-domain features that remain effective across both GAN and diffusion-generated images.

Spectral Decay The rate at which the magnitude of an image's frequency spectrum decreases with increasing spatial frequency, often following a power-law distribution in natural images. *Example:* GAN-generated images may show steeper spectral decay than real images, indicating reduced high-frequency detail.

Upsampling A process in image generation that increases the spatial resolution of feature maps or images, often using techniques such as transposed convolution, interpolation, or sub-pixel convolution. *Example:* In many GAN architectures, upsampling layers expand low-resolution feature maps to full-size images.

VGG Activations Feature representations extracted from intermediate layers of the VGG convolutional neural network, often used to capture texture and structural information in images. *Example:* Activations from the conv1_2 layer were used to compare structural characteristics of real, GAN, and diffusion-generated images.

Welch's t-test A statistical test used to determine whether two samples have significantly different means, without assuming equal variances between the groups. *Example:* Applied to compare feature values between real and synthetic images in each domain.

Wilcoxon Signed-Rank Test A non-parametric statistical test used to compare two related samples, measuring whether their population mean ranks differ. *Example:* Used to assess the significance of performance differences between two training setups in cross-architecture deepfake detection.