



UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE INGENIERÍA



Image Forensics in the Age of Generative Models: Detecting AI-Generated Images through Color Correlations

TESIS PRESENTADA A LA FACULTAD DE INGENIERÍA DE LA
UNIVERSIDAD DE LA REPÚBLICA POR

Juan Pablo Sotelo Silva

EN CUMPLIMIENTO PARCIAL DE LOS REQUERIMIENTOS
PARA LA OBTENCIÓN DEL TÍTULO DE
MAGISTER EN CIENCIA DE DATOS Y APRENDIZAJE AUTOMÁTICO.

DIRECTORES DE TESIS

Pablo Musé..... Universidad de la República
Marina Gardella École Normale Supérieure Paris-Saclay

TRIBUNAL

Quentin Bammey..... Télécom Paris
Rafael Grompone École Normale Supérieure Paris-Saclay
Lara Raad..... Universidad de la República

DIRECTOR ACADÉMICO

Pablo Musé..... Universidad de la República

Montevideo
17 Abril 2026

Image Forensics in the Age of Generative Models: Detecting AI-Generated Images through Color Correlations, Juan Pablo Sotelo Silva.

ISSN 1688-2806

This thesis was typeset in L^AT_EX using the iietesis class (v1.1).

Total pages: 118.

Compiled on martes 30 junio, 2026.

<http://iie.fing.edu.uy/>

It is of the highest importance in the art of detection to be able to recognize, out of a number of facts, which are incidental and which vital.

Es de suma importancia en el arte de la detección ser capaz de reconocer, entre una serie de hechos, cuáles son accidentales y cuáles fundamentales.

SHERLOCK HOLMES

This page has been intentionally left blank.

Acknowledgments

This is probably the most difficult section to write. Although pursuing a master's degree and completing a thesis require a great deal of effort, it is not a journey one undertakes alone; in one way or another, the people around us also become part of that process. For that reason, I would like to thank everyone who accompanied me along this path.

To Agustina, my partner, for being by my side on the front line. For being there during those afternoons in front of the computer, the long nights of study, and the weekends devoted to this work. For being my support, for encouraging me, and also for helping me stay grounded in the moments when I needed it most.

To my family, Rafael, Carolina, Ignacio, Rodrigo, Pilar, and Milagros, for their unconditional support, not only throughout this process, but throughout my entire life. For encouraging me to keep growing and learning. For showing me that, in the face of adversity, one can find the strength to keep moving forward and overcome difficulties.

To my grandparents, Tatá, Naná, Abu, and Nonó, to whom I dedicate this thesis, who have always asked me about my work and who, without fully understanding what I do, have always been proud of me. To Tatá in particular, who has been both a mentor and a friend to me. Thank you for your unconditional support and affection.

To Marina and Pablo, my advisors. Beyond supervising this thesis, they have been true mentors throughout my academic journey. From my undergraduate thesis up to now, they have consistently made time to contribute to my education, to meet with me, discuss ideas, and challenge the way I think. I am grateful for their guidance and for the trust they placed in me.

To Quentin Bammey, Rafael Grompone, and Lara Raad, for kindly accepting to serve on the defense committee for this thesis.

To Pento, where I work and continue to grow professionally, for strengthening my knowledge, shaping the way I work, and providing me with tools that have helped me continue growing and improving in my career.

To the School of Engineering at Universidad de la República, for the opportunity to pursue this master's degree, for the education it has given me throughout these years, and for being my academic home, one that I am proud to belong to.

To ANII, for the scholarship (POS_NAC_2023_1_178226) that supported this master's degree, and to ClusterUY, for the computing resources that made the experiments in this work possible.

And to everyone else who has been by my side throughout this process, thank you.

Agradecimientos (Español)

Probablemente, esta sea la sección más difícil de escribir. Si bien hacer una maestría y escribir una tesis demandan un gran esfuerzo, no es un camino que se recorra en soledad, sino que quienes nos rodean, de una forma u otra, también acompañan ese proceso. Por eso, me gustaría agradecer a todos aquellos que me acompañaron en este recorrido.

A Agustina, mi pareja, por estar a mi lado en la primera línea de batalla. Por acompañarme en esas tardes frente a la computadora, en las noches largas de estudio y en los fines de semana dedicados a este trabajo. Por ser mi sostén, por alentarme y también por centrarme en esos momentos en los que lo precisaba.

A mi familia, Rafael, Carolina, Ignacio, Rodrigo, Pilar y Milagros, por brindarme su apoyo incondicional, no solo en este proceso, sino a lo largo de toda mi vida. Por impulsarme a seguir creciendo y aprendiendo. Por mostrarme cómo, frente a la adversidad, uno puede encontrar la fuerza para seguir adelante y superar las dificultades.

A mis abuelos, Tatá, Naná, Abu y Nonó, a quienes dedico esta tesis, que siempre me han preguntado por mi trabajo y que, sin entender del todo a qué me dedico, siempre se han sentido orgullosos de mí. A Tatá, en especial, que ha sido un mentor y un amigo para mí. Gracias por tu apoyo y cariño incondicional.

A Marina y Pablo, mis tutores, que más allá de supervisar esta tesis, han sido verdaderos mentores a lo largo de mi recorrido académico. Desde la tesis de grado hasta ahora, se han hecho el tiempo de forma constante para aportar a mi formación, reunirse conmigo, discutir ideas y desafiar mi forma de pensar. Estoy agradecido por su guía constante y por la confianza que depositaron en mí.

A Quentin Bammey, Rafael Grompone y Lara Raad, por haber aceptado integrar el tribunal de defensa de esta tesis.

A Pento, donde me desarrollo profesionalmente, por fortalecer mis conocimientos, moldear mi forma de trabajo y brindarme herramientas que me han ayudado a seguir creciendo y mejorando en mi carrera.

A la Facultad de Ingeniería de la Universidad de la República, por la oportunidad de cursar esta maestría, por la formación que me ha dado en todos estos años y por ser mi casa de estudios, de la que me enorgullece formar parte.

A la ANII, por la beca (POS_NAC_2023_1_178226) que apoyó esta maestría y a ClusterUY, por los recursos de cómputo que habilitaron los experimentos de este trabajo.

Y a todos los demás que estuvieron a mi lado a lo largo de este proceso, gracias.

To my grandparents: Tatá, Naná, Abu and Nonó.

Para mis abuelos: Tatá, Naná, Abu y Nonó.

This page has been intentionally left blank.

Abstract

Modern generative models can produce photorealistic images that are difficult to distinguish from real photographs, posing major challenges for digital forensics and undermining trust in images as faithful depictions of reality. This thesis addresses that problem from three complementary perspectives: examining how real and synthetic images are produced, surveying the current landscape of AI-generated image detection, and developing a new color-based forensic method.

We first study the contrast between the physical camera pipeline and learned generative pipelines. Camera imaging pipelines systematically leave low-level traces in natural photographs, whereas generative models synthesize images through learned mappings that need not reproduce those physically grounded statistics. This asymmetry motivates the forensic search for signals tied to image formation rather than to semantic content.

Building on this foundation, we survey AI-generated image detection methods organized by evidence family: spatial-domain discriminators, frequency and residual cues, diffusion self-consistency, color-centric methods, frozen-representation detectors, and real-world forensic systems. The survey shows that no single evidence family is universally reliable and identifies inter-channel color dependencies as a comparatively underexplored gap.

We then investigate inter-channel color-space correlations as a forensic signal. Controlled perturbation experiments reveal that LPIPS responds inconsistently to changes in inter-channel dependence across color spaces. By computing local inter-channel correlation distributions for real photographs and multiple diffusion-era generators across RGB, Lab, HSV, and YUV, we demonstrate that these statistics carry generator-dependent signatures whose discriminative strength varies with the chosen representation. We translate these findings into CHROMA, a detection pipeline that augments a ResNet-50 backbone with Lab inter-channel correlation maps. Under a cross-generator benchmark, CHROMA achieves competitive performance with substantially more complex methods, while ablations confirm that the benefit is color-space-dependent and that even limited multi-generator supervision yields large improvements on previously weak generators.

The thesis also documents Spectral-Flow, an exploratory frequency-domain anomaly detection approach that produced negative results, as acquisition-dependent variability prevented cross-camera generalization, but whose failure directly motivated the pivot toward the color-correlation cues at the core of CHROMA.

Keywords: AI-generated image detection, deepfake detection, inter-channel correlations, color spaces, perceptual metrics, diffusion models, generative models.

Resumen (Español)

Los modelos generativos actuales pueden producir imágenes fotorrealistas muy difíciles de distinguir de fotografías reales, lo que plantea desafíos importantes para el análisis forense y debilita la confianza en las imágenes como representaciones fieles de la realidad. Esta tesis aborda ese problema desde tres perspectivas complementarias: estudia cómo se producen las imágenes reales y las sintéticas, releva el estado del arte en detección de imágenes generadas por IA y desarrolla un nuevo método de detección basado en el color.

En primer lugar, analizamos las diferencias entre la cadena física de formación de imagen de una cámara y los procesos de generación aprendidos. Los procesos de formación de imagen de las cámaras dejan de forma sistemática trazas de bajo nivel en las fotografías naturales, mientras que los modelos generativos producen imágenes mediante transformaciones aprendidas que no necesariamente reproducen esas estadísticas con base física. Esta asimetría motiva la búsqueda de señales forenses asociadas al proceso de formación de la imagen, y no sólo a su contenido semántico.

Sobre esa base, relevamos los métodos de detección de imágenes generadas por IA organizándolos según la familia de evidencia que explotan: detectores en el dominio espacial, señales frecuenciales y residuales, autoconsistencia en modelos de difusión, métodos centrados en el color, detectores basados en representaciones preentrenadas y sistemas pensados para escenarios de aplicaciones en el mundo real. Esta revisión muestra que ninguna familia de evidencia es universalmente confiable e identifica las dependencias intercanal de color como un aspecto comparativamente poco explorado.

Luego estudiamos las correlaciones entre canales en distintos espacios de color como señal forense. Experimentos controlados de perturbación muestran que LPIPS responde de forma inconsistente a cambios en la dependencia intercanal según el espacio de color considerado. Al calcular distribuciones locales de correlación intercanal en fotografías reales y en imágenes generadas por múltiples modelos de difusión en RGB, Lab, HSV y YUV, mostramos que estas estadísticas contienen firmas asociadas al generador cuya capacidad discriminativa varía con la representación elegida. A partir de estos hallazgos proponemos CHROMA, un sistema de detección que complementa una arquitectura ResNet-50 con mapas de correlación intercanal en Lab. En una evaluación de generalización entre generadores, CHROMA alcanza un rendimiento competitivo frente a métodos considerablemente más complejos, mientras que los experimentos de ablación confirman que el beneficio depende del espacio de color y que incluso una supervisión limitada con múltiples generadores produce mejoras importantes sobre generadores que antes resultaban difíciles de detectar.

La tesis también documenta Spectral-Flow, un enfoque exploratorio de detección de anomalías en el dominio frecuencial que produjo resultados negativos, ya que la variabilidad asociada al proceso de adquisición impidió la generalización entre cámaras. Sin embargo, ese resultado motivó directamente el giro hacia las señales de correlación de color que constituyen la base de CHROMA.

Palabras clave: detección de imágenes generadas por IA, detección de deepfakes, correlaciones intercanal, espacios de color, métricas perceptuales, modelos de difusión, modelos generativos.

Preface

This thesis was developed as part of the Master’s program in Data Science and Machine Learning (*Magister en Ciencia de Datos y Aprendizaje Automático*) at the Instituto de Ingeniería Eléctrica, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay, under the supervision of Pablo Musé (Universidad de la República) and Marina Gardella (Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli).

The work presented here addresses the problem of distinguishing camera-captured photographs from AI-generated images, drawing on image processing, camera imaging pipelines, and machine learning. It includes both a comprehensive survey of the detection landscape, organized by forensic cue family, and the development of a novel detection method. The rapid evolution of generative models throughout this research, from early diffusion systems to widely available text-to-image products, continuously reshaped the landscape we were studying. This made the problem both challenging and stimulating: any forensic approach had to contend with a moving target.

The thesis proposes CHROMA, a method that exploits inter-channel color-space correlations as forensic cues. The approach is grounded in how camera pipelines shape the statistical dependence between color channels, a process that generative models have no intrinsic reason to replicate. The survey and the experimental analyses reported here aim to contribute to a better understanding of how different forensic cue families complement each other, and how color statistics in particular can strengthen existing detection tools.

This work was partially supported by Agencia Nacional de Investigación e Innovación (ANII, Uruguay) grant POS_NAC_2023_1_178226. The experiments presented in this thesis were carried out using ClusterUY [64].

Prefacio (Español)

Esta tesis fue desarrollada en el marco de la maestría en Ciencia de Datos y Aprendizaje Automático en el Instituto de Ingeniería Eléctrica, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay, bajo la supervisión de Pablo Musé (Universidad de la República) y Marina Gardella (Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli).

El trabajo presentado aborda el problema de distinguir fotografías capturadas por cámara de imágenes generadas por IA, combinando procesamiento de imágenes, cadenas de formación de imagen en cámara y aprendizaje automático. Incluye tanto un relevamiento del estado del arte en detección, organizado por familia de trazas, como el desarrollo de un método de detección novedoso. La rápida evolución de los modelos

generativos durante esta investigación, desde los primeros sistemas de difusión hasta productos comerciales de texto a imagen, fue modificando continuamente el escenario bajo estudio. Esto hizo que el problema fuera a la vez desafiante y estimulante: cualquier enfoque forense debía enfrentarse a un objetivo en movimiento.

La tesis propone CHROMA, un método que explota correlaciones intercanal en espacios de color como trazas forenses. El enfoque se fundamenta en cómo las cadenas de procesamiento de cámara moldean la dependencia estadística entre canales de color, un proceso que los modelos generativos no tienen razón intrínseca para replicar. El relevamiento y los análisis experimentales reportados buscan contribuir a una mejor comprensión de cómo distintas familias de trazas se complementan entre sí, y cómo las estadísticas de color en particular pueden fortalecer las herramientas de detección existentes.

Este trabajo fue parcialmente financiado por la Agencia Nacional de Investigación e Innovación (ANII, Uruguay), subsidio POS_NAC_2023_1_178226. Los experimentos presentados en esta tesis fueron realizados utilizando ClusterUY [64].

Juan Pablo Sotelo Silva
Montevideo, 2026

Table of Contents

Acknowledgments	III
Abstract	VII
Preface	IX
1. Introduction	1
1.1. The detection landscape and why it matters	2
1.2. Color as an underexplored forensic axis	2
1.3. Objectives and hypothesis	3
1.4. Thesis contributions	3
1.5. Thesis organization	4
2. From Sensors to Generators: How Digital Images Are Produced	7
2.1. Digital image formation and camera pipelines	7
2.1.1. Raw acquisition	8
2.1.2. Demosaicing	9
2.1.3. Color correction	9
2.1.4. Image compression	10
2.2. Generative models	10
2.2.1. GANs	11
2.2.2. Variational autoencoders and latent-variable models	12
2.2.3. Normalizing flows	13
2.2.4. Diffusion and score-based models	14
2.2.5. Generative Models as Commercial Tools	16
2.2.6. Editing pipelines (inpainting, image-to-image, upscalers)	17
2.2.7. Training objectives and perceptual losses	18
2.3. Chapter summary	21
3. Detection of Artificially Generated Images: State of the Art	23
3.1. Problem setting and evaluation	23
3.1.1. Threat model and distribution shift	24
3.1.2. Benchmarks and evaluation protocols	24
3.1.3. Evidence sources and robustness strategies	24
3.2. Discriminative detectors and generalization	26
3.2.1. Spatial detectors and transferable cues	26
3.2.2. Training strategies for open-world generalization	27
3.2.3. Scaling generator diversity	31
3.2.4. Frozen representations for open-world detection	31
3.3. Forensic evidence beyond RGB appearance	32

Table of Contents

3.3.1. Frequency and residual cues	32
3.3.2. Diffusion self-consistency and reconstruction error	38
3.3.3. Color statistics and inter-channel cues	40
3.4. Real-world generalization: continual updates, localization, and explanation	42
3.4.1. Continual updates and online learning	42
3.4.2. Localization and explanation	45
3.5. Chapter summary	46
4. CHROMA	47
4.1. Introduction	47
4.2. Color fundamentals for forensics	48
4.2.1. Color representations	48
4.2.2. What shapes inter-channel correlations	50
4.3. Inter-Channel Correlations as Forensic Cues	51
4.3.1. LPIPS Sensitivity to Inter-Channel Dependence	52
4.3.2. Cross-Generator Distributions of Inter-Channel Correlations	53
4.4. From Forensic Cues to Actual Detection	55
4.4.1. Model and Input Representations	56
4.4.2. Evaluation Data and Protocol	57
4.4.3. Effect of Correlation-Augmented Inputs	57
4.4.4. Comparison to Prior Detectors	62
4.5. Chapter summary	64
5. Conclusions and Future Work	65
5.1. Summary of contributions and answers to the guiding questions	65
5.2. Limitations	66
5.3. Future work	67
5.3.1. Near-term directions	67
5.3.2. Medium-term directions	68
5.3.3. Long-term and exploratory directions	68
A. Spectral-Flow: An Exploratory Frequency-Domain Anomaly Detector	71
A.1. Motivation and hypothesis	71
A.2. Architecture overview	72
A.2.1. Stage 1: Spectral energy extraction	72
A.2.2. Stage 2: Autoencoder with contrastive shaping	74
A.2.3. Stage 3: Normalizing flow on frozen latents	74
A.3. Latent space analysis and negative results	74
A.4. Lessons learned	76
B. Datasets and Benchmarks Appendix	79
B.1. Synthbuster Dataset Details	79
B.1.1. Prompt Set	79
B.1.2. Uncropped Sample Images	79
C. Color Science Background	83
C.1. Perceptual and colorimetric foundations	83
C.1.1. Human color perception	83
C.1.2. Color spaces, device dependence, and gamut	84
C.1.3. CIE colorimetry: tristimulus values and chromaticity	84
C.1.4. Transfer functions and linear light	84

Table of Contents

C.2. Color spaces used in this work	85
C.2.1. RGB	85
C.2.2. CIELAB	85
C.2.3. HSV/HSL	86
C.2.4. YCbCr and luminance–chrominance families	86
References	87
List of Tables	95
List of Figures	96

This page has been intentionally left blank.

Chapter 1

Introduction

Modern generative models can synthesize photorealistic images that are, for most practical purposes, visually indistinguishable from camera-captured photographs. Text-to-image diffusion systems, now widely available through products such as Stable Diffusion, DALL-E, Midjourney, and Adobe Firefly, have turned high-fidelity image generation into a routine capability accessible to non-expert users [1, 59, 67, 83, 92]. The barrier to producing convincing synthetic imagery has dropped from specialized technical expertise to a text prompt, and the volume of AI-generated content circulating online continues to grow rapidly.

This accessibility carries significant societal risks. Synthetic images can be weaponized for disinformation campaigns, where fabricated photographs lend false credibility to misleading narratives in news, social media, and political discourse. They enable new forms of fraud, from forged identity documents and insurance claims to non-consensual intimate imagery. In journalism and legal proceedings, the evidentiary value of a photograph depends on the assumption that it records a real event; when that assumption can no longer be taken for granted, the integrity of visual evidence is undermined. More broadly, the mere awareness that any image *could* be synthetic erodes public trust in visual media as a whole, a phenomenon sometimes referred to as the “liar’s dividend,” in which even authentic imagery can be dismissed as fabricated [11, 95].

These threats motivate the development of reliable, automated forensic tools that can distinguish camera-captured photographs from AI-generated or AI-manipulated images. The challenge is fundamentally technical: when generated images become indistinguishable to the human eye, detection systems must find evidence that lies below the threshold of perception.

A digital photograph is the output of a complex physical measurement and processing chain, in which raw sensor data is captured by a color filter array (CFA), then demosaicing is applied, then white balance, color correction, tone mapping, and compression before becoming a stored image file. Each of these stages imprints systematic, largely content-independent traces on the resulting pixels [23, 54, 75]. Generative models, by contrast, learn a mapping from noise and conditioning signals to pixel values (or latent representations) through training objectives that are optimized for perceptual quality [24, 32, 83]. They have no sensor, no CFA, no physical illuminant, and no device-specific rendering chain. The traces they leave are shaped by architecture, loss functions, and training data rather than by optics and electronics.

This fundamental asymmetry between physical acquisition and learned synthesis motivates the forensic approach followed in this thesis. If camera pipelines and generative models produce systematically different low-level statistics, and if these differences

Chapter 1. Introduction

survive perceptual optimization, then they can serve as evidence for distinguishing real from generated images. The critical question is *where* to look and *how* to measure those differences.

1.1. The detection landscape and why it matters

The rapid advancement of generative models has produced a substantial body of work on AI-generated image detection. Existing detectors exploit multiple evidence families, including spatial artifacts, frequency/residual signatures, reconstruction self-consistency, color statistics, and frozen foundation representations [4, 9, 12, 15, 16, 20, 26, 52, 66, 98, 99]. The literature also shows that the training strategy (augmentation, generator diversity, bias control, and continual updates) can influence robustness as strongly as backbone design does [21, 26, 28, 70].

A central lesson from this body of work is that no single evidence family is universally reliable. Spatial detectors can overfit to texture shortcuts; spectral cues are fragile under resizing and recompression; reconstruction-based methods inherit the biases of their inversion model; and frozen-representation detectors trade generator-specific sensitivity for broad coverage. Different cue families are partially complementary, which motivates the search for additional, orthogonal forensic signals that can be fused with existing approaches to improve robustness.

Understanding the structure of this landscape, its major research lines, the strengths and failure modes of each evidence family, and the gaps that remain, is therefore a prerequisite for identifying where new forensic signals can make the largest contribution. This thesis devotes a full chapter (Chapter 3) to surveying the state of the art in depth, organized by evidence type and learning paradigm. That survey is not merely a literature review: it provides the analytical foundation for identifying the specific gap that the proposed method addresses, and its conclusions directly inform the design choices made in subsequent chapters.

1.2. Color as an underexplored forensic axis

Within this landscape, *color-centric cues* occupy a distinctive but comparatively underexplored position. Prior work has shown that chrominance residuals, palette structure, and color-distribution regularity carry forensic information [40, 50, 94]. These methods are typically lightweight and interpretable, and they are grounded in the observation that generators optimized with perceptual objectives tend to match luminance fidelity more closely than chrominance fidelity. However, most existing color-based approaches rely on marginal per-channel statistics or handcrafted relational transforms. The systematic role of inter-channel *joint* correlations, that is, how channels co-vary within local neighborhoods, and their dependence on the choice of color-space representation have remained largely unexplored.

This gap is significant because inter-channel dependencies are a direct consequence of camera imaging. Demosaicing couples neighboring pixels across channels through local interpolation; the color correction matrix (CCM) explicitly mixes channels via a 3×3 linear transform; white balance applies per-channel gains that, while preserving inter-channel correlation in isolation, determine the operating point on which the CCM and subsequent nonlinearities act; and tone mapping applies nonlinear functions that reshape correlation structure. Together, these operations produce characteristic inter-channel coupling patterns that are tied to the physical rendering chain. Generative models, whose pixel values emerge from learned transformations rather than from a

1.3. Objectives and hypothesis

sensor measurement, have no intrinsic mechanism to reproduce this coupling, unless it is present in the training data and effectively captured by the optimization objective.

A key insight that motivates this thesis is that widely used perceptual similarity metrics, notably LPIPS [106], do not uniformly penalize departures from camera-like inter-channel correlation structure. If the loss functions that guide generative training are insensitive to certain chromatic joint statistics, then generators can produce images that look perceptually convincing while still deviating from real photographs in measurable, color-space-dependent ways. These deviations constitute exploitable forensic traces.

1.3. Objectives and hypothesis

This thesis pursues two complementary goals. The first is *analytical*: to provide a structured understanding of the current detection landscape, mapping the major evidence families, their strengths, failure modes, and complementarities, so as to identify where new forensic signals are most needed. The second is *constructive*: to develop and evaluate a concrete detection method grounded in an underexplored cue family, namely inter-channel color-space correlations.

These goals are guided by the following hypothesis and three guiding questions.

Hypothesis. AI-generated images and camera-captured photographs differ systematically in their inter-channel statistical structure across color representations. These differences persist despite high perceptual realism and can be exploited as forensic evidence for distinguishing real from generated images.

Guiding questions.

1. *What are the major lines of research in AI-generated image detection, and what gaps remain?* We aim to provide a comprehensive, evidence-type-organized survey of the state of the art that identifies the specific limitations and complementarities of existing cue families, and that situates color-centric forensics within this broader landscape.
2. *Are perceptual similarity metrics sensitive to changes in inter-channel dependence in a consistent way across color spaces?* This question tests the first part of our hypothesis: whether LPIPS, as a representative perceptual objective, exhibits uneven sensitivity to correlation-structure perturbations when expressed in different color parameterizations.
3. *Do explicit inter-channel correlation descriptors improve AI-generated image detection under realistic cross-generator evaluation?* This question tests whether the gaps identified above translate into actionable forensic evidence, and whether augmenting a standard detection backbone with correlation maps yields measurable gains under a rigorous benchmark.

1.4. Thesis contributions

The contributions of this thesis are fourfold:

1. **Comprehensive survey of the detection landscape.** We provide a structured review of the state of the art in AI-generated image detection, organized by

Chapter 1. Introduction

evidence family: spatial-domain discriminators, frequency and residual cues, diffusion self-consistency approaches, color-centric methods, frozen-representation detectors, and real-world generalization concerns including continual updates, localization, and explanation. This survey maps the major lines of research, characterizes their strengths and failure modes, and identifies inter-channel joint correlations and their color-space dependence as a specific underexplored gap.

2. **Sensitivity analysis of perceptual metrics.** We provide controlled-perturbation evidence that LPIPS responds inconsistently to changes in inter-channel dependence across color spaces: it nearly collapses the distinction between correlated and decorrelated noise in the Lab color space, and reverses the expected ordering in RGB. This indicates that perceptual objectives may leave gaps in the extent to which generators are constrained to match a camera-like correlation structure.
3. **Characterization of generator-specific correlation signatures.** By computing local inter-channel correlation distributions for real photographs and multiple diffusion-era generators across RGB, Lab, HSV, and YUV, we find generator-dependent signatures whose discriminative strength varies substantially with the color representation. RGB and Lab provide the strongest cross-source separability, while YUV correlations can be nearly indistinguishable between certain generators and real images.
4. **Correlation-augmented detection.** We translate these analyses into a practical detection pipeline: augmenting a standard ResNet-50 backbone with Lab inter-channel correlation maps improves cross-generator performance over RGB-only inputs under a modest training budget, while remaining below the strongest large-scale baselines. The ablation study further indicates that the benefit is color-space-dependent and that modest multi-generator supervision can substantially improve previously weak generators, albeit with persistent failure cases on some commercial systems.

These contributions form the basis of a paper accepted at the International Conference on Pattern Recognition (ICPR 2026): “CHROMA: Detecting AI-Generated Images through Inter-Channel Color-Space Correlations”.

1.5. Thesis organization

The remainder of this thesis is organized as follows:

Chapter 2 establishes the technical foundations. We detail the camera pipeline, from raw acquisition to compression, and survey the principal families of generative models (GANs, VAEs, flows, diffusion models, and commercially available systems), contrasting the physical measurement chain of cameras with the learned computational pipelines of generators.

Chapter 3 provides a comprehensive survey of the state of the art in AI-generated image detection, organized by evidence type: spatial-domain discriminative detectors, frequency and residual cues, diffusion self-consistency approaches, color-centric methods, frozen-representation detectors, and systems addressing real-world generalization through continual updates, localization, and explanation. It reviews the main detection paradigms, highlights their complementarities, and identifies the specific gap, namely the underexploration of inter-channel joint correlations and their color-space dependence, that the proposed method addresses.

Chapter 4 develops the CHROMA framework. We first establish the color-space foundations needed for correlation-based forensics, including color representations,

1.5. Thesis organization

transfer functions, and inter-channel dependence. We then analyze LPIPS sensitivity to inter-channel dependence perturbations across color spaces, characterize cross-generator correlation distributions and quantify separability via Wasserstein distances, and finally evaluate correlation-augmented detection under a standard cross-generator benchmark, including ablation studies over color spaces, trainable transforms, and multi-generator training regimes.

Chapter 5 synthesizes the findings, discusses limitations (including the failure cases on certain commercial generators), and outlines directions for future work, including data scaling, pixel-level localization via spatially resolved correlation maps, robustness under post-processing, and fusion with complementary forensic evidence.

Appendix A documents Spectral-Flow, an exploratory frequency-domain anomaly detector that preceded the inter-channel correlation approach. Motivated by the spectral and residual cues surveyed in Chapter 3, this experiment attempted to detect generated images by learning a normalizing-flow density model over spectral-residual features of real photographs, rather than training a binary classifier. The approach did not yield competitive detection performance: the learned notion of “normality” proved too sensitive to acquisition conditions, failing to generalize across camera pipelines. We include this negative result because it highlights a fundamental limitation of one-class spectral approaches and because it motivated the search for alternative traces, leading to the discriminative color-correlation cues that became the central contribution of this thesis.

This page has been intentionally left blank.

Chapter 2

From Sensors to Generators: How Digital Images Are Produced

This chapter provides the technical foundation for the forensic problem addressed in this thesis: distinguishing camera-captured photographs from images synthesized by generative models. We first detail the digital image formation process and the standard camera pipeline, identifying how each stage, from raw acquisition to final compression, imprints characteristic artifacts on natural photographs. We then survey the generative models that constitute the current threat landscape, emphasizing how their synthesis pipelines differ from camera capture and which architectural choices shape low-level statistics. Together, these two accounts frame the asymmetry between physical acquisition and learned synthesis that forensic methods seek to exploit.

2.1. Digital image formation and camera pipelines

The study of digital image forensics relies on the fact that a photograph is not merely a collection of pixels, but the result of a complex physical measurement, subject to stochastic noise sources, followed by a sequence of processing steps. This sequence, known as the camera pipeline or Image Signal Processor (ISP), varies between manufacturers but follows a broadly consistent architecture. Crucially, each stage of this pipeline leaves systematic, often content-independent traces that serve as a “digital fingerprint” of the acquisition device.

Figure 2.1 sketches the acquisition-to-storage story we will refer back to: raw acquisition at the sensor (where each pixel measures a single color component under the Color Filter Array layout), reconstruction via demosaicing, a set of color and tone adjustments that map sensor measurements to a displayable color space, and finally storage in a compressed file format. Many additional ISP blocks exist in practice; we focus on these core stages and introduce further details only when they are directly relevant to the forensic cues studied later [23, 54, 75]. The remainder of this section walks through the main stages with an emphasis on intuition and on the kinds of artifacts they can leave behind.



Figure 2.1: Simplified camera pipeline showing the main stages from sensor acquisition to compressed image storage.

2.1.1. Raw acquisition

Raw acquisition is the step where incident light is converted into an electrical signal and then digitized into a sensor measurement (typically stored in a camera-native RAW format). In the dominant single-sensor camera design, this RAW measurement is already sampled through a *color filter array* (CFA): each pixel location records only one color component (R, G, or B) according to the CFA layout [34]. That structure is not an incidental implementation detail: it is a design choice motivated by sensor cost and by human vision (e.g., Bayer dedicates more samples to green because luminance detail is largely carried there).¹

This is the most “physical” part of the pipeline: exposure time, aperture, and sensor gain (ISO) determine how many photons are collected and how strongly the resulting charge is amplified before analog-to-digital conversion.

Crucially for forensics, this measurement is noisy. Shot noise is signal-dependent (its variance grows with intensity), while electronic read noise is closer to signal-independent; both vary from capture to capture and depend on the acquisition settings. In addition to these random components, sensors also exhibit fixed pixel-to-pixel response differences, often discussed as PRNU: a camera-specific, multiplicative pattern. While conceptually distinct from shot/read noise, PRNU manifests as a noise-like signal in noise residuals. More structured sensor artifacts can also be present at the RAW level (e.g., dark-signal non-uniformity, row/column noise or banding, and black-level offsets), further tying the measurement statistics to a particular device and operating regime.

A common statistical model of RAW measurements combines these effects, typically using a Poisson component for photon shot noise and a Gaussian component for read noise, together with quantization and offset terms. Because the variance of the Poisson component depends on the signal level, while the Gaussian component and gain parameters depend on the sensor and its electronics, the resulting noise statistics vary with both the camera hardware and the capture settings [2, 54]. Although later ISP stages may attenuate or reshape these patterns, acquisition noise, fixed-pattern behavior, and gain characteristics can still leak into the final image and serve as a weak but characteristic fingerprint of real capture.

The Bayer array shown in Figure 2.2 is the most common CFA design, but other layouts exist (e.g., X-Trans), and different layouts can imply different artifacts.

¹There are exceptions to the single-sensor CFA pipeline (e.g., three-sensor prism cameras or layered sensors), but CFA-based sampling remains the dominant design in mainstream consumer imaging.

2.1. Digital image formation and camera pipelines

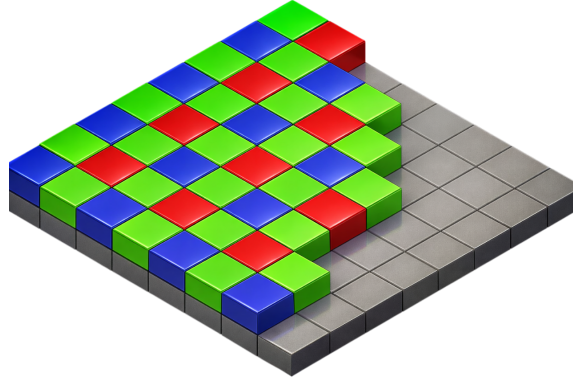


Figure 2.2: Bayer color filter array (CFA) sampling pattern (example). Single-sensor CFAs measure only one color per pixel and therefore require demosaicing to reconstruct full RGB at each pixel.

2.1.2. Demosaicing

To obtain a full-color image, cameras estimate the missing color components at each pixel (demosaicing). In its simplest form, demosaicing interpolates missing values from nearby samples; more advanced variants try to avoid “blending across edges” and to suppress color moiré and zipper artifacts. Regardless of the exact algorithm, demosaicing necessarily couples neighboring pixels (through local interpolation); and many practical demosaicing methods also exploit cross-channel correlations (e.g., using green as a structural guide), which can induce subtle inter-channel dependencies and CFA-periodic artifacts [55].

This coupling is useful to keep in mind later: it can create predictable spatial and inter-channel regularities that are often easier to characterize statistically than high-level semantic content. For this reason, CFA/demosaicing traces frequently appear in forensic pipelines, and inconsistencies in these traces have been used as evidence of manipulation [5, 73].

2.1.3. Color correction

After demosaicing, the image is still expressed in a camera-dependent sensor space. To produce something that looks “right” on screens, cameras apply a combination of color and tone processing: white balance (WB) (to compensate for the scene illuminant), a mapping from sensor RGB to a target color encoding (often approximated as a 3×3 color correction matrix, CCM), and typically nonlinear steps such as per-channel tone curves and gamma-like transfer functions [75].

In a first-order approximation, color correction can be modeled as a per-pixel linear transform, $\mathbf{x}_{\text{out}}(p) = \mathbf{M} \mathbf{x}_{\text{in}}(p)$, so that each output channel is a weighted combination of the input RGB values and the inter-channel correlations are directly affected. In practice, however, this linear stage is often embedded in a broader rendering stack that may include additional nonlinear mappings (e.g., channel-wise curves and, in some systems, LUT-like behavior). Consequently, the induced dependencies can be intensity- and content-dependent rather than purely global and linear. Even so, the overall color-rendering process leaves a characteristic footprint in inter-channel correlations and in the relative behavior of luminance-like and chrominance-like components, which later chapters exploit.

Pipeline stage	What it does	Potential forensic impact
Raw acquisition	Captures CFA-sampled sensor measurements (RAW) under gain + noise	Noise + gain behavior; CFA sampling grid and signal-dependent variance
Demosaicing	Reconstructs missing color components from CFA-sampled RAW	Interpolation artifacts; induced spatial/cross-channel correlations
Color correction	Maps sensor RGB to a target space via a CCM and tone processing	Explicit channel mixing; changes inter-channel dependence
Compression (JPEG/Web)	Block/transform coding + quantization (often with chroma subsampling)	Quantization / blocking; alters fine-scale statistics

Table 2.1: Summary of key camera-pipeline stages and examples of how they can affect low-level image statistics relevant to forensic analysis.

2.1.4. Image compression

Finally, images are typically stored and shared in a compressed format such as JPEG. In the common baseline JPEG pipeline, RGB is converted to a luminance–chrominance representation (typically YCbCr), the chrominance components may be stored at lower spatial resolution than the luminance component (often 4:2:0 chroma subsampling), and each component is then coded using an 8×8 block DCT transform followed by quantization and entropy coding. These operations are designed for compactness, but they also reshape the image: high-frequency details are attenuated, block boundaries can become visible, and rounding effects create distinctive quantization footprints.

Forensics often has to work “through” this final layer. Compression can weaken traces created earlier in the pipeline, but it can also add its own artifacts that help explain inconsistencies (e.g., double compression) or that change the statistics a detector should expect from genuine photographs [97].

While JPEG dominates consumer photography, other formats appear in practice. TIFF is often used as a high-fidelity container (possibly uncompressed or losslessly compressed), PNG is lossless and common on the web, and WebP is increasingly used in web pipelines with both lossless and lossy modes [14,25,27]. In addition, HEIF/HEIC is common in smartphone ecosystems, and AVIF is increasingly used on the web as a modern codec/container option. These choices influence which acquisition artifacts survive into the stored file and which new compression footprints are introduced.

Together, these operations progressively reshape the sensor measurements and introduce statistical dependencies in the image. Such transformations are central to forensic analysis because they can leave characteristic traces in low-level statistics. Table 2.1 summarizes the main pipeline stages and describes their potential forensic impact.

2.2. Generative models

The preceding section characterizes the acquisition pipeline of real photographs, where the sensor, CFA sampling, ISP processing, and compression collectively leave systematic traces. In contrast, modern generative models synthesize images through a learned computational pipeline (often a neural mapping from noise and conditioning

2.2. Generative models

signals to pixels or to a latent representation), and therefore do not naturally inherit camera-specific acquisition artifacts unless such traces are explicitly simulated or introduced by subsequent post-processing. This difference in “upstream” mechanisms motivates the generative-model overview below and helps clarify which low-level cues are expected to be present in real captures but absent, weakened, or qualitatively different in synthesized images. Generative modeling aims to learn a probability distribution $p_\theta(\mathbf{x})$ over high-dimensional observations (e.g., images) such that new samples $\tilde{\mathbf{x}} \sim p_\theta$ resemble draws from the (unknown) data distribution p_{data} . Contemporary deep generative models can be broadly grouped into (i) *implicit* models that provide a sampling procedure without an explicit tractable likelihood (notably GANs), and (ii) *likelihood-based* or *score-based* models that define (exactly or approximately) a probabilistic data-generation process (notably diffusion and score-based models). These families differ in training objectives, sampling cost, controllability, and failure modes; nevertheless, they share a common view of generation as learning a structured mapping from noise (and optional conditioning information) to samples.

2.2.1. GANs

Generative Adversarial Networks (GANs) formulate generation as a two-player minimax game between a generator G_θ and a discriminator D_ϕ [24]. The generator transforms a latent variable $\mathbf{z} \sim p(\mathbf{z})$ (commonly Gaussian or uniform) into a synthetic sample $G_\theta(\mathbf{z})$, while the discriminator predicts whether an input is real or generated. Figure 2.3 illustrates this adversarial training setup.

The original GAN objective is

$$\min_{\theta} \max_{\phi} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} [\log D_\phi(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [\log (1 - D_\phi(G_\theta(\mathbf{z})))] . \quad (2.1)$$

Under idealized assumptions, this objective corresponds to minimizing a divergence between p_{data} and the model distribution induced by G_θ [24]. In practice, GAN training can be unstable due to non-convex/non-concave dynamics, sensitivity to hyperparameters, and phenomena such as mode dropping (often discussed as mode collapse); additionally, practitioners often adopt the non-saturating generator loss to improve gradient signal early in training.

A major line of work improves stability by modifying the divergence or distance being optimized. Wasserstein GAN (WGAN) replaces the discriminator with a *critic* and optimizes an approximation to the Wasserstein-1 (Earth-Mover’s) distance, yielding smoother gradients when the supports of p_{data} and p_θ do not overlap [3]. Enforcing the required Lipschitz constraint is central; WGAN-GP introduces a gradient penalty that empirically improves stability and sample quality relative to weight clipping [29]. Architectural advances (e.g., convolutional inductive biases in DCGAN [77]) and large-scale training strategies (e.g., BigGAN [8]) further improve fidelity and diversity in class-conditional settings.

Conditioning mechanisms extend GANs to controllable generation. Conditional GANs incorporate side information (e.g., class labels or structured inputs) into both G_θ and D_ϕ [61]. A prominent example is image-to-image translation with conditional adversarial networks (pix2pix), which learns a mapping from an input image domain to an output domain while simultaneously learning an adversarial loss tailored to the task [37]. In parallel, style-based generator designs (StyleGAN, StyleGAN2) reorganize the generator around learned intermediate latent representations and scale-specific modulation, enabling finer control over attributes and yielding high visual quality [44, 45]. Beyond sampling, GANs also support *latent-space* operations (e.g., interpolation and attribute directions), which are frequently used to enable editing workflows in the GAN paradigm.

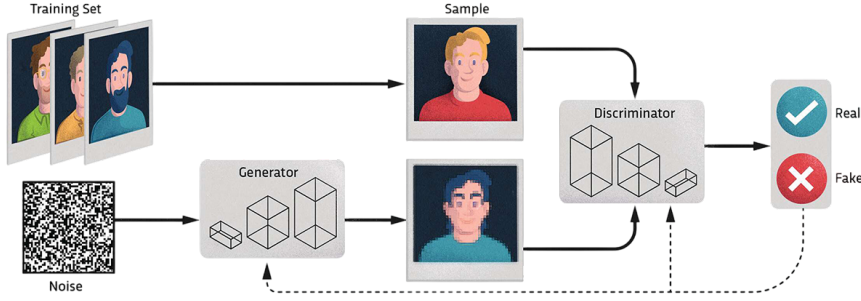


Figure 2.3: Schematic overview of a Generative Adversarial Network (GAN): the generator produces synthetic samples, while the discriminator learns to distinguish them from real data. Extracted from [6].

2.2.2. Variational autoencoders and latent-variable models

Variational autoencoders (VAEs) provide a likelihood-based approach to deep generative modeling by introducing latent variables that capture structured variation in the data [47, 81]. The model assumes a prior over latent codes $p(\mathbf{z})$ (typically standard normal) and a conditional likelihood or decoder $p_\theta(\mathbf{x} | \mathbf{z})$, e.g.,

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad p_\theta(\mathbf{x} | \mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_\theta(\mathbf{z}), \sigma^2 \mathbf{I}), \quad (2.2)$$

where $\boldsymbol{\mu}_\theta$ is a neural network. Exact maximum-likelihood learning is intractable because it requires marginalizing over $\mathbf{z}$. VAEs address this via variational inference with an *encoder* or *inference network* $q_\phi(\mathbf{z} | \mathbf{x})$, which approximates the intractable posterior and is amortized across data points [47, 81].

Training maximizes a variational lower bound (evidence lower bound, ELBO) on the log-likelihood:

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z} | \mathbf{x})} [\log p_\theta(\mathbf{x} | \mathbf{z})] - \text{KL}(q_\phi(\mathbf{z} | \mathbf{x}) \| p(\mathbf{z})) \triangleq \mathcal{L}_{\text{ELBO}}(\theta, \phi; \mathbf{x}), \quad (2.3)$$

which decomposes into a reconstruction term and a regularizer that encourages the approximate posterior to stay close to the prior. The reparameterization trick expresses samples from $q_\phi(\mathbf{z} | \mathbf{x})$ as differentiable transformations of noise (e.g., $\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$), enabling low-variance gradient estimates through stochastic backpropagation [47, 81].

From a forensic standpoint, the main relevance of the VAE framework is architectural: the encoder–decoder organization introduced here reappears as the autoencoder backbone of latent diffusion models (Section 2.2.4), where the decoder’s reconstruction fidelity directly shapes the low-level statistics of the final generated image.

From a modeling perspective, VAEs strike a balance between expressivity and tractability. The decoder $p_\theta(\mathbf{x} | \mathbf{z})$ provides an explicit (often factorized) likelihood, while

2.2. Generative models

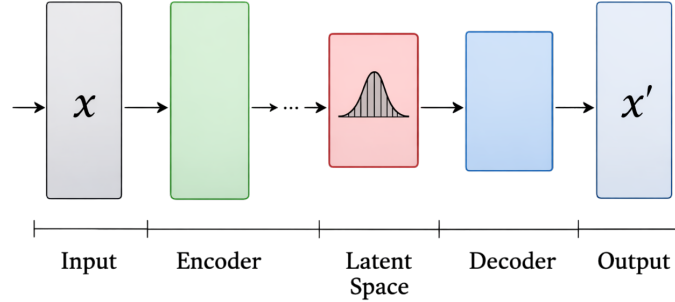


Figure 2.4: Canonical Variational Autoencoder (VAE) architecture: an encoder maps the input to a latent representation, a latent variable is sampled from the approximate posterior, and a decoder reconstructs the image from that latent code. Extracted from [102].

the encoder $q_\phi(\mathbf{z} | \mathbf{x})$ yields fast approximate inference. The KL term shapes the structure of the latent space: stronger regularization (e.g., via a weighting factor $\beta > 1$ in so-called β -VAEs) encourages more disentangled and factorized latent representations at the cost of reconstruction fidelity [31]. Such disentangled factors can support controllable manipulation of attributes in latent space, interpolation between examples, and compact representations for downstream tasks.

Early VAE-based image models often produced comparatively blurry samples when trained with simple Gaussian decoders and pixelwise reconstruction terms, reflecting a tendency to average over multiple plausible high-frequency configurations. Subsequent work introduced more expressive decoders and alternative parameterizations of the latent variables. Vector-quantized VAEs (VQ-VAEs) replace continuous latent variables with discrete codes drawn from a learned codebook, enabling powerful hybrids where an autoregressive or transformer model operates in a lower-dimensional discrete latent space while the VQ-VAE handles mapping to pixel space [78, 96]. Such latent-variable schemes underpin many recent high-fidelity models and are closely related to the autoencoder backbones used in latent diffusion systems [83]. This encoder–latent–decoder organization, shown schematically in Figure 2.4, is the canonical architectural template underlying VAEs.

2.2.3. Normalizing flows

Normalizing flows are a likelihood-based family of generative models that construct an expressive density by applying a sequence of *invertible* transformations to a simple base distribution [69, 80]. As illustrated in Figure 2.5, a simple base distribution $p_0(\mathbf{z}_0)$ (e.g., a standard Gaussian) is progressively transformed through a chain of bijections $f_1, f_2, \dots, f_K$ into a complex target distribution $p_K(\mathbf{z}_K)$. Let $\mathbf{z} \sim p_Z(\mathbf{z})$ denote a base variable and define a bijection $f_\theta : \mathbf{z} \mapsto \mathbf{x}$; by the change-of-variables formula, the induced data density is

$$p_\theta(\mathbf{x}) = p_Z(f_\theta^{-1}(\mathbf{x})) \left| \det \frac{\partial f_\theta^{-1}(\mathbf{x})}{\partial \mathbf{x}} \right|, \quad (2.4)$$

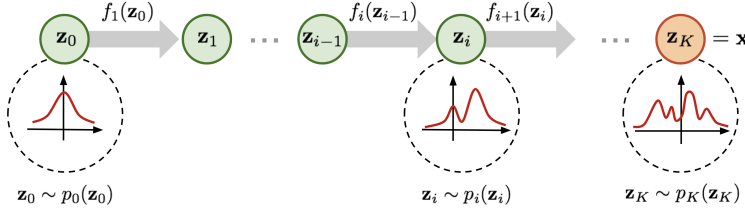


Figure 2.5: A normalizing flow transforms a simple base distribution $p_0(z_0)$ into a complex target distribution $p_K(z_K) = p(x)$ by composing a sequence of invertible mappings $f_1, \dots, f_K$. Each intermediate variable z_i follows a progressively more expressive distribution. Adapted from [101].

where the Jacobian determinant accounts for the local volume change of the transformation. In practice, f_θ is implemented as a composition of simple differentiable bijections $f_\theta = f_K \circ \dots \circ f_1$, so that both inversion and the log-determinant Jacobian can be computed efficiently layer by layer. Flow architectures therefore trade off expressivity and tractability by designing layers with structured Jacobians (triangular, block-diagonal, or otherwise easy to evaluate), such as affine coupling layers in RealNVP [19] and invertible 1×1 convolutions in Glow [46]. A key advantage over many other deep generative models is that maximum-likelihood training is straightforward (exact log-likelihood is available), while sampling and latent inference are also exact via forward and inverse passes through the network.

Flows for anomaly and out-of-distribution detection. Because flows provide tractable likelihoods, a natural baseline for anomaly detection is to score inputs by their log-likelihood under the model, flagging low-likelihood samples as anomalous. This idea is appealing in domains where anomalies correspond to low density under the training distribution (e.g., sensor faults or rare patterns). However, for high-dimensional natural images, likelihood-based out-of-distribution (OOD) detection is subtle: Nalisnick et al. [62] showed that deep generative models, including flows, can assign *higher* likelihood to semantically out-of-distribution datasets than to the training data (e.g., a model trained on CIFAR-10 assigns higher likelihood to SVHN). Kirichenko et al. [48] traced this failure to the fact that coupling-layer flows learn local pixel correlations and generic image-to-latent transformations rather than semantic structure, allowing them to compress any structured image efficiently. One explanation involves the mismatch between regions of high probability density and the *typical set* in high dimensions; typicality-based tests [63] and likelihood-ratio methods that correct for background statistics [79] have been proposed to address these confounds. In practice, flows can still be useful for anomaly detection when paired with such corrections, when applied to learned semantic representations rather than raw pixels, or in lower-dimensional or more structured domains where the likelihood-as-anomaly-score assumption holds more reliably.

2.2.4. Diffusion and score-based models

Diffusion probabilistic models define generation as the reversal of a gradual noising process. Denoising Diffusion Probabilistic Models (DDPMs) construct a forward Markov chain that adds Gaussian noise to data over T steps:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad (2.5)$$

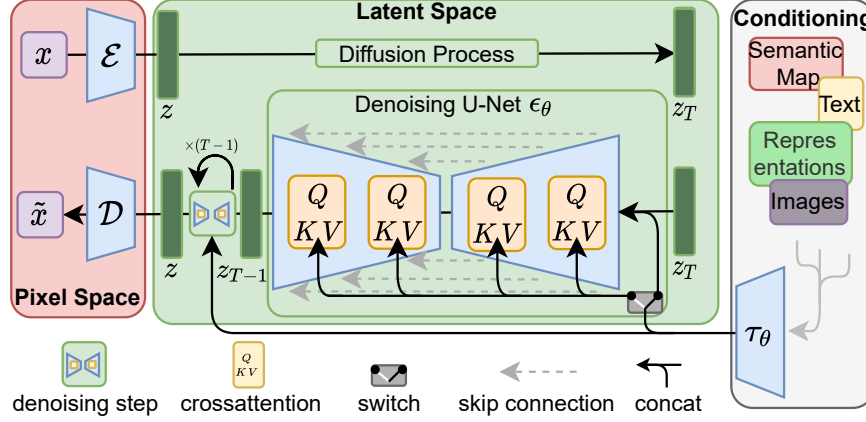


Figure 2.6: Latent diffusion model (LDM) pipeline, as used in Stable Diffusion: the image is first compressed into a lower-dimensional latent representation by an autoencoder, diffusion sampling is performed in that latent space, and the final latent is decoded back to pixel space. Extracted from [83].

where $\{\beta_t\}_{t=1}^T$ is a variance schedule [32]. The generative model learns a reverse process $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$, typically parameterized by a neural network that predicts either the added noise, the denoised sample, or a related quantity. Sampling begins from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively denoises to obtain $\mathbf{x}_0$. DDPMs connect closely to denoising score matching: learning to denoise at multiple noise levels corresponds to learning gradients of log-densities of perturbed data distributions [32, 90].

Score-based generative models explicitly target the *score function* $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ (or its noise-conditioned variants) and synthesize samples by running Langevin-type dynamics using an estimated score field [90]. A unifying perspective casts both diffusion and score-based models in continuous time via stochastic differential equations (SDEs): a forward SDE transforms data into noise, and a reverse-time SDE, defined using the score, maps noise back into data [91]. This framework yields multiple samplers, including predictor-corrector methods and deterministic probability-flow ODE variants, offering trade-offs between speed and fidelity [91].

A practical limitation of diffusion-family models is the computational cost of iterative sampling. Several approaches accelerate inference without changing the training objective. Denoising Diffusion Implicit Models (DDIM) define a non-Markovian implicit process that can generate samples in fewer steps while using the same trained denoiser [88]. For conditional generation, *guidance* methods modulate the reverse dynamics to trade off diversity and adherence to conditions. Classifier-free guidance (CFG) combines conditional and unconditional score estimates within a single model family, enabling strong conditioning without a separate classifier [33]. Finally, latent diffusion models move the diffusion process to a learned latent space (typically via an autoencoder), substantially reducing compute while preserving high-resolution synthesis and enabling flexible conditioning through cross-attention [83] (see Figure 2.6).

Consistency models

Consistency models aim to enable high-quality generation in one or a few function evaluations by learning a family of denoisers that are *self-consistent* across noise levels (or time indices) [89]. In contrast to diffusion models that require iterative reverse-time

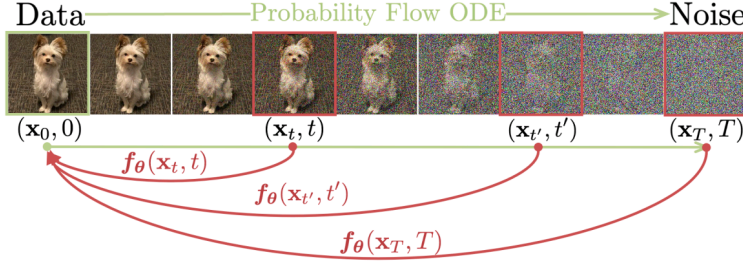


Figure 2.7: Conceptual overview of a consistency model: the same network is trained so that different noisy versions of a sample are mapped to a consistent prediction of the clean data, enabling generation in one or a few function evaluations. Extracted from [89].

sampling over many discretization steps, a consistency model parameterizes a mapping

$$\mathbf{x}_0 \approx f_\theta(\mathbf{x}_t, t), \quad (2.6)$$

where $\mathbf{x}_t$ denotes a noisy version of a sample at time (or noise level) t , and f_θ predicts a corresponding clean (or less noisy) estimate. The core idea is to enforce that predictions obtained from different noise levels agree with each other when the noisy inputs are related through the forward noising process. Concretely, for two time indices $s < t$, one may construct $(\mathbf{x}_s, \mathbf{x}_t)$ such that $\mathbf{x}_t$ is obtained by adding additional noise to $\mathbf{x}_s$, and impose a *consistency constraint* of the form

$$f_\theta(\mathbf{x}_t, t) \approx f_\theta(\mathbf{x}_s, s), \quad (2.7)$$

typically using an ℓ_2 -type regression loss on the predicted $\mathbf{x}_0$ (or an equivalent parameterization). By training f_θ to produce approximately invariant predictions across a schedule of noise levels, the model learns a stable estimator that can be evaluated at a single (large) noise level to directly obtain a sample, or can be run for a small number of steps for improved fidelity.

Consistency models can be learned in two principal ways [89]. First, *consistency training* learns f_θ directly via the self-consistency objective using paired noisy states. Second, *consistency distillation* trains f_θ to match the outputs of a stronger (typically slower) generative teacher, such as a diffusion model sampler, yielding a student that inherits sample quality while substantially reducing sampling cost. From a continuous-time viewpoint, consistency models are closely connected to probability-flow ODE formulations of diffusion/score models, and can be interpreted as learning an efficient approximation to the denoising map induced by such dynamics [89].

In practice, the primary contribution of consistency models is the reduction of sampling complexity from dozens–hundreds of denoising steps to one–few evaluations, thereby shifting the performance trade-off between sample quality and latency. However, the design of the time/noise schedule, the target parameterization (predicting $\mathbf{x}_0$, noise, or velocity-like variables), and the choice between direct training and distillation materially influence stability and fidelity [89].

This idea of a denoiser whose predictions remain approximately consistent across noise levels is illustrated conceptually in Figure 2.7.

2.2.5. Generative Models as Commercial Tools

The maturation of generative modeling has led to productized systems that expose generation and editing capabilities through interactive interfaces and APIs. In the

2.2. Generative models

commercial setting, model development is closely coupled with constraints on usability, latency, safety, and licensing.

First, *prompt-driven* conditioning has become a dominant interaction paradigm. Text-to-image systems accept natural-language prompts and internally convert them into conditioning signals (often embeddings) that steer sampling; these products typically provide iterative refinement workflows (re-rolling samples, varying seeds, adjusting guidance/strength parameters) to help users navigate the stochasticity of generation. For example, OpenAI has released DALL-E 3 as an integrated image generation capability in ChatGPT environments [67]. In addition, OpenAI exposes `gpt-image-1` as a natively multimodal image model via the Images API, supporting programmatic text-to-image generation and instruction-based image editing conditioned on one or more reference images [68]. Similarly, Stability AI announced Stable Diffusion as a text-to-image model designed to run efficiently on consumer hardware, contributing to a broad ecosystem of local and hosted tools [92]. Adobe positions Firefly as a generative creative environment integrated with design workflows and multiple model providers [1]. Midjourney’s documentation illustrates a product orientation toward prompt engineering and user-facing controls for generation behavior, including image prompts and textual rendering constraints [59, 60].

Second, commercial systems commonly incorporate *guardrails* and policy layers that mediate requests and outputs. While implementation details vary, these layers may include prompt filtering, constraint-based decoding, content moderation, and provenance-related mechanisms. The presence of such layers implies that the externally observed behavior of a commercial tool reflects not only the underlying generative model but also the surrounding safety and product infrastructure [67].

Third, productization encourages modularity: the core generator is frequently complemented by specialized models for upscaling, inpainting, background removal, or style adaptation, yielding end-to-end pipelines rather than single-model solutions. This modular view aligns with the broader research trend of chaining generative components (e.g., base synthesis plus super-resolution) [83, 85].

2.2.6. Editing pipelines (inpainting, image-to-image, upscalers)

Beyond unconditional sampling, modern generative models support *editing* tasks where the goal is to transform a given input while preserving selected content and introducing plausible modifications. Three common pipeline primitives are inpainting, image-to-image translation, and super-resolution.

Inpainting. Inpainting fills regions specified by a binary mask while keeping unmasked regions consistent with the input context. Diffusion-based inpainting can be achieved by conditioning the reverse denoising iterations on known pixels: at each step, the algorithm enforces agreement with the observed (unmasked) region while sampling the unknown region from the generative prior [53]. Latent diffusion models also report strong performance on inpainting and related inverse problems through conditioning in latent space and flexible attention-based control [83]. From the score/SDE perspective, inverse problems (including inpainting) can be framed as sampling from a posterior defined by a prior score model and an observation operator, solvable with modified reverse-time dynamics [91].

Image-to-image translation (img2img). Image-to-image translation maps an input image to an output image under a specified transformation (e.g., structural-to-photographic, domain transfer, or guided stylization). GAN-based approaches such as pix2pix learn

Chapter 2. From Sensors to Generators: How Digital Images Are Produced

a conditional generator and adversarially trained loss function to model the conditional distribution of outputs given inputs [37]. Diffusion/SDE approaches can perform image inpainting by *partial noising* of an input image followed by denoising under a generative prior, which naturally balances faithfulness to the input and realism: SDEdit formalizes this as adding noise to the input according to an SDE and then simulating the reverse process to project the result onto the model’s data manifold [57]. This strategy avoids explicit GAN inversion and is broadly applicable across guidance types [57].

Upscalers (super-resolution). Super-resolution increases image resolution while synthesizing plausible high-frequency detail. GAN-based super-resolution (e.g., ESRGAN) leverages adversarial and perceptual losses to recover sharp textures, emphasizing visually pleasing reconstructions over pixelwise fidelity [100]. Diffusion-based super-resolution treats the high-resolution image as a conditional generative target given a low-resolution input; SR3 adapts diffusion models to iterative refinement conditioned on the low-resolution observation, producing strong perceptual quality and competitive human preference scores [85]. In practice, many systems employ cascaded pipelines: a base generator produces an initial image (possibly at moderate resolution), followed by one or more super-resolution stages to reach the target resolution [85].

2.2.7. Training objectives and perceptual losses

The visual realism of contemporary generative models is tightly linked to the training objectives used to fit $p_\theta(\mathbf{x})$ to p_{data} . Although the architectures of GANs, autoencoders, autoregressive models, and diffusion/score-based models differ substantially, their objectives can be grouped into three broad families: (i) likelihood- and reconstruction-based losses, (ii) adversarial losses, and (iii) feature- or perception-oriented losses. These objectives implicitly define which aspects of an image are matched most faithfully (global structure, fine textures, color statistics, etc.) and how realism emerges over the course of training.

Likelihood and reconstruction losses. Likelihood-based models, such as variational autoencoders (VAEs) and many autoregressive or diffusion variants, maximize (approximate) log-likelihood or a variational lower bound on it [32, 47, 81]. For continuous images, this is often implemented as an ℓ_2 regression loss on either pixels or noise variables. For example, DDPMs minimize a sum of mean-squared errors between the predicted and true noise at multiple diffusion timesteps [32], while VAEs pair a reconstruction term with a Kullback–Leibler (KL) regularizer on the latent variables [47, 81]. Purely pixelwise losses (e.g., ℓ_1 or ℓ_2 in RGB space) encourage accurate recovery of low-frequency content and global layout, but tend to average over plausible high-frequency configurations when the conditional distribution $p(\mathbf{x} \mid \text{latent})$ is multimodal. This can lead to overly smooth textures if used in isolation, motivating the addition of complementary objectives.

Adversarial losses and learned divergences. Adversarial training interprets generation as minimizing a learned discrepancy between p_θ and p_{data} [24]. Different discriminator designs and objectives correspond to different divergences or distances (e.g., Jensen–Shannon divergence in the original GAN, Wasserstein distance in WGAN [3], hinge losses in later variants). In practice, adversarial losses are often combined with reconstruction terms to balance sharpness and fidelity. For instance, image-to-image translation models such as pix2pix optimize a weighted sum of an ℓ_1 reconstruction loss and a conditional adversarial loss, where the discriminator operates at the patch

2.2. Generative models

or image level to enforce local realism [37]. Multi-scale and patch-based discriminators further encourage realistic textures and edges at different spatial scales, contributing to the high perceptual quality characteristic of modern GAN and diffusion-based super-resolution systems [85].

As training progresses, the adversarial game tends to first remove grossly unrealistic artifacts detectable by a relatively weak discriminator, and then refines higher-frequency details as the discriminator becomes more capable. The decreasing discriminator and generator losses over optimization therefore loosely correspond to a progression from coarse semantic plausibility to fine-grained perceptual realism, although the dynamics can be highly nonlinear and model-dependent.

Perceptual feature-space losses. To better align training objectives with human visual perception, *perceptual losses* compare images in the feature space of a pretrained network rather than at the pixel level. A common choice is to use intermediate activations of a classification network such as VGG-16/19 [87]. Johnson et al. [41] showed that optimizing an ℓ_2 loss between feature maps of generated and target images, combined with optional style or total-variation terms, yields reconstructions and stylizations that are substantially sharper and more visually convincing than those obtained with pixelwise losses alone. Such feature-space losses encourage the model to match semantic structure, spatial arrangements, and mid-level texture statistics captured by the pretrained encoder, while retaining flexibility in exact pixel values.

Beyond hand-crafted feature choices, learned perceptual similarity metrics explicitly calibrate feature distances to human judgments. LPIPS (Learned Perceptual Image Patch Similarity) learns a linear combination (or small network) on top of deep features to predict human-rated similarity between image patches, resulting in a metric that correlates strongly with human perception [106]. LPIPS is widely used as an evaluation metric and, in some cases, is directly optimized as part of the training loss. In these settings, the generator is driven to produce outputs that are not only close in feature space to reference images but also aligned with human notions of “looking the same” up to small permissible variations. This substantially improves the perceived quality of textures (e.g., hair, grass, and fabric) and local contrast.

In the examples of Figure 2.8, the ranking induced by MSE differs from the ranking induced by LPIPS. The blurred image remains close in average pixel error but is perceptually degraded (loss of high-frequency structure), whereas the darkened image produces a larger pixel shift while preserving most spatial content and therefore a lower LPIPS. This mismatch illustrates the core motivation for perceptual objectives in generative pipelines: pixelwise losses are informative for low-level fidelity, but feature-space losses better capture perceived similarity.

Perceptual objectives have also been extended to multi-modal settings. Joint text-image embedding spaces, such as those learned by CLIP [76], are widely used either as conditioning representations (e.g., frozen text encoders) or as auxiliary signals for guidance and evaluation. However, many modern text-to-image diffusion systems are trained primarily with denoising objectives conditioned on text embeddings, rather than by directly maximizing a CLIP cosine-similarity loss between generated images and text. In combination, likelihood-based, adversarial, and perceptual components provide complementary signals: the former anchor the model to the data distribution, while the latter shape fine-grained appearance and ensure that small changes in the objective correspond to perceptible differences in image quality.

Temporal structure of objectives. For iterative generative processes, the loss is naturally indexed by a time or noise variable t . In diffusion and score-based models, the



Figure 2.8: Empirical comparison between pixel and perceptual distances (all values computed against pope.png, sourced from [58]). Gaussian blur and dark color-correction examples are self-generated transformations. The face-swap example is produced with gpt-image-1 [68]. Gaussian blur yields very low MSE but relatively high LPIPS, while a mostly global illumination change yields higher MSE with lower LPIPS. A semantic edit (face swap) increases LPIPS strongly even with moderate MSE, illustrating why perceptual losses are useful complements to pixelwise criteria.

training objective is an expectation over timesteps, with per-step losses weighted according to the noise schedule [32, 90]. Early in training, the model learns to denoise at higher noise levels (where the task is easier and dominated by coarse structure), and later refines its performance at lower noise levels that control fine detail. Similarly, in consistency and distillation frameworks, the student network is trained to match teacher outputs across a range of times or noise scales [89]. These temporal organizations of the loss effectively implement a curriculum over difficulty and scale, contributing to the gradual emergence of globally coherent layouts followed by increasingly realistic local details as optimization proceeds.

Overall, the combination and scheduling of training objectives, including pixelwise reconstruction, probabilistic or score-based losses, adversarial terms, and perceptual or feature-space similarities, play a central role in the ability of modern generative models to synthesize images that are both statistically aligned with p_{data} and subjectively perceived as highly realistic.

2.3. Chapter summary

This chapter established the two image-production paradigms whose statistical contrast underpins the forensic approach of this thesis. First, we traced how the camera pipeline, from raw acquisition through demosaicing, color correction, and compression, imprints systematic, largely content-independent traces on every photograph; these traces are tied to physical sensors, interpolation algorithms, and device-specific rendering choices that generative models have no intrinsic reason to reproduce. Second, we surveyed the principal families of generative models (GANs, VAEs, flows, diffusion/score-based models, and commercially available systems), highlighting that their synthesis pipelines are driven by learned mappings and training objectives rather than by a physical measurement chain, so the low-level statistics they produce are shaped by architecture and loss design rather than by optics and electronics.

Together, these accounts motivate a forensic strategy that targets low-level statistical differences between the two paradigms. Chapter 3 surveys how existing detection methods exploit such differences and identify the remaining gaps.

This page has been intentionally left blank.

Chapter 3

Detection of Artificially Generated Images: State of the Art

The recent evolution of image generators, from GAN-based synthesis to diffusion and foundation-model pipelines, has turned AI-generated and AI-edited imagery into a routine practice, with such content being widely disseminated across online media.

A persistent challenge in the literature is the trade-off between *in-distribution performance* and *open-world generalization*. In response, recent work seeks more stable evidence. This includes spectrum and noise-residual properties linked to architectural periodicities, as well as diffusion self-consistency cues derived from inversion and reconstruction error. Other approaches rely on physically motivated color and chrominance statistics, or representation-centric methods that leverage frozen foundation encoders [15, 16, 50, 52, 66, 94, 98]. Because the generator landscape evolves continuously, some work on real-world generalization reframes detection as a continual-update problem [21]. A related line of research broadens the task beyond binary classification to include localization and explanation [35].

This chapter organizes the state of the art by the *type of evidence* and the *learning paradigm* used to extract it. We begin with spatial-domain discriminative detectors and training strategies aimed at generalization, then cover frequency and residual cues, diffusion-mechanism reconstruction approaches, color-centric cues, frozen-representation detectors, and finally real-world forensic systems that incorporate continual updates, localization, and natural-language explanations.

Supplementary dataset and benchmark material referenced in the evaluation discussion is collected in Appendix B.

3.1. Problem setting and evaluation

This section fixes the practical setting for real-world use of generated-image detectors. We summarize the threat model and expected sources of distribution shift, outline common benchmark and evaluation conventions, and introduce a cue-centric view of detection methods that will structure the rest of the chapter.

3.1.1. Threat model and distribution shift

A useful way to frame the problem is not simply to ask whether a detector can identify generated images, but to specify *which* forms of synthetic or manipulated content it must remain reliable on at real-world testing time. In practice, this requires a broader threat model in which the detector is expected to generalize beyond the generators seen during training, remain stable under common post-processing, and cope with partially edited images rather than only fully synthetic ones. Concretely, the detector must often handle:

Open-world generators: test images may come from *unseen* architectures, training corpora, or model families (e.g., GANs versus diffusion), so evaluation must emphasize generalization rather than performance on a fixed set of known generators [26, 70].

Post-processing: resizing, JPEG recompression, blur, noise, cropping, and platform-specific delivery pipelines can attenuate or overwrite the forensic traces on which a detector relies [21, 26].

Edits and partial synthesis: many realistic cases involve inpainting or localized modifications, which shifts the task from pure image-level classification toward *tamper localization* and mixed-content analysis [21, 35].

As a consequence, strong evaluation protocols explicitly separate train and test generators, control for trivial file-format or preprocessing cues, and report performance at operationally meaningful false-alarm regimes rather than only aggregate accuracy [26].

3.1.2. Benchmarks and evaluation protocols

Results are highly sensitive to dataset construction and protocol choices. Common best practices include:

Matching semantics and preprocessing: ensuring real and fake sets are comparable to avoid learning trivial acquisition or encoding differences [9, 26].

Benchmarking across families and time: evaluating on diverse GAN and diffusion sources, including closed commercial systems when possible [4, 21, 98].

Operational metrics: reporting AUC/EER alongside low-FPR operating points and fixed-threshold robustness under perturbations, reflecting screening use cases [26].

These considerations also motivate newer benchmarks that emphasize diffusion-era realism, cross-domain shift, and mixed real/synthetic content with localization and explanation labels [4, 35, 40].

3.1.3. Evidence sources and robustness strategies

Across the literature, approaches can be grouped by the primary cues they exploit:

Spatial-domain artifacts (textures, boundaries, local inconsistencies): learned with CNN/ViT backbones, often improved via patch-based analysis and augmentation to reduce reliance on semantics [9, 99].

Noise-residual and spectral signatures: periodic peaks, autocorrelation structure, and power-spectrum discrepancies linked to generator architecture and upsampling, sometimes framed as modeling the “real” distribution and treating synthetic as out-of-distribution [15, 43, 52].

3.1. Problem setting and evaluation

Mechanism-tied self-consistency (diffusion inversion/reconstruction): evidence derived from how an image behaves under diffusion inversion and reconstruction, emphasizing generator-mechanism invariances rather than spatial-domain artifacts [12, 98].

Color-centric statistics: chrominance disparities and inter-channel relations motivated by camera imaging vs direct RGB synthesis, including palette and distribution effects for diffusion-era images [40, 50, 94].

Representation-centric signals: rather than targeting explicit forensic artifacts, these approaches rely on frozen foundation encoders (e.g., CLIP) and operate in their embedding space, under the assumption that real and AI-generated images occupy systematically different regions of that space; detection is then implemented with lightweight decision rules such as linear probes, nearest-neighbor retrieval, or distance-to-distribution measures applied to fixed embeddings [16, 66].

A recurring empirical observation is that these cue families are often *partially complementary*, motivating simple fusion or ensembles when robustness is prioritized over minimal complexity.

Beyond architecture choice, generalization is strongly influenced by how the detector is trained:

Augmentation for real-world conditions: injecting JPEG, blur, resizing, and other transformations during training to match degradations often found in real-world applications [21, 26, 99].

Scaling generator diversity (data-centric robustness): increasing the number of distinct generators during training can be a key factor for cross-generator generalization, sometimes surpassing the gains that can be obtained through architectural novelty [70].

Ensembles and orthogonalization: training multiple learners on deliberately different splits to capture diverse forensic traces and reduce single-artifact failure modes [56].

Real-only approaches: learning a compact model of real-image statistics and rejecting outliers as synthetic, aiming to reduce dependence on any specific generator family [52].

Semantic alignment: generating fakes *from* real images via diffusion conditioning so real/fake pairs match in content and preprocessing, reducing shortcut learning and improving calibration under unseen generators [28].

Disentanglement and invariance objectives: explicitly separating content from (common vs specific) forgery fingerprints to avoid method-specific shortcuts [104].

Finally, detection is increasingly framed as a *system* problem:

- **Continual learning:** updating detectors as new generators appear, while maintaining stable performance on prior sources [21].
- **Localization:** identifying edited regions for inpainting and partial manipulations, often with pixel-level supervision or surrogate mixing strategies [21, 35].
- **Interpretability:** providing heatmaps, spectral masks, reconstruction-error maps, or natural-language rationales to support human decision-making [9, 35, 51].

These directions reflect the practical reality that forensic tools are used in broader moderation and verification pipelines, where actionable evidence and model-update strategies matter as much as peak offline accuracy.

3.2. Discriminative detectors and generalization

3.2.1. Spatial detectors and transferable cues

One starting point for open-world detection is the hypothesis that many generators leave shared low-level artifacts that a standard discriminative backbone can learn, provided the training protocol avoids shortcut cues. Wang et al. [99] challenge this view: contrary to prior observations that detectors often fail under cross-architecture or cross-dataset shift, their central hypothesis is that contemporary generators share systematic, low-level “fingerprints” that can be learned with a standard discriminative backbone when the training protocol is designed to avoid dataset bias and to emulate real-world image handling. Concretely, they train a standard discriminative CNN detector using synthetic images from a single generator, paired with real images chosen to be semantically and visually comparable, so as to reduce dataset shortcuts unrelated to the generation process. To evaluate generalization beyond the training generator, they introduce the *ForenSynths* benchmark and test generalization across diverse generator architectures and synthesis tasks under controlled real-fake pairing. A second methodological emphasis is robustness to downstream distribution shift: during training they simulate common post-processing operations such as blur and JPEG recompression alongside standard geometric augmentations, and they evaluate per-generator datasets separately to expose how content and post-processing affect generalization.

This optimistic generalization narrative motivates a more adversarial question: how robust are these detectors once the evaluation reflects real distribution shift and social-media-style processing. More recent in-the-wild evidence reinforces this concern: Karageorgiou et al. [42] show that current detectors already struggle on deceptive synthetic images circulating online, and that performance often degrades further as images accumulate reposting- and platform-induced transformations throughout their online lifespan. In a more controlled but already critical setting, Gragnaniello et al. [26] provide an experimentally grounded assessment of whether GAN-generated images are truly “easy” to detect in practice, arguing that much of the optimistic narrative stems from favorable and unrealistic assumptions such as well-aligned train/test data, absence of social-media processing, and availability of samples from the target generator. Rather than introducing a new detector, their main contribution is a *fair, controlled comparison* of representative detector families under realistic distribution shifts, with an emphasis on isolating which design choices actually remain stable once evaluation conditions depart from the training setup. To ensure comparability, all methods are trained on the *same* large-scale training set adopted from Wang et al. [99], pairing real images with GAN-generated samples so that performance differences reflect detector design rather than data advantages. Generalization is then evaluated on *unseen* generator architectures and tasks under a protocol that avoids “training-set reuse” confounds by separating the real sources used for training and testing. A central methodological emphasis is robustness to real-world conditions: the authors explicitly stress that resizing and JPEG compression can destroy the high-frequency evidence exploited by many detectors, motivating the evaluation protocol to include these perturbations and to use screening-oriented criteria such as performance at low false-alarm regimes rather than relying only on aggregate accuracy measured under clean, balanced, and otherwise favorable test conditions. Finally, they also study which design choices actually improve robustness and generalization. Using a strong CNN baseline trained with heavy augmentation, they show that both pre-training and augmentations that mimic post-processing play an important role.

Beyond robustness, it is also important to understand which image regions and

3.2. Discriminative detectors and generalization

cues actually drive transferable decisions, especially when localization is a downstream requirement. Chai et al. [9] aim to identify *which* cues make GAN-generated and manipulated images detectable *in a way that generalizes* across generators, datasets, and training conditions. The core methodological decision is to use a *patch-based* detector with an explicitly *limited receptive field*: a fully-convolutional classifier produces local real/fake predictions over an image grid, which can be aggregated to an image-level score while simultaneously yielding a spatial heatmap for localization and interpretation. With this setup, their initial analyses show that strong predictive patches are often concentrated in *background and hair-transition regions*, suggesting that early transferable artifacts are frequently boundary/texture inconsistencies outside the central face. They then train a generator adversarially against these patch-based predictions (i.e., adding a loss that pushes generated images to reduce patch-level “fakeness”), and, after this attack, observe a *focus shift* in newly trained patch classifiers: attention moves away from background cues and becomes more concentrated on *facial features and intra-face textures*. They further emphasize careful control of *spurious dataset cues*: apparent “generalization” can arise from preprocessing or file-format mismatches between real and synthetic images, so they enforce matched resizing/processing pipelines to encourage reliance on generator-induced artifacts rather than trivial encoding differences.

3.2.2. Training strategies for open-world generalization

A complementary set of methods targets open-world generalization by changing how detectors are trained and what representations they operate on, with the goal of reducing dependence on brittle, generator-specific artifacts.

One pragmatic approach is to build diversity into the detector itself, so that different learners capture partially different traces and failures do not align. Mandelli et al. [56] explicitly target the open-world case where test images originate from a previously unseen generator, such as StyleGAN3 excluded from training. Their main contribution is a two-part ensemble strategy: (i) *orthogonal training* of multiple CNN detectors so that different learners capture partially different forensic traces, and (ii) a *conservative aggregation* rule aimed at reducing missed detections on synthetic images. All detectors operate on local patches, but each CNN is trained on a deliberately different data split, so orthogonality arises from variation in semantic content, post-processing and compression conditions, and the set of generators seen during training. At inference, an input image is decomposed into many patches and evaluated by each CNN separately. Within each CNN, patch-level scores are first combined using an “any-patch” rule: if any patch is judged synthetic, that CNN assigns the image a synthetic label; otherwise it assigns a real label. In other words, the conservative rule is applied *within* each detector, not across the ensemble. This is motivated by the idea that synthetic evidence may appear only locally, whereas the real class is comparatively more stable and easier to model globally. The final ensemble decision is then obtained by averaging the image-level scores produced by the different CNNs. Under this design, diversity across detectors remains useful because the CNNs are not required to agree on the same traces or fail on the same images; rather, they contribute complementary evidence that improves robustness to unseen generators and common post-processing.

Another family of ideas aims to suppress semantics by moving away from raw RGB inputs and into transformed domains that emphasize process-related traces. Tan et al. [93] propose using gradients from a frozen pretrained model as detection inputs, with the goal of replacing a strongly data-dependent representation by one tied to a fixed *transformation model*. Concretely, a pretrained CNN is used to transform each image I into a gradient map rather than processing the RGB pixels directly. If $l = M(I)$

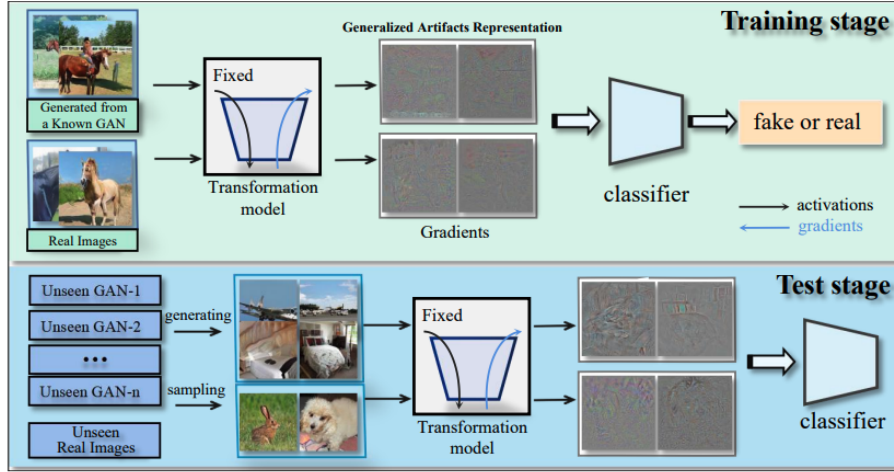


Figure 3.1: Overview of Learning on Gradients (LGrad) [93]. A frozen pretrained CNN is used as a transformation model: instead of feeding the RGB image directly to the detector, the method computes the gradient of the model output with respect to the input image and uses this gradient map as the artifact representation. The resulting normalized gradients are then passed to a binary classifier that predicts whether the image is real or GAN-generated.

denotes the output of the frozen model, the representation is defined as $G = \partial \sum(l) / \partial I$: the gradient of the summed output with respect to the input image. In practice, G can be read as a pixelwise sensitivity map showing which input locations most influence the model response. The detector is then trained on normalized versions of these gradient maps, under the hypothesis that they retain generation-related traces while reducing dependence on image content. Experimentally, they emphasize generalization protocols where the detector is trained using fakes from a *single known* GAN (ProGAN) and then evaluated under multiple “train on one setting, test on another” shifts that change either the semantic category, the generator family, or both. Using Class Activation Maps (CAM) [107], they observe a systematic difference between real and generated images: activations for real images concentrate on semantically meaningful regions such as faces, whereas activations for GAN-generated images predominantly appear in background regions. This is broadly consistent with the earlier observations of Chai et al. [9], who also report that transferable evidence often concentrates outside the central face. In the present case, the authors interpret this pattern as evidence that gradients capture high-level inconsistencies introduced by generative processes, rather than only low-level texture artifacts, supporting their claim that gradients act as a generalized representation of generation traces, as illustrated by the full LGrad pipeline in Figure 3.1.

Instead of changing the input domain, a different approach is to enforce invariances inside the representation by explicitly separating transferable traces from method-specific ones. UCF targets cross-forgery generalization in face deepfake detection by learning a representation that emphasizes *method-shared* forgery evidence while suppressing both content and method-specific shortcuts [104]. The key takeaway is that the detector’s decision is driven by a “common” forgery component intended to generalize across manipulations and datasets, rather than by cues that are predictive only for particular forgery methods. Figure 3.2 summarizes both the UCF architecture and its latent-space behavior under the FF++ analysis protocol: the top panel illustrates the

3.2. Discriminative detectors and generalization

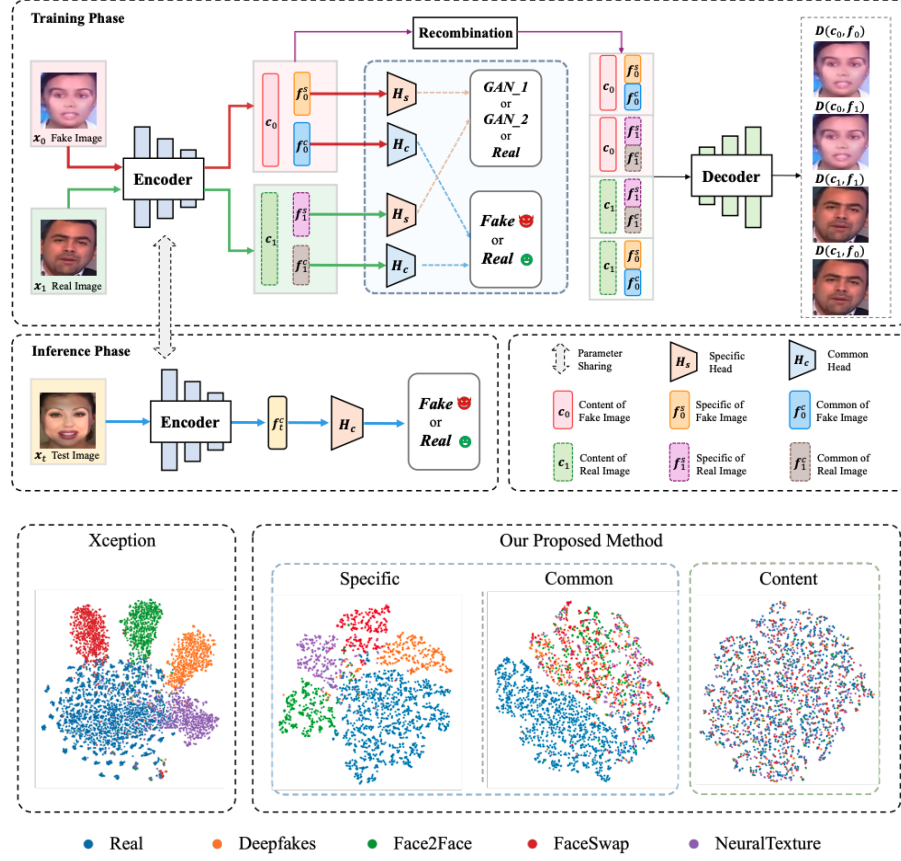


Figure 3.2: UCF architecture and latent-space behavior under the FF++ analysis protocol [104]. Top: overview of UCF (Uncovering Common Features) for generalizable deepfake detection. Bottom: t-SNE visualization on FF++ comparing latent features from baseline Xception and UCF; the baseline representation forms method-dependent clusters, while in UCF the specific branch remains method-discriminative, the common branch captures method-shared forgery evidence, and the content branch emphasizes forgery-irrelevant factors.

interaction between content, specific, and common forgery features during training and inference, while the bottom panel reports the paper’s t-SNE analysis showing that baseline Xception features cluster by forgery method (method-specific behavior), whereas UCF separates specific/common/content factors so that the common branch captures method-shared forgery cues and the content branch retains forgery-irrelevant information.

Relatedly, contrastive and multi-branch training can be used to impose explicit class structure on the embedding space, aiming to stabilize decisions across generator families. The DeepFeatureX-SN framework [72] addresses this with multi-encoder contrastive training that separates real, GAN, and diffusion sources more robustly. Its key methodological idea is to replace a single monolithic feature extractor with a *tri-partite* set of specialized feature extractors, each trained to emphasize discriminative traces for one reference class (REAL, GAN, or DM). Each Base Model is implemented as a Siamese CNN trained with contrastive learning on paired samples, encouraging

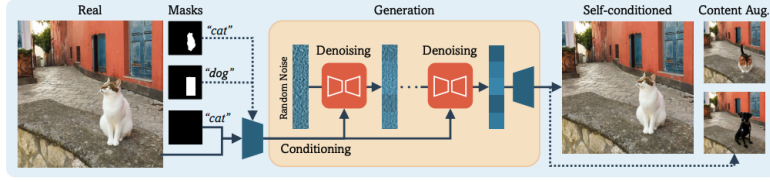


Figure 3.3: B-Free: bias-free training paradigm. Synthetic images are generated from self-conditioned transformations of real images and augmented with additional diffusion-based edits to avoid semantic biases; this yields training data where the only systematic difference between classes is the act of AI generation, improving generalization and calibration [28].

embeddings from the same class to collapse while enforcing separation from the other classes; after this stage, the three Base Models are frozen and used as fixed feature extractors. Given an input image I , the framework concatenates the three embeddings ($\Phi_{\text{REAL}}(I)$, $\Phi_{\text{GAN}}(I)$, $\Phi_{\text{DM}}(I)$) into a single descriptor, which is then processed by a lightweight classifier composed of stacked 1D convolutions followed by a final linear head to output three classes (REAL/GAN/DM), with a weighted cross-entropy objective to mitigate class imbalance.

A notable practical emphasis is placed on *data diversity and bias control*: the authors assemble a large multi-source corpus in a uniform lossless format (PNG) spanning numerous GAN and diffusion generators and a wide range of semantic categories and resolutions, and they define evaluation splits that explicitly separate *seen* vs *unseen* generative architectures for generalization testing. Moreover, they probe real-world shifts by testing under post-processing and geometric transformations, including JPEG compression (notably, their generalization test sets are JPEG-compressed while training is performed on raw images).

A recurring lesson from early generalization studies is that apparent detector robustness can be an artifact of the *training set construction* rather than of any truly generator-related forensic evidence: if real and synthetic images differ systematically in semantics, resolution, format, or preprocessing, a detector can achieve strong benchmark scores by exploiting these *spurious cues*.

This observation is turned into an explicit *training paradigm* in B-Free [28]. Instead of pairing real images with text-prompted generations (which can remain semantically misaligned), B-Free generates each fake *from the real image itself* through a self-conditioned diffusion pipeline, so that each real-fake pair is content-matched by construction, as summarized in Figure 3.3. This design preserves semantic alignment by construction while still injecting model-specific generation traces; importantly, the authors show that performing diffusion denoising/regeneration (not just autoencoder reconstruction) yields stronger and broader forensic discrepancies, including low-frequency inconsistencies that improve generalization to unseen generators.

To avoid overfitting to narrowly paired reconstructions, B-Free adds *content augmentation* through additional diffusion-based edits on top of these self-conditioned samples. The resulting corpus is then used to fine-tune a pretrained vision transformer detector end-to-end under a controlled sampling protocol intended to preserve subtle forensic cues. In ablations, this combination of self-conditioned pairing and augmentation is the key factor behind both higher cross-generator balanced accuracy and improved score calibration, supporting the paper’s central claim that *bias-free pairing is not only dataset hygiene, but a detector-training primitive* [28].

3.2. Discriminative detectors and generalization

3.2.3. Scaling generator diversity

At a larger scale, generalization can also be driven by the diversity of synthetic sources seen during training, sometimes more than by detector architecture. Park and Owens [70] focus on the core failure mode of prior fake-image detectors: poor generalization to images produced by previously unseen generators. Their main contribution is data-centric rather than architectural: they introduce *Community Forensics*, a training corpus that scales the *number of distinct generators* by systematically collecting community-created models, arguing that limited generator diversity is the primary bottleneck for generalization. Concretely, they download thousands of open-source text-to-image latent diffusion models from a model-sharing community, and complement this with a smaller curated set of open-source generators spanning different architectures (e.g., GANs and pixel-space diffusion) as well as several commercial systems. Generated images are prompted using captions drawn from multiple real-image datasets to diversify scene content; images are saved in their original formats (typically lossless PNG) and, where possible, avoided additional resizing/compression to reduce spurious format cues. For detection, they deliberately use standard discriminative backbones (ViT and ConvNeXt with CLIP- or ImageNet-style pretraining) trained end-to-end as binary classifiers, coupled with an augmentation pipeline that extends common post-processing simulation (blur and JPEG compression) to broader “in-the-wild” transformations such as padding, resizing, rotation, and shear. Their experiments are structured as scaling and ablation studies: they vary the number of training generators while controlling the image budget, compare diffusion-only training to more architecturally diverse subsets, test freezing versus finetuning pretrained backbones, and analyze whether strict semantic pairing between the prompt source data and generated images is necessary.

3.2.4. Frozen representations for open-world detection

This section examines a shift from training dedicated forensic backbones to *reusing frozen foundation representations* for generated-image detection. The common premise is that large pretrained encoders (especially CLIP-like vision-language models) provide a broad, semantically rich feature space that is less tied to any single fake domain, so detection can be implemented with lightweight heads, nearest-neighbor rules, or small reference sets rather than end-to-end optimization on synthetic artifacts. This design is motivated by open-world conditions in which only limited fake sources are available at training time and test images come from unseen generator families.

Within this paradigm, a key design choice is to keep the feature extractor fixed and move the decision rule into a generic embedding space. Ojha et al. [66] show that universal detection becomes feasible when the feature extractor is frozen and only lightweight decision rules are learned in a broad embedding space, even when training sees fakes from a single generator. Their main finding is that the standard paradigm of training an end-to-end real-vs-fake classifier tends to learn an *asymmetric* decision rule: it strongly keys on artifacts of the *training* fake domain, while treating the “real” class as a *sink* that absorbs anything not matching those artifacts, which causes many unseen-family fakes to be misclassified as real. To address this, they propose performing real-vs-fake classification *without learning a fake-specific feature space*: images are embedded using a frozen pretrained foundation encoder, and detection is implemented with lightweight decision rules in the fixed embedding space rather than end-to-end optimization on synthetic artifacts.

When only small reference sets are feasible, semantic pairing becomes a practical tool for reducing content confounds while still operating in the same frozen embedding space. Concretely, in the canonical paired setup of Cozzolino et al. [16], a small set

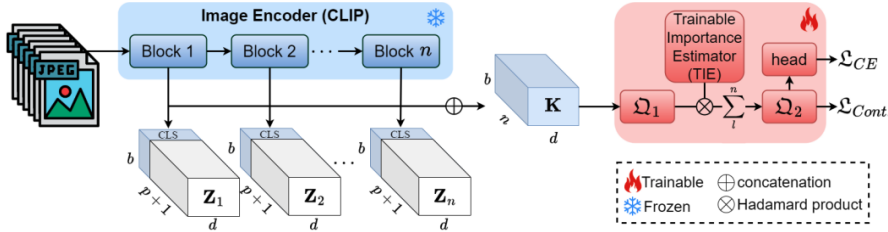


Figure 3.4: RINE architecture [49]. The method aggregates intermediate representations from a frozen CLIP image encoder using a lightweight, trainable weighting and classification module to improve generalization.

of real images is matched with caption-coupled synthetic counterparts produced by a single text-driven generator, so each real/fake pair shares the same textual description. They keep CLIP (ViT-L/14) frozen as an embedding function and train only a lightweight linear separator in CLIP feature space (using next-to-last-layer image features), then stress-test out-of-distribution generalization across unseen generator families and under social-media-like post-processing with a fixed operating point (no test-time calibration; threshold-dependent accuracy uses a fixed threshold of 0.5). Finally, they report evidence that CLIP-based decisions are at least partially orthogonal to classical low-level forensic traces, which motivates simple fusion with trace-based detectors as a practical way to improve reliability under unseen generators and post-processing.

A natural refinement is to ask which layers of a frozen foundation model best expose forensic evidence, rather than assuming the last-layer embedding is optimal for generalization. Koutlis and Papadopoulos [49] revisit the “frozen foundation model + lightweight head” paradigm by arguing that last-layer CLIP features are often too semantic for forensic generalization, and that synthetic-image detection should recover more low-level evidence from *intermediate* ViT blocks. Their method, RINE (Representations from Intermediate Encoder-blocks), reads out and aggregates intermediate representations from a frozen CLIP-like ViT with a lightweight trainable module, as illustrated in Figure 3.4; the central claim is that *which layer* is read out can matter as much as the downstream classifier choice for open-world generalization [49]. The main takeaway is that intermediate-layer readouts can expose more transferable evidence than the final embedding, even when the encoder itself remains frozen.

3.3. Forensic evidence beyond RGB appearance

3.3.1. Frequency and residual cues

Frequency-domain cues arise because many generators impose structured constraints on pixel formation. In GAN and diffusion pipelines, upsampling and decoding operations can introduce periodic correlations (visible as peaks or directional structure in average spectra/autocorrelations) that differ from camera-acquired imagery. Importantly, these cues are not guaranteed to be stable under post-processing: resizing can shift spectral peaks, and JPEG/WebP compression can introduce dominant periodic artifacts that partially mask generator traces. For this reason, modern spectral methods either (i) explicitly train with compression-aware protocols, (ii) learn masked spectral fingerprints that expose which frequency regions matter, or (iii) model real-image spectral statistics and treat synthetic images as out-of-distribution.

3.3. Forensic evidence beyond RGB appearance

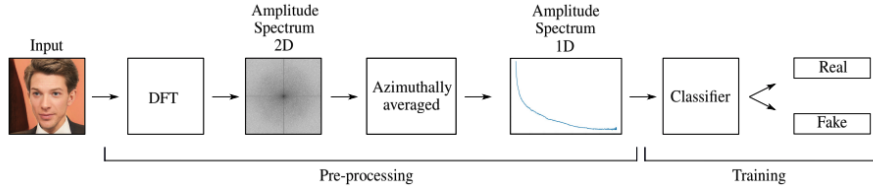


Figure 3.5: Overview of the processing pipeline in [20]. The method contains two main blocks: a feature-extraction block using DFT and azimuthal averaging, and a training block where a classifier uses the transformed features to determine whether a face is real or fake. Input images are transformed to grayscale before DFT.

Architecture-level analyses tie these anomalies to specific upsampling operators and their characteristic spectral effects [20]. One influential GAN-era explanation is that common generator upsampling operators (*transposed convolution* and interpolation-plus-convolution) systematically distort the spectral distribution of synthesized images [20]. Their analysis uses azimuthal integration of the DFT power spectrum, that is, averaging spectral power over angles at a fixed distance from the frequency origin, to obtain a compact 1D radial spectral profile. They then show, both theoretically and empirically, that transposed convolution tends to introduce high-frequency replicas while interpolation-based up-sampling tends to suppress high frequencies. In this setting, simple classifiers on spectral descriptors (e.g., SVM or even unsupervised clustering) already separate real and generated faces with very high accuracy on the evaluated benchmarks; Figure 3.5 illustrates the full processing pipeline, from DFT-based feature extraction to final real/fake classification.

Beyond detection, the same paper proposes a spectral regularization term for GAN training, added to the generator objective to align generated and real spectral profiles, and reports that this regularization, together with using larger convolutional filters after the upsampling step, can reduce spectral mismatch and improve training stability.

To move from mechanism to a usable forensic feature space, residual-based pipelines explicitly suppress content and model the remaining noise statistics. Liu et al. [52] shift the focus from semantics to the *imaging trace* left in the high-frequency residual. Their key contribution is a representation they call the *Learned Noise Pattern* (LNP): instead of working directly with RGB pixels, the method first applies a learned denoising module to suppress scene content and isolate a residual intended to capture camera- and processing-related noise statistics. The authors argue that real photographs, despite wide semantic diversity, share more consistent residual statistics, while generated images often introduce *structured artifacts* (e.g., periodic patterns) that become more evident once the image is mapped into this noise space.

To make such artifacts easier to detect and compare, the LNP is further analyzed in the frequency domain. Concretely, the method computes a Fourier-domain descriptor from the LNP (using the magnitude/amplitude of the discrete Fourier transformation), yielding a compact spectral representation in which generator-induced periodicities tend to appear as distinctive peaks or regular structures. A detector is then trained using these LNP-based features (and, in particular, their spectral signatures) to discriminate real images from synthesized ones; the primary experimental protocol follows a standard *train-on-one-generator, test-on-others* setup to evaluate generalization, highlighting that the learned noise representation and its spectrum can generalize beyond the specific generator seen during training.

In addition to the main supervised detector, the paper also reports an auxiliary one-class experiment using a classical novelty detector, specifically an OC-SVM, to

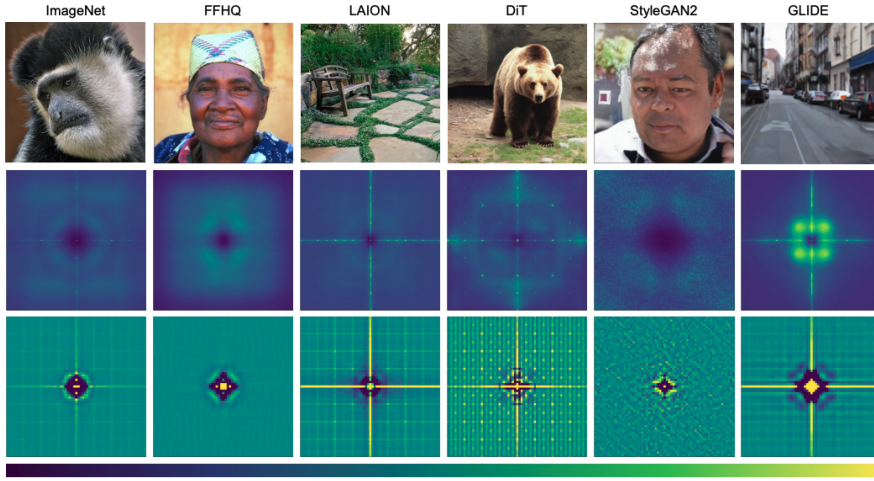


Figure 3.6: Examples of real and synthetic images drawn from popular datasets, as analyzed in [15]. Real datasets (left): ImageNet [18], FFHQ [44], LAION [86]. Synthetic datasets (right): DiT [71], StyleGAN2 [45], GLIDE [65]. For each dataset, the figure shows the average power spectra (middle) and autocorrelation (bottom) of noise residuals. Real images embed some traces of JPEG artifacts, while generated images often exhibit anomalous periodic patterns in the spatial domain.

illustrate a simple but important point: when the input is the LNP (or its spectrum), even simple one-class modeling can separate real images from generated ones more effectively than when using the original RGB images.

Measurement studies then ask which residual and spectral properties persist across generator families, beyond any single model or detector. Corvi et al. [15] aim to clarify *which low-level properties of synthetic images are stable and forensically informative* across modern generator families, motivated by the rapid shift from GANs to VQ-based and diffusion-based models. The distinctive contribution is an analysis framework (rather than a new detector) showing that *no contemporary generator is artifact-free*: beyond GANs, diffusion models and VQ-GAN/transformer pipelines also leave characteristic “artificial fingerprints” that become evident when semantic content is suppressed and second-order statistics are examined. Methodologically, they extract a noise residual (via denoising) and characterize each source using the *average autocorrelation* in the spatial domain and the corresponding *average power spectrum* in the Fourier domain, complemented by derived *radial* and *angular* spectral summaries to capture systematic discrepancies between real and synthetic distributions in the mid-high frequencies, as shown in Figure 3.6.

A key observation is that many generator-specific traces manifest as spectral peaks and periodic autocorrelation patterns whose structure is strongly linked to architectural choices (e.g., upsampling factors in latent diffusion), suggesting an architecture-driven origin rather than semantic bias. They further emphasize two confounders that are especially important in real-world settings: (i) common post-processing (notably resizing and JPEG recompression) can partially or completely obscure generation traces and introduce new dominant artifacts, and (ii) biases in the generator’s *training data* (e.g., pervasive JPEG grid artifacts) can be *transferred* into generated images, potentially misleading fingerprint-based analyses unless the data provenance is controlled.

In the diffusion era, benchmarks and simple baselines make these cues operatio-

3.3. Forensic evidence beyond RGB appearance

nal and comparable across modern generators. Bammey [4] targets the problem of detecting *diffusion-model* generated images and makes two contributions that are particularly useful for the community: (i) a benchmark tailored to the diffusion era, designed to reduce common evaluation confounds (e.g., mismatched content distributions and uncontrolled compression), and (ii) a lightweight spectral baseline that is easy to reproduce and strong enough to serve as a reference point.

On the benchmark side, the paper builds a dataset covering several widely used generators, including Stable Diffusion (v1.3, v1.4, v2, and SDXL), Glide, Midjourney (v5), DALL-E (2 and 3), and Adobe Firefly, together with high-quality *real* photographs. Real images are sourced from uncompressed photo repositories (notably RAISE and Dresden), and the *RAISE-1k* subset is used as a *content guide* to align the semantic diversity of real and generated images. For each real image, an automatic description is obtained (via Midjourney descriptor and CLIP Interrogator), then manually refined into a photorealistic text prompt. Importantly, the original photos are *not* used for image-to-image conditioning; they only guide the prompt construction so that the synthetic images remain independently generated while preserving category balance.

On the baseline detector side, the proposed method operationalizes the observation that diffusion images can exhibit distinctive *frequency artefacts* at specific periodicities. The pipeline first applies a simple high-pass residual (a cross-difference filter) to amplify these artefacts, then computes the FFT magnitude of this residual and extracts a small set of peak magnitudes at targeted periods (notably 2, 4, and 8) across directions and color channels. These tabular features are finally fed to a standard classifier (histogram-based gradient boosting trees), yielding an interpretable and inexpensive detector. Since JPEG compression can also introduce energy at similar periodicities, the method explicitly incorporates *JPEG-aware* handling by training separate models for different JPEG quality factors and selecting the appropriate model at inference using an estimated quantization table; Figure 3.7 shows representative images from each source in the resulting benchmark.

Beyond tabular baselines, interpretable spectral fingerprinting aims to expose which frequency regions drive a detector decision. The MaskSim framework [51] makes the decision process auditable by learning which frequencies matter and exposing a reference pattern as a generator fingerprint, as illustrated by the pipeline flowchart in Figure 3.8. Its main goal is to replace opaque end-to-end classifiers with a model that (i) *learns which frequencies* are discriminative for a given generator and (ii) exposes these cues through a trainable *spectral mask* and a *reference spectrum* (a generator “fingerprint”). Operationally, images are transformed into a frequency representation, modulated by a learnable mask that emphasizes informative regions, and compared to a learnable reference spectrum that serves as the generator fingerprint; the resulting similarity score is then mapped to a real/fake decision. This design makes it possible to inspect which frequency bands drive detection and to reason about when those traces may be suppressed by post-processing or altered by rescaling.

A different approach is to model the real-image spectrum itself and treat synthetic images as out-of-distribution, while also supporting arbitrary input resolutions. Karageorgiou et al. [43] propose *SPAI* for AI-generated image detection under two real-world constraints: (i) detectors should generalize to unseen generators whose artifacts differ from those observed during training, and (ii) detectors should operate on any-resolution imagery without discarding subtle forensic evidence through aggressive resizing. The key conceptual shift is to model an invariant reference, namely the spectral distribution of real images, and treat synthetic images as out-of-distribution samples with respect to this learned real-image prior. SPAI learns this prior through *masked spectral learning*, a self-supervised frequency reconstruction pretext task trai-

Chapter 3. Detection of Artificially Generated Images: State of the Art



Figure 3.7: Sample images from the Synthbuster benchmark [4]. Each row corresponds to a generator, and each column corresponds to a unique semantic prompt, identified by the RAISE-1k image ID. Supplementary prompt details and uncropped samples are provided in Appendix B (Section B.1).

3.3. Forensic evidence beyond RGB appearance

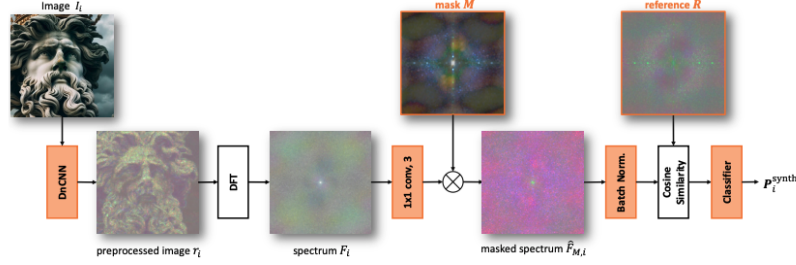


Figure 3.8: Flowchart of MaskSim [51]. The method learns a spectral mask and a reference spectrum to expose which frequency regions act as generator fingerprints and to make decisions auditable.

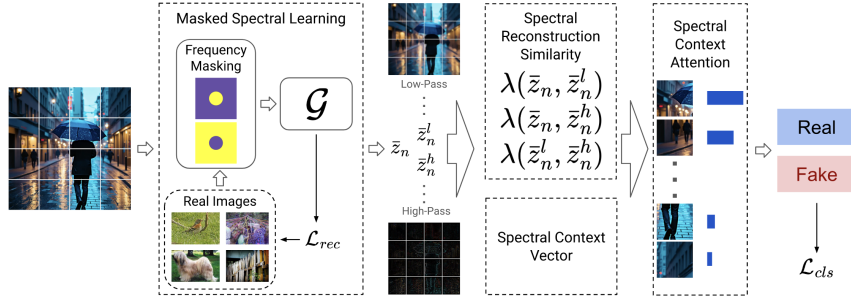


Figure 3.9: Overview of the SPAI approach [43]. A real-image spectral prior is learned via masked spectral learning and then used for out-of-distribution scoring, while patch-level processing enables any-resolution detection without aggressive resizing.

ned exclusively on real images. Low- or high-frequency components are removed using a radial mask in the Fourier domain and transformed back to the spatial domain, producing spectrally degraded inputs. A transformer is then trained to reconstruct the original image, while supervision is imposed in the frequency domain by minimizing reconstruction discrepancies between the Fourier spectra of the reconstructed and original images. This procedure implicitly captures the spectral statistics characteristic of real imagery.

At inference time, SPAI treats synthetic images as out-of-distribution by scoring how consistently the learned real-spectrum model represents an input under spectral perturbations, rather than classifying RGB content directly. To support any-resolution inference, evidence is computed locally and aggregated across an image so that subtle spectral deviations are not washed out by forced resizing, as illustrated in Figure 3.9, which summarizes the full pipeline from masked spectral learning of a real-spectrum prior to out-of-distribution scoring and any-resolution aggregation.

Although SPAI demonstrates strong cross-generator generalization, its performance ultimately depends on the preservation of spectral information. When images undergo heavy post-processing, screenshots, or re-acquisition through display pipelines, the original frequency statistics may be partially destroyed, reducing detection reliability.

3.3.2. Diffusion self-consistency and reconstruction error

This section covers detectors that do not primarily search for generic texture artifacts, but instead probe whether an image is *consistent with the internal mechanics of diffusion generation*. The common strategy is to run an image through a diffusion-related reconstruction process and treat the resulting residual (or transformed residual) as forensic evidence. Intuitively, if an image was produced by a diffusion pipeline, it should often be easier for a compatible diffusion model to invert and reconstruct than a natural camera photograph, which can lead to measurable differences in reconstruction self-consistency.

Process-aware detectors therefore generate evidence by forcing an image through a diffusion inversion and reconstruction cycle, rather than classifying RGB content directly. Wang et al. [98] propose diffusion reconstruction error as a mechanism-tied alternative to RGB decision boundaries when generalization across unseen diffusion models is the priority. Their key claim is that a detector should not classify images from pixels alone; instead, it should measure whether an image is *self-consistent under a diffusion inversion–reconstruction cycle*. Concretely, an input image is inverted and reconstructed with a pretrained diffusion model, and a reconstruction-error map is treated as the primary forensic evidence rather than RGB content, as shown in Figure 3.10. The forensic intuition is mechanism-tied: diffusion-generated images are expected to lie closer to the manifold implicitly modeled by the inversion/reconstruction pipeline, so they should reconstruct with lower and differently distributed error than real camera images. From this perspective, the method is best understood as a *representation shift* as much as a classifier: its main contribution is not a heavy new backbone, but a preprocessing transformation that suppresses semantic shortcuts and amplifies generation-mechanism traces. Framed this way, DIRE is an early and influential example of “process-aware” detection, where evidence is produced by forcing an image through a generative model’s dynamics and then classifying the residual.

After introducing the method, the paper supports it with *DiffusionForensics*, a benchmark tailored to diffusion-era evaluation. The benchmark combines real images (*LSUN-Bedroom*, *ImageNet*) with synthetic images from diverse diffusion families, including *unconditional* (ADM, DDPM, iDDPM, PNDM), *conditional* (ADM on ImageNet), and *text-to-image* systems (Latent Diffusion, Stable Diffusion v1/v2, VQ-Diffusion). A useful audit perspective is that DIRE’s strengths and risks are both tied to this design choice. Its strength is improved cross-model robustness relative to pure RGB detectors, especially under common post-processing (e.g., JPEG compression and blur), because the decision is based on reconstruction behavior rather than brittle texture artifacts. Its limitation is *dependency on the chosen inversion/reconstruction model and settings*: if future generators, editing pipelines, or severe transformations distort the alignment between test images and the reconstruction model, residual separability can shrink.

Subsequent work refines this idea by guiding reconstruction error toward informative frequency bands and by reducing reconstruction cost. FIRE refines reconstruction-error detection by explicitly emphasizing informative mid-frequency bands and improving robustness under common post-processing perturbations [12]. The central intuition is that diffusion models exhibit consistent weaknesses when reconstructing certain *mid-band* frequency components of real photographs. Rather than feeding a raw reconstruction residual directly to a classifier, FIRE explicitly guides the detector toward these frequency regions, aiming to reduce content bias and emphasize harder-to-reconstruct signals. At a high level, FIRE compares reconstruction behavior *before* and *after* suppressing informative mid-band frequency components, and uses the resulting reconstruction-error evidence to improve robustness under distribution shift. To make this practical at scale, FIRE replaces a full diffusion denoising reconstruction

3.3. Forensic evidence beyond RGB appearance

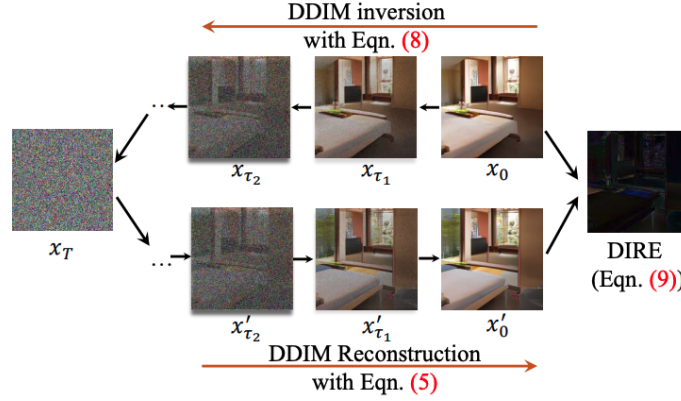


Figure 3.10: Illustration of the process of computing DIRE given an input image x_0 [98]. An image is inverted and reconstructed with a diffusion model, and the resulting reconstruction-error map is used as forensic evidence.

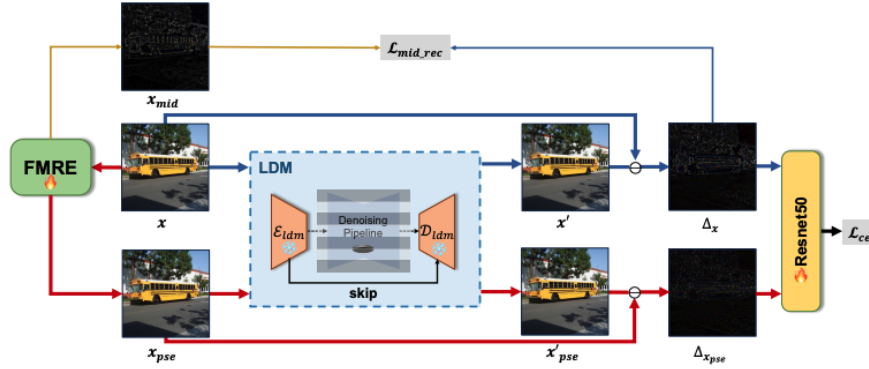


Figure 3.11: Overview of FIRE (*Frequency-guided Reconstruction Error*) adapted from Chu et al. [12]. FIRE emphasizes reconstruction-error evidence tied to informative mid-frequency components to improve robustness under common post-processing.

pipeline with reconstruction through the latent diffusion autoencoder, substantially reducing computational cost while preserving informative error signals [12]. The paper evaluates FIRE on the *DiffusionForensics* benchmark (using the LSUN-Bedroom and ImageNet subsets) and on a self-collected dataset built to reflect newer text-to-image systems. The latter uses real images sampled from LAION and generated images from recent/open/closed diffusion models (including DALL-E 3, Kandinsky 3, Midjourney, SDXL, and Segmind Vega), with prompts sourced from Midjourney. The authors also explicitly test robustness under perturbations common in online media pipelines, including JPEG compression, crop-resize, blur, and additive noise, and incorporate related augmentations during training. Figure 3.11 provides an overview of the FIRE pipeline.

Figure 3.12 highlights the FMRE mask-learning module used by FIRE: a shared encoder feeds two decoder branches that estimate the mid-band mask and its complementary mask, which are then used to construct frequency-guided reconstruction

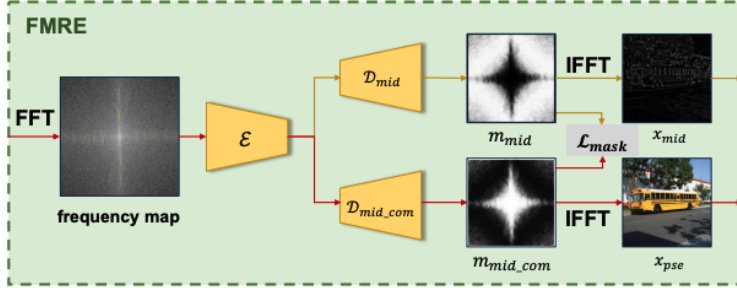


Figure 3.12: FMRE architecture (shared encoder with two decoders) used in FIRE to predict the mid-band mask and its complementary mask.

evidence.

3.3.3. Color statistics and inter-channel cues

This section reviews detectors that treat *color* as the primary forensic signal rather than relying mainly on semantic content or large end-to-end backbones. The shared hypothesis is that image generators can match global perceptual realism while still leaving measurable inconsistencies in chrominance relations, palette organization, and full color-distribution statistics. Compared with texture-centric approaches, these methods are typically more interpretable and often easier to audit under domain shift, because their evidence can be traced to explicit channel transforms, residual maps, and low-dimensional statistical descriptors. The discussion moves from early GAN-era chrominance residual analysis to diffusion-era palette and inter-channel cues, and finally to cross-domain color-distribution modeling.

A physically motivated line of work exploits discrepancies between *camera imaging* and *GAN-based RGB synthesis* by moving into chrominance and inter-channel domains. Li et al. [50] contribute an interpretable, low-dimensional feature pipeline built on the observation that, while GAN outputs may closely match real images in RGB space, they exhibit more consistent statistical mismatches in chrominance-related channels such as H , S , C_b , and C_r , especially after suppressing semantics in a residual domain. Concretely, they (i) transform an input image into multiple color spaces, (ii) compute first-order differential residuals (horizontal/vertical) on selected chrominance channels, (iii) apply truncation to bound residual values, and (iv) encode residual statistics via co-occurrence matrices (with fixed offsets and a small order), concatenating these into a compact feature vector that is finally classified with a conventional learner rather than an end-to-end CNN.

A second emphasis is evaluation under realistic *mismatch* conditions: beyond the matched train/test case, they explicitly test (a) mismatched real-image sources (same semantic type but different datasets), (b) mismatched semantic content (e.g., training on faces and testing on bedrooms, and vice versa), and (c) mismatched GAN models, to probe cross-domain generalization rather than only in-domain separability. They also consider a *model-unaware* setting where no fake samples are available, training a one-class classifier using only real images to flag generated images as outliers.

To support these scenarios, they curate large-scale real datasets spanning faces and bedrooms (e.g., CelebA/LFW, LSUN-Bedroom, CelebA-HQ/FFHQ) and generate synthetic counterparts from multiple GAN families (including DCGAN, WGAN-GP, ProGAN, StyleGAN, BigGAN, and CocoGAN), enabling systematic analysis across

3.3. Forensic evidence beyond RGB appearance

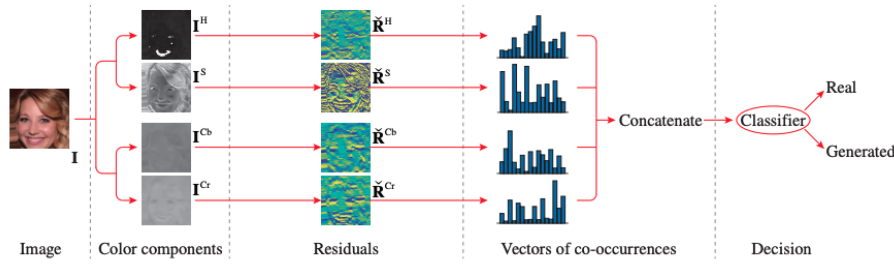


Figure 3.13: Pipeline of chrominance-disparity-based detection adapted from Li et al. [50]. The input image is decomposed into selected color components (e.g., H, S, C_b, C_r), transformed into residual maps, encoded via co-occurrence statistics, and finally classified as real or generated.

resolution and content diversity; Figure 3.13 summarizes the chrominance-residual and co-occurrence pipeline used in this setting.

Building on this residual-statistics family, diffusion-era work revisits color cues through palette structure and inter-channel relationships. Uhlenbrock et al. [94] study *color statistics* as an interpretable forensic cue for detecting diffusion-generated and inpainted images. Their motivation is that many universal deep detectors, while effective, are difficult to interpret and often require frequent retraining as new generators appear; in contrast, they argue that simple color-based cues can provide a more transparent and data-efficient alternative. The central hypothesis is that modern image generators optimize perceptual quality with a stronger emphasis on *luminance* than on *chrominance*, which leaves systematic discrepancies in cross-channel relationships. The paper supports this claim by analyzing the behavior of perceptual loss (LPIPS) under perturbations in different color components, showing higher sensitivity to luminance-oriented channels than to chroma channels, which motivates focusing on color transforms and inter-channel relations as potentially more stable detection cues.

Beyond standard color spaces (HSV, YCbCr, $L^*a^*b^*$), the authors explore *relational* RGB transforms obtained by reordering each pixel’s channels according to intensity into upper, median, and lower components (U, M, L), and then forming ratios such as L/U , L/M , and M/U . These transforms are used both for qualitative forensic inspection (where inpainted regions may become visually salient in residual maps) and for quantitative detection. For the detector itself, the paper proposes a lightweight handcrafted pipeline: selected color channels are log-transformed, high-pass residuals are extracted using a diagonal Laplacian filter (chosen to emphasize non-isotropic differences), residuals are quantized and truncated, and co-occurrence histograms are computed from many residual-based statistics rather than learned deep features. These histogram features are then classified with a Random Forest, a design that deliberately favors interpretability and low training cost over end-to-end learned representations.

Figure 3.14 illustrates the paper’s qualitative forensic angle: relational inter-channel transforms can suppress semantic content and amplify chrominance residual patterns, making synthetic or inpainted regions more visually salient.

A complementary direction models color distribution regularity directly, emphasizing cross-domain protocols where semantic shortcuts are discouraged. Jia et al. [40] propose a color-centric detector motivated by a generalization failure repeatedly observed in AI-generated image forensics: many detectors perform well in-domain but degrade under cross-domain shift because they overfit to semantic/style cues rather than to generation traces. Their key methodological idea is to characterize *color distribution regularity* as a forensic signal, based on the observation that synthetic images

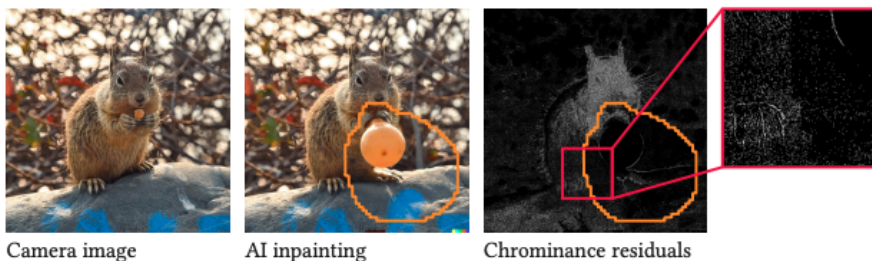


Figure 3.14: Example from Uhlenbrock et al. [94] showing how relational color-channel transforms improve the visual saliency of manipulated content. Left: natural camera image. Middle: AI inpainting example (edited region outlined). Right: residual representation after relational inter-channel processing, where the manipulated region and its boundary become substantially more visible.

often exhibit less uniform RGB color distributions (e.g., more saturated or peakier histogram behavior) than real images.

To make this discrepancy measurable, the paper introduces a lightweight color-distribution feature extraction process that deliberately perturbs and partially restores the image. Concretely, the input is repeatedly corrupted with Gaussian noise and low-bit quantization, and then reconstructed by averaging multiple noisy-quantized realizations; the resulting *color difference map* (the difference between the original image and the restored version) acts as a proxy for color-distribution uniformity, with larger restoration-induced deviations tending to indicate synthetic content, as illustrated qualitatively in Figure 3.15. The proposed detector, *CoD*, processes both the original RGB image and the color difference map through separate feature-extraction branches, fuses the resulting representations, and performs classification with a compact CNN backbone, emphasizing efficiency while retaining robustness under distribution shift.

The method is evaluated in a deliberately challenging cross-domain setting introduced by the same work, *FakeART*, which is designed to reduce semantic shortcuts and test generalization to newer generators. In this benchmark, training is performed on natural photographs (e.g., COCO/LSUN) with synthetic counterparts from earlier generators, while evaluation shifts to a different visual domain (artworks, e.g., WikiArt) and more recent diffusion systems. This setup makes the benchmark itself an important contribution, but in the context of this chapter the central takeaway is that a simple color-distribution feature can materially improve cross-domain robustness when paired with a lightweight detector, as illustrated in Figure 3.16 by representative cross-model outputs and their RGB-channel distribution statistics.

3.4. Real-world generalization: continual updates, localization, and explanation

3.4.1. Continual updates and online learning

Real-world settings require updating detectors over time and extending outputs beyond a single binary score, especially when images can be partially synthesized or heavily post-processed.

In an online setting where generators arrive continuously, the practically relevant question is not only cross-model generalization in isolation, but how detectors evolve

3.4. Real-world generalization: continual updates, localization, and explanation

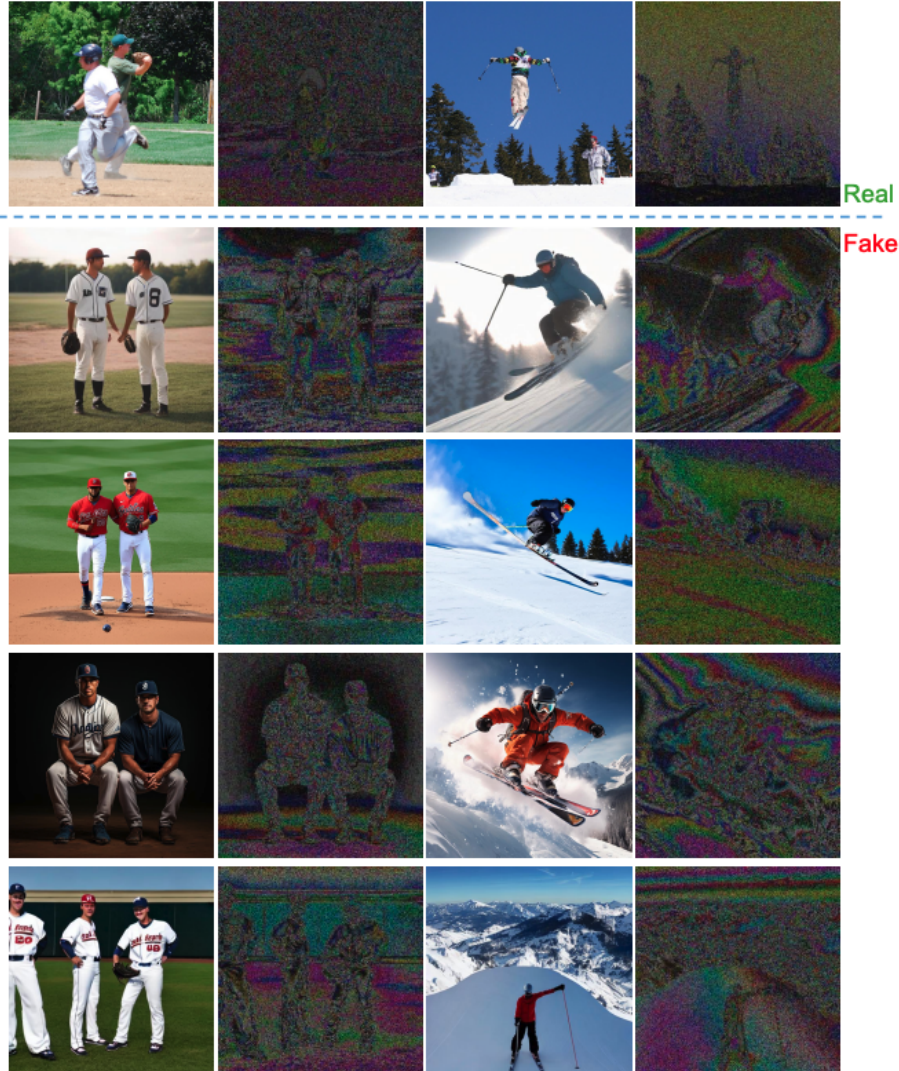


Figure 3.15: Qualitative examples of original images and corresponding color difference maps (visualized with amplification) from Jia et al. [40]. The examples illustrate that generated images tend to produce stronger restoration-induced color deviations than real images.

over time and retain robustness as new sources are added. Epstein et al. [21] address this with a protocol and dataset construction that simulates real-world conditions: they curate a chronological dataset of images from 14 generative sources (spanning early diffusion models, latent-diffusion variants, and closed commercial systems) and train a single binary detector *incrementally* by adding each generator in historical release order while keeping a fixed pool of “real” images throughout. The detector itself is intentionally standard (an ImageNet-pretrained ResNet-50) and the emphasis is placed on training practice, notably using augmentations that reflect real-world post-processing (random crops plus degradations such as blur, grayscale, and watermark simulation), to support forward generalization to models not yet seen in training.

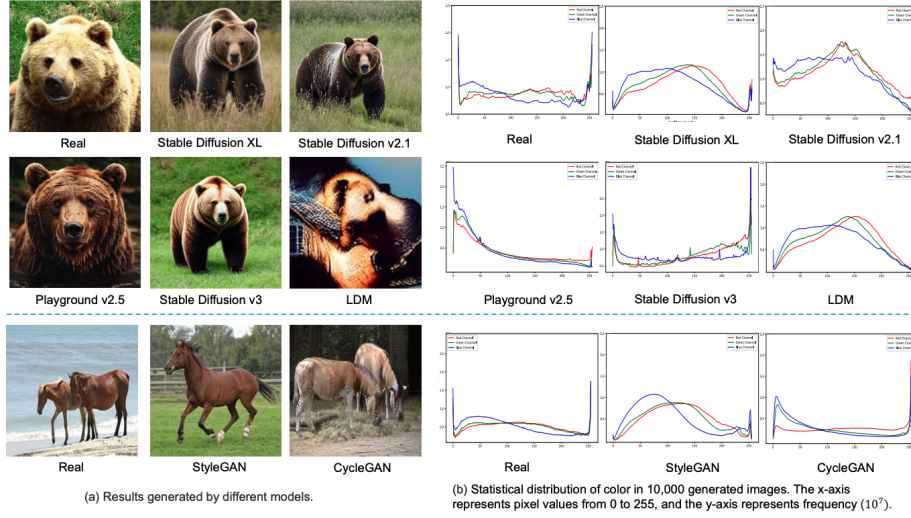


Figure 3.16: Examples and color-distribution statistics from Jia et al. [40]. Left: representative images from real data and multiple generative models. Right: corresponding RGB-channel histogram profiles over large sample sets, highlighting model-dependent distribution behaviors that motivate color-centric forensic cues.

Beyond whole-image classification, they extend the same online perspective to *pixel-level* detection for generative inpainting, motivated by mixed real/synthetic images: they synthesize inpainted training data when an inpainting API is available (random masks over real images, with original pixels copied back outside the mask to isolate generated regions) and train a fully-convolutional segmentation model (FCN with a ResNet-50 backbone) using a loss suitable for class imbalance. For the more constrained case where inpainting examples are unavailable, they propose using CutMix on whole images as a surrogate supervision signal to induce pixel-wise predictions, analyzing how much localization capability can be obtained from whole-image supervision alone and how incorporating previously seen inpainting models improves generalization.

Continual learning also appears in sequential domain settings, where the main risk is forgetting earlier forensic traces while adapting to new ones. Cheng et al. [10] explicitly target long-horizon updates and catastrophic forgetting as new forgery domains arrive over time. Rather than treating generalization as a one-shot unseen-domain problem, they argue that standard fine-tuning causes decision boundaries to drift toward the most recent forgery domain and progressively degrades performance on earlier ones.

Their proposed framework, *SUR-LID*, combines two complementary components. First, *Sparse Uniform Replay* (SUR) selects a replay memory intended to better preserve the *global feature distribution* of prior tasks, instead of relying on purely random rehearsal or highly redundant exemplars. Second, the *Latent-space Incremental Detector* (LID) imposes structure in the feature space so that newly learned forgery domains can be incorporated while maintaining separation and alignment with previously learned ones. In this sense, the method is not only a memory-based continual learner, but also a *feature-organization* strategy: replay preserves past support, while latent-space constraints stabilize inter-task discrimination as new fake domains are introduced (the paper frames this as a “stacking brick by brick” process).

Chapter 3. Detection of Artificially Generated Images: State of the Art

suggests that future real-world forensic systems, especially for social-media moderation and trust-and-safety workflows, should be judged not only by classification accuracy, but also by their ability to localize suspicious regions and communicate actionable, human-interpretable evidence.

3.5. Chapter summary

This chapter surveyed the state of the art in AI-generated image detection, organized by the type of forensic evidence exploited and the learning paradigm used to extract it.

Several cross-cutting observations emerge from this survey. First, no single evidence family is universally reliable: spatial-domain detectors generalize well when augmentation and data diversity are carefully managed, but can overfit to texture shortcuts; spectral and residual cues expose architecture-linked periodicities, yet are fragile under resizing and recompression; reconstruction-based methods inherit the biases of their inversion model; and frozen-representation detectors trade generator-specific sensitivity for broad semantic coverage. Second, training strategy often matters as much as architecture: augmentation that simulates real-world conditions, bias-free real-fake pairing, generator-diversity scaling, and continual-update protocols each contribute substantially to cross-generator robustness. Third, these cue families are partially complementary, which motivates fusion rather than reliance on any single signal.

Within this landscape, *color-centric cues* occupy a distinctive but still comparatively underexplored position. Existing color-based methods (Section 3.3.3) have demonstrated that chrominance residuals, palette structure, and color-distribution regularity carry forensic information, particularly because generators optimized with perceptual objectives tend to match luminance fidelity more closely than chrominance fidelity. However, prior work has largely focused on marginal per-channel statistics or handcrafted relational transforms, and the systematic role of inter-channel *joint* correlations, their dependence on the choice of color space, and their interaction with perceptual training objectives remain open questions.

Chapter 4 addresses this gap directly: it analyzes how perceptual metrics respond to inter-channel dependence, characterizes generator-specific correlation signatures across multiple color representations, and translates these findings into a lightweight detection pipeline that augments a standard backbone with inter-channel correlation maps computed in a carefully chosen color space.

Chapter 4

CHROMA

4.1. Introduction

The state of the art in generated-image detection (Chapter 3) indicates that robust forensic decisions rarely arise from a single, universally reliable cue family. Spatial-domain discriminators can achieve high accuracy in matched settings, yet their evidence may be brittle under open-world generators and post-processing. Frequency-domain and reconstruction-based approaches improve robustness by targeting mechanism- or architecture-tied traces, but they remain sensitive to resizing/compression and may fail for generator families whose spectral artifacts are weak or intentionally suppressed. In this landscape, *color* provides a complementary and comparatively lightweight axis: it is directly linked to camera imaging pipelines (e.g., demosaicing, white balance, color correction, and tone mapping) and has been repeatedly observed to exhibit systematic discrepancies between natural and synthetic images, including diffusion-era outputs and inpainting [40, 50, 94]. However, much of the existing color-centric literature relies on marginal statistics (e.g., per-channel histograms or handcrafted residual co-occurrences), which may overlook a critical aspect of natural image formation: *how channels co-vary*.

This chapter develops the hypothesis that *inter-channel dependencies* constitute a stable forensic trace that is (i) physically motivated by camera processing, (ii) partially orthogonal to texture- and spectrum-based artifacts, and (iii) imperfectly constrained by common perceptual objectives used in modern generative pipelines. In particular, perceptual similarity metrics such as LPIPS [106] are widely used both as evaluation criteria and as training losses for reconstruction and editing tasks, yet they operate on RGB feature representations and need not penalize changes in chromatic *joint statistics* uniformly across color parameterizations. If perceptual objectives underweight certain correlation-structure perturbations, then modern generators can match human-perceived similarity while still deviating from camera-like channel coupling, leaving exploitable traces.

Motivated by this perspective, we introduce CHROMA, a correlation-aware view of generated-image forensics. The chapter is organized as follows. We first introduce the color-space foundations needed for our analysis (Section 4.2), summarizing the color representations and inter-channel dependence concepts that underpin the forensic features we study. We then address two empirical questions: (i) *Are perceptual metrics sensitive to channel-dependence changes in a consistent way across color spaces?* (ii) *Do explicit inter-channel correlation descriptors improve detection under realistic*

Chapter 4. CHROMA

cross-generator evaluation, when added to a standard spatial backbone? Concretely, we first conduct controlled perturbation experiments showing that LPIPS responses to matched-power noise can vary substantially depending on the color space in which the perturbation is expressed, and that its sensitivity to *decorrelating* channels is uneven across representations (Section 4.3). We then characterize how local inter-channel correlation distributions differ between real photographs and diffusion-era generators, and quantify separability via distributional distances across color spaces. Finally, we translate these analyses into a simple detection recipe: we compute correlation maps in selected color spaces and use them as additional input channels to a fixed ResNet-50 backbone, isolating the effect of the representation change (Section 4.4). Under the cross-generator protocol proposed by Cozzolino et al. [16], this correlation augmentation yields consistent gains for several generator families under a modest training budget, while also exposing limitations on certain commercial systems, reinforcing the broader conclusion from Chapter 3 that robust detection benefits from complementary evidence streams and data diversity [70].

Overall, CHROMA positions inter-channel correlation structure as a practical forensic signal: it is lightweight to compute, grounded in natural image formation, and complementary to the spatial and spectral evidence emphasized in prior detectors, while also making explicit the regimes where color-dependence cues succeed and where they break. Part of the work presented in this chapter has been accepted at the International Conference on Pattern Recognition (ICPR 2026). The source code is publicly available at <https://github.com/JPSoteloSilva/CHROMA>.

4.2. Color fundamentals for forensics

The forensic features studied in this chapter rely on inter-channel correlations computed in specific color representations. Before developing these features, we briefly introduce the color representations we use and the mechanisms that shape inter-channel correlations differently in camera-captured and generated images. A comprehensive treatment of the underlying color science, including all transform equations and derivations, is provided in Appendix C.

Color is not a physical property of objects in isolation, but a perceptual response that depends on the spectral power distribution of the light stimulus, the observer’s photoreceptor responses, and the surrounding viewing conditions [36]. Human vision projects the high-dimensional spectrum onto a three-dimensional signal via three cone types (L, M, S), so that different spectra can produce identical responses, a phenomenon known as *metamerism*. A practical consequence is that numerically identical pixel triplets can appear different depending on illumination, adaptation, and surround, as illustrated by the well-known checker-shadow illusion (Figure 4.1) [103]. This context dependence means that coordinate-based color comparisons are meaningful only when their assumptions (reference white, transfer function, viewing conditions) are made explicit.

4.2.1. Color representations

A *color space* provides a coordinate system for describing colors numerically. Different parameterizations serve different goals: hardware compatibility, efficient compression, or perceptual interpretability [22, 75]. A critical distinction for forensics is that pixel values stored in common image formats are typically *not* proportional to scene radiance; nonlinear transfer functions (gamma encoding) compress the signal before storage, and statistics computed on encoded values can differ systematically

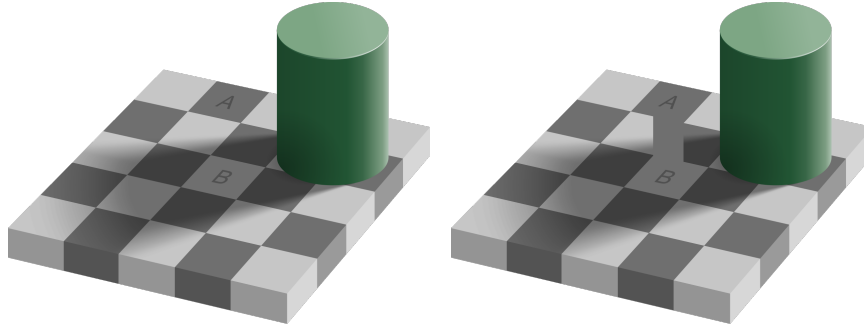


Figure 4.1: Checker shadow illusion [103]. Left: tiles A and B have identical pixel intensity but appear different due to the surrounding illumination context. Right: linking the two regions reveals they share the same value. This illustrates why equal pixel values do not guarantee equal perceived lightness, and why color coordinates must be interpreted together with their viewing and encoding assumptions.

from those computed in linear light. Unless stated otherwise, we treat pixel values as encoded and note explicitly when linearization is applied.

In this chapter we use four color-space families, each decomposing the three-channel signal differently and producing distinct inter-channel correlation structures:

- **RGB** (additive, device-dependent): the native representation of camera and display pipelines. An (R, G, B) triplet is ambiguous without specifying the standard (primaries, white point, transfer function). We distinguish between linear RGB (proportional to light) and gamma-encoded RGB (as stored in files).
- **CIELAB** (perceptually motivated, hereafter **Lab**): derived from the device-independent CIE XYZ system relative to a reference white, with axes that correspond more closely to perceived lightness (L^*) and opponent-color directions (a^* , b^*) than raw RGB. We also use the Lab-derived chroma C_{ab}^* and hue angle h_{ab} .
- **HSV/HSL** (intuitive reparameterizations): these spaces decompose RGB into hue, saturation, and a brightness/lightness axis. They remain device-dependent and are not perceptually uniform, but they are convenient for visualization and heuristic reasoning.
- **YCbCr / YUV** (luminance–chrominance separation): a family of transforms that separate a luma component (Y') from two chroma components. This decomposition, historically motivated by broadcasting and compression, makes it explicit which statistics are tied to intensity and which to chromatic variation.¹

The choice of color space is not a minor implementation detail: it determines how inter-channel covariance is distributed among the resulting components, and therefore which correlation patterns are visible and which are masked. This dependence is central to the analyses that follow.

¹YUV and YCbCr are related but distinct transforms: both share essentially the same luma, but they differ in chroma scaling and offset [75]. YUV originates from the analog PAL/SECAM standard and produces signed, zero-centered chroma channels, whereas YCbCr is the digital counterpart defined by ITU-R BT.601 [38] and BT.709 [39], which biases chroma so that all channels fall in $[0, 1]$. In the experiments of this thesis we use Kornia’s `rgb_to_yuv` [82], which implements the YUV convention with signed chroma; we refer to this representation as “YUV” throughout. When discussing JPEG compression or formal standards we use “YCbCr.”

4.2.2. What shapes inter-channel correlations

In real cameras, inter-channel dependence arises both locally (CFA sampling and demosaicing couple channels through interpolation) and globally (the CCM mixes channels across the entire image, while WB gains set the operating point on which it acts). These mechanisms produce characteristic correlation patterns that are constrained by a device’s ISP design choices. The following paragraphs detail the principal mechanisms and then contrast them with the situation in generated images.

In real images

Several stages of the camera pipeline introduce or reshape inter-channel dependencies. We describe the principal mechanisms in pipeline order, followed by a note on how the choice of color representation itself affects measured correlations.

Demosaicing. Since each pixel in a CFA-sampled image records only one color component, demosaicing must estimate the missing two channels by interpolating from spatially neighboring samples. This local interpolation couples the R, G, and B values at each pixel through a spatial filter whose support and coefficients depend on the algorithm [55]. The coupling creates two forensically relevant effects: (i) CFA-periodic inter-channel correlations whose spatial structure reflects the specific CFA layout and demosaicing kernel, and (ii) local cross-channel dependencies that are absent in images that were never CFA-sampled, such as those produced by generative models. These CFA-induced traces have been exploited for forgery detection [73] and for camera model identification via demosaicing residual features. For a full description of the demosaicing stage, see Section 2.1.2 in Chapter 2.

White balance and color correction. The perceived color of a surface depends on the spectral power distribution (SPD) of the illuminant: the same reflectance under different light sources yields different sensor responses and therefore different channel values [36]. In camera pipelines, *white balance* compensates for the scene illuminant so that neutrals are rendered without a dominant tint. In its simplest form, white balance is a diagonal gain $\text{diag}(g_R, g_G, g_B)$ applied in approximately linear camera space: it scales each channel independently but does not mix them, and therefore leaves the correlation between any two channels unchanged. In practice, however, white balance is followed by a 3×3 *color correction matrix* (CCM), which maps camera RGB to a target color encoding (often as part of a camera-to-XYZ and XYZ-to-output chain) [75]. Because the CCM has non-zero off-diagonal entries it *does* mix channels, and it is this cross-channel mixing that directly reshapes inter-channel correlations and the distribution of chromaticity. The diagonal WB gains determine the operating point on which the CCM acts: different gain vectors shift the input distribution of the CCM, so the combined effect of WB + CCM on the final correlation structure is illuminant-dependent even though the WB step alone is correlation-preserving. Implementations may further interleave clipping, tone mapping, and gamut mapping, adding non-linear channel interactions that also modify correlation structure.

Compression and chroma subsampling. JPEG compression reshapes inter-channel statistics through three successive operations: (i) conversion from RGB to YCbCr, which is a linear channel-mixing transform that explicitly redistributes inter-channel covariance into luminance and chrominance components; (ii) chroma subsampling (commonly 4:2:0), which reduces the spatial resolution of the Cb and Cr channels

4.3. Inter-Channel Correlations as Forensic Cues

relative to Y , attenuating high-frequency chrominance detail and altering local luma-chroma correlation; and (iii) block-wise DCT quantization with channel-dependent quantization tables (typically coarser for chrominance), which introduces structured quantization noise that differentially affects luminance and chrominance statistics [97]. These compression artifacts are shared by both real and generated images when the latter are JPEG-compressed after synthesis; however, in real photographs the compression interacts with pre-existing pipeline traces (CFA periodicity, CCM-induced correlations), producing joint statistics that differ from those of a generated image compressed with identical settings. For a full description of the compression pipeline, see Section 2.1.4 in Chapter 2.

In generated images

Generative models synthesize images through learned computational pipelines (noise-to-pixel or noise-to-latent mappings) that have no physical sensor, no CFA, and no device-specific white-balance or color-correction chain. At a computational level, the output of a generator is the result of a deep succession of linear and nonlinear operations (matrix multiplications, convolutions, normalization layers, and element-wise activations) that jointly produce all three color channels. Unlike the camera pipeline, where each stage has a physically grounded role and the inter-channel coupling follows from specific optical and electronic design choices, the channel interactions in a generator emerge from weight matrices and activation functions that were fitted to minimize a training loss. Consequently, the inter-channel correlation structure in generated images is not shaped by the camera mechanisms described above, but rather by the training data distribution and the optimization objective. Any camera-like correlation patterns present in the output must have been learned implicitly from the training corpus, rather than imposed by a consistent physical rendering chain. This means that the correlation structure can vary across generator architectures and training datasets in ways that are not tied to a specific device or ISP [15, 50].

Furthermore, the training objectives commonly used in modern generative pipelines are largely agnostic to inter-channel joint statistics in specific color parameterizations. Perceptual losses such as LPIPS operate on deep features extracted from RGB images and do not uniformly penalize departures from camera-like correlation structure across all color representations [94, 106]. As a result, generators can produce images that are perceptually convincing, matching human-perceived similarity in luminance and texture, while still deviating from real photographs in color-space-dependent correlation statistics. This gap is the central forensic opportunity exploited by the analyses that follow.

4.3. Inter-Channel Correlations as Forensic Cues

Within natural images, inter-channel dependencies provide a compact description of how color components co-vary. Having established which mechanisms shape these dependencies and why they differ between camera-captured and generated images (Section 4.2.2), we now study them empirically as a forensic trace. We first analyze how a standard perceptual similarity metric reacts to controlled perturbations that modify channel dependence, and later examine how correlation distributions differ between real and generated images across multiple color spaces.

4.3.1. LPIPS Sensitivity to Inter-Channel Dependence

Perceptual similarity metrics such as LPIPS [106] are deeply intertwined with modern generative modeling. In addition to being reported at evaluation time, perceptual feature losses are routinely used when adapting or fine-tuning diffusion pipelines for inverse problems and image editing (e.g., reconstruction, inpainting, or image-to-image tasks), where matching human-perceived similarity cannot be captured by pixelwise criteria alone. However, LPIPS is computed from the deep features of RGB images, and its sensitivity to chromatic structure depends on how a perturbation manifests when mapped back to RGB.

Since LPIPS operates in RGB feature space, and following the color-relation perspective of [94], we investigate whether it consistently reflects changes in inter-channel dependence across color spaces. We design controlled perturbations that either preserve channel coupling or explicitly decorrelate channels in RGB, Lab, HSV, and YUV, and measure the induced LPIPS distances. If LPIPS is uneven in this respect, then perceptual objectives may fail to enforce correlation structure uniformly, motivating the use of correlation-based descriptors for detection.

To analyze LPIPS sensitivity to inter-channel dependence, we apply controlled, low-amplitude perturbations to RGB images and measure the induced perceptual distance. Let x denote a natural RGB image. We construct two perturbed versions of x by adding noise on a subset of pixels comprising a fraction ρ of the image. In the first case, we add *channel-correlated* noise by sampling a single noise field n with i.i.d. entries and adding it identically to the three channels:

$$x_c^{\text{corr}} = x_c + n, \quad c \in \{R, G, B\}.$$

In the second case, we add *decorrelated* noise by sampling three independent noise fields n_R, n_G, n_B with the same marginal distribution and adding them channel-wise:

$$x_c^{\text{decorr}} = x_c + n_c, \quad c \in \{R, G, B\}.$$

In both settings, we use zero-mean Gaussian noise with standard deviation $\sigma = 0.005$ in normalized $[0, 1]$ RGB intensities, yielding visually imperceptible perturbations, and we vary ρ to control the fraction of affected pixels. We then compute the LPIPS distance between each perturbed image and its corresponding original in RGB space, obtaining LPIPS responses as a function of ρ under correlated and decorrelated noise.

This procedure is repeated over the full RAISE-1k dataset [17]. For each ρ , we report the mean LPIPS across all images, yielding dataset-level response curves.

We extend the same analysis to alternative color spaces, including Lab, YUV, and HSV. Given an RGB image, we convert it to a target color space, apply correlated and decorrelated noise to its channels as above, and transform the perturbed result back to RGB before computing LPIPS. To ensure comparability across color spaces, the noise variance in each space is adjusted so that, after conversion back to RGB, the induced perturbations have the same effective magnitude $\sigma = 0.005$ as the RGB baseline. The resulting dataset-averaged curves are summarized in Figure 4.2.

Figure 4.2 reveals that LPIPS responds very differently to perturbations that are matched in *effective RGB power* but injected in different color spaces. First, for a fixed coverage ρ , the LPIPS magnitude varies substantially across color spaces: YUV consistently yields the largest distances, whereas HSV correlated noise produces the smallest LPIPS response of all conditions, remaining nearly flat across ρ . This indicates that perturbations preserving HSV channel coupling are almost invisible to LPIPS. RGB and Lab fall in between, with their relative ordering depending on whether the injected noise is correlated or decorrelated. Since the perturbations are power-matched after reconstruction to RGB, this spread cannot be explained by the magnitude of

4.3. Inter-Channel Correlations as Forensic Cues

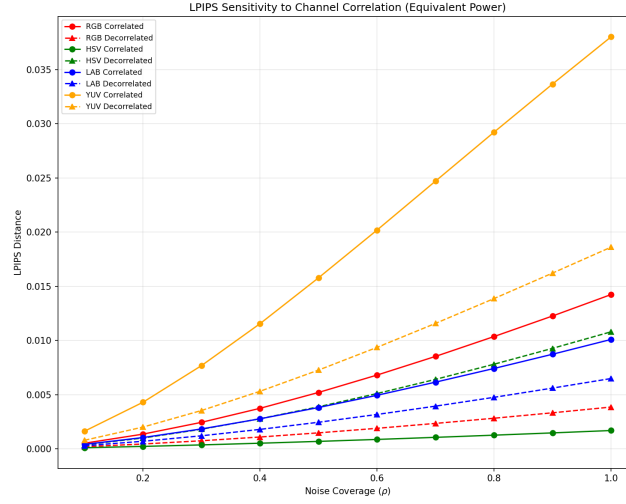


Figure 4.2: RAISE-1k-averaged LPIPS distance between an original RGB image and its perturbed versions as a function of noise coverage ρ for different color spaces (RGB, Lab, HSV, YUV) under correlated (solid) and decorrelated (dashed) channel noise with matched perturbation power.

noise, but rather by how LPIPS weights the resulting distortions once they are induced through different channel parameterizations. In other words, LPIPS is not invariant to reparameterizations that redistribute variation across luminance and chrominance components.

Second, and more importantly for our motivation, the metric is *inconsistent* in how it treats changes in channel dependence. The gap between correlated and decorrelated noise is pronounced in some spaces (most clearly YUV, and to a lesser extent HSV), but it nearly collapses in Lab. In RGB, the ordering is notably *reversed*: correlated noise yields a higher LPIPS distance than decorrelated noise, meaning that perturbations that preserve RGB channel coupling are perceived as *more* distorted than those that disrupt it. This means that two perturbations that explicitly differ in whether they preserve or disrupt inter-channel structure can be assigned almost the same perceptual distance, or even an inverted ranking, depending on the coordinate system in which the perturbation is expressed.

These behaviors suggest that LPIPS can under-emphasize certain chromatic *joint-statistics* changes, particularly those expressed in perceptually meaningful coordinates such as Lab, while being more sensitive to others, such as YUV. Consequently, when generators are optimized using such perceptual objectives, matching inter-channel correlation structure is not uniformly encouraged across color spaces. This motivates studying inter-channel correlation statistics directly: if these dependencies are only weakly constrained by perceptual losses, they may persist as systematic, generator-dependent discrepancies that can be exploited as forensic evidence for discriminating between real and generated images.

4.3.2. Cross-Generator Distributions of Inter-Channel Correlations

Given that LPIPS reacts unevenly to perturbations that alter inter-channel dependence across color spaces, we next examine whether these correlation statistics themselves exhibit systematic, source-dependent differences that can be exploited for

Chapter 4. CHROMA

detection. For each image and each color space, we construct local inter-channel correlation maps by computing the Pearson correlation coefficient over sliding $k \times k$ windows, so that each map value captures channel dependence within a local neighborhood. In our experiments, we set $k = 7$.

For each color space, we compute three such maps corresponding to the three channel pairs, and then aggregate correlation values over all pixels and images in a dataset to obtain one empirical distribution per source. Figure 4.3 shows the *log-density* estimates for each channel pair, comparing real images from RAISE-1k [17] to synthetic images generated by DALL-E, GLIDE, Midjourney v5, Firefly, and Stable Diffusion from Synthbuster [4]. Importantly, Synthbuster was designed to align the semantic content of its synthetic images with the real-image reference set, thereby reducing scene-dependent variation and enabling the attribution of distribution shifts to the generative process rather than to differences in depicted content.

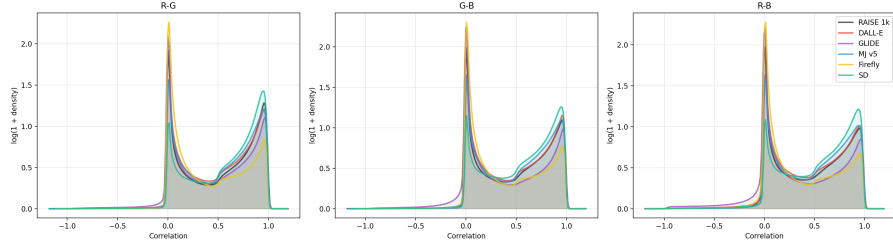
Several patterns are visible in Figure 4.3. For RGB and HSV, the curves share a dominant mode and largely overlap, with generator-dependent differences appearing mainly in the tails and in the relative mass around the peak. In Lab, the separation is more pronounced overall: all three channel-pair correlations show generator-specific shifts and shape changes. This suggests that Lab correlations capture deviations that are less apparent in RGB/HSV. In contrast, YUV correlations are tightly concentrated around a narrow region for most sources, leading to smaller visible shifts across generators.

To summarize these differences quantitatively, Figure 4.4 reports pairwise Wasserstein distances between the correlation-feature distributions for each color space; for a given color space, the reported value is the average of the three pairwise inter-channel Wasserstein distances. The matrices confirm that separability depends strongly on the representation: RGB and Lab show consistently larger off-diagonal distances than HSV and YUV, and they also exhibit a clearer generator-specific structure. For example, in RGB the largest gaps arise for pairs involving Firefly and Stable Diffusion (Firefly-SD = 0.293), while Lab also yields substantial separation for several generator pairs (e.g., GLIDE vs. Firefly/SD), reflecting the broader and more shifted Lab log-density curves. Notably, across all four color spaces, GLIDE is consistently the most distant generator from all other sources, suggesting that its generative process induces a particularly distinctive correlation footprint. By comparison, YUV distances remain systematically smaller across the board, consistent with the tighter concentration of its correlation distributions; in particular, the RAISE-SD distance in YUV is only 0.005, meaning that Stable Diffusion’s inter-channel correlations in YUV are nearly indistinguishable from those of real photographs.

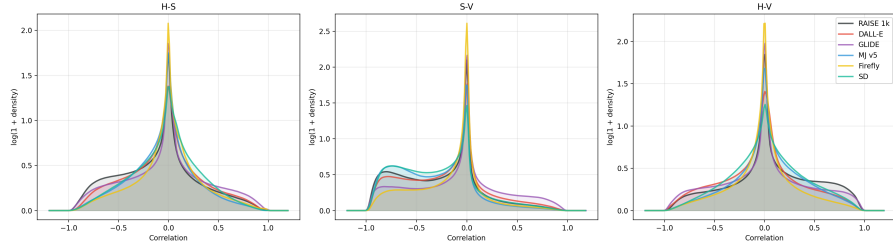
From a forensic perspective, the matrices also reveal that some generators (DALL-E, Midjourney v5) consistently exhibit small Wasserstein distances to RAISE 1k across multiple color spaces, indicating that their inter-channel correlation statistics are closer to those of real images. This proximity foreshadows the difficulty of detecting these sources using correlation cues alone and motivates the combination with other forensic evidence.

Taken together, these results suggest that inter-channel correlation statistics go beyond separating real and synthetic images, encoding *generator-dependent* signatures whose strength depends on the chosen color space. Since RGB and Lab provide the most consistent separation, we focus on correlation features derived from these spaces in the experiments that follow.

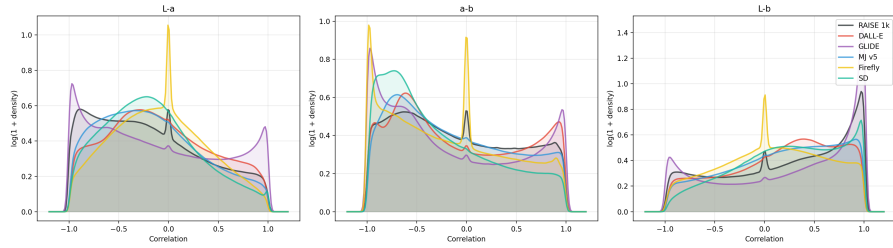
4.4. From Forensic Cues to Actual Detection



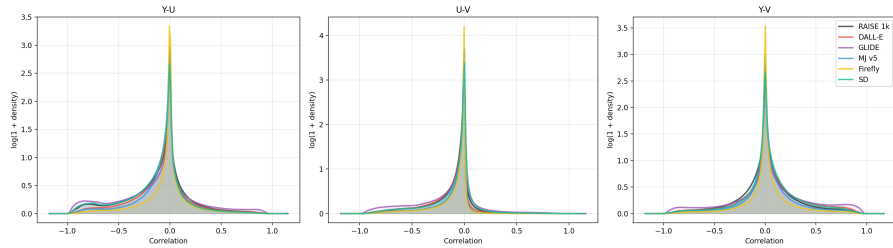
(a) RGB Inter-Channel Correlations



(b) HSV Inter-Channel Correlations



(c) Lab Inter-Channel Correlations



(d) YUV Inter-Channel Correlations

Figure 4.3: Log-density estimates of pairwise inter-channel correlation features across color spaces (RGB, HSV, Lab, YUV). Each row corresponds to one color space and shows the three channel-pair correlations. We report real images from RAISE-1k [17] and synthetic images from multiple generators in Synthbuster [4].

4.4. From Forensic Cues to Actual Detection

In this section, we evaluate the hypothesis that inter-channel correlation maps provide complementary evidence beyond standard RGB appearance cues, and that

Chapter 4. CHROMA

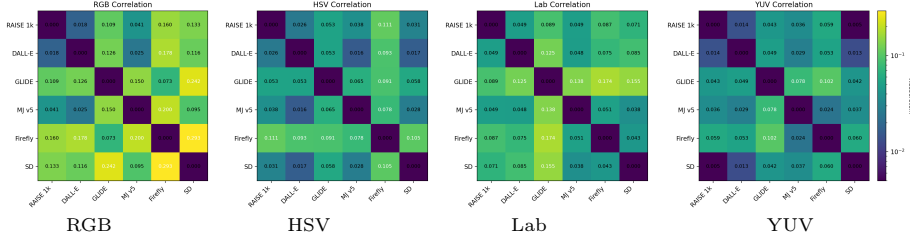


Figure 4.4: Mean Wasserstein distance matrices between inter-channel correlation feature distributions across sources, shown for multiple color spaces (RGB/HSV/Lab/YUV). Real images are from RAISE-1k [17] and synthetic images are from Synthbuster [4].

such cues can be exploited under a modest training budget. To isolate the effect of the input representation, we keep the backbone fixed and compare models trained with standard RGB inputs to models augmented with inter-channel correlation maps. We consider correlation features computed in the color spaces that exhibited the strongest cross-source separation (RGB and Lab, based on the analysis in Section 4.3), both individually and in combination.

4.4.1. Model and Input Representations

Backbone. We use ResNet-50 [30] as a simple and widely adopted CNN baseline. The network is initialized from ImageNet [18] pre-training and fine-tuned for binary real-vs-generated classification with a single linear classification head.

Correlation maps. Let $x \in \mathbb{R}^{H \times W \times 3}$ denote an RGB image. For any color space $\mathcal{S}$ with three channels, we define $\Phi_{\mathcal{S}}(x) \in \mathbb{R}^{H \times W \times 3}$ as the stack of pairwise inter-channel correlation maps computed in $\mathcal{S}$.

Fixed input variants. We evaluate four fixed input representations:

- **RGB:** $x \in \mathbb{R}^{H \times W \times 3}$.
- **RGB+Corr(RGB):** $\text{concat}(x, \Phi_{\text{RGB}}(x)) \in \mathbb{R}^{H \times W \times 6}$.
- **RGB+Corr(Lab):** $\text{concat}(x, \Phi_{\text{Lab}}(x)) \in \mathbb{R}^{H \times W \times 6}$.
- **RGB+Corr(RGB)+Corr(Lab):** $\text{concat}(x, \Phi_{\text{RGB}}(x), \Phi_{\text{Lab}}(x)) \in \mathbb{R}^{H \times W \times 9}$.

Trainable color transformations. Our earlier analyses show that the choice of color parameterization can materially change separability (Section 4.3.2) and downstream detection performance (Table 4.2). Motivated by this, we also test whether a detector can *learn* a useful color reparameterization for correlation extraction, rather than relying only on fixed transforms. Concretely, we introduce a trainable, per-pixel mapping $T_{\theta} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ implemented with 1×1 convolutions, and compute correlation maps in the resulting learned 3-channel space, i.e., $\Phi_{T_{\theta}}(x)$. We consider two instantiations: (i) a *Lab-initialized* linear 1×1 transform whose coefficients are initialized from the fixed Lab conversion used in our pipeline and then fine-tuned end-to-end; and (ii) a small *five-layer* 1×1 convolutional stack to allow a more flexible learned mixing of RGB channels before computing correlations.

Adapting the first layer. To accommodate $C \in \{3, 6, 9\}$ input channels, we replace the first convolution of ResNet-50 with a C -channel variant. We initialize its weights

4.4. From Forensic Cues to Actual Detection

Source	# images	Origin
GLIDE	700	[4]
Stable Diffusion 1.3	700	[4]
Stable Diffusion 1.4	700	[4]
Stable Diffusion 2.0	700	[4]
Stable Diffusion XL	700	[4]
StyleGAN2-FFHQ	588	[7]
DALL-E 2 (COCO-Val)	1000	[16]
DALL-E 3 (COCO-Val)	1000	[16]
Core train split	10000	[16]
Total	16088	

Table 4.1: Training data composition for the limited multi-generator regime used in ResNet-50 experiments.

from the original RGB filters by copying/averaging them across the additional channels, while keeping all other layers unchanged. This isolates the effect of the input representation, as it does not introduce architectural changes.

4.4.2. Evaluation Data and Protocol

We follow the benchmark protocol proposed by Cozzolino et al. [16], evaluating on its fixed test suite, spanning GAN-based, diffusion-based, and commercial text-to-image generators, and reporting per-generator AUC to quantify cross-generator generalization under a common testbed. Training uses a *10k* split, which contains 10k synthetic images generated by Latent Diffusion. This setting establishes an *in-domain* reference (Latent Diffusion) and enables measuring how detectors trained on a single generator generalize both to the seen source and to *unseen* generators at test time.

The test set consists of 18 generators and is constructed with generator-specific real/synthetic sources: for each generator, the real images are drawn from a reference dataset associated with that generator (e.g., LSUN, FFHQ), and the corresponding synthetic set is produced by the target generator under evaluation. For text-to-image systems, the protocol uses captioned real images (RAISE) and generates synthetic counterparts from the same textual descriptions, yielding semantically aligned real/fake pairs for evaluation across commercial tools.

Motivated by Section 4.3.2, where different generators exhibit distinct correlation statistics, we also evaluate a *limited multi-generator* training regime that augments the core **10k** split introduced by Cozzolino et al. [16] with a small number of additional synthetic samples from diverse sources, while keeping the backbone and training procedure unchanged. Concretely, we add **700** images each from **GLIDE** and five **Stable Diffusion** variants (**SD 1.3/1.4/2.0/XL**) from Synthbuster [4], **588** images from **StyleGAN2-FFHQ** [7], and **1000** images each from **DALL-E 2** and **DALL-E 3** (COCO-Val) following [16]. The resulting training set comprises a total of **16,088** images, as detailed in Table 4.1. This experiment tests whether modest multi-generator supervision improves cross-generator robustness by increasing coverage of generator-specific modes in correlation space.

4.4.3. Effect of Correlation-Augmented Inputs

Table 4.2 summarizes three related comparisons under a fixed ResNet-50 backbone and training budget: whether making inter-channel correlation structure explicit im-

Chapter 4. CHROMA

proves detection, whether the color transform should remain fixed or be learned, and how these conclusions change between single-generator training and the limited multi-generator supervision regime (†). We report per-generator AUC under the protocol of Cozzolino et al. [16].

Do correlation-augmented inputs help? The main conclusion is nuanced rather than absolute: correlation augmentation can provide complementary forensic evidence, but only when the representation is well chosen, and it is not uniformly beneficial across generators. Among the fixed variants, augmenting RGB with Lab correlation maps yields the strongest and most consistent gains on many GAN and diffusion families, indicating that inter-channel dependence in a perceptual color space captures traces that RGB intensities alone do not. Under single-generator training, RGB+Corr(Lab) raises AUC on several sources (e.g., StyleGAN2: 70,3 % $\rightarrow$ 78,9 %; SDXL: 93,3 % $\rightarrow$ 97,3 %; Ediff-I: 73,3 % $\rightarrow$ 83,5 %), and under limited multi-generator training RGB+Corr(Lab)[†] achieves the highest overall average (85,6 %), improving on RGB[†] (83,2 %) by +2,4 pp and on its single-generator counterpart (73,9 %) by +11,7 pp. At the same time, the benefit is clearly representation-dependent: adding RGB correlation maps alone typically reduces AUC and does not reliably help when combined with Lab correlations. The failure cases are equally informative. Adobe Firefly remains a severe outlier for the correlation-based variants, including the trainable ones: several models fall far below chance on Firefly (e.g., 3,8 % for RGB+Corr(Lab)[†] in Table 4.2), implying that the learned decision rule becomes effectively inverted for that generator. DALL-E 3 shows a similar, though less extreme, tendency: Lab correlations systematically reduce its AUC relative to RGB alone (e.g., 43,6 % $\rightarrow$ 9,0 % in the single-generator setting, and 37,1 % $\rightarrow$ 20,2 % in the multi-generator regime). These cases suggest that correlation cues learned from the training generators can become misleading for some commercial pipelines whose color processing may better approximate real-camera statistics.

Trainable transformations: learning the color reparameterization. Given the strong dependence on the chosen color space, we also evaluate the trainable transformations introduced in Section 4.4.1. The Lab-initialized trainable mapping achieves performance close to the best fixed Lab-correlation variants, but does not substantially outperform the handcrafted Lab representation. The deeper 1×1-convolution stack performs slightly below, indicating that extra flexibility does not automatically translate into better cross-generator generalization.

This result deserves closer examination, because the trainable transform is *initialized* at the fixed Lab conversion and therefore starts from a representation that already provides strong forensic separation (Section 4.3.2). One might expect that end-to-end fine-tuning could only improve upon this starting point by adapting the color mixing to the detection objective. Instead, we observe a slight but consistent degradation on unseen generators, which we attribute to a form of *overfitting to the training-set generator distribution*.

To concretely illustrate the extent of this drift, we report the post-training parameter values of the Lab-initialized transform, using the same notation as the standard sRGB-to-Lab pipeline (Appendix C). The learned linearization is

$$C_{\text{lin}} = \begin{cases} \left(\frac{C' + 0,0637}{1,0660} \right)^{2,382}, & C' > 0,0404, \\ \frac{C'}{12,564}, & \text{otherwise,} \end{cases} \quad (4.1)$$

to be compared with the standard sRGB parameters (0,055, 1,055, 2,4, 0,04045, 12,92).

4.4. From Forensic Cues to Actual Detection

The learned RGB-to-XYZ matrix and white point are

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0,4271 & 0,3576 & 0,1601 \\ 0,2104 & 0,6934 & 0,0848 \\ 0,0116 & 0,1338 & 0,9483 \end{bmatrix} \begin{bmatrix} R_{\text{lin}} \\ G_{\text{lin}} \\ B_{\text{lin}} \end{bmatrix}, \quad (X_n, Y_n, Z_n) = (0,9272, 0,9910, 1,0585), \quad (4.2)$$

versus the standard sRGB/D65 matrix and white point (0,9505, 1,0000, 1,0888). The learned Lab-stage parameters become

$$L^* = 114,09 f\left(\frac{Y}{Y_n}\right) - 15,81, \quad (4.3)$$

$$a^* = 491,92 \left[f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right) \right] - 0,022, \quad (4.4)$$

$$b^* = 196,77 \left[f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right) \right] - 0,022, \quad (4.5)$$

with

$$f(t) = \begin{cases} t^{0,3385}, & t > 0,00872, \\ 7,691 t + 0,1339, & \text{otherwise,} \end{cases} \quad (4.6)$$

to be compared with the standard Lab constants (116, -16, 500, 200, 0) and (1/3, 0,008856, 7,787, 16/116), respectively.

The learned values stay close to their standard Lab counterparts across all stages of the pipeline, confirming that the optimization does not discover a radically different color space but rather applies small, coordinated perturbations to the Lab parameterization.

From a regularization perspective, the fixed Lab transform acts as an *implicit inductive bias*: it constrains the correlation computation to a perceptually meaningful coordinate system whose structure is grounded in color science, independently of any particular generative model. Releasing this constraint introduces a classic bias-variance trade-off. The deeper 1×1 stack amplifies this effect: its greater capacity makes it easier to fit training-set correlations, yet the resulting representation departs further from the principled Lab geometry that confers cross-generator robustness, yielding even lower average AUC.

This finding carries a practical implication: under limited and low-diversity training data, the most effective strategy for correlation-based detection is to *fix* the color transform to a well-chosen representation and let the backbone learn to exploit the resulting correlation maps, rather than jointly optimizing the color parameterization. The fixed transform effectively provides a strong prior that the small training set cannot improve upon; attempting to learn a “better” mapping risks overfitting to the narrow generator coverage available during training.

Single-generator vs. limited multi-generator training. The third pattern in Table 4.2 is that moving from single-generator training to the limited multi-generator regime (†) substantially improves cross-generator robustness, especially for the correlation-augmented variants. The clearest example is ScoreSDE, which rises from ~20 % AUC under single-generator training to 97,5 % under the multi-generator Lab-correlation variant, a gain of roughly 75 pp. This supports the view developed in Section 4.3.2: correlation statistics exhibit generator-dependent modes, and even modest exposure to additional generators improves coverage of those modes, producing better generalization to unseen sources. At the same time, the remaining weak cases show that extra generator diversity is not a complete solution. Several commercial tools remain difficult even in the multi-generator setting, suggesting that within-generator source shifts

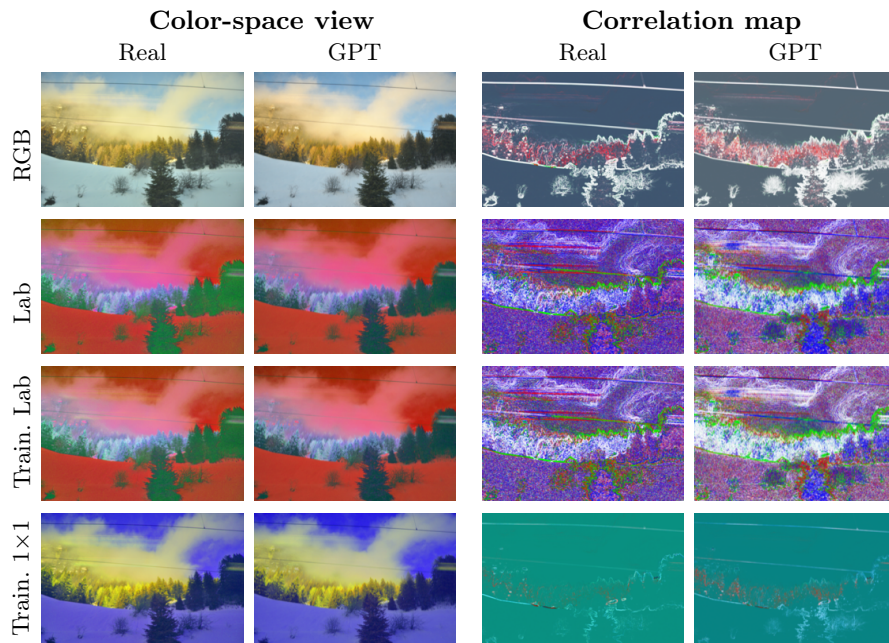


Figure 4.5: Real photograph vs. GPT [68] recreation under the color-space transforms and correlation maps used in this work. Each row shows one representation; left pair: color-space view, right pair: corresponding local correlation map.

and out-of-distribution processing pipelines can still alter low-level statistics enough to induce domain mismatch or negative transfer. Since these systems are black boxes, we cannot determine which aspects of their generation pipeline make them closer to real-camera statistics, but the results indicate that improving robustness on such cases will likely require more than input augmentation alone.

Figure 4.5 provides a qualitative complement to these quantitative trends. Although the real photograph and its GPT-generated recreation remain visually very similar in the image domain, their differences become more evident when examined through local correlation maps. In the left pair of columns, the RGB images are virtually indistinguishable, and the transformed color-space views (Lab, Train. Lab, Train. 1×1) reveal only subtle differences that do not appear to carry sufficient discriminative information on their own. By contrast, the right pair of columns shows that the corresponding correlation maps expose clearer structural differences, concentrated around textured regions and object boundaries such as the tree line, fine vegetation structure, and high-contrast contours. The snow-covered areas also exhibit noticeably different responses in the Lab-based representations, where the real and generated images diverge more clearly. This visual example helps explain why correlation maps, particularly in Lab, can add useful evidence even when the underlying color-space images look nearly identical.

Input variant	GAN family					Diffusion family									Commercial tools				AVG
	Pro GAN	Style GAN2	Style GAN3	Style GAN-T	Giga GAN	Score SDE	ADM	GLIDE	Latent Diff.	Stable Diff.	DeepFL IF	Ediff-I	DiT	SDXL	DALL- E2	DALL- E3	Midj.	Adobe Firef.	
RGB	88.9	70.3	82.6	94.7	76.5	22.2	71.7	96.2	88.2	99.1	90.5	73.3	93.9	93.3	57.1	43.6	88.0	<u>49.0</u>	76.6
RGB + Corr(RGB)	65.1	47.6	51.6	10.5	54.1	16.9	58.3	66.2	62.8	87.5	79.8	59.3	70.5	69.9	42.3	58.1	49.9	69.2	56.6
RGB + Corr(Lab)	88.1	78.9	82.5	99.3	74.5	20.1	76.9	98.0	93.5	99.6	92.1	83.5	96.0	<u>97.3</u>	52.8	9.0	78.1	9.3	73.9
RGB + Corr(RGB) + Corr(Lab)	88.3	73.2	81.6	96.9	73.2	22.0	73.4	95.1	89.7	98.7	90.9	77.3	94.1	94.2	52.0	15.5	83.7	8.8	72.7
RGB [†]	91.0	68.5	58.0	<u>99.9</u>	77.5	<u>96.8</u>	81.0	100.0	94.4	100.0	<u>97.6</u>	59.3	99.1	100.0	100.0	37.1	99.8	36.9	<u>83.2</u>
RGB + Corr(RGB) [†]	89.4	71.8	59.2	99.5	71.7	92.1	81.4	99.8	93.8	100.0	<u>97.6</u>	65.8	94.7	100.0	100.0	<u>44.5</u>	<u>98.6</u>	30.0	82.8
RGB + Corr(Lab) [†]	96.1	89.4	86.5	100.0	82.7	97.5	84.8	99.6	97.4	<u>99.9</u>	98.8	87.7	98.3	100.0	100.0	20.2	97.7	3.8	85.6
RGB + Corr(RGB) + Corr(Lab) [†]	97.4	<u>85.4</u>	<u>86.9</u>	100.0	<u>84.4</u>	96.3	<u>82.5</u>	99.6	95.2	99.8	98.8	<u>87.2</u>	<u>99.0</u>	100.0	52.0	15.5	83.7	8.8	81.8
RGB + Corr(Lab Trainable) [†]	<u>97.0</u>	81.2	87.4	100.0	85.0	97.5	80.5	99.7	<u>95.8</u>	<u>99.9</u>	98.9	83.0	<u>99.0</u>	100.0	<u>99.9</u>	12.7	98.1	4.9	<u>84.5</u>
RGB + Corr (1x1 Trainable) [†]	89.8	74.8	64.5	99.1	77.6	90.8	<u>82.5</u>	<u>99.9</u>	95.0	100.0	97.3	64.5	96.8	100.0	<u>99.9</u>	25.7	96.6	21.5	82.0

Table 4.2: Ablation over input representations (AUC, %) under the benchmark protocol of Cozzolino et al. [16]. † indicates training with additional generator data; gray marks in-domain generators. **Bold** and underline denote the best and second-best results per column.

4.4.4. Comparison to Prior Detectors

Table 4.3 presents a comparison between our correlation-augmented ResNet-50 variants and representative detectors under the benchmark protocol of Cozzolino et al. [16]. While many competing approaches rely on substantially larger training budgets, specialized preprocessing, or ensembles, our objective is to quantify how far a lightweight representation change can go when paired with a standard CNN backbone.

The primary conclusion from this comparison is a limit: CHROMA is not reliable as a stand-alone universal detector. Although the method is strong on many GAN/diffusion generators, it shows severe failures on some commercial systems, so aggregate averages must be interpreted together with worst-case behavior.

Concretely, Adobe Firefly (3,8 %) and DALL-E 3 (20,2 %) are well below chance for RGB+Corr(Lab)[†], which means the learned decision is effectively *reversed* on those sources (Table 4.2). This is operationally unsafe without complementary checks. The contrast with Cozzolino et al. [16] is substantial: the same generators are detected at 87,2 % and 92,9 % AUC by that method, indicating that complementary cues (e.g., spectral signals) can succeed where color-correlation evidence fails. Conditioned on that limitation, CHROMA remains strong under constrained training for many sources. In particular, RGB+Corr(Lab)[†] is best or second-best on ScoreSDE (97,5 %), where several prior methods are much lower (Wang et al. 32,9 %, Corvi et al. 65,0 %, LGrad 80,6 %), and reaches 97,7 % on Midjourney. The best overall average in the table is still Cozzolino et al. 10k+ (90,0 %), which exceeds CHROMA (85,6 %) by 4,4 pp. Taken together, the supported claim is narrower: correlation cues are a useful and efficient complementary stream, but robust generalization should rely on cue fusion rather than on color-correlation evidence alone [74].

Relatedly, recent work shows that robustness in synthetic image detection is driven by generator diversity and training scale, often more so than by architectural complexity [70]. In this context, CHROMA’s results suggest that correlation cues can be learned efficiently under limited supervision, but they do not by themselves resolve hard commercial outliers. A realistic path forward is therefore twofold: increase training diversity and combine correlation cues with complementary forensic evidence.

Method	GAN family					Diffusion family									Commercial tools				AVG
	Pro GAN	Style GAN2	Style GAN3	Style GAN-T	Giga GAN	Score SDE	ADM	GLIDE	Latent Diff.	Stable Diff.	DeepFl. IF	Ediff-I	DiT	SDXL	DALL-E2	DALL-E3	Midj.	Adobe Firef.	
Wang et al.	100	<u>96.5</u>	<u>98.5</u>	98.9	66.6	32.9	64.3	48.5	59.2	41.5	78.0	64.9	58.6	54.3	64.8	10.9	40.2	84.8	64.6
PatchFor.	92.3	84.5	91.8	91.2	64.7	83.3	74.8	96.2	78.1	62.4	62.7	78.7	83.1	68.4	41.9	52.7	57.8	49.4	73.0
Grag. et al.	100	99.8	97.5	98.8	82.8	92.1	74.7	62.8	91.9	52.5	69.9	69.6	65.3	58.0	58.3	2.4	43.1	63.5	71.3
Mand. et al.	96.2	93.8	100.0	92.6	61.8	99.8	56.5	40.5	70.0	36.8	47.2	65.0	59.1	27.0	14.5	14.7	24.3	36.7	57.6
Liu et al.	100	99.8	98.4	98.5	98.2	95.4	82.5	76.5	<u>97.6</u>	77.4	72.2	98.7	88.0	31.1	70.4	0.2	40.7	11.8	74.3
Corvi et al.	79.4	73.7	50.0	97.1	63.4	65.0	80.7	91.9	100.0	100.0	99.9	85.7	100.0	100.0	69.4	<u>60.8</u>	100.0	98.0	84.2
LGrad	100	91.2	83.8	81.8	82.2	80.6	76.9	66.1	81.1	61.5	68.8	74.1	56.2	57.2	58.6	37.9	56.3	40.6	69.7
Ojha et al.	100	93.9	92.3	98.2	<u>96.0</u>	58.4	86.7	80.8	85.7	89.5	92.9	80.6	77.8	85.1	95.2	36.4	66.2	<u>97.5</u>	84.1
DIRE-1	50.6	56.9	47.8	<u>99.9</u>	74.1	44.3	<u>75.7</u>	71.4	68.7	39.4	<u>98.9</u>	<u>99.1</u>	<u>99.6</u>	47.1	44.7	47.6	51.0	57.4	65.2
DIRE-2	54.2	52.5	43.0	99.6	76.0	41.0	70.1	70.1	69.3	46.9	97.0	98.2	98.3	42.8	41.0	49.6	47.8	43.0	63.3
NPR	100	85.6	77.0	96.4	88.7	91.1	<u>86.3</u>	79.3	90.2	64.5	91.6	80.1	78.4	76.7	39.5	48.7	77.0	32.1	76.8
Cozzolino et al. 10k+	93.4	87.1	87.6	<u>99.9</u>	78.5	89.2	79.9	99.7	<u>84.7</u>	91.3	97.9	99.4	94.0	<u>90.1</u>	86.3	92.9	81.7	87.2	90.0
RGB + Corr(Lab) [†]	96.1	89.4	86.5	100.0	82.7	<u>97.5</u>	84.8	<u>99.6</u>	97.4	<u>99.9</u>	98.8	87.7	98.3	100.0	100.0	20.2	97.7	3.8	<u>85.6</u>
RGB + Corr(RGB) + Corr(Lab) [†]	<u>97.4</u>	85.4	86.9	100.0	84.4	96.3	82.5	<u>99.6</u>	95.2	99.8	98.8	87.2	99.0	100.0	52.0	15.5	83.7	8.8	81.8
RGB + Corr(Lab Trainable) [†]	97.0	81.2	87.4	100.0	85.0	<u>97.5</u>	80.5	99.7	95.8	<u>99.9</u>	<u>98.9</u>	83.0	99.0	100.0	<u>99.9</u>	12.7	<u>98.1</u>	4.9	84.5

Table 4.3: Comparison with prior methods in terms of AUC (%) under the benchmark protocol of Cozzolino et al. [16]. Results are reported across GAN-based generators, diffusion-based generators, and commercial tools. Gray entries indicate the in-domain generator(s) associated with each method's training setup as reported by the protocol, whereas the remaining columns measure cross-source generalization. **Bold** marks the best overall average and underline marks the second-best.

4.5. Chapter summary

This chapter first introduced the color-space foundations needed for correlation-based forensics, including the distinction between encoded and linearized values, the role of white balance and color correction in shaping inter-channel dependence, and the color representations used throughout. It then investigated inter-channel color-space correlations as forensic evidence for AI-generated image detection and translated the findings into a practical detection pipeline.

The analysis proceeded in two stages. First, we showed that LPIPS, a perceptual metric widely used in generative training, responds inconsistently to perturbations that alter inter-channel dependence: depending on the color space in which the perturbation is expressed, the metric can nearly collapse the distinction between correlated and decorrelated noise, or even reverse the expected ordering. This unevenness implies that generators optimized with perceptual objectives are not uniformly pressured to match camera-like correlation structure across all color parameterizations, leaving exploitable gaps. Second, we characterized these gaps empirically by computing local inter-channel correlation distributions for real photographs and multiple diffusion-era generators across RGB, Lab, HSV, and YUV, finding that correlation statistics carry generator-dependent signatures whose strength varies substantially with the color space: RGB and Lab provide the strongest cross-source separation, while YUV correlations are nearly indistinguishable between some generators and real images.

The detection experiments confirmed that these analytical findings translate into measurable gains. Augmenting a standard ResNet-50 with Lab correlation maps consistently improved cross-generator AUC relative to RGB-only inputs, obtaining competitive performance with substantially more complex pipelines. The ablation study further revealed that the benefit is color-space-dependent (RGB correlations alone are detrimental), that trainable color transforms do not outperform fixed Lab under limited data, and that even modest multi-generator supervision yields dramatic improvements on previously undetectable generators. At the same time, the results exposed clear limitations: certain commercial tools (Adobe Firefly, DALL-E 3) consistently yielded inverted or near-chance detection, indicating that inter-channel correlation evidence alone is insufficient for robust real-world use and that complementary cue families remain necessary.

Chapter 5 synthesizes these contributions, discusses their limitations in broader context, and outlines directions for future work, including data scaling, localization via spatially resolved correlation maps, and fusion with complementary forensic signals.

Chapter 5

Conclusions and Future Work

This thesis investigated the detection of AI-generated images through the lens of inter-channel color-space correlations. Starting from the observation that camera imaging pipelines and generative models produce systematically different low-level color statistics under the datasets and protocols tested, we developed CHROMA, a forensic framework that makes these differences explicit and actionable. The central hypothesis, that widely used perceptual training objectives do not uniformly penalize departures from camera-like inter-channel correlation structure, was confirmed for the majority of generator families and color-space representations evaluated, but qualified by the finding that certain commercial systems (Adobe Firefly, DALL-E 3) do not exhibit exploitable correlation discrepancies under the learned decision boundary.

5.1. Summary of contributions and answers to the guiding questions

The thesis was guided by three questions, formulated in Chapter 1. We began by establishing the two image-production paradigms (Chapter 2): the camera pipeline, from CFA sampling and demosaicing to white balance, color correction, and compression, which systematically shapes inter-channel dependencies in natural photographs; and the synthesis pipelines of modern generative models (GANs, diffusion, and commercial tools), whose training objectives and architectural choices need not reproduce camera-like chromatic coupling. Building on this contrast, Chapter 4 established the color-space foundations needed for correlation-based forensics and developed the CHROMA framework. Below we summarize the four contributions of this thesis, organized around the questions they address.

Comprehensive survey of the detection landscape. *What are the major lines of research in AI-generated image detection, and what gaps remain?* The comprehensive survey presented in Chapter 3 identified six major evidence families: spatial-domain discriminative detectors, frequency and residual cues, diffusion self-consistency approaches, color-centric methods, frozen-representation detectors, and systems addressing real-world generalization through continual updates, localization, and explanation. A central finding is that no single evidence family is universally reliable; each exhibits characteristic failure modes under cross-generator or cross-condition evaluation. Within this landscape, inter-channel *joint* correlations, and in particular their dependence on

Chapter 5. Conclusions and Future Work

the choice of color-space representation, were identified as specifically underexplored. This survey provides the analytical foundation on which the experimental contributions rest, and directly motivated the work in Chapter 4.

Sensitivity analysis of perceptual metrics. *Are perceptual similarity metrics sensitive to changes in inter-channel dependence in a consistent way across color spaces?* No. Through controlled perturbation experiments on the RAISE-1k dataset, we showed that LPIPS responds inconsistently to changes in inter-channel dependence across color parameterizations. It nearly collapses the distinction between correlated and decorrelated noise in Lab, and reverses the expected ordering in RGB, where perturbations that *preserve* channel coupling are rated as more distorted than those that disrupt it. This finding has direct implications for generative training: if perceptual objectives do not uniformly penalize departures from camera-like correlation structure, generators can produce perceptually convincing images while still exhibiting exploitable inter-channel discrepancies. These results confirm the first part of our hypothesis: perceptual objectives used in generative training do not uniformly enforce camera-like correlation structure, leaving color-space-dependent gaps that generators need not close.

Characterization of generator-specific correlation signatures. By computing local inter-channel correlation maps across RGB, Lab, HSV, and YUV and measuring distributional distances between real photographs (RAISE-1k) and synthetic images from multiple diffusion-era generators (Synthbuster), we showed that inter-channel correlations carry *generator-dependent* signatures. RGB and Lab consistently provided the strongest separability, while YUV proved nearly indistinguishable between certain generators and real images. These results established that the choice of color representation is not a minor implementation detail but a critical design decision for correlation-based forensics.

Correlation-augmented detection. *Do explicit inter-channel correlation descriptors improve AI-generated image detection under realistic cross-generator evaluation?* Yes, with qualification. We translated the preceding analyses into a practical detection pipeline by augmenting a standard ResNet-50 backbone with inter-channel correlation maps computed in Lab. Under the cross-generator benchmark of Cozzolino et al. [16], the RGB+Corr(Lab) variant achieved an average AUC of 85,6% with the multi-generator training regime, competitive with substantially more complex pipelines, using only $\sim 16k$ training images and a standard CNN backbone. The ablation study further showed that: (i) Lab correlations consistently improve over RGB-only inputs across most generator families; (ii) RGB correlations alone are detrimental, confirming that the utility of correlation cues is color-space-dependent; (iii) trainable color transformations do not outperform fixed Lab under limited training data, as the learned parameters stay close to their Lab counterparts and the additional flexibility leads to overfitting; and (iv) even modest multi-generator supervision yields dramatic improvements on previously undetectable generators (e.g., ScoreSDE rising from $\sim 20\%$ to 97,5% AUC). However, the cue fails for Adobe Firefly and DALL-E 3, where the learned decision is effectively inverted, indicating that correlation evidence alone cannot serve as a universal classifier.

5.2. Limitations

Several limitations should be acknowledged.

5.3. Future work

First, inter-channel correlation cues are not universally effective. Adobe Firefly and DALL-E 3 consistently yielded AUC values well below 50% for Lab-correlation variants, indicating that the learned decision is effectively *inverted* for these generators. This suggests that some commercial pipelines may better approximate natural camera color statistics, or differ from diffusion-family generators in ways that cause negative transfer under the learned decision boundary. The contrast with methods such as Cozzolino et al. [16], which detect these same generators at 87%–93% AUC using complementary spectral cues, underscores that correlation evidence alone is insufficient as a universal classifier.

Second, our experiments were conducted under the specific evaluation protocol of Cozzolino et al. [16], which, while rigorous, does not cover all real-world conditions. The robustness of correlation cues under heavy post-processing (e.g., aggressive JPEG recompression, social-media re-encoding, or screenshot-based re-acquisition) was not systematically evaluated and remains an open question.

Third, training data diversity was limited by design: the multi-generator regime used only ~16k images from a handful of generator families. While this constraint was intentional, designed to test how far a lightweight representation change can go, it also means that the method’s ceiling under richer supervision remains unexplored.

Fourth, the Firefly/DALL-E 3 inversion carries practical consequences: a forensic system relying solely on correlation-based scores risks systematic false negatives on these generators, or, if the score polarity is not calibrated per source, systematic false positives on real images. In any real-world pipeline, correlation evidence should therefore be combined with complementary cue families and accompanied by calibration or abstention mechanisms that account for generator-dependent score distributions.

5.3. Future work

The findings of this thesis open several promising directions for future research, which we organize by expected impact and timeline.

5.3.1. Near-term directions

Scaling training data diversity. The ablation study revealed that even a modest expansion of the training set from a single generator (10k images) to a limited multi-generator corpus (~16k images) yielded dramatic improvements in cross-generator robustness, with some generators jumping from near-random to near-perfect detection. Recent work has shown that generator diversity in the training set is often a more powerful lever for generalization than architectural complexity [70]. Scaling the training corpus to include a substantially larger and more diverse set of generators, spanning both open-source models and commercial tools across multiple generation paradigms and post-processing pipelines, is therefore a natural and high-impact next step. Given that CHROMA already benefits substantially from relatively modest increases in training diversity, such scaling could yield substantial gains without requiring architectural changes.

Fusion with complementary forensic evidence. The uneven performance across generator families reinforces a recurrent finding in the forensic literature: no single cue is universally reliable [74]. The generators where CHROMA struggles (notably Adobe Firefly and DALL-E 3) are precisely those where spectral and representation-based methods succeed, and vice versa. This complementarity motivates simple fusion

Chapter 5. Conclusions and Future Work

strategies, such as score-level combination of correlation-based, spectral, and frozen-representation detectors, that could substantially improve coverage across the generator landscape while retaining the interpretability and efficiency of individual cue families.

Robustness under post-processing. Real-world images are rarely encountered in their original form: social-media platforms apply resizing, recompression, format conversion, and other transformations that can attenuate forensic traces. Systematically evaluating and improving the robustness of correlation-based cues under these conditions is essential for practical use. Training with augmentations that simulate common laundering pipelines, in line with prior approaches [21, 26], and evaluating under standardized post-processing benchmarks would clarify the operational limits of inter-channel correlation evidence.

5.3.2. Medium-term directions

Correlation maps as input to segmentation architectures for localization. A compelling extension is to move from image-level detection to pixel-level *localization* of AI-generated content. The correlation maps computed in this thesis are spatially resolved, making them natural candidates as input channels for dense prediction architectures such as U-Net [84]. Concretely, a U-Net (or similar encoder-decoder architecture) could receive correlation maps as additional input channels, alongside or in place of RGB, and be trained to predict a pixelwise mask indicating which regions of an image are synthetically generated. This would be particularly valuable for detecting inpainting and localized edits, where only a portion of the image has been manipulated, and the correlation structure in the edited region may differ systematically from the surrounding camera-captured content. The spatial locality of our correlation descriptors and their demonstrated sensitivity to generator-specific traces suggest that such an approach could provide interpretable, fine-grained localization evidence.

Learned color spaces under richer supervision. Our experiments showed that trainable color transformations overfit under limited data, with the learned parameters converging to slight perturbations of the Lab initialization. However, this finding may be specific to the low-data regime. Under substantially larger and more diverse training corpora, learned color representations could potentially discover more informative parameterizations that capture generator-specific correlation structures not accessible through any fixed color space. Revisiting trainable transforms with the data-scaling strategies advocated by Park and Owens [70] is a natural follow-up.

5.3.3. Long-term and exploratory directions

Extension to video and temporal consistency. As AI-generated video becomes increasingly prevalent, extending correlation-based analysis to the temporal domain is a timely direction. Temporal consistency of inter-channel correlations across frames could provide additional forensic evidence, since generative video models may exhibit frame-to-frame correlation fluctuations that differ from those of camera-captured video, where the physical imaging pipeline imposes consistent chromatic behavior across a shot.

Adversarial robustness and counter-forensics attacks. While inter-channel correlation cues are rooted in physical camera processes, their vulnerability to targeted ad-

5.3. Future work

versarial attacks deserves investigation. An adversary aware of the correlation-based decision rule could potentially craft perturbations that match camera-like correlation statistics while preserving generative content. Understanding the difficulty of such attacks, and developing defenses through adversarial training or ensemble strategies, is important for assessing the long-term viability of color-based forensic evidence.

Revisiting spectral anomaly detection with stronger representations. The Spectral-Flow experiment documented in Appendix A showed that learning a one-class density model over spectral-residual features of real images is fundamentally limited by the diversity of acquisition conditions: a feature extractor trained from scratch cannot distinguish “a different kind of real” from “generated.” However, the core idea of anomaly-based detection, which requires no generator-specific training data and could therefore remain effective as the generative landscape evolves, remains appealing.

A natural first reaction is to replace the from-scratch encoder with a frozen, pretrained backbone (e.g., CLIP or DINOv2) that already encodes broad visual priors. However, to the best of our knowledge, no pretrained encoders exist for spectral-residual or Fourier-domain inputs: available foundation models operate on RGB images and have not been trained on the kind of frequency-domain representations that Spectral-Flow requires. Fine-tuning such models on spectral inputs is possible but would sacrifice the very generality that makes pretrained representations attractive.

A more promising direction may therefore be to keep the from-scratch encoder but address the root cause of its failure: insufficient diversity in the training data. Our experiment trained on datasets that, despite their size, were dominated by particular camera families and compression standards. Training on a substantially larger and more diverse corpus of real images, spanning many cameras, resolutions, compression levels, and post-processing pipelines, could expand the learned normality region to better reflect the true variability of real-world photography. Equally important is the training strategy itself: exploring data augmentation that explicitly simulates the transformations that most affect spectral residuals (resizing, JPEG and WebP recompression at varied quality factors, re-encoding through social-media pipelines, and mild sharpening or denoising) could teach the encoder to produce representations that are invariant to these benign perturbations while remaining sensitive to generator-induced artifacts. Combining richer data with augmentation-aware training and the inter-channel correlation evidence developed in this thesis could yield a multi-cue anomaly detector that leverages both frequency-domain and color-domain forensic axes without requiring labeled generated data.

In summary, this thesis has shown that inter-channel color-space correlations constitute a lightweight, physically grounded, and partially orthogonal forensic signal for AI-generated image detection. While not a universal solution, CHROMA demonstrates that carefully chosen color representations can extract complementary evidence from images at minimal computational cost, and that the interplay between perceptual training objectives and color-space geometry creates exploitable gaps in current generative models. For practitioners building forensic pipelines today, inter-channel correlation evidence is best used as a complementary signal alongside spectral and representation-based detectors rather than as a standalone decision criterion. As the generative landscape continues to evolve, we expect that combining correlation-aware analysis with diverse training data, complementary cue families, and spatially resolved architectures will be a productive path toward robust and interpretable forensic systems.

This page has been intentionally left blank.

Appendix A

Spectral-Flow: An Exploratory Frequency-Domain Anomaly Detector

This appendix documents an exploratory line of work that preceded the inter-channel correlation approach developed in Chapter 4. Motivated by the frequency and residual cues surveyed in Section 3.3.1 of Chapter 3, we investigated whether generated images could be detected through anomaly detection rather than binary classification: instead of training a model to distinguish real from fake, we attempted to learn what “normal” (real) images look like in a spectral-residual feature space and flag anything that deviates as potentially generated.

The experiment did not yield competitive detection performance. We include it here for completeness, because the failure itself is informative: it highlights fundamental difficulties in defining a universal model of normality over spectral-residual features, and it motivated the pivot toward the discriminative color-correlation cues that became the central contribution of this thesis. Section A.1 presents the motivation and hypothesis, Section A.2 describes the architecture, Section A.3 analyzes the negative results, and Section A.4 summarizes the lessons learned.

A.1. Motivation and hypothesis

The core idea behind Spectral-Flow is straightforward. Most existing detectors are binary classifiers: they require examples of both real and generated images, and their performance depends heavily on which generators were seen during training. We wanted to take a different approach and build an anomaly detector that only needs to learn what *real* images look like. If we could learn a good representation of normality, then any image that deviates from it would be flagged as potentially generated, without ever having seen a specific generator during training. This is appealing because new generators appear constantly, and a model that does not depend on generator-specific examples would, in principle, remain effective as the landscape evolves.

The choice of spectral residuals as the input representation was motivated by prior work surveyed in Chapter 3. Noise-residual pipelines suppress scene semantics by subtracting a denoiser output from the image and then analyze the resulting residual in the Fourier domain [15, 52]. When averaged over many images, the power spectra

Appendix A. Spectral-Flow: An Exploratory Frequency-Domain Anomaly Detector

of these residuals reveal architecture-dependent peaks tied to the upsampling ratios of generators [15], and even lightweight spectral baselines achieve reasonable detection of diffusion-generated images [4]. In parallel, recent work models the spectral distribution of real images as a prior and treats synthetic images as out-of-distribution with respect to this learned reference [43]. These results suggested that spectral residuals carry enough discriminative information to support a one-class detection approach.

Our hypothesis combined these observations into a concrete pipeline with two training phases. In the first phase, we train a convolutional autoencoder on the spectral-energy tensors of real images, so that the encoder learns a compact latent representation of what real spectral residuals look like. In the second phase, we freeze the encoder and fit a normalizing flow exclusively on those real-image latents, learning their density. At inference time, any new image is passed through the frozen preprocessing and encoder, and the normalizing flow scores it: real images should receive high likelihood (low anomaly score), while generated images, whose spectral residuals differ from the learned normal distribution, should receive low likelihood (high anomaly score).

This formulation relies on a critical assumption: that the spectral-residual features of real images occupy a well-defined, compact region of the latent space, and that this region generalizes across acquisition conditions, cameras, and post-processing pipelines. As we show in Section A.3, this assumption does not hold in practice.

A.2. Architecture overview

Spectral-Flow is an end-to-end pipeline organized in three stages, illustrated in Figure A.1: a spectral energy extractor that transforms raw images into normalized spectral-energy tensors, an autoencoder that learns a latent representation of these tensors, and a normalizing flow that models the density of real-image latents. We describe each stage below.

A.2.1. Stage 1: Spectral energy extraction

The preprocessing stage transforms a single image into a spectral-energy tensor that amplifies generator-specific frequency cues while suppressing scene content. Following the residual-extraction protocol of Corvi et al. [15], we compute the noise residual

$$r = x - D(x; \sigma), \quad (\text{A.1})$$

where x is the input RGB image and $D(\cdot; \sigma)$ is a pre-trained, frozen DnCNN denoiser [105] with $\sigma = 1$. Operating in the residual domain suppresses high-level scene semantics and exposes low-level periodicities and autocorrelation regularities that are characteristic of the generator pipeline [15].

Patch extraction. The residual is split into overlapping local patches $\{r_u = P_u(r)\}_{u=1}^U$ using a sliding window with patch size $p_H \times p_W$ and stride $s_H \times s_W$. Cropping to local regions mitigates spectral leakage and improves stationarity: each patch behaves as a more stationary signal than the full residual, yielding cleaner Fourier representations. Additionally, averaging across multiple patches decorrelates content-specific energy and increases the effective sample count for estimating spectral statistics.

Dihedral augmentations and group averaging. For each patch r_u , we generate eight transformed copies via the dihedral group $G \subset D_4$ consisting of four rotations ($0^\circ, 90^\circ, 180^\circ, 270^\circ$) and their horizontal flips. This is a key design choice: for any rotation R and reflection

A.2. Architecture overview

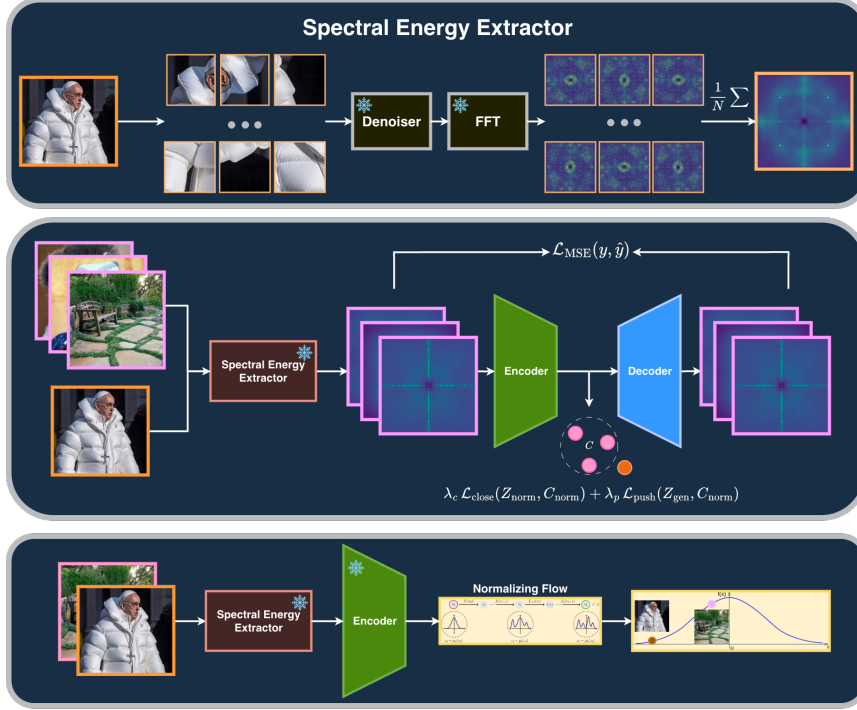


Figure A.1: Architecture of the Spectral-Flow pipeline. **Top:** the spectral energy extractor computes a noise residual via a frozen DnCNN denoiser, extracts overlapping patches, applies dihedral augmentations (rotations and flips), and averages their Fourier power spectra to produce a scale-normalized spectral-energy tensor. **Middle:** a convolutional autoencoder is trained on spectral-energy tensors of real images with an MSE reconstruction loss, augmented by a contrastive term that compacts normal latents and repels generated-image latents. **Bottom:** a normalizing flow is fitted on frozen encoder latents from real images; at inference, the negative log-likelihood under the flow serves as an anomaly score.

H , the spectral energy satisfies $E(\omega; Rr) = E(R^\top \omega; r)$ and $E(\omega; Hr) = E(H^\top \omega; r)$. Averaging over the group orbit therefore attenuates orientation-specific content patterns (which are arbitrarily oriented) while preserving isotropic or group-symmetric periodicities, precisely the kind of periodic structure that generator architectures introduce through their upsampling stages. Under the assumption of independent patches, the signal-to-variance ratio (SVR) of generator-induced peaks scales as $\sqrt{8U}$; in practice the improvement is smaller due to patch overlap and spatial correlation, but the variance reduction is still substantial.

Fourier energy and normalization. For each transformed patch $g \cdot r_u$, we compute the orthonormal 2-D FFT per channel, take the squared magnitude, apply `fftshift`, and average across all $8U$ augmented patches:

$$\hat{E}(\omega) = \frac{\frac{1}{8U} \sum_{u,g} |\mathcal{F}\{g \cdot r_u\}(\omega)|^2}{C \cdot \left(\frac{1}{|\Omega|} \sum_{\omega'} \hat{E}(\omega') + \varepsilon \right)}, \quad (\text{A.2})$$

with $C = 1/4$ and $\varepsilon = 10^{-8}$. The denominator performs scale-invariant normalization by the global mean energy, yielding a representation that is independent of absolute intensity and robust to contrast variations across patches. This normalization is

Appendix A. Spectral-Flow: An Exploratory Frequency-Domain Anomaly Detector

consistent with the observation that diffusion-image frequency detectors benefit from careful spectral normalization [4]. The complete single-image preprocessing pipeline can be summarized as:

$$x \xrightarrow{D} r \xrightarrow{\text{patch}} \{r_u\}_u \xrightarrow{\text{flip/rot}} \{g \cdot r_u\} \xrightarrow{\mathcal{F}, |\cdot|^2, \text{mean}} \hat{E}.$$

A.2.2. Stage 2: Autoencoder with contrastive shaping

A convolutional autoencoder maps the spectral-energy tensor $\hat{E}$ to a latent vector of dimension $d = 512$ via a residual encoder backbone followed by global average pooling and a linear projection. A decoder with transposed convolutions reconstructs the input; the primary training loss is the mean-squared error between the reconstruction and the original spectral-energy tensor, computed over real images only.

To improve the geometry of the latent space, we augment the reconstruction objective with a contrastive term that exploits generated images when they are available during training. The contrastive loss has two components: a compactness term that penalizes normal latents whose Euclidean distance to the batch centroid exceeds a margin m_{norm} , pulling them inward, and a separation term that penalizes abnormal latents whose distance to the normal centroid falls below a margin m_{abn} , pushing them outward. Optionally, a lightweight classifier head operates on *detached* latents with a cross-entropy loss, providing an auxiliary discrimination signal without leaking gradients into the encoder.

The combined objective encourages normal latents to cluster compactly while abnormal latents are pushed to the periphery. This is the geometry that the downstream normalizing flow will attempt to exploit for anomaly scoring.

A.2.3. Stage 3: Normalizing flow on frozen latents

After autoencoder training, the entire upstream pipeline (preprocessing and encoder) is frozen. A normalizing flow, constructed as a sequence of invertible transformations (e.g., rational quadratic spline couplings interleaved with permutations), is then fitted exclusively on the encoder latents of real images by maximizing log-likelihood.

At inference, the anomaly score for a new image is the negative log-probability of its latent under the flow: real images, whose latents should fall in the high-density region learned from training, receive low anomaly scores, while generated images receive high anomaly scores. The raw scores can be calibrated into probabilities via empirical CDF, conformal p-values, or Gaussian typicality, supporting threshold-based decision-making without additional post-processing.

A.3. Latent space analysis and negative results

With the architecture in place, we now examine the structure of the learned latent space. Figure A.2 shows a UMAP projection of the encoder latents for 2,400 images: 600 real photographs drawn from three datasets (ImageNet, FFHQ, and RAISE-1k) and 1,800 synthetic images from nine generators spanning the Synthbuster benchmark.

The visualization reveals two problems that together explain the failure of the anomaly-detection approach. First, and most critically, the real images themselves do not form a single compact cluster. ImageNet, FFHQ, and RAISE-1k samples are scattered across the central region of the projection, reflecting the fact that different real-image datasets, with their different cameras, resolutions, and post-processing pipelines, produce distinct spectral-residual profiles. A normalizing flow fitted on these

A.3. Latent space analysis and negative results

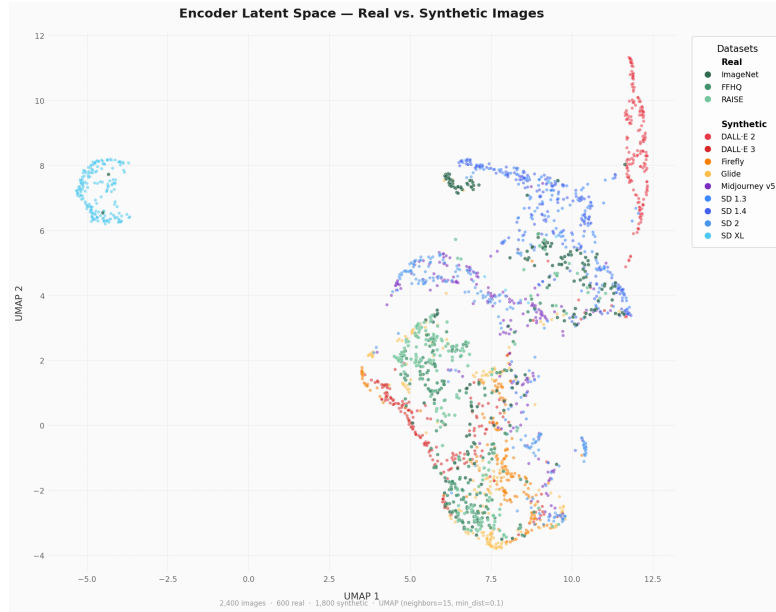


Figure A.2: UMAP projection of Spectral-Flow encoder latents for 2,400 images (600 real, 1,800 synthetic). Real-image sources (ImageNet, FFHQ, RAISE) are shown in shades of green; synthetic generators are shown in distinct colors. The real images do not form a single compact cluster: they are scattered across the central region and overlap heavily with several generators. Meanwhile, some generators (notably SDXL, DALL-E 3) form isolated clusters far from the real images, while others (Glide, Midjourney, earlier SD versions) are interspersed with real data. This fragmented geometry makes density-based anomaly detection unreliable: a normalizing flow fitted on the real-image latents cannot define a coherent normality boundary.

latents would need to cover a fragmented, multi-modal region rather than a tight high-density core, making it difficult to define a coherent boundary between “normal” and “abnormal.”

Second, the generator latents do not occupy a single, well-separated region either. Some generators form isolated clusters: SDXL is pushed to the upper-left corner, far from any real data, and DALL-E 3 occupies its own region in the upper-right. For these generators, a density-based approach could in principle work, since their latents clearly fall outside the support of the real-image distribution. However, other generators, including Glide, Midjourney v5, and earlier Stable Diffusion versions, are heavily interspersed with the real-image latents in the central region. For these sources, the spectral-residual representation simply does not separate real from generated: their frequency-domain characteristics, as captured by the encoder, are too similar to those of the real images they were trained to approximate.

This combination is fatal for anomaly detection. A normalizing flow must assign high likelihood to real-image latents and low likelihood to generated-image latents, but when both populations overlap in the latent space, no density threshold can achieve reliable separation. The model either accepts the interspersed generators as normal (high false negatives) or rejects the scattered real-image clusters as anomalous (high false positives). We identify three interrelated causes for this latent-space geometry.

Appendix A. Spectral-Flow: An Exploratory Frequency-Domain Anomaly Detector

Feature extractor trained from scratch. The convolutional encoder backbone is trained from scratch on spectral-energy inputs. Unlike pretrained backbones (e.g., CLIP or ImageNet-trained ResNets) that have internalized broad visual priors from hundreds of millions of diverse images, our encoder learns features that are tightly coupled to the statistical properties of its training data. The resulting representations are biased toward the specific cameras, resolutions, and post-processing pipelines present in the training set rather than capturing universal spectral characteristics of real photography.

Training data bias. Even large-scale datasets do not span the full space of real-image spectral characteristics. Different cameras produce different sensor noise profiles; different image processing pipelines (demosaicing algorithms, white balance, sharpening, compression) leave distinct traces in the noise residual and its spectrum. Figure A.2 makes this concrete: the three real-image datasets (ImageNet, FFHQ, RAISE-1k), despite all being genuine photographs, do not converge to the same latent region. Instead, they are scattered across the central area of the projection, reflecting the heterogeneity of their acquisition conditions. A density model fitted on this fragmented real-image distribution will inevitably have porous support, with some generators falling inside the high-density regions and some real images falling outside them.

Fragility of spectral-residual cues. The noise-residual spectrum is sensitive to a wide range of benign transformations. Resizing shifts spectral peaks, JPEG compression introduces periodic energy at quantization frequencies, changes in lighting or exposure affect the noise floor, and even minor post-processing (sharpening, denoising, tone mapping) can reshape the residual’s spectral profile [15]. As a consequence, the spectral-energy representation of “normality” is not a stable, well-localized region but a broad, acquisition-dependent manifold. Density modeling over such a space is inherently unreliable: the normalizing flow learns a narrow mode that does not encompass the diversity of real-world conditions.

These findings are consistent with known limitations of likelihood-based anomaly detection in high-dimensional spaces. Nalisnick et al. [62] demonstrated that deep generative models, including normalizing flows, can assign *higher* likelihood to semantically out-of-distribution data than to training data, and Kirichenko et al. [48] traced this to the tendency of coupling-layer flows to learn local correlations and generic transformations rather than semantically meaningful structure. In our setting, the flow models the statistics of spectral-energy tensors from a narrow slice of the real-image distribution and cannot distinguish between “a different kind of real” and “generated.”

A.4. Lessons learned

The Spectral-Flow experiment clarifies a fundamental limitation of anomaly-detection approaches for generated-image forensics. The space of real images, when represented through spectral residuals, is too diverse and acquisition-dependent to be captured by a compact density model. The difficulty is not in the normalizing flow itself but in the upstream representation: a feature extractor trained from scratch on a limited dataset produces latent codes where distributional shift (images from a different camera or pipeline) is indistinguishable from anomaly (images from a generator).

This negative result informed the direction of the thesis in two concrete ways. First, it reinforced the conclusion from the state-of-the-art survey (Chapter 3) that spectral and residual cues, while forensically informative in controlled settings, are fragile under the distribution shifts encountered in real-world use. Second, and more importantly, it

A.4. Lessons learned

motivated the search for forensic signals that are more robust to acquisition variability. The inter-channel color-space correlations developed in Chapter 4 represent such a signal: rather than attempting to model a universal distribution of “normality,” Chroma operates in a discriminative setting where the model learns *what distinguishes* real from generated images through the statistical relationships between color channels. These cues arise from the physics of the image-formation process (demosaicing, white balance, color correction) and are less sensitive to the specific camera or compression pipeline, making them a more stable forensic axis than the spectral residuals explored here.

This page has been intentionally left blank.

Appendix B

Datasets and Benchmarks

Appendix

This chapter groups supplementary dataset and benchmark material referenced throughout the thesis, with a primary focus on the evaluation setting summarized in Section 3.1.2 of Chapter 3. The same benchmark sources (notably RAISE-1k and Synthbuster) are also used in the empirical analyses of Chapter 4. It is organized to support future expansion with additional datasets while keeping per-dataset details self-contained.

B.1. Synthbuster Dataset Details

B.1.1. Prompt Set

Table B.1 lists the specific prompts used to generate the sample images shown in Figure 3.7. Each row corresponds to a unique scene (identified by the real image ID from RAISE-1k), and the prompt was used to generate synthetic variations across multiple models.

B.1.2. Uncropped Sample Images

Figures B.1 and B.2 display the full, uncropped versions of the sample images shown in Figure 3.7.

Appendix B. Datasets and Benchmarks Appendix



Figure B.1: Uncropped sample images from the Synthbuster benchmark [4] (Part 1/2).

B.1. Synthbuster Dataset Details

Image ID	Prompt
r152d9ff5t	Mountains covered in snow and trees with a blue sky, spring day, big, beautifully bright white, idaho, corduroy, white cloud, family.
r03ed2669t	A village with a townhouse by a river, in the style of religious building, the building are light yellow and beige clay with grey rooves, view from the other side of the river, aerial photography.
r01170470t	Windmill on a lake, in the style of dutch tradition, floating wooden building, green and brown, shiny, sunny sky, 22mf/10.
r18d1abact	Many sheep grazing on a grassy hill near a lake, realistic photo of a town, Austria.
r09245347t	Wooden ceiling of the church with view to an altar, warm lighting.

Table B.1: Prompts used for the qualitative comparison in Figure 3.7. The Image ID corresponds to the original filename in the RAISE-1k dataset [17], and the prompt set follows Synthbuster [4].

Appendix B. Datasets and Benchmarks Appendix

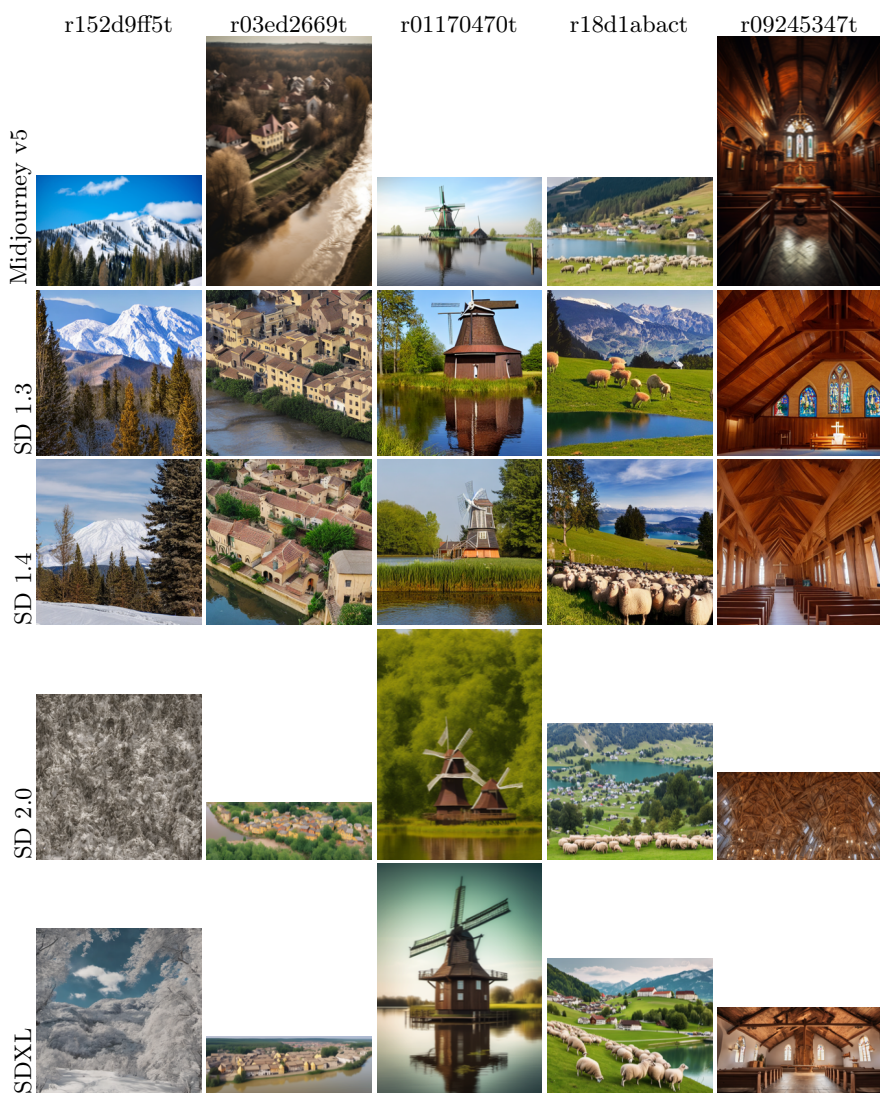


Figure B.2: Uncropped sample images from the Synthbuster benchmark [4] (Part 2/2).

Appendix C

Color Science Background

This appendix provides the color-science foundations used in the forensic analyses of Chapter 4. Section C.1 reviews human color perception and the colorimetric framework that underpins all color representations. Section C.2 then details the specific color spaces and transforms employed in the experiments, with full derivations.

C.1. Perceptual and colorimetric foundations

C.1.1. Human color perception

Color is not a physical property of objects in isolation, but a perceptual response to a light stimulus characterized by its spectral power distribution (SPD). Human vision samples this high-dimensional signal through three types of photoreceptor cells called *cones* (L, M, and S, sensitive primarily to long, medium, and short wavelengths, respectively), effectively projecting the full spectrum onto a three-dimensional response; different SPDs can therefore produce identical cone responses, a phenomenon known as *metamerism* [36]. The brain further decomposes these three-channel signals into perceptual attributes such as brightness, hue, and saturation, whose precise values depend on illumination, adaptation, and surround conditions.

Following CIE-adopted terminology, we will use the following common descriptors [13, 36]:

- **Brightness:** sensation by which an area appears to exhibit more or less light.
- **Hue:** sensation by which an area appears similar to (or a mixture of) red, yellow, green, and blue.
- **Colourfulness:** sensation of the degree to which an area appears more or less chromatic.
- **Lightness:** brightness relative to a reference white in the scene.
- **Chroma:** colourfulness relative to the brightness of a reference white.
- **Saturation:** colourfulness relative to the area’s own brightness.

A practical implication is that “the same” numerical triplet can correspond to different appearances depending on illumination, adaptation, surround, and viewing conditions. Consequently, purely coordinate-based comparisons can be misleading unless their assumptions (reference white, transfer function, viewing conditions) are made explicit. A famous illustration of this context dependence is Adelson’s checker-shadow

Appendix C. Color Science Background

illusion (Figure 4.1 in Chapter 4) [103]: tiles with identical pixel intensity appear different due to their surrounding illumination context, demonstrating that equal pixel values do not guarantee equal perceived lightness.

C.1.2. Color spaces, device dependence, and gamut

A *color space* is a coordinate system for describing colors numerically so that they can be communicated and transformed between devices and programs. The proliferation of color spaces is not redundancy: different parameterizations are convenient for different goals (hardware compatibility, efficient storage/compression, or perceptual interpretability) [22, 75].

We distinguish three levels of description: (i) a *colorimetric space/model* (e.g., XYZ, Lab) that is defined relative to a standard observer; (ii) an *RGB color space*, defined by primaries and a white point (e.g., sRGB/Rec. 709 with D65); and (iii) an *encoding*, which includes a transfer function (OETF/EOTF), quantization, and range conventions. Correlation features depend strongly on (iii), so throughout this work we state explicitly whether statistics are computed on encoded or linearized values.

A key distinction is *device dependence*. Many RGB encodings are defined relative to specific primaries, a white point, and a transfer function; therefore an (R, G, B) triplet is ambiguous without those parameters. In contrast, device-independent representations aim to specify color relative to a standard observer/model so that it can be converted consistently across capture and display pipelines.

Any given device or encoding can represent only a subset of visible colors, called its *gamut*; out-of-gamut colors are clipped or compressed, which can alter color statistics.

C.1.3. CIE colorimetry: tristimulus values and chromaticity

The CIE XYZ system represents a stimulus by three tristimulus values (X, Y, Z) , where Y is commonly treated as a luminance-related component. Chromaticity coordinates

$$x = \frac{X}{X + Y + Z}, \quad y = \frac{Y}{X + Y + Z}, \quad z = 1 - x - y, \quad (\text{C.1})$$

remove the overall intensity scale and separate chromaticity from magnitude; this separation underpins both luminance–chrominance representations (e.g., YCbCr) and perceptually motivated spaces (e.g., Lab).

XYZ is device-independent, while RGB is defined by a concrete set of primaries, white point, and transfer function. For a given RGB standard the relation is deterministic:

$$(R', G', B') \xrightarrow[\text{inverse transfer}]{\text{linearization}} (R, G, B) \xrightarrow[3 \times 3 \text{ matrix}]{} (X, Y, Z), \quad (\text{C.2})$$

where (R', G', B') are encoded values and (R, G, B) are linear-light values. The specific sRGB linearization and matrix are given in Section C.2.1.

C.1.4. Transfer functions and linear light

Pixel values stored in common image formats are often *not* proportional to scene radiance or display luminance. Historically, CRT displays were well-approximated by a power-law response $L \approx V^\gamma$ (often discussed around $\gamma \approx 2.5$), while broadcast and imaging conventions adopted related, though not identical, system transfer exponents (often discussed in the 2.2–2.8 range depending on standard and viewing assumptions), motivating gamma correction and, more generally, nonlinear transfer functions [22, 75].

For forensic analysis, the operational takeaway is simple: many operations (linear filtering, mixing, energy-based statistics) are only physically meaningful in linear light.

C.2. Color spaces used in this work

This section details the four color-space families used in the experiments of Chapter 4. Each offers a different decomposition of the three-channel signal, and their inter-channel correlation structures differ accordingly.

C.2.1. RGB

RGB is the natural representation for many camera and display pipelines. However, the interpretation of an (R, G, B) triplet depends on the RGB standard (primaries/white point) and on the transfer function. We distinguish between *linear* RGB (approximately proportional to light) and *gamma-encoded* RGB (as commonly stored in images).

For sRGB (D65), linearization is

$$C = \begin{cases} \frac{C'}{12,92}, & C' \leq 0,04045, \\ \left(\frac{C' + 0,055}{1,055} \right)^{2,4}, & C' > 0,04045, \end{cases} \quad (\text{C.3})$$

applied channel-wise to $C' \in \{R', G', B'\}$. Then XYZ is obtained by

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0,4124564 & 0,3575761 & 0,1804375 \\ 0,2126729 & 0,7151522 & 0,0721750 \\ 0,0193339 & 0,1191920 & 0,9503041 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \quad (\text{C.4})$$

The inverse transform is also linear in linear light:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 3,2404542 & -1,5371385 & -0,4985314 \\ -0,9692660 & 1,8760108 & 0,0415560 \\ 0,0556434 & -0,2040259 & 1,0572252 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \quad (\text{C.5})$$

followed by clipping to $[0, 1]$ (if necessary) and forward sRGB encoding

$$C' = \begin{cases} 12,92 C, & C \leq 0,0031308, \\ 1,055 C^{1/2,4} - 0,055, & C > 0,0031308. \end{cases} \quad (\text{C.6})$$

XYZ thus serves as a device-independent reference, while RGB provides standard-dependent coordinates linked to XYZ through a fixed linear transform in linear light and the corresponding transfer functions. When converting between standards with different white points (e.g., D50 vs D65), a chromatic adaptation transform (e.g., Bradford) is typically applied [36].

C.2.2. CIELAB

CIELAB (Lab) is derived from XYZ relative to a reference white (X_n, Y_n, Z_n) :

$$L^* = 116 f\left(\frac{Y}{Y_n}\right) - 16, \quad (\text{C.7})$$

$$a^* = 500 \left[f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right) \right], \quad (\text{C.8})$$

$$b^* = 200 \left[f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right) \right], \quad (\text{C.9})$$

Appendix C. Color Science Background

with

$$f(t) = \begin{cases} t^{1/3}, & t > 0.008856, \\ 7.787t + \frac{16}{116}, & \text{otherwise.} \end{cases} \quad (\text{C.10})$$

Lab is widely used because its axes relate more directly to perceived lightness and opponent-color directions than raw RGB [22, 36]. We also use the Lab-derived chroma and hue angle $C_{ab}^* = \sqrt{a^{*2} + b^{*2}}$ and $h_{ab} = \text{atan2}(b^*, a^*)$.

C.2.3. HSV/HSL

HSV/HSL spaces reparameterize RGB into hue, saturation, and a brightness/lightness-like axis. They are useful for visualization and heuristic reasoning, but they remain device-dependent and are not guaranteed to be perceptually uniform; moreover, multiple competing definitions exist, so the exact transform must be specified for reproducibility [22]. For the common HSV definition from normalized RGB, let

$$C_{\max} = \max(R', G', B'), \quad C_{\min} = \min(R', G', B'), \quad \Delta = C_{\max} - C_{\min}. \quad (\text{C.11})$$

Then

$$V = C_{\max}, \quad S = \begin{cases} 0, & C_{\max} = 0, \\ \frac{\Delta}{C_{\max}}, & C_{\max} > 0, \end{cases} \quad (\text{C.12})$$

and hue is

$$H = \begin{cases} 0, & \Delta = 0, \\ 60^\circ \left(\frac{G' - B'}{\Delta} \bmod 6 \right), & C_{\max} = R', \\ 60^\circ \left(\frac{B' - R'}{\Delta} + 2 \right), & C_{\max} = G', \\ 60^\circ \left(\frac{R' - G'}{\Delta} + 4 \right), & C_{\max} = B'. \end{cases} \quad (\text{C.13})$$

C.2.4. YCbCr and luminance–chrominance families

Many standards define transforms that separate a luminance-like component (Y or Y') from two chroma components. In many practical encodings (including common YCbCr), the transform is applied to nonlinear, gamma-encoded primaries so the resulting luma is Y' rather than linear luminance Y . This separation is historically motivated by broadcasting and compression, but it is also useful in forensics because it makes it explicit which statistics are tied to intensity and which are tied to chromatic variation. Ignoring digital scaling/offset constants, a common normalized YCbCr form is

$$Y' = K_R R' + K_G G' + K_B B', \quad (\text{C.14})$$

$$C_b = \frac{B' - Y'}{2(1 - K_B)}, \quad C_r = \frac{R' - Y'}{2(1 - K_R)}, \quad (\text{C.15})$$

with $K_G = 1 - K_R - K_B$ (e.g., BT.601: $K_R = 0.299$, $K_B = 0.114$). In practice, implementations also apply range/offset conventions (e.g., limited-range 8-bit video with offsets 16 and 128) and use coefficients tied to a specific standard (BT.601/709/2020). We use YCbCr definitions consistent with the decoding pipeline used in experiments and report that choice.

References

- [1] Adobe. Adobe firefly. Product page, 2026.
- [2] Cecilia Aguerrebere, Julie Delon, Yann Gousseau, and Pablo Musé. Study of the digital camera acquisition process and statistical modeling of the sensor raw data. Technical Report hal-00733538, HAL, 2013.
- [3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, 2017.
- [4] Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 01 2024.
- [5] Quentin Bammey, Rafael Grompone von Gioi, and Jean-Michel Morel. Forgery detection by internal positional learning of demosaicing traces. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 328–338, January 2022.
- [6] Peter Bentley. How do machine learning gans work? <https://www.sciencefocus.com/future-technology/how-do-machine-learning-gans-work>, Jan 11 2023. [Accessed 12th Feb, 2026].
- [7] Kshitiz Bhargava. Stylegan-stylegan2 deepfake face images. Kaggle dataset.
- [8] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. arXiv:1809.11096.
- [9] Lucy Chai, David Bau, Jaesik Lim, Steven W. Chan, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European Conference on Computer Vision (ECCV)*, pages 103–120. Springer, 2020.
- [10] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Li Hao, Jiaxin Ai, Qin Zou, Chen Li, and Zhongyuan Wang. Stacking brick by brick: Aligned feature isolation for incremental face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2411.11396.
- [11] Robert Chesney and Danielle Keats Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6):1753–1820, 2019.
- [12] Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12830–12839, 2025.

References

- [13] CIE. International lighting vocabulary (ilv). CIE standard / online vocabulary. Accessed: 2020.
- [14] Aldus Corporation. Tiff revision 6.0, June 3, 1992.
- [15] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 973–982, 2023.
- [16] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4356–4366, 2024.
- [17] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. RAISE: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference*, pages 219–224, 2015.
- [18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [19] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations*, 2017. arXiv:1605.08803.
- [20] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7890–7899, 2020.
- [21] David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 382–392, 2023.
- [22] Adrian Ford and Alan Roberts. Colour space conversions, 1998. Revision dated 2001-07-20; distributed by Charles Poynton.
- [23] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Prentice Hall, 3 edition, 2008.
- [24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.
- [25] Google. An image format for the web. <https://developers.google.com/speed/webp>. [Accessed 9th April, 2024].
- [26] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.
- [27] PNG Development Group. Png (portable network graphics) specification, version 1.2. <http://www.libpng.org/pub/png/spec/1.2/png-1.2.pdf>. [Accessed 9th April, 2024].
- [28] Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18685–18694, 2025.

- [29] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017. arXiv:1704.00028.
- [30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [31] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β -VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.
- [32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020. arXiv:2006.11239.
- [33] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. arXiv:2207.12598.
- [34] Gerald C. Holst. *CCD Arrays, Cameras, and Displays*. SPIE Press, 1998.
- [35] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28831–28841, 2025.
- [36] R. W. G. Hunt. *The Reproduction of Colour*. Wiley, 6 edition, 2004.
- [37] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.
- [38] ITU-R. Recommendation BT.601-7: Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios. International Telecommunication Union, 2011.
- [39] ITU-R. Recommendation BT.709-6: Parameter values for the HDTV standards for production and international programme exchange. International Telecommunication Union, 2015.
- [40] Zexi Jia, Chuanwei Huang, Yeshuang Zhu, Hongyan Fei, Xiaoyue Duan, Zhiqiang Yuan, Ying Deng, Jiawei Zhang, Jinchao Zhang, and Jie Zhou. Secret lies in color: Enhancing AI-generated images detection with color distribution analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13445–13454, 2025.
- [41] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.
- [42] Dimitrios Karageorgiou, Quentin Bammey, Valentin Porcellini, Bertrand Goupil, Denis Teyssou, and Symeon Papadopoulos. Evolution of detection performance throughout the online lifespan of synthetic images. In *Computer Vision – ECCV 2024 Workshops*, pages 400–417, 2024.
- [43] Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstathios Gavves. Any-resolution ai-generated image detection by spectral learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18706–18717, 2025.

References

- [44] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. arXiv:1812.04948.
- [45] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. arXiv:1912.04958.
- [46] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, 2018.
- [47] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.
- [48] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems*, 2020.
- [49] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, 2024.
- [50] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang. Identification of deep network generated images using disparities in color components. *Signal Processing*, 174:107616, 2020.
- [51] Yanhao Li, Quentin Bammey, Marina Gardella, Tina Nikoukhah, Jean-Michel Morel, Miguel Colom, and Rafael Grompone Von Gioi. Masksim: Detection of synthetic images by masked spectrum similarity analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3855–3865, 2024.
- [52] Yichi Liu, Xuefeng Wei, Shunquan Wang, and Yun Q. Shi. Detecting generated images by real images. In *European Conference on Computer Vision (ECCV)*, pages 87–103. Springer, 2022.
- [53] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, 2022. arXiv:2201.09865.
- [54] H. Maître. *From Photon to Pixel: The Digital Camera Handbook*. ISTE Limited, 2017.
- [55] Henrique S. Malvar, Li-wei He, and Ross Cutler. High-quality linear interpolation for demosaicing of bayer-patterned color images. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 485–488. IEEE, 2004.
- [56] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting gan-generated images by orthogonal training of multiple cnns. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3091–3095. IEEE, 2022.
- [57] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. arXiv:2108.01073.

- [58] Midjourney. The pope drip. Wikimedia Commons media file (**File:Pope Francis in puffy winter jacket.jpg**), May 2023. AI-generated image of Pope Francis wearing a puffy winter jacket. Public domain on Commons (AI-generated, no human author). Accessed 28th February, 2026.
- [59] Midjourney. Midjourney documentation: Image prompts. Documentation page, 2026.
- [60] Midjourney. Midjourney documentation: Text in images. Documentation page, 2026.
- [61] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.
- [62] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don’t know? In *International Conference on Learning Representations*, 2019. arXiv:1810.09136.
- [63] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. *arXiv:1906.02994*, 2019.
- [64] Sergio Nesmachnow and Santiago Iturriaga. Cluster-uy: Collaborative scientific high performance computing in uruguay. In Moisés Torres and Jaime Klapp, editors, *Supercomputing*, pages 188–202, Cham, 2019. Springer International Publishing.
- [65] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022.
- [66] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023.
- [67] OpenAI. Dall-e 3 is now available in chatgpt plus and enterprise. OpenAI product note, 2023.
- [68] OpenAI. GPT Image models: gpt-image-1, 2025. OpenAI API documentation (GPT Image family includes **gpt-image-1**). Accessed: 2025-12-27.
- [69] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.
- [70] Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8245–8257, 2025.
- [71] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. arXiv:2212.09748.
- [72] Orazio Pontorno, Luca Guarnera, and Sebastiano Battiato. DeepFeatureX-SN: Generalization of deepfake detection via contrastive learning. *Multimedia Tools and Applications*, 84(39):47721–47740, 2025.
- [73] Alin Popescu and Hany Farid. Exposing digital forgeries in color filter array interpolated images. *IEEE Transactions on Signal Processing*, 53(10):3948–3959, 10 2005.

References

- [74] Alin C. Popescu and Hany Farid. Statistical tools for digital forensics. In *Proceedings of the 6th International Conference on Information Hiding*, IH'04, page 128–147, Berlin, Heidelberg, 2004. Springer-Verlag.
- [75] Charles Poynton. *Digital Video and HD: Algorithms and Interfaces*. Morgan Kaufmann, 2 edition, 2012.
- [76] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.
- [77] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016. arXiv:1511.06434.
- [78] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [79] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2019. arXiv:1906.02845.
- [80] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015. PMLR 37.
- [81] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic back-propagation and approximate inference in deep latent gaussian models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1278–1286, 2014.
- [82] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for PyTorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3674–3683, 2020.
- [83] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. arXiv:2112.10752.
- [84] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.
- [85] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023. arXiv:2104.07636.
- [86] Christoph Schuhmann, Richard Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Razvan Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Christopher Mullis. LAION-400m: Open dataset of CLIP-filtered 400 million image-text pairs. In *NeurIPS 2021 Workshop on Data-Centric AI*, 2021. arXiv:2111.02114.

- [87] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
- [88] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. arXiv:2010.02502.
- [89] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32211–32252, 2023.
- [90] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, pages 11895–11907, 2019.
- [91] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. arXiv:2011.13456.
- [92] Stability AI. Stable diffusion launch announcement. Stability AI news post, 2022.
- [93] Chengxu Tan, Yue Zhao, Lei Zhang, Haoqiang Song, Long Ma, Yu Guo, Shiqi Wang, and Weisi Lin. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12105–12114, 2023.
- [94] Lea Uhlenbrock, Davide Cozzolino, Denise Moussa, Luisa Verdoliva, and Christian Riess. Did you note my palette? unveiling synthetic images through color statistics. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec '24*, page 47–52, New York, NY, USA, 2024. Association for Computing Machinery.
- [95] Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1), 2020.
- [96] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [97] Gregory K. Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.
- [98] Ruihan Wang, Chen Ni, Xintao Wang, Liming Jiang, and Ziwei Liu. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22405–22415, 2023.
- [99] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8695–8704, 2020.
- [100] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018. arXiv:1809.00219.

References

- [101] Lilian Weng. Flow-based deep generative models. <https://lilianweng.github.io/posts/2018-10-13-flow-models/>, 2018. Lil’Log [Accessed 12th Feb, 2026].
- [102] Wikipedia contributors. Variational autoencoder. https://en.wikipedia.org/wiki/Variational_autoencoder, 2024. [Accessed 12th Feb, 2026].
- [103] Wikipedia contributors. Checker shadow illusion. https://en.wikipedia.org/wiki/Checker_shadow_illusion, 2025. [Accessed 28th February, 2026].
- [104] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22412–22423, 2023.
- [105] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.
- [106] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
- [107] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. arXiv:1512.04150.

List of Tables

2.1.	Summary of key camera-pipeline stages and examples of how they can affect low-level image statistics relevant to forensic analysis.	10
4.1.	Training data composition for the limited multi-generator regime used in ResNet-50 experiments.	57
4.2.	Ablation over input representations (AUC, %) under the benchmark protocol of Cozzolino et al. [16]. † indicates training with additional generator data; gray marks in-domain generators. Bold and <u>underline</u> denote the best and second-best results per column.	61
4.3.	Comparison with prior methods in terms of AUC (%) under the benchmark protocol of Cozzolino et al. [16]. Results are reported across GAN-based generators, diffusion-based generators, and commercial tools. Gray entries indicate the in-domain generator(s) associated with each method's training setup as reported by the protocol, whereas the remaining columns measure cross-source generalization. Bold marks the best overall average and <u>underline</u> marks the second-best.	63
B.1.	Prompts used for the qualitative comparison in Figure 3.7. The Image ID corresponds to the original filename in the RAISE-1k dataset [17], and the prompt set follows Synthbuster [4].	81

This page has been intentionally left blank.

List of Figures

2.1. Simplified camera pipeline showing the main stages from sensor acquisition to compressed image storage.	8
2.2. Bayer color filter array (CFA) sampling pattern (example). Single-sensor CFAs measure only one color per pixel and therefore require demosaicing to reconstruct full RGB at each pixel.	9
2.3. Schematic overview of a Generative Adversarial Network (GAN): the generator produces synthetic samples, while the discriminator learns to distinguish them from real data. Extracted from [6].	12
2.4. Canonical Variational Autoencoder (VAE) architecture: an encoder maps the input to a latent representation, a latent variable is sampled from the approximate posterior, and a decoder reconstructs the image from that latent code. Extracted from [102].	13
2.5. A normalizing flow transforms a simple base distribution $p_0(\mathbf{z}_0)$ into a complex target distribution $p_K(\mathbf{z}_K) = p(\mathbf{x})$ by composing a sequence of invertible mappings $f_1, \dots, f_K$. Each intermediate variable $\mathbf{z}_i$ follows a progressively more expressive distribution. Adapted from [101].	14
2.6. Latent diffusion model (LDM) pipeline, as used in Stable Diffusion: the image is first compressed into a lower-dimensional latent representation by an autoencoder, diffusion sampling is performed in that latent space, and the final latent is decoded back to pixel space. Extracted from [83].	15
2.7. Conceptual overview of a consistency model: the same network is trained so that different noisy versions of a sample are mapped to a consistent prediction of the clean data, enabling generation in one or a few function evaluations. Extracted from [89].	16
2.8. Empirical comparison between pixel and perceptual distances (all values computed against <code>pope.png</code> , sourced from [58]). Gaussian blur and dark color-correction examples are self-generated transformations. The face-swap example is produced with <code>gpt-image-1</code> [68]. Gaussian blur yields very low MSE but relatively high LPIPS, while a mostly global illumination change yields higher MSE with lower LPIPS. A semantic edit (face swap) increases LPIPS strongly even with moderate MSE, illustrating why perceptual losses are useful complements to pixelwise criteria.	20

List of Figures

3.1. Overview of Learning on Gradients (LGrad) [93]. A frozen pretrained CNN is used as a transformation model: instead of feeding the RGB image directly to the detector, the method computes the gradient of the model output with respect to the input image and uses this gradient map as the artifact representation. The resulting normalized gradients are then passed to a binary classifier that predicts whether the image is real or GAN-generated.	28
3.2. UCF architecture and latent-space behavior under the FF++ analysis protocol [104]. Top: overview of UCF (Uncovering Common Features) for generalizable deepfake detection. Bottom: t-SNE visualization on FF++ comparing latent features from baseline Xception and UCF; the baseline representation forms method-dependent clusters, while in UCF the specific branch remains method-discriminative, the common branch captures method-shared forgery evidence, and the content branch emphasizes forgery-irrelevant factors.	29
3.3. B-Free: bias-free training paradigm. Synthetic images are generated from self-conditioned transformations of real images and augmented with additional diffusion-based edits to avoid semantic biases; this yields training data where the only systematic difference between classes is the act of AI generation, improving generalization and calibration [28]. . .	30
3.4. RINE architecture [49]. The method aggregates intermediate representations from a frozen CLIP image encoder using a lightweight, trainable weighting and classification module to improve generalization.	32
3.5. Overview of the processing pipeline in [20]. The method contains two main blocks: a feature-extraction block using DFT and azimuthal averaging, and a training block where a classifier uses the transformed features to determine whether a face is real or fake. Input images are transformed to grayscale before DFT.	33
3.6. Examples of real and synthetic images drawn from popular datasets, as analyzed in [15]. Real datasets (left): ImageNet [18], FFHQ [44], LAION [86]. Synthetic datasets (right): DiT [71], StyleGAN2 [45], GLIDE [65]. For each dataset, the figure shows the average power spectra (middle) and autocorrelation (bottom) of noise residuals. Real images embed some traces of JPEG artifacts, while generated images often exhibit anomalous periodic patterns in the spatial domain.	34
3.7. Sample images from the Synthbuster benchmark [4]. Each row corresponds to a generator, and each column corresponds to a unique semantic prompt, identified by the RAISE-1k image ID. Supplementary prompt details and uncropped samples are provided in Appendix B (Section B.1). 36	
3.8. Flowchart of MaskSim [51]. The method learns a spectral mask and a reference spectrum to expose which frequency regions act as generator fingerprints and to make decisions auditable.	37
3.9. Overview of the SPAI approach [43]. A real-image spectral prior is learned via masked spectral learning and then used for out-of-distribution scoring, while patch-level processing enables any-resolution detection without aggressive resizing.	37
3.10. Illustration of the process of computing DIRE given an input image x_0 [98]. An image is inverted and reconstructed with a diffusion model, and the resulting reconstruction-error map is used as forensic evidence. 39	

3.11. Overview of FIRE (<i>Frequency-guided Reconstruction Error</i>) adapted from Chu et al. [12]. FIRE emphasizes reconstruction-error evidence tied to informative mid-frequency components to improve robustness under common post-processing.	39
3.12. FMRE architecture (shared encoder with two decoders) used in FIRE to predict the mid-band mask and its complementary mask.	40
3.13. Pipeline of chrominance-disparity-based detection adapted from Li et al. [50]. The input image is decomposed into selected color components (e.g., H, S, C_b, C_r), transformed into residual maps, encoded via co-occurrence statistics, and finally classified as real or generated.	41
3.14. Example from Uhlenbrock et al. [94] showing how relational color-channel transforms improve the visual saliency of manipulated content. Left: natural camera image. Middle: AI inpainting example (edited region outlined). Right: residual representation after relational inter-channel processing, where the manipulated region and its boundary become substantially more visible.	42
3.15. Qualitative examples of original images and corresponding color difference maps (visualized with amplification) from Jia et al. [40]. The examples illustrate that generated images tend to produce stronger restoration-induced color deviations than real images.	43
3.16. Examples and color-distribution statistics from Jia et al. [40]. Left: representative images from real data and multiple generative models. Right: corresponding RGB-channel histogram profiles over large sample sets, highlighting model-dependent distribution behaviors that motivate color-centric forensic cues.	44
3.17. SIDA framework adapted from Huang et al. [35]. The assistant-style model jointly predicts authenticity class, localizes tampered regions when applicable, and produces a natural-language rationale.	45
4.1. Checker shadow illusion [103]. Left: tiles A and B have identical pixel intensity but appear different due to the surrounding illumination context. Right: linking the two regions reveals they share the same value. This illustrates why equal pixel values do not guarantee equal perceived lightness, and why color coordinates must be interpreted together with their viewing and encoding assumptions.	49
4.2. RAISE-1k-averaged LPIPS distance between an original RGB image and its perturbed versions as a function of noise coverage ρ for different color spaces (RGB, Lab, HSV, YUV) under correlated (solid) and decorrelated (dashed) channel noise with matched perturbation power.	53
4.3. Log-density estimates of pairwise inter-channel correlation features across color spaces (RGB, HSV, Lab, YUV). Each row corresponds to one color space and shows the three channel-pair correlations. We report real images from RAISE-1k [17] and synthetic images from multiple generators in Synthbuster [4].	55
4.4. Mean Wasserstein distance matrices between inter-channel correlation feature distributions across sources, shown for multiple color spaces (RGB/HSV/Lab/YUV). Real images are from RAISE-1k [17] and synthetic images are from Synthbuster [4].	56
4.5. Real photograph vs. GPT [68] recreation under the color-space transforms and correlation maps used in this work. Each row shows one representation; left pair: color-space view, right pair: corresponding local correlation map.	60

List of Figures

A.1. Architecture of the Spectral-Flow pipeline. Top: the spectral energy extractor computes a noise residual via a frozen DnCNN denoiser, extracts overlapping patches, applies dihedral augmentations (rotations and flips), and averages their Fourier power spectra to produce a scale-normalized spectral-energy tensor. Middle: a convolutional autoencoder is trained on spectral-energy tensors of real images with an MSE reconstruction loss, augmented by a contrastive term that compacts normal latents and repels generated-image latents. Bottom: a normalizing flow is fitted on frozen encoder latents from real images; at inference, the negative log-likelihood under the flow serves as an anomaly score.	73
A.2. UMAP projection of Spectral-Flow encoder latents for 2,400 images (600 real, 1,800 synthetic). Real-image sources (ImageNet, FFHQ, RAI-SE) are shown in shades of green; synthetic generators are shown in distinct colors. The real images do not form a single compact cluster: they are scattered across the central region and overlap heavily with several generators. Meanwhile, some generators (notably SDXL, DALL-E 3) form isolated clusters far from the real images, while others (Glide, Midjourney, earlier SD versions) are interspersed with real data. This fragmented geometry makes density-based anomaly detection unreliable: a normalizing flow fitted on the real-image latents cannot define a coherent normality boundary.	75
B.1. Uncropped sample images from the Synthbuster benchmark [4] (Part 1/2).	80
B.2. Uncropped sample images from the Synthbuster benchmark [4] (Part 2/2).	82

This is the last page.
Compiled on martes 30 junio, 2026.
<http://iie.fing.edu.uy/>